Understanding variation in nucleotide diversity across the mouse genome

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

Ded - i - ca - tion

Noun

1. The quality of being dedicated or committed to a task or purpose.
   “His dedication to his duties“
2. The action of dedicating a church or other building.
   “The dedication and unveiling was attended by some 5,000 people“

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 was hugely helpful and friendly, it's just a shame he left Edinburgh. Enormous thanks to Dan Halligan for helping me out even after he left the lab.

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

In no particular order, thanks to my friends Rasmus, Luiz, Stevie, Andres, Nathan and Billy for palling around with me in Edinburgh and Jaz and Michael for palling around in Vancouver. If you are reading this and are not listed, but think that you should be, I'm sorry.

Thanks to the Booker gang for their support, particularly Mum and Dad. Stella, you nailed it. NCSJTN83.

Throughout my PhD, Arya has looked after me very well and even married me. Thanks, pal.

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:

- 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

Declaration i

Dedication ii

Acknowledgements iii

Publications v

Contents vi

List of Figures xii

List of Tables xiii

1 Introduction 3

1.1 Key concepts 3

1.1.1 Genetic Drift 4

1.1.2 Mutation 4

1.1.3 Selection 4

1.1.4 Migration 4

1.1.5 Recombination 4

1.2 Models of selective sweeps 4

1.3 Using Models of Selective Sweeps to Estimate Positive Selection Parameters 5

1.3.1 The Correlation Between Diversity and the Rate of Recombination 6

1.3.2 Correlations Between Neutral Diversity and Non-Neutral Divergence 7

1.3.3 Patterns of Diversity Around the Targets of Selection 9

1.4 Fitting genome wide patterns 11

1.5 State of research 11

1.6 Aims of the thesis 11

2 The recombination landscape in wild house mice inferred using population genomic data 12

2.1 Abstract 12

2.2 Introduction 13

2.3 Materials and Methods 16

2.3.1 Polymorphism data for *Mus musculus castaneus* 16

2.3.2 Inferring phase and estimating switch error rates 17

2.3.3 Estimating recombination maps and validation of the approach . 18

2.3.4 Recombination rate estimation for *M. m. castaneus* 20

2.3.5 Broad scale comparison to previously published maps 22

2.3.6 Fine-scale recombination rate variation in wild *M. m. castaneus* 24

2.3.7 Examining the correlation between nucleotide diversity and recombination rate 25

2.4 Results 26

2.4.1 Phasing SNPs and estimating the switch error rate 26

2.4.2 Simulations to validate LDhelmet for the population sample of *M. m. castaneus* 27

2.4.3 Recombination rates in the *M. m. castaneus genome* 28

2.4.4 Comparison of the *M. m. castaneus* map to maps constructed using inbred lines 29

2.4.5 Analysis of fine-scale recombination rates in wild *M. m. castaneus* 32

2.4.6 Correlations between recombination rate and properties of protein coding genes in *M. m. castaneus* 34

2.5 Discussion 34

3 Understanding the factors that shape patterns of nucleotide diversity in the house mouse genome 43

3.1 Abstract 44

3.2 Introduction 44

3.3 Materials and Methods 48

3.3.1 Samples and polymorphism data 48

3.3.2 Functional elements in the murid genome 49

3.3.3 The site frequency spectrum around functional elements 50

3.3.4 Overview of DFE-alpha analysis 50

3.3.5 Inference of the uSFS and the DFE 51

3.3.6 Two methods for inferring the rates and effects of advantageous mutations based on the uSFS 53

3.3.7 Forward-in-time simulations modelling background selection and selective sweeps 54

3.4 Results 61

3.4.1 Inferring the unfolded site frequency spectrum 61

3.4.2 Estimating the frequencies and strengths of deleterious and advantageous mutations 63

3.4.3 Forward-in-time population genetic simulations 65

3.4.4 i) Patterns of nucleotide diversity around functional elements in simulated populations 65

3.5 Discussion 71

3.5.1 Estimating selection parameters based on the uSFS 72

3.5.2 Patterns of diversity and Tajimas D around functional elements . 73

3.5.3 Rates of nucleotide substitutions in simulations 74

3.5.4 Do our results provide evidence for strongly selected advantageous mutations? 74

3.5.5 Limitations of the study 77

3.6 Conclusions 79

4 Estiamting parameters of selective sweeps from patterns of genetic diversity in house mice 80

4.1 Introduction 80

4.2 Materials and Methods 83

4.2.1 Simulations 83

4.2.2 Analysis of the uSFS 84

4.2.3 Model of Recurrent Sweeps with Background Selection 85

4.2.4 Analysis of Mouse Data 86

4.2.5 Estimates of B 88

4.3 Results 89

4.3.1 Estimating selection parameters from the uSFS of simulated data 89

4.3.2 Patterns of genetic diversity around protein-coding exons and conserved non-coding elements 91

4.3.3 Diversity expected in the absence of selection, π_0 93

4.3.4 Parameters of selective sweep obtained from patterns of nucleotide diversity . . . 94

4.4 Discussion . . . 95

4.4.1 Analysis of the uSFS . . . 96

4.4.2 Estimating parameters of positive selection from the uSFS versus patterns of diversity . . . 97

4.4.3 The relative contribution of adaptive substitutions in protein-coding and regulatory regions to fitness change in mice . . . 99

4.5 Conclusions . . . 101

4.6 Acknowledgements . . . 102

5 General Discussion 103

5.1 Soft and partial selective sweeps in mice . . . 104

5.2 The interaction between natural selection and demographic history . . . 104

5.3 Making use of more of the available data . . . 106

5.4 Moving beyond mice . . . 108

Bibliography 110

Appendices 117

A Booker *et al.* 2017 - BMC Biology 118

B Recombination in wild mice 129

B.1 Supplementary Material . . . 129

B.2 Booker *et al.* 2017 - Genetics . . . 135

C Understanding patterns of genetic diversity 149

C.1 Supplementary Methods . . . 149

C.1.1 Demographic correction 149

C.1.2 Divergence correction 150

C.1.3 Ancestral effective population size, Ne-anc Model B 151

List of Figures

2.1 The effect of switch errors on recombination rate inference 28

2.2 Comparison of LD-based and pedigree-based recombination maps 31

2.3 Broad-scale correlations between recombiantion maps for *Mus musculus castaneus* and *Mus musculus domesticus* 32

4.1 The 91

B.2 The effect of switch errors on recombination rate inference 133

B.3 The effect of switch errors on recombination rate inference 134

List of Tables

2.1 Summary of recombination rates per chromosomes 30

2.2 Correlations between recombination rate and genetic diversity 34

B.1 Summary of recombination rates per chromosomes 130

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). 131

B.3 The total number of SNPs in the dataset, and the number of SNPs after applying filters. 131

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. 133

ThesisAbstract

In this thesis I will describe several attempts to understand the ways in which evolutionary forces have shaped patterns of genetic diversityu across the genome of the house mouse. I will start by introducing core concepts in evolutionary genetics and describing how recombination an selection interact to shape patterns of genetic diversity. I will then describe three large projects in which I examine aspects of the evolutionary biology of house mice.

1. **Recombination rate variation.** In this chapter I estiamte the landscape of recombination rate vatiation in wild house mice

2. **Understanding the factors that shape patterns of genetic diversity.** In this chapter I estiamte the distribution of fitness effects for new mutations from the site frequency spectrum. These estiamtes are then used in forward-in-time simulations to dissect the causes of troughs in genetic diversity around funcitonal elements.

3. **Estimating sweep parameters from patterns of genetic diversity.** In this chapter I use an explicit model of seletion at linked sites to estiamte the frequeency and strength of positively selected mutations occuring in different functional elements in the mouse genome.

Finally, I discuss the implications of my findings from my three chapters and end by suggesting further avenues for enquiry.

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.

In this introduction, I will start by briefly describing the main population genetic processes that will be discussed in this thesis. I will then describe how population genetic models can be used to make inferences from genomic data. I will end by describing the state of research in population genomics, with particular regard to natural selection, and describe the main aims of this thesis.

1.1 Key concepts in evolutionary genetics

Evolution is a population genetic process. Although population genetics does not perhaps capture all aspects of evolution (such as cultural, behavioural and psychological evolution), population genetics provides a framework for understanding how evolution occurs at the molecular level.

Broadly speaking, there are five processes that govern the evolutionary process: genetic drift, mutation, selection, migration and recombination.

1.1.1 Genetic Drift

1.1.2 Mutation

1.1.3 Selection

1.1.4 Migration

1.1.5 Recombination

1.2 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.

Hard/classic sweeps

The most well-studied model of sweeps. A new advantageous mutation rapidly increases in frequency to eventual fixation (shown in [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 [B]). A similar process, also termed a soft sweep, can occur if an advantageous mutation arises by multiple, distinct mutation events (shown in [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 [A]). The term partial sweep describes the situation wherein a sweeping allele becomes effectively neutral at a certain frequency in its trajectory (shown in [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).

1.3 Using Models of Selective Sweeps to Estimate Positive Selection Parameters

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, 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.

1.3.1 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 $\lambda 2N_e s = 5.37 \text{ x } 10^{-8}$, 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., 1995). 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.

1.3.2 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*).

1.3.3 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 above), 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. (2012) 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. (2012) 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; 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.

1.4 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*.

1.5 State of research

1.6 Aims of the thesis

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 many evolutionary processes. The landscape of recombination has been studied previously in the house mouse, *Mus musculus*, and it is known that the different sub-species exhibit different suites of recombination hotspots. However, it is not established whether broad-scale variation in the rate of

recombination is conserved between the sub-species or whether hotspots identified in laboratory strains reflect the diversity of hotspots locations in natural populations. In this study, we construct a fine-scale recombination map for the Eastern house mouse sub-species, *M. m. castaneus*, using 10 individuals sampled from its ancestral range. We perform simulations to assess how accurately recombination rates are inferred considering phasing errors. We use a novel approach to quantify phase error, which we estimate to affect 0.5% of heterozygous SNPs in our data. We use LDhelmet to construct recombination maps for each autosome. We find that the spatial distribution of recombination rate is strongly positively correlated between our castaneus map and a map constructed using inbred lines of mice derived predominantly from *M. m. domesticus*. However, despite this high similarity we find that potential recombination hotspots in wild mice show little overlap with the locations of double-strand breaks in wild-derived strains of laboratory mice, though the greatest overlap is with a strain derived from wild *M. m. castaneus*. Finally, we also find that levels of genetic diversity in *M. m. castaneus* are positively correlated with the rate of recombination, consistent with pervasive natural selection acting in the genome. Our study suggests that recombination rate variation is conserved at broad scales between two sub-species of *M. musculus*, though not at fine scales.

2.2 Introduction

In many species, rates of crossing-over are not uniformly distributed across chromosomes, and understanding this variation and its causes is important for many aspects of molecular evolution. Experiments in laboratory strains or managed populations examining the inheritance of markers through pedigrees have allowed direct estimation of rates of crossing-over in different regions of the genome. Studies of this kind are impractical for many wild populations, where pedigree structures are largely unknown (but see Johnston et al. 2016). In mice, there have been multiple 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 sub-species, *Mus musculus domesticus* (Yang et al., 2011). Recombination rate variation in laboratory strains may not, therefore, reflect natural rates and patterns in wild mice of different sub-species. 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 population's pedigree and many genotyped individuals (but see Wang et al. 2017).

To understand variation in recombination rates, patterns of linkage disequilibrium (LD) in a sample of individuals drawn from a population can be used. Coalescent-based methods have been developed that use such data to indirectly estimate recombination rates at very fine scales (Hudson, 2001; McVean et al., 2002, 2004; Auton and McVean, 2007; Chan et al., 2012). The recombination rates estimated in this way reflect variation in crossing-over rates in populations ancestral to the extant population, 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 (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 PRDM9 binding motif observed in humans.

Great ape species have different alleles of the PRDM9 gene (Schwartz et al., 2014) and relatively little hotspot sharing (Winckler et al., 2005; Stevison et al., 2015). Correlations between the broad-scale recombination landscapes of the great apes are, however, relatively strongly positive (Stevison et al., 2011, 2015). This suggests that, while hotspots evolve rapidly, the overall genetic map changes more slowly. Indeed, multiple closely related species pairs with different hotspot locations show correlations between recombination rates at broad scales (Smukowski and Noor, 2011), as do 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* sub-species complex began to split into the present-day sub-species around 350,000 years ago (Geraldes et al., 2011). In this time, functionally distinct alleles of the PRDM9 gene and different suites of hotspots have evolved in the sub-species (Smagulova et al., 2016). In addition, between members of the *M. musculus* sub-species complex, there is also variation in recombination rates at relatively broad scales for multiple regions of the genome (Dumont et al., 2011), and recombination rates can be polymorphic between *M. m. domesticus* individuals (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 mouse autosomes. The recombination rate map they constructed is similar to a genetic map generated using crosses by Cox et al. (2009). Both studies were conducted using the classical inbred lines, whose ancestry is largely *M. m. domesticus* (Yang et al., 2011), and their estimated recombination rate landscapes may therefore reflect that of *M. m. domesticus* more than other members of the *M. musculus* sub-species complex.

In this study, we construct a recombination map for the house mouse sub-species *M. m. castaneus*. We used the genome sequences of 10 wild-caught individuals of *M. m. castaneus* from the species expected 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 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 map of recombination rates and compared this to previously published genetic maps for *M. musculus*. We show that variation in recombination rates in *M. m. castaneus* is very similar to rate variation estimated in the classical inbred strains, at broad scales. However, we find little correspondence in fine-scale recombination rate variation between *M. m. castaneus* and previously reported rate This suggests that, at broad scales, recombination rates have been relatively highly conserved since the sub-species began to diverge.

2.3 Materials and Methods

2.3.1 Polymorphism data for *Mus musculus castaneus*

We analyzed the genomes of 10 wild-caught *M. m. castaneus* individuals sequenced by 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 among the highest levels of genetic diversity among the *M. musculus* sub-species (Baines and Harr, 2007). In addition, the individuals sequenced represent a single population cluster and showed little evidence for substantial inbreeding Halligan et al. (2010). (Halligan et al., 2013) sequenced individual genomes to high coverage using multiple libraries of Illumina paired-end reads, which were mapped 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 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 purposes of estimating recombination rates, variable sites were filtered on the basis of several conditions: Insertion/deletion polymorphisms were excluded, because the method used to phase variants (see below) cannot process these sites. We also excluded sites with more than two alleles and those that failed the Samtools Hardy-Weinberg equilibrium test (p ¡ 0.002).

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 estimate haplotypes, heterozygous SNPs called in *M. m. castaneus* were phased using read-aware phasing in ShapeIt2 (Delaneau et al., 2013). ShapeIt2 uses sequencing reads that span multiple heterozygous variants, phase-informative reads (PIRs), and LD to phase variants at the level of whole chromosomes. Incorrectly phased heterozygous SNPs, termed switch errors, may upwardly bias estimates of the recombination rate, because they appear identical to legitimate crossing-over events. To assess the impact of incorrect phasing on our recombination rate inferences, we quantified the switch error rate as follows. The population sample of *M. m. castaneus* comprised of 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-chromosome of the three males (samples H12, H28 and H34 from Halligan et al. (2013)) to make three datasets of pseudo-females, which are female-like, but in which 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 used by Halligan et al. (2013), as outlined above, 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 the switch error rates produced by ShapeIt2 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 not included in 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 choose filter parameters to apply to the autosomal data that generated a low switch error rate in ShapeIt2, while maintaining a high number of heterozygous SNPs. We obtained 20 phased haplotypes for each of the 19 mouse autosomes. With these, we estimated the recombination rate landscape for *M. m. castaneus*.

2.3.3 Estimating recombination maps and validation of the approach

LDhelmet (v1.7; Chan et al. 2012) generates a sex-averaged map of recombination rates from a sample of haplotypes that are 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 that are best supported between adjacent SNPs. LDhelmet appears to perform well for species with large effective population size (N_e) and has been shown to be robust to the effects of selective sweeps, which may be prevalent and reduce diversity in and around functional elements of the *M. m. castaneus* genome (Halligan et al., 2013). The underlying model of LDhelmet relies on the assumption that populations are at recombination-drift equilibrium. We assume this to be the case for our sampled population, however violation of this may result in biased recombination rate estimates. However, the analyses conducted by 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 gives reliable recombination maps, we validated and parametrized LDhelmet using simulated datasets.

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 leads to the most accurate estimates of the recombination rate in chromosomes that have levels of diversity and base content similar to *M. m. castaneus*. Chromosomes with constant values of $\rho = 4N_e r$ 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 for all simulations, as this is close to the genome-wide average for our data, based on pairwise differences. Simulations were run for 10,000 generations to achieve equilibrium levels of polymorphism, at which time 10 diploid individuals were sampled from the population. Each simulation was repeated 20 times, resulting in 10Mbp of sequence for each value of ρ. 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). We inferred recombination rates from the simulated data in windows of 4,400 SNPs with a 200 SNP overlap between windows, following (Chan et al., 2012). We analyzed the simulated data using LDhelmet with block penalties of 10, 25, 50 and 100. The default parameters of LDhelmet are tuned to analyse *Drosophila melanogaster* data (Chan et al., 2012). Since the *D. melanogaster* population studied by Chan et al. (2012) has comparable levels of genetic diversity to *M. m. castaneus* we used the defaults for all other parameters, other than the block penalty and estimate of θ.

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 chosen SNP filter set (see Results). We then inferred the recombination rates, as above, for the simulated population using these error-prone data. We assessed the effect of switch errors on recombination rate inference by comparing estimates based on the simulated data both with and without switch errors. It is worth noting that there is the potential for switch errors to undo crossing-over events, reducing inferred recombination rates, if they affect heterozygous SNPs that are breakpoints of recombinant regions.

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

We used LDhelmet (Chan et al., 2012), to estimate recombination rates for each of the *M. m. castaneus* autosomes. It is well established that autosomal recombination rates differ between the sexes in *M. musculus* (Cox et al., 2009; Liu et al., 2014). A drawback of LD-based approaches is that they give sex-averaged recombination rates.

We used both *M. famulus* and *R. norvegicus* as outgroups to assign ancestral alleles to polymorphic sites. LDhelmet incorporates both 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 outgroups shared the same allele, we assigned that allele as ancestral and these sites were then used to populate the mutation matrix, following Chan et al. (2012). This approach ignores the possibility of both 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. when 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 to also include the possibility of ancestral allele misinference. At unresolved sites (i.e., if the outgroup alleles did not agree 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 analyzed a total of 44,835,801 SNPs in LDhelmet to construct recombination maps for each of 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 a conservatively estimated broad-scale recombination map. 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 $\sim$ 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 recombination 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

The recombination rate map inferred with a block penalty of 100 for *M. m. castaneus* was compared with two previously published genetic 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/). Hereafter, this map shall be referred to 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). Hereafter, this map shall be referred to as the Brunschwig map. Both the Brunschwig and Cox maps were constructed using far fewer markers than the present study, $\sim 500,000$ and $\sim 10,000$ SNPs, respectively and both maps 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 analyzed by Cox et al. (2009), the mean genome-wide ancestry attributable to *M. m. domesticus*, *M. m. musculus* and *M. m. castaneus* is 94.8%, 5.0% and 0.2%, respectively (data downloaded from the Mouse Phylogeny Viewer (Wang et al., 2012) http://msub.csbio.unc.edu). Values for

all classical strains, 60 of which were analyzed by Brunschwig et al. (2012), are similar (Yang et al., 2011).

Recombination rates in the Brunschwig map and our *castaneus* map were inferred in terms of the population recombination rate ($\rho = 4Ner$), units that are not directly convertible to centimorgans (cM), but were converted to cM/Mb for comparison purposes using frequency weighted means, as follows. Both LDhat and LDhelmet give estimates of ρ (per Kbp and bp, respectively) between pairs of adjacent SNPs. To account for differences in the physical distance between adjacent SNPs when calculating cumulative ρ, we used the number of bases between a pair of SNPs to weight that pairs contribution to the sum. By setting the total map distance for each chromosome to be equal to those found by Cox et al. (2009), we scaled the cumulative ρ at each analysed SNP position to cM values.

At the level of whole chromosomes, we compared mean recombination rates from the *castaneus* map with several previously published maps. The frequency-weighted mean recombination rates (in terms of ρ) for each of the chromosomes from the *castaneus* and Brunschwig maps were compared with the cM/Mb values obtained by Cox et al. (2009) as well as independent estimates of the per chromosome recombination rates from Jensen-Seaman et al. (2004). Pearson correlations were calculated for each comparison. Population structure in the inbred line data analyzed by Brunschwig et al. (2012) may have elevated LD, thus downwardly biasing estimates of ρ. To investigate this, we divided the frequency-weighted mean recombination rates per chromosome from the *castaneus* and Brunschwig maps by the rates given in Cox et al. (2009) to obtain estimates of effective population size.

At the Mbp scale, we compared variation in recombination rates across the autosomes in the different maps using windows. We calculated Pearson correlations between the frequency weighted-mean recombination rates (in cM/Mb) in non-overlapping windows for the *castaneus*, Cox and Brunschwig maps. The window size

considered may affect the correlation between maps, so we calculate Pearson correlations in windows of 1Mbp to 20Mbp in size. 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 in wild *M. m. castaneus*

To assess the distribution of fine-scale recombination rates in *M. m. castaneus* we used Gini coefficients and Lorenz curves. Applied to genetic maps, Gini coefficients and Lorenz curves have been used as a quantitative measure of the extent of heterogeneity of recombination rates in a genome (e.g. Kaur and Rockman 2014). Using our recombination maps generated using a block penalty of 10, we constructed 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 obtain the locations of potential recombination hotspots we adapted a script used by Singhal et al. (2015). We divided the genome into non-overlapping windows 2Kbp wide and, using the maps we generated using a block penalty of 10, classified all windows where the recombination rate was at least 5x greater than the recombination rate in the surrounding 80Kbp as potential hotspots. After identification, we merged all hotspots that were located directly next to one another.

To ask whether the fine-scale recombination rate variation in *M. m. castaneus* is like that reported for inbred lines, we compared the locations of putative hotspots in our data to the locations of DSBs reported by Smagulova et al. (2016). In their study, Smagulova et al. (2016) generated sequencing reads corresponding to the locations of DSBs in inbred strains of mice representing each of the principle M. musculus subspecies as well as *M. m. molossinus*, an inter-sub-specific hybrid of *M. m. castaneus* and *M. m. musculus*. Their reads were mapped to the mm10 genome so to compare the

locations of we converted the coordinates of DSBs to mm9 using the UCSC LiftOver tool (https://genome.ucsc.edu/cgi-bin/hgLiftOver), using the default parameters. The locations of putative hotspots present in our dataset were compared to the locations of DSBs in each of the strains studied by Smagulova et al. (2016) using BedTools v2.17.0 Quinlan and Hall (2010). To determine the amount of overlap between our list of hotspots and each of the lists of DSBs expected by chance, we approximated the null distribution of hotspot sharing using a randomization approach. For each of the inbred strains analyzed by Smagulova et al. (2016), we randomized the locations of our putative hotspots (using BedTools shuffle with the chrom option) and obtained the number of overlapping hotspots and DSB locations. For each comparison, this procedure was repeated 1000 times, per inbred strain, and the maximum number of null overlaps was taken as an approximate 0.1% significance threshold.

2.3.7 Examining the correlation between nucleotide diversity and recombination rate

There is evidence that natural selection is pervasive in the protein-coding genes and conserved non-coding elements in the murid genome (Halligan et al., 2010, 2011, 2013). Directional selection acting on selected sites within exons may reduce diversity at linked neutral sites through the processes of background selection and/or selective sweeps. These processes have the largest effect in regions of low recombination, and can therefore generate positive correlations between diversity and the recombination rate, as has been observed in multiple species (Cutter and Payseur, 2013). We used our *castaneus* map to examine the relationship between nucleotide diversity and recombination rates 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). We calculated the frequency-weighted mean recombination rate, from the block penalty 100 map, and the GC content for each gene. Using the approximate *castaneus* reference,

described above, and the outgroup alignment, we obtained the locations of 4-fold degenerate synonymous sites. 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, exhibiting a spurious excess of diverged sites, where ¿ 80% of sites were missing. We also excluded five genes that were diverged at all non-CpG prone 4-fold sites, as 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 local recombination rates in protein coding genes with nucleotide diversity and divergence. Variation in the mutation rate across the genome may influence genome-wide analyses of nucleotide polymorphism, so we also examined the correlation between the ratio of nucleotide diversity and divergence from *R. norvegicus* at neutral sites and the rate of recombination. We used non-parametric Kendall rank correlations for all comparisons.

All analyses were conducted using Python scripts, except 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. (2015).

2.4 Results

2.4.1 Phasing SNPs and estimating the switch error rate

In order to infer recombination rates from our sample of individuals, we required phased SNPs. Taking advantage of the high sequencing depth of the sample generated by Halligan et al. (2013), we phased SNPs using ShapeIt2, an approach that makes use of both LD and sequencing reads to resolve haplotypes. We phased each of the mouse autosomes, giving a total of 44,835,801 SNPs for estimation of recombination rates (Table S3).

By constructing pseudo-female individuals, we quantified the switch error rate incurred when inferring phase from our data. After filtering of variants, ShapeIt2 achieved low switch error rates for all parameter combinations tested (Table S1). We chose a set of filters (GQ ¿ 15, QUAL ¿ 30) that resulted in a mean switch error rates across the three pseudo-females of 0.46% (Table S1) and filtered out, on average, 44% of the available SNPs (Table S3). More stringent filtering resulted in slightly lower mean switch error rates, but also resulted in the removal of many more variants from the dataset (Table S1), thus reducing power to resolve recombination rates in downstream analyses.

2.4.2 Simulations to validate LDhelmet for the population sample of *M. m. castaneus*

We assessed the performance of LDhelmet when applied to our dataset by simulation. In the absence of switch errors, LDhelmet accurately infers the average recombination rate down to values of ρ/bp = 2 x 10^{-4} (Figure 1). Below this value, LDhelmet overestimated the scaled recombination rate for the simulated populations (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 ρ/bp was $<$ 2 x 10^{-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). Given that the simulations incorporated the mutation rate matrix (Table S2) and mutation rate ($\theta = 4N_e\mu$) estimated for *M. m. castaneus* we concluded that LDhelmet is applicable to the dataset of 10 *M. m. castaneus* individuals sequenced by Halligan et al. (2013).

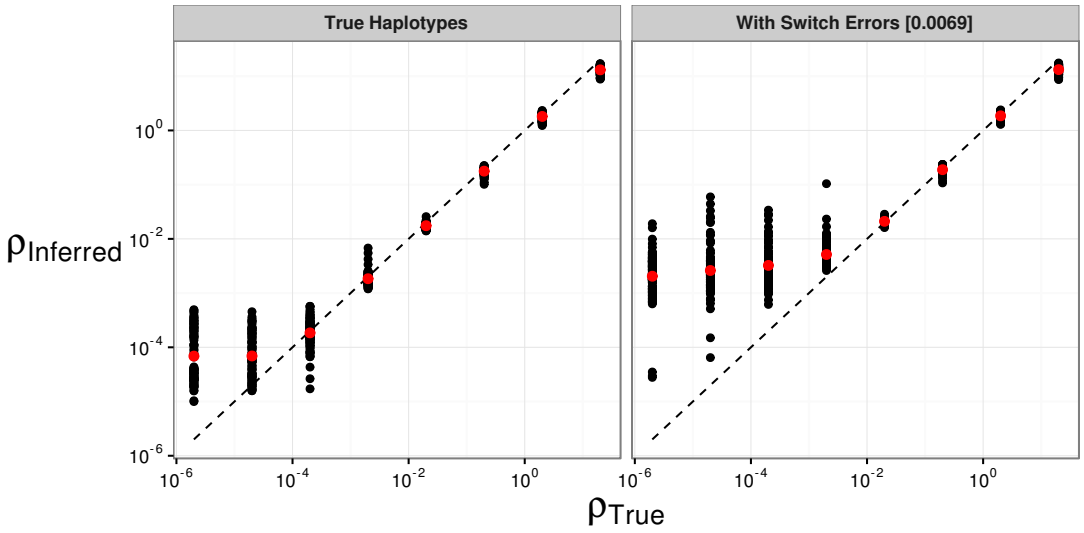

Figure 2.1: The effect of switch errors on the mean recombination rate inferred using LDhelmet with a block penalty of 100. Each black point represents results for a window of 4000 SNPs, with 200 SNPs overlapping between adjacent windows, using sequences simulated in SLiM for a constant value of ρ/bp. 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

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

A recombination rate map for each *M. m. castaneus* autosome was constructed using LDhelmet. We analyzed a total of 44,835,801 phased SNPs across the 19 mouse autosomes and the X-chromosome. From the map constructed using a block penalty of 100, the frequency weighted mean value of ρ/bp for all autosomes was 0.009. This value is greater than the lower detection limit suggested by both the simulations with and without switch errors (Figure 1). For the X-chromosome, the frequency-weighted mean rate was 0.0026, which is closer to the lower detection limit, but still above it (Figure 1). Because of this, the lower SNP density and smaller number of alleles used for inference, results for the X-chromosome may be more error-prone than for the autosomes.

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 for whole chromosomes provides

us with a baseline comparison for which we have an a priori expectation: 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 in mice and that the X-chromosome, which only undergoes recombination in females, should have the lowest rate. Both expectations have been met in previous studies of recombination in *M. musculus* (Jensen et al., 2007; **?**). Indeed, we find that the frequency-weighted mean recombination rates for chromosome 19 and the X-chromosome are the highest and lowest, respectively (Table 1). We also found that the frequency-weighted mean recombination rates for each of the chromosomes we analyzed were highly correlated with the direct estimates given in Jensen et al. (2007) (Pearson correlation = 0.59, $p = 0.005$) and Cox et al. (2009) (Pearson correlation = 0.68, $p = 0.001$), excluding the X-chromosomes does not substantially change the correlation results. These correlations suggest that our analysis captures real variation in recombination rates at the scale of whole chromosomes in the *M. m. castaneus* genome.

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

We compared the intra-chromosomal variation in recombination rates between our *castaneus* map and previously published maps. Figure 2 shows the variation in recombination rates across the largest and smallest autosomes in the mouse genome, chromosomes 1 and 19, respectively. It is clear that the *castaneus* and Cox maps are very similar (see also Figure S2 showing a comparison of all autosomes). Correlation coefficients between the maps are >0.8 for window sizes of 8Mbp and above (Figure 3), though the correlations are noisier when considering chromosomes separately (Figure S3). Although the broad-scale correlation between the *castaneus* and Cox maps is high (Figure 3), there were several regions of the genome that substantially differ, for

Table 2.1: 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).

Chromosome	Cox cM/Mb	Freq. Weighted Mean	N_e Estimate	Freq. Weighted Mean	N_e Estimate
1	0.50	0.0079	395,000	0.000015	745
2	0.57	0.0088	386,000	0.000015	653
3	0.52	0.0083	400,000	0.000014	693
4	0.56	0.0091	408,000	0.000020	889
5	0.59	0.0090	382,000	0.000015	646
6	0.53	0.0089	421,000	0.000015	728
7	0.58	0.0100	429,000	0.000019	801
8	0.58	0.0094	404,000	0.000014	610
9	0.61	0.0096	394,000	0.000018	749
10	0.61	0.0096	392,000	0.000023	928
11	0.70	0.0102	365,000	0.000019	689
12	0.53	0.0089	420,000	0.000019	897
13	0.56	0.0095	426,000	0.000014	629
14	0.53	0.0084	395,000	0.000013	632
15	0.56	0.0083	371,000	0.000024	1,080
16	0.59	0.0091	386,000	0.000017	721
17	0.65	0.0087	335,000	0.000052	2,020
18	0.66	0.0098	371,000	0.000021	785
19	0.94	0.0122	323,000	0.000026	681
X	0.48	0.0026	137,000	-	-
Mean	-	0.0092	-	0.000020	-

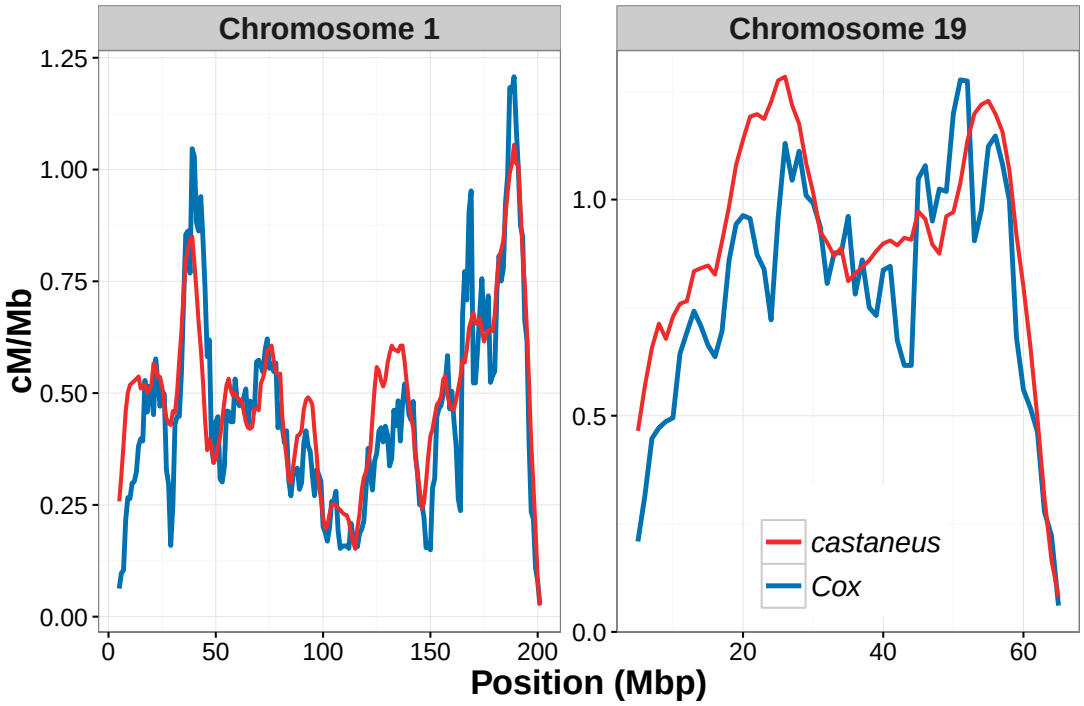

Figure 2.2: Comparison of sex-averaged recombination rates for chromosomes 1 and 19 of *M. m. castaneus* inferred by LDhelmet (red) with rates estimated in the pedigree-based study of Cox et al. (2009) (blue). Recombination rates were scaled to units of centimorgans per megabase for the *castaneus* map by setting the total map length of each chromosome to the corresponding map length of Cox et al. (2009).

example in the center of chromosome 9 (Figure S2). 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 using an LD-based approach with a sample of 60 inbred mouse strains and a relatively low SNP density. Population structure in the lines used by Brunschwig et al. (2012) or the sub-species from which they were derived would elevate LD, resulting in downwardly-biased chromosome-wide values of ρ. This is also reflected in the N_e values estimated from the frequency-weighted average recombination rates for each chromosome. The estimates of Ne are substantially different between the *castaneus* and Brunschwig maps, i.e. the *castaneus* estimates are consistently $\sim 500\times$ higher (Table 1). The estimates of Ne from the *castaneus* map are in broad agreement with the estimates of Ne based on polymorphism data (Geraldes et al. 2008; Geraldes et al. 2011). The lower SNP density used to construct the Brunschwig map would also likely result in a lower resolution recombination map.

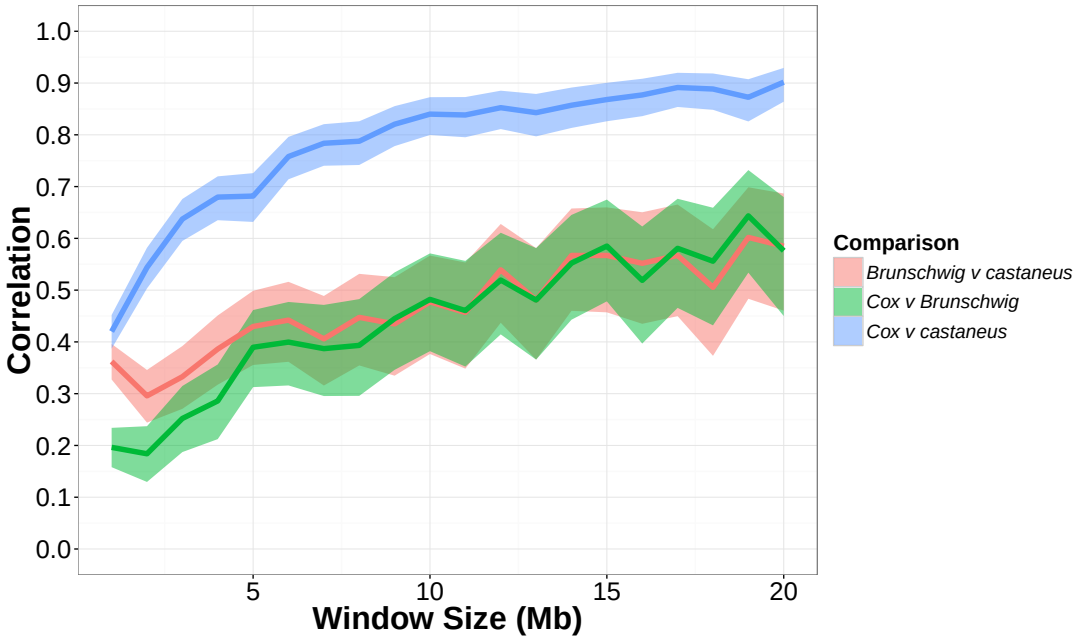

Figure 2.3: Pearson correlation coefficients between the recombination map inferred for *M. m. castaneus*, the Brunschwig et al. (2012) map and the Cox et al. (2009) map. Correlations were calculated in nonoverlapping windows of varying size across all autosomes. Confidence intervals (95%) are indicated by shading

2.4.5 Analysis of fine-scale recombination rates in wild *M. m. castaneus*

To locate potential recombination hotspots in wild *M. m. castaneus* we ran LDhelmet at a lower block penalty. As expected, the lower block penalty introduced more fine-scale variation into the recombination map; for example, see Figure S4. We used this fine-scale variation to locate 39,972 potential recombination hotspots in wild *M. m. castaneus* across the autosomes and X-chromosome. On average, there was 15 hotspots per Mbp across for all chromosomes tested. The total number of putative hotspots we identified is more than double the 15,061 DSB locations identified for CAST, a wild-derived strain representing *M. m. castaneus*, by Smagulova et al. (2016). In classical inbred lines, a total of 47,073 recombination hotspots were previously identified using a coalescent-based approach by Brunschwig et al. (2012), though they did not analyze the X-chromosome in their study.

To obtain a measure of the heterogeneity of recombination rates in the genome, we constructed Lorenz curves and calculated their Gini coefficients (Figure S5). In the context of a genetic map, Gini coefficients close to zero represent more uniform distributions of crossing-over rates and values closer to one indicates that recombination events are restricted to a small number of locations in a genome. Using the map constructed with a block penalty of 10, the mean Gini coefficient for across all autosomes was found to be 0.78. Our estimate is in line with that of Kaur and Rockman (2014), who reported a median Gini coefficient of 0.77 for chromosome 1 in inbred mice using a high-density map of crossing over locations observed in a crossing study (Paigen et al. 2008). The Lorenz curve for the X-chromosome was distinct from the autosomes (Figure X), however, with a Gini coefficient of 0.95, which is similar to the upper limit of the confidence interval around the estimate of Kaur and Rockman (2014).

We compared the locations of our potential recombination hotspots to the positions of DSBs reported by Smagulova et al. (2016). We found only a small overlap between the locations of potential recombination hotspots inferred for wild-caught mice and the locations of DSBs observed in the wild-derived inbred strains analyzed by Smagulova et al. (2016) (Table S4). The inbred strain CAST, representing M. m. castaneus, had the greatest amount of overlap, with 12.2% of DSB locations overlapping a putative hotspot and 4.1% after correcting for the number of overlaps expected seen by chance (Table S4). The second greatest overlap was with PWD, a strain that represents *M. m. musculus* (Table S4). All strains representing *M. m. domesticus* (13R, B6 and C3H) showed less than 1% overlap after correction. Note that our estimates of the null expectation are likely conservative, as false positives due to, for example, switch errors, present in our set of putative hotspots will inflate the probability of chance overlaps.

2.4.6 Correlations between recombination rate and properties of protein coding genes in *M. m. castaneus*

By examining the correlation between genetic diversity and recombination rate, we determined whether our map captures variation in N_e across the genome. We found that recombination rates at autosomal protein coding genes are significantly and positively correlated with levels of neutral genetic diversity, at all sites regardless of base context and at non-CpG-prone sites only (Table 2). Divergence from the rat at 4-fold sites was also significantly and positively correlated with recombination rate when analyzing all sites. However, for non-CpG-prone sites we found a small negative correlation (Table 2). There was also a significant and positive relationship between recombination rate and a genes GC content ($\tau = 0.125$, p < 2.2 x 10^{-16}). The correlation between recombination rate and neutral diversity divided 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.

	Correlation Coefficient	
	Non-CpG Prone Sites	All Sites
Nucleotide diversity (π)	0.090	0.20
Divergence from rat (d_{rat})	-0.038	0.062
Corrected diversity (π/d_{rat})	0.10	0.18

Table 2.2: Correlation coefficients between recombination rate and pairwise nucleotide diversity and divergence from the rat at fourfold degenerate sites for protein coding genes

2.5 Discussion

By constructing fine-scale maps of the recombination rate for M. m. castaneus, we have shown that there is a high degree of similarity between the recombination landscape for wild-caught mice and their laboratory counterparts, at relatively broad

scales. Our maps capture variation in the recombination rate, similar to that observed in a more traditional linkage map, at the level of both whole chromosomes and genomic windows of varying size. However, we found that a relatively small proportion of DSB locations identified in wild-derived strains by Smagulova et al. (2016) overlapped with the recombination hotspots we identified, suggesting that at the fine-scale recombination rates are highly variable between, and even within, sub-species. We discuss potential reasons for this below.

Recombination landscapes inferred using coalescent approaches, as in this study, reflect ancestral variation in recombination rates. We show that this ancestral variation is highly correlated with contemporaneous recombination rates in inbred mice representing *M. m. domesticus*, suggesting that the broad-scale variation in recombination rate has not evolved dramatically since the sub-species began to diverge, around 350,000 years ago (Geraldes et al. 2011). At a finer scale, however, we have shown that there is considerable variation in the locations of recombination hotspots within the *M. m. castaneus* sub-species. Our findings reflect results in hominids and the great-apes, which suggest that, although the locations of recombination hotspots are strongly diverged between species, broad-scale patterns of recombination rate are relatively conserved (Lesecque et al. 2014; Stevison et al. 2015). However, there do seem to be multiple relatively large regions of the genome that distinguish *M. m. castaneus* and *M. m. domesticus*. For example, we observe peaks in recombination rate for *M. m. castaneus* on chromosomes 4, 5, 14 and 15 that are not present in the Cox map (Figure S2). Since present-day populations of *M. m. domesticus* exhibit karyotype variation (Gimenez et al. 2017), is seems plausible that chromosomal translocations or fusions in ancestral populations may have affected our rate estimates. The application of traditional mapping approaches to *M. m. castaneus* individuals could potentially help elucidate this.

The correlation between the *castaneus* and Cox maps for the X-chromosome seems to be weaker than for autosomes of similar physical length (e.g Chromosomes 2 and

3) (Figure), perhaps suggesting that the genetic map of the X-chromosome evolves faster than the autosomes. However, the X-chromosome has substantially reduced SNP density (Table S3) and recombination rates were inferred using 17 alleles rather than the 20 used for each of the autosomes making comparisons between these correlations slightly problematic. Nevertheless, these results are potentially consistent with those of Dumont et al. (2011), who found that there are significant differences in genetic length between *M. m. castaneus* and *M. m. musculus* (when crossed to *M. m. domesticus*) in multiple regions of the genome, particularly on the X-chromosome.

A recent study by Stevison et al. (2015) reported that correlations between recombination rates declined with genetic divergence between great ape species. For example, between European humans and gorillas, genetic divergence is $\sim 1.4\%$, while the Spearman-rank correlation between their respective recombination maps, at the 1Mbp scale, is $\sim$ 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, at the 1Mbp scale (Note, Pearson correlations are shown in Figure 3). This perhaps suggests that recombination rate differences have accumulated faster between *M. m. castaneus* and *M. m. domesticus* than it has between great apes. However, it should be noted that the comparisons performed by Stevison et al. (2015) were between recombination maps constructed with similar volumes of data for each species, using identical methods, which is not the case for the comparison we make between our maps and those 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 sub-species of house mice, as well as sister species, using LD-based methods would help elucidate the time-scale of recombination rate evolution in wild mice.

We investigated how the landscape of fine-scale recombination rates inferred for wild *M. m. castaneus* compares to that of wild-derived laboratory mice. There was only a small amount of overlap between the locations of DSBs in wild-derived strains

and our lists of putative hotspots. The greatest overlap was with inbred strains derived from *M. m. castaneus* (Table S4). We found that 12% (or 4% above null expectation) of DSB locations reported for CAST, by Smagulova et al. (2016), overlapped with a hotspot we inferred for *M. m. castaneus*. Such a low proportion is striking, suggesting that even within house mouse sub-species there is a great variation in the locations of recombination hotspots. Though, this is perhaps this is not surprising when considering that PRDM9 determines the locations of the vast majority of recombination hotspots in mice (Brick et al. 2012) and that even a single amino acid change to the zinc-finger array of that protein can result in dramatic shifts in the landscape of DSBs (Smagulova et al. 2016). Furthermore, in wild M. musculus there is a large diversity of PRDM9 alleles in each of the different sub-species (Kono et al. 2014) so the locations of DSBs in the CAST strain, observed by Smagulova et al. (2016), may represent only a small proportion of the diversity of hotspot locations in wild mice. Despite the small overlap, the similarity of the mean Gini coefficient for our map and the estimate for M. musculus given by Kaur and Rockman (2014), suggests that the distributions of recombination rates in wild mice and inbred lines are similarly heterogeneous. Interestingly, Smagulova et al. (2011), showed that there is a high correlation between a genetic map constructed using DSBs mapped in inbred mice, using the same approach as Smagulova et al. (2016), and the Cox map. We have shown that our *castaneus* map is highly correlated to the Cox map despite little overlap between the locations of DSBs in domesticus-derived strains the locations of hotspots are highly different between our study and DSB maps for different sub-species. These results perhaps suggest that the binding motifs of the different PRDM9 alleles in the sub-species have been in broadly similar genomic regions, resulting in recombination rates evolving rapidly at finer-scales, but more slowly at broader scales. An analysis of recombination rates in sister species of mice, or other murid rodents, would be useful in understanding the causes of rate variation in this system.

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), giving us more power to capture variation in the recombination rate. 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. By dividing the mean estimated /bp values (inferred using LDhelmet) for each chromosome by the corresponding recombination rate estimated from crosses (Cox et al. 2009), we showed that Ne estimates from the Brunschwig map are much lower than estimates based on our map (Table 1). This is consistent with the presence of elevated LD between the SNPs in the inbred lines analyzed by Brunschwig et al. (2012). It should be noted, however, that the estimates of Ne will be biased, as = 4Ne is a parameter in both LDhat and LDhelmet. In spite of this potential bias, the differences in Ne estimated from the Brunschwig and *castaneus* maps shown in Table 1 are striking, given that the effective population sizes of *M. m. domesticus* and *M. m. castaneus* are expected to be 150,000 and 350,000, respectively (Geraldes et al. 2008). The Brunschwig map does, however, capture true variation in recombination rates, because 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 inferred 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, variation in the rate and locations of hotspots were still accurately detected in their study.

We obtained an estimate of the switch error rate, taking advantage of the hemizygous sex chromosomes of males present in our sample. This allowed us to assess the extent by which switch errors affected our ability to infer recombination rates in

M. m. castaneus. It should be noted, however, that our inferred switch error rate may not fully represent that of the autosomes. This is because multiple factors influence the ability to phase variants using ShapeIt2 (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. As the sex-averaged recombination rate for the X-chromosome is expected to be $\frac{3}{4}$ that of the autosomes, it likely has elevated LD, and thus there will be higher power to infer phase. In contrast, the level of X-linked nucleotide diversity in *M. m. castaneus* is approximately one half that of the autosomes (Kousathanas et al. 2014), and thus there would be a higher probability of phase informative reads on the autosomes. While it is difficult to assess whether the switch error rates we estimated from the X-chromosome analysis will be the same as on the autosomes, the analysis allowed us to explore the effects of different SNP filters on the error rate.

By simulating the effect of switch errors on estimates of the recombination rate, we inferred the range over which /bp is accurately estimated in our data. 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 introduced switch errors at random into the simulation data and estimates of /bp obtained from these datasets reflect this detection threshold; below 2x10-3 /bp, we found that LDhelmet consistently overestimates the recombination rate in the presence of switch errors (Figure 1; Figure S1). This highlights a possible source of bias affecting LD-based recombination mapping studies using inferred haplotypes, suggesting that error in phase inference needs to be carefully considered before attempting to estimate recombination rates and/or recombination hotspots using LD-based approaches.

Consistent with studies in a variety of organisms, 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 recombination (Table 2). Cai et al. (2009) found evidence suggesting that recombination may be mutagenic, though insufficient to account for the correlations they observed between recombination and diversity. The Kendall correlation between /drat and recombination rate of 0.20 for all 4-fold sites, a value that 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 inferred for regions of the genome where the true 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 /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 Ne across the genome. This indicates that a recombination mediated process influences levels of genetic diversity. Previously, Halligan et al. (2013) showed that there are troughs 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 genome of *M. m. castaneus*. In contrast, a previous study by Geraldes et al. (2011) examining the correlation between levels of polymorphism and recombination rate in wild mice found that *M. m. musculus* exhibited a significant correlation between diversity and recombination while for both *M. m. castaneus* and *M. m. domesticus* the relationship was non-significant. Using genome-wide data, we found a fairly weak, but significant, positive correlation for *M. m. castaneus* so perhaps the Geraldes et al. (2011) study was underpowered as it only analyzed 27 autosomal loci. However, it should be noted that both the measure of recombination rate we used and neutral genetic diversity are compounded with effective population size, so part of the positive correlation we detected could be driven by random fluctuation in Ne across the genome.

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 been demonstrated previously that the turnover of hotspots has led to rapid evolution of fine-scale rates of recombination in the *M. musculus* sub-species complex (Smagulova et al., 2016) and our results suggest that even within sub-species, hotspot locations have diverged. On a broad scale, however, our results suggest that the recombination landscape is very strongly conserved between, at lease, *M. m. castaneus* and *M. m. domesticus*. In addition, our estimate of the switch-error rate implies that phasing errors leads to upwardly biased estimates of the recombination rate when the true recombination 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).

To further our understanding of the evolution of the rate of recombination in the house mouse we need to directly compare sub-species. The comparison of our results and previously published maps indicates that there is broad-scale agreement in recombination rates between *M. m. castaneus* and *M. m. domesticus*. In this study, we have assumed that inbred lines derived from *M. m. domesticus* reflect natural variation in recombination rates in that sub-species, though this is not necessarily the case. Furthermore, previous studies have shown that recombination rates in *M. m. musculus* are perhaps the most distinct of the sub-species: The overall rate of crossing-over is higher in *M. m. musculus* males is higher than in the other sub-species (Dumont and Payseur, 2011) and there is also evidence of recombination rate modifiers of large effect segregating within *M. m. musculus* (Dumont et al., 2011). Despite these predictions, the hotspots we detected in our study and those of Smagulova et al. (2016) show more overlap with *M. m. musculus* than with *M. m. domesticus*. Samples of natural

populations, like the one studied here, could be used to more clearly elucidate the variation in recombination rate landscape specific to the different sub-species. A broad survey of this kind would most efficiently be generated using LD-based approaches.

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 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 to 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, however, can explain the troughs in diversity around CNEs. This is not the case for protein-coding exons, where observed dips in diversity cannot be explained by parameter estimates obtained from the uSFS. We discuss the extent to which our results provide evidence of sweeps playing a role in shaping patterns of nucleotide diversity and the limitations of using the uSFS for obtaining inferences of the frequency and effects of advantageous mutations.

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 levels of 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; Charlesworth 2012; 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, because the proportion of neutral amino acid substitutions that occur 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), the results of the Sattath test may be difficult to interpret in species with genomes that exhibit 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.

There are a number of methods that estimate the frequency and strength of advantageous mutations from models of the effects 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. However, the estimate of the deleterious per site mutation rate they obtained far exceeded published values of the point mutation rate in D. melanogaster. They, reasonably, attributed this discrepancy to the effects of selection at linked sites in addition to those they had explicitly modelled. The selection parameters estimated by Elyashiv, et al. (2016) were inferred from nucleotide

diversity only. There is information in the distribution of allele frequencies, the site frequency spectrum (SFS), however, 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 the effects of 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 (Ne) 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. However, Halligan, et al. (2013) found that troughs in diversity around recent nonsynonymous and synonymous substitutions in M. m. castaneus were similar. Taken at face value, this could be taken as evidence that SSWs are infrequent, but, in addition, Halligan, et al. (2013) found that there are also troughs in diversity around randomly chosen synonymous or nonsynonymous sites that are similar to those observed around substitutions. These results, therefore, suggest that selection at linked sites affects nucleotide diversity across large portions of the genome, making the analysis of patterns of diversity around substitutions difficult to interpret. Our understanding of the forces that have shaped patterns of diversity in the house mouse and mammals in general is, thus, somewhat unclear.

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 parameterise 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. (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 sequencing 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. (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. 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 the distance to 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 (pa) and the strength of selection acting on them (Nesa).

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.

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. (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 2statistic, but the numbers of sites in the ith and n-ith 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

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 Nes (where s is the difference in fitness between homozygotes for ancestral and derived alleles, assuming semi-dominance). For a population where Ne = 1,000 and s = 0.05, for example, the strength of selection is therefore approximately equivalent to that of a population where Ne = 10,000 and s = 0.005. By scaling parameter values according to the population size of the simulations (Nsim), we modelled the much larger M. m. castaneus population (Ne 500,000 (Geraldes, et al. 2011) in a computationally tractable way.

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, =4Ne, to set the mutation rate () 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 = 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 0.0083/4Nsim.

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 =4Ner. 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 Nsim 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 Nsim= 1,000 diploid individuals were allowed to evolve for 10,000 generations (10Nsim) 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 (/drat) 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.

where sim(i) and obs(i) are the diversity values from simulations and M. m. castaneus, respectively, in window i around a particular class of functional element and nw 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 (Krat) is 15% at non-CpG-prone sites far from protein-coding exons. Under neutrality, divergence is expected to be equal to 2T, where T is the time in generations since the two-species shared a common ancestor and 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 4Ne) and since Krat = 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 10Nsim 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. (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 (Kfam) is 3.4%. In the simulations, given that the mutation rate was 2.075x10-6, 8,193 generations are sufficient to approximate the M. m. castaneus lineage since its split with M. famulus Kfam. 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

To investigate genetic variation around functional elements in house mice, we analysed the genomes of 10 wild-caught individuals that had been 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 1). This is the expected pattern 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 1), further indicating the action of purifying selection in those classes of sites. 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 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).

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 outgroup species (Keightley, et al. 2016). This method attempts to correct for biases that are inherent in parsimony methods.

A populations 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).

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 S1). For example, 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 1).

3.4.2 Estimating the frequencies and strengths of deleterious and advantageous mutations

We inferred 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 2). The first, as described by Schneider, et al. (2011), makes use of 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.

We performed a comparison of different DFE models, including discrete distributions that have 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 S3). 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 S4). This presumably reflects a lack of power. These best-fitting models were identified whether we estimated the DFE under Model A or Model B. Parameter estimates pertaining to the dDFE were also similar between Models A and B (Table 2).

In our current 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; Halligan, et al. 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 ¡ —Nes— ¡ 100 (Table 2). This contrasts with previous studies using the folded SFS which found an appreciable proportion of mutations in the 1 ¡ —Nes— ¡ 100 range (Halligan, et al. 2013; Kousathanas and Keightley 2013). Because this difference may have an effect on the reductions in diversity caused by background selection, 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 (results below).

For all classes of functional sites, we inferred that moderately positively selected mutations are fairly frequent under both Models A and B (Table 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 average scaled selection strengths of advantageous mutations were fairly similar (Table 2), i.e. Nes 8, implying that s is on the order of 10-5 (assuming Ne = 500,000; (Geraldes, et al. 2011)). We found that estimates of the frequency of advantageous mutations (pa) obtained under Model B for 0-fold sites and UTRs were 3 times higher than those obtained under Model A. Confidence intervals overlapped, however (Table 2). 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 very similar (Table 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.

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

Using the selection parameter estimates obtained from DFE-alpha (Table 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 () at all positions by the mean at distances greater than 75Kbp and 4Kbp away from exons and CNEs, respectively. These distances were chosen as they are the approximate values beyond which remains constant.

Simulations incorporating only deleterious mutations predicted a chromosome-wide reduction in genetic diversity. Around exons and CNEs, diversity plateaued at values that were 94% of the neutral expectation (Figures S2-3). However, simulations involving only BGS did not fully predict the observed troughs in diversity around functional elements. 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 2-3). Similar predictions were obtained for Models A or B (Figures 2-3) or for the gamma dDFEs inferred by Halligan et al. (Halligan, et al. 2013) (Figure S4). Our simulations incorporating deleterious mutations suggest, then, that while BGS affects overall genetic diversity across large portions of the genome, but positive selection presumably also makes a substantial contribution to the dips in diversity around functional elements.

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 2; Table 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 shallower and narrower than the one generated by deleterious mutations alone (Figure 2; Table 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 2; Table 3). In simulations that incorporated both advantageous and deleterious mutations, the troughs in diversity around exons were not as large as those observed in M. m. castaneus (Figure 2; Table 3). However, 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 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 2).

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, can explain patterns of diversity observed in M. m. castaneus (Figure 3; Table 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 similar to the ones generated by deleterious mutations alone (Figure 3; Table 3). Although both processes are required to explain the patterns observed in mice, our simulations suggest that BGS makes a bigger contribution to the overall reduction in neutral diversity than SSWs (Figure S3). The troughs in diversity around CNEs in our simulations were slightly shallower than those observed in the mouse genome (Figure 3), perhaps suggesting that we failed to detect infrequent, strongly selected advantageous mutations in CNEs or that we slightly underestimated the true frequency of advantageous mutations occurring in those elements.

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. 1995; Kim 2006). SSWs and BGS generate troughs in diversity at linked sites (Figures 2-3), 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 4). This likely reflects a genome-wide process, such as population size change, that we have not modelled.

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 4). However, under Model B, the simulations exhibit a substantial dip in Tajimas D, which is not observed in the real data (Figure 4).

In the case of CNEs, we observed a trough in Tajimas D in the real data (Figure 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 2). In the case of 0-fold sites, for example, pa was 3x higher under Model B than Model A (Table 2). We then 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 (dsel/dneut) between the simulated and observed data. In simulations that assumed the estimates of selection parameters obtained under Model A, dsel/dneut values were similar to those observed in M. m. castaneus for all classes of selected sites (Table 4). Under Model B, however, the simulations predicted substantially more substitutions at nonsynonymous sites and UTRs than were seen in the real data (Table 4). This suggests that, under Model B, pa for 0-fold sites and UTRs may be overestimated.

iv) Re-estimating the DFE from simulated data

BGS and SSWs both perturb allele frequencies at linked neutral sites, and this 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 S2). 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. Simulations assumed the selection parameters obtained under either Model A or B, and in each case, the 3-epoch model gave the best fit to the data. The estimated demographic parameters inferred were somewhat different between simulations assuming Model A or Model B selection parameters, but in each case a population bottleneck followed by an expansion was inferred (Table S5). This is an interesting observation, since our simulations assumed a constant population size, but selection at linked sites appears to distort the neutral site uSFS, and a demographic history is estimated as the one inferred from the real data (Table S5).

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 S5). 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 (pa) was overestimated (Table S5). A possible explanation for this is that the demographic correction we applied to the uSFS for selected sites (see Supplementary Methods) may not fully capture the effects of selection at linked sites. SSWs increase the proportions of high frequency derived alleles (Kim 2006), and it is

possible that their contribution to the uSFS for selected sites was partially unaccounted for, creating the appearance of more frequent advantageous mutations in the uSFS.

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. One method that has been suggested as a means of teasing the two processes apart takes 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.

In our simulations, we measured diversity around substitutions occurring on a time-scale that is 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 the selection parameters estimated under Model A or Model B (Figure 5). However, the troughs in diversity around substitutions were deeper in the simulations assuming Model B (Figure 5), reflecting the higher frequency of advantageous mutations (Table 2). In

the immediate vicinity of nonsynonymous substitutions, diversity was lower than the corresponding value for synonymous substitutions (Figure 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 that can 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.

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 background selection (BGS) and/or SSWs (Halligan, et al. 2013). In this paper, we analysed the genome 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 parameterise 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 exhibit reduced nucleotide diversity, reduced nucleotide divergence and an excess of low frequency variants (Table 1; Figure 1), consistent with the action of natural selection (Halligan, et al. 2010; Halligan, et al. 2013). The estimates of the DFEs included substantial proportions of strongly deleterious mutations (Table 2). In addition, the best-fitting models also included a single class of advantageous mutations. Additional classes were not statistically supported, however. 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 best-fitting DFEs fit the data very well (Figure S5), 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) used simulations to show that selection parameters can be accurately estimated from the uSFS, whilst ignoring between-species divergence, if pa is sufficiently high. In our analysis of 0-fold sites and UTRs, Model B gave a significantly better fit and higher estimates of the frequency of advantageous mutations (pa) than Model A (Table 2). For CNEs, however, Models A and B did not significantly differ in fit, and the selection parameter estimates were very similar (Table 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 this 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 in the time since the accumulation of

between-species divergence 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. In addition, we that found that the selection parameters obtained fail to explain the patterns of diversity around protein-coding exons, whereas they explain the patterns of diversity around CNEs, so we think the latter explanation 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 patterns of diversity around functional elements. We found that BGS does not fully explain the troughs in diversity observed around either protein-coding exons or CNEs (Figures 2-3). These results are consistent with Halligan, et al. (2013).

Our simulations suggest that BGS and SSWs both produce genome-wide reductions in neutral diversity (Figures S3-4), 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 2-3). Around protein-coding exons, the combined effects of advantageous and deleterious mutations generated a shallower trough in diversity than the one observed (Figure 2). 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 (Figure 3).

There is an overall excess of rare variants in M. m. castaneus relative to neutral

expectation, as indicated by a strongly negative Tajimas D at putatively neutral sites (Table 1) and in the regions surrounding exons and CNEs (Figure 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 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 4). Secondly, we observed a pronounced dip in Tajimas D around simulated exons, which is not present in the real data (Figure 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 rate and frequency of advantageous mutations based on the uSFS relies on the presence of advantageous 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 pa may be overestimated when analysing the uSFS, if the true value of pa is low.

Advantageous mutations with large effects have shorter sojourn times than those with milder effects (Fisher 1930; Kimura and Ohta 1969). If strongly selected advantageous mutations are infrequent, it is therefore unlikely that they would be observed to be segregating. This could explain why the estimated selection parameters fail to predict the deep troughs in diversity around exons that we observe in the real data (Figure 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), utilised 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) = 4.5 x 10^{-3} and the scaled strength of selection ($N_e s_a$) = 11.5 for 0-fold nonsynonymous sites. The second study, by Campos, et al. (2017), estimated pa = 2.2 x 10^{-4} and $N_e s_a$ = 241, based on the correlation between synonymous site diversity and nonsynonymous site divergence. Although the individual parameter estimates differ substantially, the compound parameter Nesapa (which approximates the rate of SSWs) was similar between the studies (0.055 and 0.052 for Campos, et al. (2017) and Keightley, et al. (2016) 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 mutations. 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. 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.

The patterns of diversity and Tajimas D around CNEs and the nucleotide divergence within CNEs in our simulated populations were similar to those observed in the M. m. castaneus data, regardless of which estimate of the DFE we used (i.e. Model A or B) (Figure 3-4; Table 3). This suggests that the four classes of mutational effects inferred provide a reasonable approximation for the full distribution of fitness effects for CNEs.

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. Nesa ¿ 1) then the rate of change of fitness due to 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 2). Given that there are more CNE nucleotides in the genome than there are 0-fold sites (Table 1), this could imply that adaptation at regulatory sites causes the greatest fitness change in mice. However, we have argued that protein-coding genes may be subject to strongly selected advantageous mutations, which were undetectable by analysis of the uSFS. If this were the case,

adaptation in protein-coding genes could make a larger contribution to fitness change than 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 (2017) 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 (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; Williamson, et al. 2014) (Table 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 modes of sweeps that occur, since the relative probabilities of a hard or 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 likely to be 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 S2). In our simulations, however, we showed that the inferred demographic history may largely be an artefact of selection at linked sites (Table S5). 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 4). This reduction is presumably caused by a demographic history or strong selection that was not included in our simulations. Less biased estimates of the demographic history of M. m. castaneus may be obtained from regions of the genome experiencing high recombination rates, located far from functional elements. Finally, mouse populations may rapidly oscillate in size (e.g. seasonally (Pennycuik, et al. 1986)). 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 very rapidly between and within M. musculus subspecies (Smagulova, et al. 2016). Assuming a single suite of recombination hotspots in simulations may produce misleading results if hotspot locations evolve faster than the rate of neutral coalescence. Recombination hotspots are an important feature of the

recombination landscape in mice and thus potential 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 can explain the patterns of diversity around CNEs, but not around protein-coding exons. We also argue that mutations with moderately advantageous effects frequently occur at 0-fold and UTR sites, but that undetectable, strongly advantageous mutations may occur in both these classes of sites. 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 using 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.

Chapter 4

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

4.1 Introduction

In the past 30 years of population genetic research it has become clear that natural selection shapes patterns of nucleotide diversity across the genomes of many species (Corbett-Detig et al., 2015; Cutter and Payseur, 2013). Because genetically linked sites do not evolve independently, selection acting at one site may have consequences for another. The consequences of selection at linked sites are intrinsically linked to the frequency and strength of selected mutations as well as, crucially, the rate of recombination (REF DUMP). Two main modes of selection at linked sites have been identified; selective sweeps caused by the spread of advantageous mutations and

background selection caused by the removal of deleterious variants. The two processes are related and can both potentially explain the positive correlations between nucleotide diversity and recombination rate reported in many species (Cutter and Payseur, 2013). However, the proportion of nonsynyonomous substitutions attributable to adaptive evolution (α) is typically high (50%) (Galtier 2016; but see Booker et al. 2017a for caveats), suggesting that selective sweeps may play a substantial role in shaping nucleotide diversity across the genomes of many species.

Selective sweeps have been subject to rigorous population genetic research (Maynard Smith and Haigh, 1974; Coop and Ralph, 2012; Hermisson and Pennings, 2005; Barton, 2000). The classic footprint of a selective sweep is a trough in nucleotide diversity at neutral sites surrounding substitutions. Reductions in nucleotide diversity caused by selective sweeps are related to the strength of selection acting on advantageous mutations as well as the frequency with which they arise. Taking advantage of this, Wiehe and Stephan (1993) used a model of selective sweeps to estimate the frequency and strength of advantageous mutations in *Drosophila melanogaster* 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 ($2N_e s \approx 5{,}000$) 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 parameters we estiamtes could explain troughs in diversity observed around protein-coding exons or conserved non-coding elements (CNEs) Using simulations, we showed that the parameters of the DFE we obtained were unable to explain the troughs in diversity around protein-coding exons, but were able to explain the troughs around conserved non-coding elements. A possible explanation for 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. occur in protein-coding regions have, on average, larger effects on fitness that those occurring in regulatory regions which may affect the power to

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.

4.2 Materials and Methods

4.2.1 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 $\rho = 4N_e r$, 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.

4.2.2 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.

4.2.3 Model of Recurrent Sweeps with Background Selection

? gave expressions for the expected neutral diversity given 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 rate of sweeps 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 fraction of new mutations occurring within a focal element that are advantageous and γ_a is the scaled selection coefficient of a new mutation ($2N_e s_a$) (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 4.5 we use $g_c = \kappa r_c$ where κ is the ratio of the crossing over rate to the gene conversion rate.

There is evidence that the distribution of fitness effects for advantageous mutations is beneficial from both theoretical Hernandez and Uricchio (2015) Griffiths REFS) and empirical studies (DATA Papers?). IT is straightforward to incorporate an exponential distribution of advantageous mutation effects to Equation 3

$$P_{sc,j} \approx \int_0^\infty f_x(\gamma) V_a \tau \gamma_a^{\frac{-4r_{i,j}}{s}} \, d\gamma \tag{4.4}$$

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).

4.2.4 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.

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 (discussed below), we perform all analyses in parallel, assuming both of these recombination maps.

Recombination proceeds via crossing-over or gene conversion, but the above formulae (Equations 4.2 and 4.4) assume that genetic distance is solely a product of the local crossing-over rate and the physical distance. We incorporated gene conversion into genetic distance by calculating $r_{i,j}$ in Equations 4.2 and 4.4 using Equation 1 from Thornton (2014)

$$r_{i,j} = d_{i,j} r_c + g_c d_g \left(1 - e^{-\frac{d_{i,j}}{d_g}} \right) \tag{4.5}$$

where $d_{i,j}$ is the physical distance between a focal neutral site and a selected site, r_c is the rate of recombination by crossing-over, g_c is the rate of non-crossing over gene conversion and d_g is the mean length of a gene conversion tract. This assumes that the distribution of gene conversion tract lengths is exponential. We assumed a mean tract

length of 144bp and that the gene conversion rate was 10.5% of the local crossing-over rate (Paigen *et al.* 2008).

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

4.2.5 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 fit Loess curves using a span of 0.2 and used the number of sites contributing to each analysis bin as weights.

4.3 Results

4.3.1 Estimating selection parameters from the uSFS of simulated data

Parameters of the DFE can be estimated directly from unfolded site frequency spectra (uSFS) if selected mutations are segregating in populations of interest (REFS). It has been repeatedly demonstrated that parameters of the DFE for deleterious mutations (dDFE) can be accurately estimated from population genetic data. It has also been shown that the parameters of advantageous mutations can also be estimated from the uSFS (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.

4.3.2 Patterns of genetic diversity around protein-coding exons and conserved non-coding elements

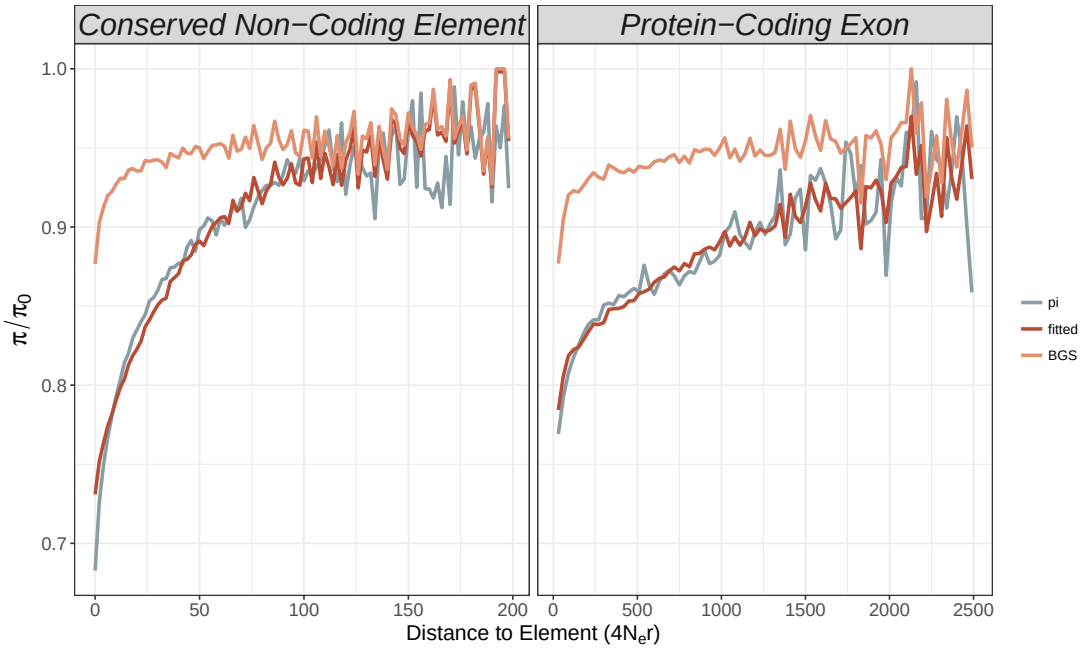

Figure 4.1:

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 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 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.

4.3.3 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?). 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 B.3. 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

4.3.4 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 4.5 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.

4.4 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.

4.4.1 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.

4.4.2 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

4.4.3 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:

$$\Delta W \propto \mu p_a n_a E(s_a)^2, \tag{4.6}$$

We parametrized Equation 4.6 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 4.6.

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.

4.5 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.

4.6 Acknowledgements

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

Chapter 5

General Discussion

The work I have done for this thesis does not, by any means, close the book on our understanding of selection in the house mouse, however, there are a number of useful insights that can be gleaned form my analyses. In this discussion I will start this describing the main insights to be take from my analyses. I will go on to describe several avenues for further research, suggested by my results.

I have assumed that all CNEs share a single DFE which is a first approximation, but is unrealistic. The method used by Halligan et al (2013) to identify the CNEs I analysed throughout my thesis successfully identifies known regulatory features in the genome. There are multiple roles played by the CNEs identified using phastCons, or similar approaches. These include the regulation of gene expression and the control of alternate splicing. It seems reasonable to expect that these two classes of elements, which have very different biochemical roles, will be subject to different DFEs. Dividing up CNEs that have been identified using phylogenetic methods into different functional categories represents a challenge. Data of the kind generated by ENCODE could be

used to assign biochemical function to CNEs identified using phylogenetic approaches, i.e. one could take the intersection of ENCODE data and phastCons elements.

5.1 Soft and partial selective sweeps in mice

The work in this thesis has relied on the assumption of hard selective sweeps. Both the DFE-alpha and polyDFE analyses used in Chapters 3 and 4 assume that there selected mutations possess a single selection, unchangeable selection coefficient. However, it could be the case that when the environment changes, alleles that were once neutral become selectively advantageous or disadvantageous. Alleles that become advantageous in this manner and subsequently fix generate soft sweeps. The reductions in diversity between soft and hard sweeps produce distinct patterns of genetic diversity (Hermisson and Pennings?). A major difference between soft sweeps and hard sweeps is in the distribution of haplotypes in the population after fixation. In the case of soft sweeps, the allele, when segregating neutrally, may be present on mut

Schrider and Kern used a machine learning approach to classify regions of the human genome as having experinced a hard or soft sweep or being linked to a hard or soft sweep. They found that the

5.2 The interaction between natural selection and demographic history

This thesis has focussed on the effects of BGS and SSWs and has assumed, expect where explicitly modelled, that the demographic history of *M. m. castaneus* has not influenced our analyses. This may influence the results of both Chapters 3 and 4 as the effects of, at least, BGS can become amplified under population size change (Hernandez paper). 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 (Table REF). It is tempting to interpret these results in light of recent human history; since mice are commensal to humans their population numbers have likely exploded in the recent past (REF?). However, as we also showed in Chapter 3, selection at linked sites caused by DFE parameters estimated from the data cause one to infer a population expansion even there is not one. *M. m. castaneus* may very well have undergone a rapid popualtion expansion in the recent past, but it is likely that the demographic parameters we inferred are highly affected by selection at linked sites.

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

Since demographic models assume neutrality, it is necessary 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, that experience high recombination rates. Regions free from conserved elements (both coding and non-coding) may be free from the effects of selection at sites, especially if the are highly

recombining. One could go a step further and fit a model of selection at linked sites to genome-wide polymorphism data (e.g. using the methods of Elyashiv or Hammer) and identify regions that have only experience minuscule 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, there is the possibility that there are BGS and SSW effects present even when there are no annotations present. An alternative strategy would be to use the machine learning methods of Schrider and Kern (2016). S/HIC builds a classifier that can discriminate between neutrally evolving sequneces and sequnqce that are influencesd by selection at linked sites by 'learning' from extensive simulations, what neutrally evolving sequences look like and identify them from a combination of numerous summary statistics. Parametrising demographic models from the neutrally evolving sequences identified using the

Neutralome, Elyashiv, S/HIC, find regions of the genome that appear to be evolving neutrally and use them to parametrise demographic models. Are demographic models too simplistic?

5.3 Making use of more of the available data

In Chapters 3 and 4 I used methods to estimate the distribution of fitness effects by analysis of the uSFS. Both methods I used, DFE-alpha and polyDFE, make use of a putatively neutral class of sites to account for distortions in the uSFS away from neutral expectation caused by processes other than the direct effects of selection. In the case of DFE-alpha, an explicit demographic model is fit to the neutral uSFS, while polyDFE uses the neutral sites' uSFS to estimate a set of nuisance parameters. However, processes other than population size change can distort the uSFS, such as selection at linked sites, so by not explicitly modelling selection at linked sites, both the demographic correction

applied in DFE-alpha analyses and the nuisance parameters in polyDFE essentially throw away information.

A substantial hurdle to population genomic research is in making use of all the available data. For example, in Chapters 3 and 4 we have analysed either the site frequency spectrum or nucleotide diversity. These are just two data summaries that be analysed in a population genetic model. As we demonstrated in Chapter 4, the SFS is a useful summary of the data that can be used to very accurately estimate the dDFE, but the uSFS is pretty poor for estimating the parameters of strongly selected advantageous mutations particularly if they are rare. In such cases, patterns of genetic diversity are perhaps more informative. Ideally one would make use of information present both the uSFS for potentially selected sites, whilst simultaneously modelling the reductions in neutral diversity caused by said selection. One possibility for such an enterprise would be in performing approximate Bayesian computation (ABC) 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 your data are an estimate of the true parameters. 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 we did in Chapter 3). Many thousands of different combinations of DFE parameters could be simulated and from these simulations, one could extract summary statistics for the site frequency spectrum, linkage disequilibrium and haplotype structure within, and in the regions surrounding, multiple classes of functional elements. The biggest difficulty in applying an analysis such as this the computational demands of the many, 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 as the evolution of whole chromosomes is typically tracked. In the original SLiM publication, Messer described how by tracking just the simulated mutations, simulations of purifying selection acting on a whole human chromosome (100Mbp long; 10^4 diploid individuals; for 10^5 generations) took just 4 days. As impressive as that is, it in infeasible that ABC could be performed 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 infeasible to perform ABC of the kind described. However, very recent advances in computational efficiency of forward-in-time simulators (Kelleher et al., 2018) may bring ABC of the kind outlined within reach.

5.4 Moving beyond mice

Recombination rates in other rodents, patterns of gentic diversity.

Mice are an excellent model organism for studies of mammalian molecular evolution. Obviously, the genomic resources available for mice (reference genome, annotations kit available for studies in mice is close to unrivalled. However, it remains to be seen whether the conclusions that we reached in this thesis can be generalised to other organisms.

Something about comparing the Nam et al(2017) analysis with the proposed analysis.

Throughout this thesis, we have examined evidence for molecular evolution at several levels; from the very broad (e.g. looking at the correlation between

recombination rate and genetic diversity in Chapter 2) to the very precise (analysis of the uSFS in Chapter 3). In this thesis we have come to a number of conclusions.

In Chapter 4 we used parameters of selection obtained by analysing patterns of genetic diversity, to try and answer one of evolutionary biology's long-standing questions: Do mutations in protein-coding or regulatory regions of the genome contribute more to adaptive evolution? We found that protein-coding regions do appear to contribute more to adaptive evolution by virtue of the increased selection pressure on protein-changing variants. It remains to be seen whether

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.

Auton, A. and McVean, G. (2007). Recombination rate estimation in the presence of hotspots. *Genome Res*, 17(8):1219–27.

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

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

Baudat, F., Buard, J., Grey, C., Fledel-Alon, A., Ober, C., Przeworski, M., Coop, G., and de Massy, B. (2010). Prdm9 is a major determinant of meiotic recombination hotspots in humans and mice. *Science*, 327(5967):836–840.

Baudat, F., Imai, Y., and de Massy, B. (2013). Meiotic recombination in mammals: localization and regulation. *Nat Rev Genet*, 14(11):794–806.

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.

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.

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.

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

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

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).

Consortium, T. . G. P., 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.

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.

Davies, B., Hatton, E., Altemose, N., Hussin, J. G., Pratto, F., Zhang, G., Hinch, A. G., Moralli, D., Biggs, D., Diaz, R., Preece, C., Li, R., Bitoun, E., Brick, K., Green, C. M., Camerini-Otero, R. D., Myers, S. R., and Donnelly, P. (2016). Re-engineering the zinc fingers of prdm9 reverses hybrid sterility in mice. *Nature*, 530(7589):171–6.

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.

Delaneau, O., Howie, B., Cox, A. J., Zagury, J. F., and Marchini, J. (2013). Haplotype estimation using sequencing reads. *Am J Hum Genet*, 93(4):687–96.

Dumont, B. L. and Payseur, B. A. (2011). Genetic analysis of genomic-scale recombiantion rate evolution in house mice. *PLoS Genet*, 7(6):11.

Dumont, B. L., White, M. A., Steffy, B., Wiltshire, T., and Payseur, B. A. (2011). Extensive recombination rate variation in the house mouse species complex inferred from genetic linkage maps. *Genome Res*, 21(1):114–25.

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.

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.

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.

Grey, C., Barthes, P., Chauveau-Le Friec, G., Langa, F., Baudat, F., and de Massy, B. (2011). Mouse prdm9 dna-binding specificity determines sites of histone h3 lysine 4 trimethylation for initiation of meiotic recombination. *PLoS Biol*, 9(10):e1001176.

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.

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

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

Hudson, R. R. (2001). Two-locus sampling distributions and their applications. *Genetics*, 159:12.

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

Jensen, J. D., Thornton, K. R., Bustamante, C. D., and Aquadro, C. F. (2007). On the utility of linkage disequilibrium as a statistic for identifying targets of positive selection in nonequilibrium populations. *Genetics*, 176(4):2371–9.

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.

Johnston, S. E., Berenos, C., Slate, J., and Pemberton, J. M. (2016). Conserved genetic architecture underlying individual recombination rate variation in a wild population of soay sheep (ovis aries). *Genetics*, 203(1):583–98.

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.

Kaur, T. and Rockman, M. V. (2014). Crossover heterogeneity in the absence of hotspots in caenorhabditis elegans. *Genetics*, 196(1):137–48.

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.

Kelleher, J., Thornton, K., Ashander, J., and Ralph, P. (2018).

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

Li, H. and Durbin, R. (2009). Fast and accurate short read alignment with burrows-wheeler transform. *Bioinformatics*, 25(14):1754–60.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., and Genome Project Data Processing, S. (2009). The sequence alignment/map format and samtools. *Bioinformatics*, 25(16):2078–9.

Liu, E. Y., Morgan, A. P., Chesler, E. J., Wang, W., Churchill, G. A., and Pardo-Manuel de Villena, F. (2014). High-resolution sex-specific linkage maps of the mouse reveal polarized distribution of crossovers in male germline. *Genetics*, 197(1):91–106.

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.

McVean, G., Awadalla, P., and Fearnhead, P. (2002). A coalescent-based method for detecting and estimating recombination from gene sequences. *Genetics*, 160:1231–1241.

McVean, G., Myers, S. R., Hunt, S., Deloukas, P., Bentley, D. R., and Donnelly, P. (2004). The fine-scale structure of recombination rate variation in the human genome. *Science*, 304.

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.

Myers, S. R., Bowden, R., Tumian, A., Bontrop, R. E., Freeman, C., MacFie, T. S., McVean, G., and Donnelly, P. (2010). Drive against hotspot motifs in primates implicates the prdm9 gene in meiotic recombination. *Science*, 327.

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

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.

Quinlan, A. R. and Hall, I. M. (2010). Bedtools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, 26(6):841–842.

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. (2017). Soft sweeps are the dominant mode of adaptation in the human genome. *Mol Biol Evol*.

Schwartz, J. J., Roach, D. J., Thomas, J. H., and Shendure, J. (2014). Primate evolution of the recombination regulator prdm9. *Nat Commun*, 5:4370.

Singhal, S., Leffler, E., Sannareddy, K., Turner, I., Venn, O., Hooper, D., Strand, A., Li, Q., Raney, B., Balakrishnan, C., Griffith, S., McVean, G., and Przeworski, M. (2015). Stable recombination hotspots in birds. *Science*, 350(6263):6.

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.

Smukowski, C. S. and Noor, M. A. (2011). Recombination rate variation in closely related species. *Heredity (Edinb)*, 107(6):496–508.

Smukowski Heil, C. S., Ellison, C., Dubin, M., and Noor, M. A. (2015). Recombining without hotspots: A comprehensive evolutionary portrait of recombination in two closely related species of drosophila. *Genome Biol Evol*, 7(10):2829–42.

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., Hoehn, K. B., and Noor, M. A. (2011). Effects of inversions on within- and between-species recombination and divergence. *Genome Biol Evol*, 3:830–41.

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*.

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.

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.

Wang, J. R., de Villena, F. P., and McMillan, L. (2012). Comparative analysis and visualization of multiple collinear genomes. *BMC Bioinformatics*, 13 Suppl 3:S13.

Wang, R. J., Gray, M. M., Parmenter, M. D., Broman, K. W., and Payseur, B. A. (2017). Recombination rate variation in mice from an isolated island. *Mol Ecol*, 26(2):457–470.

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.

Winckler, W., Myers, S. R., Richter, D. J., Onofrio, R. C., McDonald, G. J., Bontrop, R. E., McVean, G., Gabriel, S. B., Reich, D., Donnelly, P., and Altshuler, D. (2005). Comparison of fine-scale recombination rates in humans and chimpanzees. *Science*, 308.

Yang, H., Wang, J. R., Didion, J. P., Buus, R. J., Bell, T. A., Welsh, C. E., Bonhomme, F., Yu, A. H., Nachman, M. W., Pialek, J., Tucker, P., Boursot, P., McMillan, L., Churchill, G. A., and de Villena, F. P. (2011). Subspecific origin and haplotype diversity in the laboratory mouse. *Nat Genet*, 43(7):648–55.

Appendices

Appendix A

Booker *et al.* 2017 - BMC Biology

Booker et al. BMC Biology (2017) 15:98
DOI 10.1186/s12915-017-0434-y

BMC Biology

REVIEW Open Access

Detecting positive selection in the genome

Tom R. Booker, Benjamin C. Jackson and Peter D. Keightley*

Abstract

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 genetics theory to quantify the rate and strength of positive selection acting in a species' genome. In this review we 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 population's haplotype structure. The application of these tests has resulted in the detection of pervasive adaptive molecular evolution in multiple species.

Neutral theory and the extent of selection

The extent to which positive selection contributes to molecular evolution has been a long-standing question in evolutionary genetics. The classic paradigm in modern evolutionary genetics has been the neutral theory, which contends that the vast majority of molecular changes are a consequence of genetic drift, positive selection playing only a minor role [1]. However, it is becoming increasingly clear that natural selection, both positive and negative, is pervasive in many genomes, to such an extent that negative selection has been proposed as a null model for explaining variation in levels of genetic diversity across the genome [2]. Indeed, the question currently asked by researchers is no longer 'is positive selection present?' but instead 'how frequent and strong is positive selection?'. Fittingly, then, a number of different approaches have been proposed to quantify the frequency and strength of positive selection using population genetic (and genomic) approaches.

The purpose of this review is to describe the different lines of evidence that have been used to determine the frequency and strength of positive selection in multiple species. We will start by discussing the McDonald-Kreitman test [3] and its extensions, which have been used to quantify the frequency of adaptive molecular evolution acting directly on protein-coding genes. We then discuss how predictions of selective sweep models (Fig. 1) can be used to estimate the parameters of positive selection indirectly, using variability at linked neutral sites. Finally, we describe how recent results from large-scale genomic datasets have challenged the bases of these methods. Note, we will not focus on the many methods to identify individual adaptive events or genome scans to detect local adaptation (for a review, see [4]), nor will we discuss experimental evolution (for reviews, see [5] and [6]).

Quantifying the frequency of positive selection—the McDonald-Kreitman test

Some of the strongest evidence for adaptive molecular evolution has come from application of the McDonald-Kreitman (MK) test [3] and methods based on it. Testing for evidence of positive selection requires a suitable null hypothesis. Under the neutral hypothesis of molecular evolution, differences accumulate by genetic drift, positive selection playing only a minor role [1]. The MK test can be used to test for positive selection by comparing within-species nucleotide diversity and between-species nucleotide divergence for sites subject to natural selection and sites assumed to be evolving neutrally. Most studies have analyzed nonsynonymous sites of protein-coding genes, using synonymous sites as a neutral reference. We will focus on such analyses here, although the MK test has also been applied to a variety of non-coding genomic elements in several species. If synonymous mutations evolve neutrally and nonsynonymous mutations are either neutral or are strongly deleterious, the ratio of the number of nonsynonymous to synonymous polymorphisms for a gene (P_n/P_s) is expected to be equal to the ratio of nonsynonymous to synonymous divergence (D_n/D_s) (although it should be noted that measures of polymorphism and divergence are not entirely independent). Strongly positively selected mutations, however, will inflate D_n, while contributing negligibly to P_n (Table 1).

* Correspondence: peter.keightley@ed.ac.uk
Institute of Evolutionary Biology, University of Edinburgh, Edinburgh EH9 3FL, UK

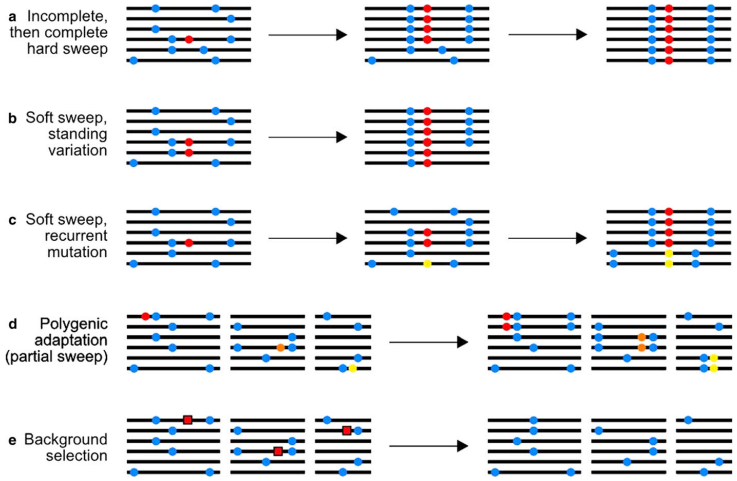

Fig. 1. Selective sweeps and background selection. Maynard Smith and Haigh [79] showed that as an advantageous mutation rises in frequency it drags with it linked neutral polymorphisms. With increasing genetic distance from the selected site, the effect is reduced, resulting in troughs in genetic diversity in surrounding regions. **a** Hard/classic sweeps - the most well-studied model of sweeps. A new advantageous mutation rapidly increases in frequency to eventual fixation. 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 [27, 79]. **b**-**c** Soft sweeps - a neutral allele segregating in a population may become favoured (due, for example, to a change in the environment). **b** The segregating allele may be associated with multiple haplotypes, and as it rises in frequency, so do the multiple haplotypes. **c** A similar process, also termed a soft sweep, can occur if an advantageous mutation arises by multiple, distinct mutation events. See [66] for a thorough review on soft sweep models. **d** 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 **a**. The term partial sweep describes the situation wherein a sweeping allele becomes effectively neutral at a certain frequency in its trajectory **d**. The magnitude of both processes' effects on linked neutral diversity depends on the frequency reached by the sweeping allele when selection is 'turned off' or on the time of sampling [33]. Partial sweeps may be common in cases of adaptation involving selection on quantitative traits [67]. **e** Background selection - as natural selection purges deleterious mutations, neutral alleles linked to the selected locus are also lost. The process of background selection is qualitatively similar to recurrent selective sweeps, since both processes reduce local genetic diversity [80] and skew the SFS towards rare variants [81, 82]. Models of background selection envisage a neutral site linked to many functional sites at different distances, such that the effects of selection at many sites accumulate to reduce diversity [83, 84]. Blue circles represent neutral alleles, red, yellow and orange circles represent advantageous alleles, and red squares represent deleterious alleles

The MK test ratios allow estimation of the fraction of nonsynonymous differences, α, driven to fixation by position selection for a set of genes or other class of sites [7]:

$$\alpha = 1 - \frac{D_S P_n}{D_n P_s}$$

Table 1 MK table for the *Adh* gene [3] showing numbers of fixed differences and polymorphic sites between and within *D. melanogaster, D. simulans* and *D. yakuba*

	Differences (*D*)	Polymorphism (*P*)
Nonsynonymous	7	2
Synonymous	17	42

Note that the ratio of fixed nonsynonymous to synonymous differences (7/17) is substantially higher than the ratio of nonsynonymous to synonymous polymorphisms (2/42), indicating that some amino acid differences are adaptive

A weakness of this approach is that it assumes the strict neutral model, where deleterious nonsynonymous mutations can be frequent, but are assumed to be strongly selected against, such that they contribute negligibly to polymorphism and divergence. If there are slightly deleterious mutations, these will tend to inflate P_n while not becoming fixed. This reduces the power to detect adaptive evolution for a given gene and potentially downwardly biases estimates of α for a group of genes. Omitting low frequency variants preferentially removes slightly deleterious mutations and can potentially reduce this bias [8, 9], but the result is sensitive to the arbitrary cut-off value chosen. More recently, approaches for estimating α have been developed that use the spectrum of allele frequencies [10–13], explicitly modeling the contribution of deleterious mutations to polymorphism and divergence. Within all of these approaches, the distribution of fitness effects (DFE) of

nonsynonymous mutations is estimated, based on the relative levels of nonsynonymous versus synonymous polymorphism and the properties of the frequency distribution of numbers of allele copies present at segregating sites (the 'site frequency spectrum' (SFS); Fig. 2).

Various models for the DFE have been assumed in these analyses, a common one being the gamma distribution. The estimated parameters of the DFE are then used to calculate the expected number of nonsynonymous differences between the species pair; the difference between the observed and the expected divergence is attributed to positively selected mutations and used to estimate α [14] (Box 1). It is possible to base inferences on the unfolded or folded SFS (Fig. 2); in the former case, polymorphisms need to be polarised using outgroup species, and it is then feasible to include advantageous mutations within the analysis [12]. It is also possible to base inferences solely on standing polymorphism, that is, to ignore the between-species divergence altogether [13, 15]. With all these different flavors of the basic method, recent demographic changes, altering the shape of both the synonymous and nonsynonymous SFSs compared to that expected under the neutral model, are incorporated in the analysis. Correcting for demographic change by allowing changes in effective population size appears to substantially correct for other genome-wide processes that distort the SFSs, such as background selection [16].

Empirical findings from applying the MK test and its derivatives

While initial results from the application of these approaches were somewhat confusing, a more consistent picture emerged as larger data sets became available. Initial results indicated that adaptive protein evolution is widespread in *Drosophila*, with α values typically as high as 40% [17], whereas estimates for humans were generally substantially lower and in some cases nonsignificantly different from zero [17].

The frequency of adaptive substitution is expected to be higher in populations of large effective size, N_e, since the probability of fixation of a newly arising advantageous mutation increases with N_e [18], and more advantageous mutations appear in large populations. However, α is not simply a function of the rate of fixation of advantageous

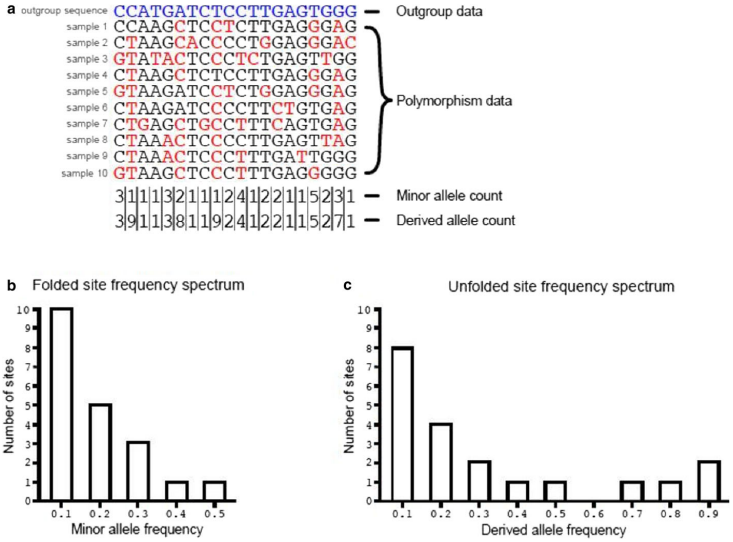

Fig. 2. The site frequency spectrum. The numbers of variants segregating at different frequencies in a population can be summarized as the site frequency spectrum (SFS). Consider the ten chromosome samples shown in **a**. Observations of a particular minor allele frequency are used to populate the folded SFS **b**. 'Unfolding' the SFS requires knowledge of whether alleles are ancestral or derived. Aligning sequenced data to an outgroup (the blue nucleotides in **a**) allows the inference of ancestral and derived states for polymorphic and diverged sites, by maximum parsimony. However, the parsimony approach makes a number of biologically unrealistic assumptions; for example, that there have been no mutations in the lineage leading to the outgroup. Because of these, a number of alternative approaches have been proposed that have been shown to be more accurate than parsimony (e.g. [15]). Various evolutionary processes can alter the SFS, including directional and balancing selection, gene conversion, population size change and migration. For example, purifying selection prevents harmful variants from rising in frequency, resulting in a skew in the SFS towards rare variants. Multiple statistics have been proposed to summarize both the folded and unfolded SFS, and these can shed light on the evolutionary process (reviewed in [4])

Box 1 Calculation of α and ω_a using estimates of the distribution of fitness effects of new mutations

Assume we are focusing on the evolution of protein-coding genes between two species, and that we have polymorphism data for a focal species. The amino acid divergence between the species (D_n) is the sum of the divergence attributable to positively selected mutations (D_a) and that attributable to the fixation of neutral and slightly deleterious mutations (D_{na}):

$$D_n = D_a + D_{na}$$

The amino acid divergence can be estimated directly from the sequence data of the two species. Methods such as DFE-alpha [11] infer D_{na} by calculating the average fixation probability of a deleterious mutation—based on the distribution of fitness effects of new deleterious mutations—estimated from the information contained in the folded nonsynonymous and synonymous site frequency spectra (Fig. 2) of the focal species. The adaptive divergence is then $D_a = D_n - D_{na}$. The estimated proportion of amino acid substitutions driven to fixation by positive selection (α) is the ratio of the adaptive divergence (D_a) and the amino acid divergence (D_n):

$$\alpha = D_a/D_n$$

An alternative and potentially more informative estimator of the frequency of adaptive molecular evolution is ω_a, the ratio of the adaptive divergence and the synonymous divergence:

$$\omega_a = D_a/D_s$$

Galtier [13] proposed a complementary statistic, ω_{na}, which gives an estimate of the rate of non-adaptive amino acid substitutions.

mutations, since the overall rate of substitution (the denominator used in the calculation of α) includes the rate of fixation of deleterious mutations (Box 1), and these are expected to fix less frequently in large populations. This implies that α should increase with N_e, even if the rate of fixation of advantageous mutations does not change. Campos et al. [19] observed a positive correlation between α and the rate of recombination for protein-coding genes in the *Drosophila melanogaster* genome. Since N_e for a genomic region is positively related to the rate of recombination [20], increased rates of fixation of advantageous mutations and decreased rates of fixation of deleterious mutations are expected in high recombination regions. Campos et al. also observed that the rate of recombination is positively correlated with ω_a, the estimated rate of advantageous substitution relative to the rate of neutral substitution (Box 1), suggesting that beneficial substitutions increase with increasing recombination rate, perhaps due to decreasing interference between selected loci [21].

Similarly, a positive correlation between the N_e for a species and ω_a was observed by Gossmann et al. [22] in an analysis of protein-coding genes from 13 eukaryotic species pairs. Evidence from a much larger study [13], however, does not support a relationship between N_e and the rate of adaptive molecular evolution. Galtier [13] studied protein-coding genes in 44 metazoan species pairs to investigate the relationships between the rate of adaptive evolution (measured using α and ω_a) and N_e. There was a positive relationship between α and N_e, but a negative relationship between the estimated rate of fixation of deleterious mutations (ω_{na}) and N_e. However, ω_a was nonsignificantly correlated with N_e, implying that the positive correlation between N_e and α is driven by variation in the fixation rate of deleterious mutations. This result also implies that adaptation of protein-coding genes may not be limited by the supply of new mutations.

Are most amino acid substitutions adaptive?

A notable conclusion from Galtier's study is that average α exceeds 50%, implying that most amino acid substitutions are adaptive in many species. Primates, notably hominids, are an exception, tending to have lower α, presumably because of their small effective population sizes, leading to the accumulation of slightly deleterious amino acid mutations. Taken at face value, Galtier's study is, therefore, a strong challenge to the neutral hypothesis of molecular evolution, as it suggests that a large proportion of amino acid changes resulted from positive selection in a variety of species. There are, however, several caveats. First, if selection is operating in the reference class of sites (in the case of protein-coding genes, selection on codon usage operating on synonymous sites), upwardly biased estimates of α are expected [23], and this kind of selection is most prevalent in species with large N_e. Second, Fay [24] highlights a number of difficulties with the MK-based approach, including local adaptation and epistasis among deleterious mutations, both of which could inflate values of α. Finally, Galtier included 'mirror species pairs' where polymorphism data were available for both species of the pair, and two estimates of α and ω_a could therefore be calculated. While estimates of these quantities were mostly in reasonable agreement, one mirror species pair from an earlier study of ours (the house mouse and brown rat) produced strikingly different estimates: α = +0.32 if polymorphism data for mice are analyzed and α = −0.29 if data from rats are analyzed [25]. The negative estimate for rats was attributed to a population bottleneck in the brown rat, increasing the frequency of slightly deleterious amino acid mutations in current rat populations. Nucleotide divergence between mice and rats accumulated over a much longer time-scale, however, and was presumably largely unaffected by this bottleneck. Similar results have been found for several plant species, where estimates of α are for the most part close to zero [26], and in some cases significantly less than zero. These examples highlight a fundamental

problem with the MK-based approach—within-species nucleotide diversity and between-species divergence can be decoupled from one another by ancient demographic events not captured by current polymorphism data, potentially undermining the ability to estimate the prevalence of adaptive molecular evolution.

Using models of selective sweeps to estimate positive selection parameters

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 analyses of this process, termed a selective sweep (Fig. 1), have 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 [27]. Thus, comparing diversity at multiple neutral loci linked to selected regions, in principle, should provide an indirect means for estimating the average rate and strength of positive selection in the genome.

If a population experiences recurrent selective sweeps, several patterns are predicted by theory. Under recurrent 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 predictions has been met in empirical studies, and each has been used to estimate parameters of positive selection using models of hard selective sweeps.

The correlation between diversity and the rate of recombination

In the late 1980s, evidence began to emerge suggesting that genetic polymorphism is reduced in genomic regions that experience restricted crossing-over [28, 29]. Soon after, Begun and Aquadro [30] 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 [31]. Begun and Aquadro pointed out that the correlation is qualitatively consistent with the action of recurrent selective sweeps. Wiehe and Stephan [32] formulated expressions, based on the correlation between nucleotide diversity and the rate of recombination, to estimate the compound parameter for the intensity of selection $\lambda 2N_e s$, 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 (the reduction in relative fitness experienced by homozygotes), assuming semi-dominance. They applied their method to the data of Begun and Aquadro [30], estimating $\lambda 2N_e s = 5.37 \times 10^{-8}$, but their method could not disentangle the individual parameters. More recently, Coop and Ralph [33] 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 and recombination observed by Begun and Aquadro [30] can also be explained by background selection, the reduction in neutral diversity caused by the removal of linked deleterious mutations (Fig. 1) [34]. The correlation between neutral diversity and the recombination rate predicted by background selection is quantitatively similar to that observed in *D. melanogaster* [35]. Indeed, recent studies suggest that background selection is a major determinant of nucleotide diversity variation at broad scales (>100 kbp) in humans [36] and *D. melanogaster* [2, 37]. It is clear, then, that background selection is a key confounding factor when attempting to make inferences about positive selection from diversity patterns.

Correlation between neutral diversity and non-neutral divergence

Under a model of recurrent sweeps, there should be a negative correlation between nucleotide divergence at selected sites and diversity at linked neutral sites. This is because rapidly evolving regions of the genome will experience more sweeps, which will reduce levels of linked neutral diversity more than slowly evolving regions. The relationship between neutral diversity and selected divergence should therefore carry information on the rate and strength of selective sweeps.

The abovementioned correlation was first described by Andolfatto [38] for the X chromosome of *D. melanogaster* using synonymous site diversity and nonsynonymous divergence, and has been subsequently reported in other *Drosophila* species [39]. Using the correlation, Andolfatto [38] estimated the compound parameter for the intensity of selection $\lambda 2N_e s = 3 \times 10^{-8}$ for the X chromosome in *D. melanogaster* (similar to the value obtained based on the correlation of synonymous site diversity and recombination rate [32]; see above). Using an estimate of α obtained from an MK-based analysis, Andolfatto [38] decomposed $\lambda 2N_e s$ into its constituent parameters and found that advantageous mutations in the protein-coding genes of *D. melanogaster* are moderately weakly selected but relatively frequent. In a similar study, Macpherson et al. [40] examined the correlation between mean neutral diversity and selected (nonsynonymous) divergence in *Drosophila simulans,* and estimated $\lambda 2N_e s$ to be $\sim 10^{-7}$. However, they used a model that 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, allowing them to obtain estimates of the λ and s parameters by regression. Although estimates of the compound parameter $\lambda 2N_e s$ are similar between the two studies, the

estimated rate and fitness effect parameters were quite different, Macpherson et al. [40] estimating that advantageous mutations are relatively rare and have large fitness effects. The discrepancies between the studies may be due to differences in biology between *D. melanogaster* and *D. simulans*, or may reflect differences in methodology. 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 [38]. 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. [40], perhaps explaining why they inferred a lower overall rate of sweeps, with higher selection coefficients (for a full description, see [41]). 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, since these will affect patterns of diversity but not necessarily divergence. Recently, Campos et al. [42] estimated positive selection parameters from the correlation between synonymous site diversity and non-synonymous divergence across the entire *D. melanogaster* genome in the presence of both background selection and gene conversion. Their parameter estimates suggest that strongly selected advantageous mutations are relatively infrequent, making up ~ 0.02% of all new mutations at nonsynonymous sites.

In summary, analysis of the correlation between neutral diversity and putatively selected divergence has suggested that advantageous mutations in *Drosophila* are either relatively frequent, but weakly selected, or rare and strongly selected. Obviously, assuming that all advantageous mutations that occur in a genome belong to a single class of fitness effects is an oversimplification of what is likely to be a complex distribution. It may be that the discrepancy between the above studies comes about because they are capturing different parts of the distribution of fitness effects for positively selected mutations. This is corroborated by recent work described below.

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 above), 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. [43], was first applied to *D. simulans*, and the above pattern was found. By fitting a hard sweep model to the shape of the diversity trough, they estimated α values of 5 and 13%, depending on whether one or two classes of beneficial mutational effects were fitted. Note that their estimates of α are substantially lower than those obtained using MK-based methods for *D. melanogaster* [38]. Sattath et al. [43] suggested that modes of selection other than hard sweeps may help explain this discrepancy. However, even when modelling two classes of beneficial mutations, they found that amino acid substitutions are driven by relatively strongly adaptive mutations ($s \sim 0.5\%$ and $s \sim 0.01\%$). Their estimates of the selection strength are therefore in broad agreement with the estimate of $s \sim 1\%$ obtained by Macpherson et al. [40], based on the correlation between synonymous diversity and non-synonymous divergence in *D. simulans*. The results from the Sattath et al. [43] analysis are consistent with the hypothesis that adaptation in protein-coding genes is fairly frequent and driven by strong, hard sweeps.

The Sattath test has since been applied in a variety of organisms, including humans [44], house mice [45], *Capsella grandiflora* [46] and maize [47]. 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 [44] and maize [47]. However, Enard et al. [48] 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. [48] found evidence for frequent adaptive substitutions in humans, particularly in regulatory sequence. To address the issues raised by Enard et al. [48], Beissinger et al. [47] 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 *C. grandiflora* [49] and *Drosophila* [50], whereas in humans, maize and wild house mice it decays over distances closer to 10,000 bp [25, 51, 52]. It may be, therefore, that the Sattath test is only applicable when there is relatively short-range LD, such that the patterns of diversity around selected substitutions are decoupled from the patterns of diversity around neutral substitutions. 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.

Fitting genome-wide variation in nucleotide diversity and divergence

Methods for estimating 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. [53] 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 in 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 ($s \sim 0.0003\%$), summing to $\alpha = 40.4\%$, which is similar to the estimate obtained using the MK test [38]. 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. [53] showed that the variation in nucleotide diversity expected under a model combining the diversity-reducing 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 likely to be pervasive, at least in *D. melanogaster*. However, several points need to be considered regarding their results. Firstly, the strength of selection on the weakly selected class of beneficial mutations in Elyashiv et al.'s study may be too weak (assuming $N_e = 10^6$ for *D. melanogaster*, $N_e s \sim 3$), such that the fixation probability of a newly arising advantageous mutation is very similar to that of a neutral allele. Such weak selection in *D. melanogaster* may not necessarily limit the frequency of hard sweeps, however, as it has been suggested that adaptation in *D. melanogaster* may be limited by current census population size rather than long-term N_e [54]. Secondly, the Elyashiv et al. [53] approach does not incorporate gene conversion, which may have a substantial impact on the effects of sweeps within genes [42]. Finally, their method overestimated the rate of deleterious mutations, though the authors suggested that this could be due to the presence of modes of adaptation other than hard sweeps in *D. melanogaster*.

Haplotype structure can reveal both soft and incomplete selective sweeps

The extent to which adaptive evolution proceeds according to the hard sweep model is the subject of ongoing study. All of the approaches to infer the strength and tempo of adaptation we have discussed, with the exception of Coop and Ralph [33], have relied on either patterns of between-species substitution or the predictions made by hard sweep models. If adaptive change is limited by the supply of new mutations, hard sweeps must be the main mode of adaptive evolution. As described above, however, adaptation does not seem to be limited by the mutation rate, so perhaps alternative modes are common. The following section will describe how information carried in the distribution of haplotypes can be used to distinguish different forms of selective sweeps.

While a favoured allele is sweeping through a population, it carries with it linked variants on the same chromosome (Fig. 1). In the hypothetical case of a hard sweep arising from a single new beneficial mutation, with no further recombination or mutation, this will result in one haplotype coming to completely dominate the population. Although this situation is extreme, it serves as an example to highlight the fact that a lack of haplotype diversity, or, equivalently, an increase in LD between alleles at different sites, can be used as an indicator of the action of positive selection. In the case of soft sweeps, more than one haplotype may be elevated to a high frequency, and in the cases of incomplete and partial sweeps, a single haplotype may be at a higher frequency than expected under null models.

Using haplotype structure to detect soft selective sweeps

The distribution of haplotypes at a locus has been analyzed to detect selection where adaptive evolution is very recent (for example [55–60]) and where it does not proceed according to the hard sweep model (for example [61–63]). Several test statistics have been proposed to analyze the distribution of haplotype frequencies in a sample (for descriptions of these see [64]). However, the power to detect selection decays quickly after a selective event ends [61]. There are several reasons for this, including the loss of ancestral haplotypes through genetic drift, recombination occurring before and after the fixation of an adaptive mutation shortening the haplotype generated by the sweep, and, finally, further mutation creating new haplotypes not associated with the initial sweep. The signatures present in the haplotype structure (for example a skew towards a small number of high frequency haplotypes) generated by positive selection persist for only $\sim 0.01\ N_e$ generations, which is an order of magnitude shorter than the persistence time of signatures in the site frequency spectrum [61, 65, 66].

Haplotype-based tests outperform diversity and site frequency spectrum-based tests at detecting soft sweeps. This is because, under the soft sweep model, several haplotypes may be carried to high frequency, resulting in characteristic signatures in a population's haplotype structure, while leaving polymorphism less affected

[61, 67]. There is now a sizeable amount of theoretical and empirical evidence suggesting that soft sweeps contribute to adaptive evolution in nature [66, 68]. For example, Garud et al. [62] introduced a set of haplotype-based statistics that together can detect both hard and soft sweeps, and discriminate between them. They applied their statistics to North American *D. melanogaster* and found evidence suggesting that soft sweeps are more common than hard sweeps. Similar results for a Zambian population were subsequently reported by Garud and Petrov [69]. However, soft sweeps arising from multiple de novo mutations require high beneficial mutation rates. In the case of soft sweeps from standing variation, even if alleles are segregating at appreciable frequencies in the population before the onset of selection, they may still be more likely to result in a hard sweep than a soft one (reviewed by [70]).

Using haplotype structure to detect incomplete or partial selective sweeps

As is the case for soft sweeps, the signatures of both incomplete and partial selective sweeps left in polymorphism data are less clear than for hard sweeps (Fig. 1). For example, haplotype-based methods have revealed footprints of incomplete sweeps around certain alleles that are known to confer resistance to malaria [56]. If polygenic traits are the target of selection, partial sweeps may be common, because selection can bring about rapid evolution by acting on standing variation at multiple loci, affecting levels of diversity at linked neutral sites [67, 71]. A haplotype-based statistic introduced by Field et al. [63] called the singleton density score ('SDS') is able to detect very recent selection, including selection operating on polygenic traits. It quantifies the extent to which selection has distorted the genealogy of sampled haplotypes, as measured by the distribution of singleton mutations around ancestral and derived alleles at a focal locus. Field et al. provide evidence of selection on multiple polygenic traits, including height, in the ancestors of British people within the last 3000 years, suggesting that partial sweeps may be a common form of adaptive evolution. However, their study relied on published catalogues of genome-wide association study hits and > 3000 sequenced genomes, resources not available for most organisms. It remains to be seen whether these findings are general across different species groups. Finally, recent theoretical work by Jain and Stephan [72] suggests that the allele frequency shifts resulting from polygenic adaptation may be too subtle to be detected using common approaches, although this depends on the number of loci underlying quantitative traits. Indeed, quantitative traits can respond to selection when loci underlying the trait have $N_e s < 1$ [73]. Biologically grounded simulations using realistic trait architectures and selection regimes are likely necessary to determine how readily polygenic adaptation can be detected using population genomic data.

Patterns of LD can thus be used to infer the action of positive selection. Hard sweeps produce distinctive patterns of LD, but this information adds little for detecting hard sweeps when information from diversity and the site frequency spectrum is available [74], although it may be useful for distinguishing selection from demographic effects [75]. Haplotype information is useful, however, when selection is ongoing and/or it does not proceed according to the hard sweep model. One drawback of haplotype-based statistics is that they are often descriptive—although they provide a means for detecting sweeps, they do not provide a direct means for parameter estimation. An exception is the estimator of Messer and Neher [76], which is based on the frequency spectrum of haplotypes that arise during a sweep, and which may outperform diversity-based estimators of the strength of selection in some circumstances, although it requires a deep population sample (at least hundreds or thousands of sequences) to provide accurate estimates.

Future directions: sweep modes and non-model organisms

Over the last ~ 30 years, much information about the action of natural selection has been leveraged from patterns of between-species substitution and within-species polymorphism. Researchers have accumulated evidence suggesting not only that adaptive evolution is frequent across a variety of species, but that it appears to be driven by strongly selected mutations. The application of recently developed tests and models to data from non-model organisms remains a challenge, however, since they variously require a population sample for very many individuals, a high quality reference genome and annotations, a genetic map and genome sequences of suitable outgroup species. Understanding the process of adaptive change in the genome across diverse taxa may therefore be challenging due to a lack of appropriate data.

A major challenge for understanding the forces of natural selection operating in the genome will be the incorporation of both soft and partial sweeps into theory and inference methods. The recent findings of Field et al. [63], Garud et al. [62] and Garud and Petrov [69] all suggest that both partial and soft sweeps may occur frequently. If modes of adaptation other than hard sweeps are common, current methods for inferring positive selection may result in systematically biased inferences. For example, a key parameter in the partial sweep model is the frequency that a beneficial mutation reaches before selection is 'switched off'. As this critical frequency decreases, the inferred rate of sweeps increases over multiple orders of magnitude [33]. This example from theory, as well as the recent empirical results from population haplotype structure, should

stimulate efforts to quantify the extent to which different sweep modes contribute to molecular evolution. To that end, Schrider and Kern have developed a machine learning approach [77] to classify region signatures of sweeps as either hard or soft. Application of their approach suggests that soft sweeps may be the dominant mode of adaptation in human evolution [78]. Estimating selection parameters based on the signatures of soft sweeps remains an open problem.

Box 2 Glossary

DFE—the distribution of fitness effects for new mutations

Folded site frequency spectrum (folded SFS)—the distribution of minor allele frequencies in a sample of nucleotide sequences

Unfolded site frequency spectrum (unfolded SFS)—the distribution of derived allele frequencies in a sample of nucleotide sequences

α—the proportion of substitutions that have been driven to fixation by positive selection, and not by other forces, such as drift

ω_α—the rate of fixation of advantageous mutations relative to rate for neutral mutations

N_e—effective population size

s—the absolute selection coefficient, the difference in fitness between homozygotes for wild-type alleles and homozygotes for mutant alleles (in diploids)

$N_e s$—the effective strength of selection, the strength of directional selection relative to random drift

LD—linkage disequilibrium, nonrandom associations of alleles at different loci

Acknowledgements
We thank Brian Charlesworth for helpful discussions and two anonymous referees for comments on the manuscript. TRB is supported by a BBSRC EASTBIO studentship. BCJ and PDK are funded by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program (grant agreement number 694212).

Authors' contributions
TRB, BCJ and PDK wrote the manuscript. All authors read and approved the final manuscript.

Competing interests
The authors declare that they have no competing interests.

Publisher's Note

Published online: 30 October 2017

References

1. Kimura M. The neutral theory of molecular evolution. Cambridge University Press; 1983.
2. Comeron J. Background selection as a baseline for nucleotide variation across the *Drosophila* genome. PLoS Genet. 2014;10(6):e1004434.
3. McDonald JM, Kreitman M. Adaptive protein evolution at the *Adh* locus in *Drosophila*. Nature. 1991;351:652–4.
4. Casillas S, Barbadilla A. Molecular population genetics. Genetics. 2017;205(3):1003–35.
5. Thurman TJ, Barrett RD. The genetic consequences of selection in natural populations. Mol Ecol. 2016;25(7):1429–48.
6. Bailey SF, Bataillon T. Can the experimental evolution programme help us elucidate the genetic basis of adaptation in nature? Mol Ecol. 2016;25(1):203–18.
7. Charlesworth B. The effect of background selection against deleterious mutations on weakly selected, linked variants. Genet Res. 1994;63(03):213.
8. Charlesworth J, Eyre-Walker A. The McDonald-Kreitman test and slightly deleterious mutations. Mol Biol Evol. 2008;25(6):1007–15.
9. Fay JC, Wyckoff GJ, Wu CI. Positive and negative selection on the human genome. Genetics. 2001;158:1227–34.
10. Boyko AR, Williamson SH, Indap AR, Degenhardt JD, Hernandez RD, Lohmueller KE, et al. Assessing the evolutionary impact of amino acid mutations in the human genome. PLoS Genet. 2008;4(5):e1000083.
11. Eyre-Walker A, Keightley PD. Estimating the rate of adaptive molecular evolution in the presence of slightly deleterious mutations and population size change. Mol Biol Evol. 2009;26(9):2097–108.
12. Schneider A, Charlesworth B, Eyre-Walker A, Keightley PD. A method for inferring the rate of occurrence and fitness effects of advantageous mutations. Genetics. 2011;189(4):1427–37.
13. Galtier N. Adaptive protein evolution in animals and the effective eopulation size hypothesis. PLoS Genet. 2016;12(1):e1005774.
14. Loewe L, Charlesworth B, Bartolome C, Noel V. Estimating selection on nonsynonymous mutations. Genetics. 2006;172(2):1079–92.
15. Keightley PD, Campos JL, Booker TR, Charlesworth B. Inferring the frequency spectrum of derived variants to quantify adaptive molecular evolution in protein-coding genes of *Drosophila melanogaster*. Genetics. 2016;203(2):975–84.
16. Messer PW, Petrov DA. Frequent adaptation and the McDonald-Kreitman test. Proc Natl Acad Sci U S A. 2013;110(21):8615–20.
17. Eyre-Walker A. The genomic rate of adaptive evolution. Trends Ecol Evol. 2006;21(10):569–75.
18. Fisher RA. The genetical theory of natural selection. Oxford University Press; 1930.
19. Campos JL, Halligan DL, Haddrill PR, Charlesworth B. The relation between recombination rate and patterns of molecular evolution and variation in *Drosophila melanogaster*. Mol Biol Evol. 2014;31(4):1010–28.
20. Charlesworth B, Charlesworth D. Elements of evolutionary genetics. Greenwood Village, Colorado: Roberts & Company; 2010.
21. Castellano D, Coronado-Zamora M, Campos JL, Barbadilla A, Eyre-Walker A. Adaptive evolution Is substantially impeded by Hill-Robertson interference in *Drosophila*. Mol Biol Evol. 2016;33(2):442–55.
22. Gossmann TI, Keightley PD, Eyre-Walker A. The effect of variation in the effective population size on the rate of adaptive molecular evolution in eukaryotes. Genome Biol Evol. 2012;4(5):658–67.
23. Matsumoto T, John A, Baeza-Centurion P, Li B, Akashi H. Codon usage selection can bias estimation of the fraction of adaptive amino acid fixations. Mol Biol Evol. 2016;33(6):1580–9.
24. Fay JC. Weighing the evidence for adaptation at the molecular level. Trends Genet. 2011;27(9):343–9.
25. Deinum EE, Halligan DL, Ness RW, Zhang YH, Cong L, Zhang JX, et al. Recent evolution in *Rattus norvegicus* is shaped by declining effective population size. Mol Biol Evol. 2015;32(10):2547–58.
26. Gossmann TI, Song BH, Windsor AJ, Mitchell-Olds T, Dixon CJ, Kapralov MV, et al. Genome wide analyses reveal little evidence for adaptive evolution in many plant species. Mol Biol Evol. 2010;27(8):1822–32.
27. Barton NH. Genetic hitchhiking. Philos Trans R Soc Lond B Biol Sci. 2000; 355(1403):1553–62.
28. Aguade M, Miyashita N, Langley CH. Reduced variation in the yello-achaete-schute region in natural populations of *Drosophila melanogaster*. Genetics. 1989;122:607–15.
29. Stephan W, Langley CH. Evolutionary consequences of DNA mismatch inhibited repair opportunity. Genetics. 1992;132:567–74.
30. Begun DJ, Aquadro CF. Levels of naturally occuring DNA polymorphism correlate with recombination rate in *Drosophila melanogaster*. Nature. 1992;356:519–20.
31. Cutter AD, Payseur BA. Genomic signatures of selection at linked sites: unifying the disparity among species. Nat Rev Genet. 2013;14(4):262–74.

32. Wiehe T, Stephan W. Analysis of a genetic hitchhiking model, and its application to DNA polymorphism data from *Drosophila melanogaster*. Mol Biol Evol. 1993;10(4):842–54.
33. Coop G, Ralph P. Patterns of neutral diversity under general models of selective sweeps. Genetics. 2012;192(1):205–24.
34. Charlesworth B, Morgan MT, Charlesworth D. The effect of deleterious mutations on neutral molecular variation. Genetics. 1993;134:1289–303.
35. Charlesworth B. Background selection and patterns of genetic diversity in *Drosophila melanogaster*. Genet Res. 1996;68:131–49.
36. McVicker G, Gordon D, Davis C, Green P. Widespread genomic signatures of natural selection in hominid evolution. PLoS Genet. 2009;5(5):e1000471.
37. Charlesworth B. The role of background selection in shaping patterns of molecular evolution and variation: evidence from variability on the *Drosophila* X chromosome. Genetics. 2012;191(1):233–46.
38. Andolfatto P. Hitchhiking effects of recurrent beneficial amino acid substitutions in the *Drosophila melanogaster* genome. Genome Res. 2007;17(12):1755–62.
39. Haddrill PR, Zeng K, Charlesworth B. Determinants of synonymous and nonsynonymous variability in three species of *Drosophila*. Mol Biol Evol. 2011;28(5):1731–43.
40. Macpherson JM, Sella G, Davis JC, Petrov DA. Genomewide spatial correspondence between nonsynonymous divergence and neutral polymorphism reveals extensive adaptation in *Drosophila*. Genetics. 2007;177(4):2083–99.
41. Sella G, Petrov DA, Przeworski M, Andolfatto P. Pervasive natural selection in the *Drosophila* genome? PLoS Genet. 2009;19(6):e1000495.
42. Campos JL, Zhao L, Charlesworth B. Estimating the parameters of background selection and selective sweeps in *Drosophila* in the presence of gene conversion. Proc Natl Acad Sci U S A. 2017;114(24):E4762–771.
43. Sattath S, Elyashiv E, Kolodny O, Rinott Y, Sella G. Pervasive adaptive protein evolution apparent in diversity patterns around amino acid substitutions in *Drosophila simulans*. PLoS Genet. 2011;7(2):e1001302.
44. Hernandez RD, Kelly JL, Elyashiv E, Melton SC, Auton A, McVean G, et al. Classic selective sweeps were rare in recent human evolution. Science. 2011;331:920–4.
45. Halligan DL, Kousathanas A, Ness RW, Harr B, Eory L, Keane TM, et al. Contributions of protein-coding and regulatory change to adaptive molecular evolution in murid rodents. PLoS Genet. 2013;9(12):e1003995.
46. Williamson RJ, Josephs EB, Platts AE, Hazzouri KM, Haudry A, Blanchette M, et al. Evidence for widespread positive and negative selection in coding and conserved noncoding regions of *Capsella grandiflora*. PLoS Genet. 2014;10(9):e1004622.
47. Beissinger TM, Wang L, Crosby K, Durvasula A, Hufford MB, Ross-Ibarra J. Recent demography drives changes in linked selection across the maize genome. Nat Plants. 2016;2(7):16084.
48. Enard D, Messer PW, Petrov DA. Genome-wide signals of positive selection in human evolution. Genome Res. 2014;24(6):885–95.
49. Josephs EB, Lee YW, Stinchcombe JR, Wright SI. Association mapping reveals the role of purifying selection in the maintenance of genomic variation in gene expression. Proc Natl Acad Sci U S A. 2015;112(50):15390–5.
50. Langley CH, Stevens K, Cardeno C, Lee YC, Schrider DR, Pool JE, et al. Genomic variation in natural populations of *Drosophila melanogaster*. Genetics. 2012;192(2):533–98.
51. The 1000 Genomes Project Consortium, Auton A, Brooks LD, Durbin RM, Garrison EP, Kang HM, et al. A global reference for human genetic variation. Nature. 2015;526(7571):68–74.
52. Chia JM, Song C, Bradbury PJ, Costich D, de Leon N, Doebley J, et al. Maize HapMap2 identifies extant variation from a genome in flux. Nat Genet. 2012;44(7):803–7.
53. Elyashiv E, Sattath S, Hu TT, Strutsovsky A, McVicker G, Andolfatto P, et al. A genomic map of the effects of linked selection in *Drosophila*. PLoS Genet. 2016;12(8):e1006130.
54. Karasov T, Messer PW, Petrov DA. Evidence that adaptation in Drosophila is not limited by mutation at single sites. PLoS Genet. 2010;6(6):e1000924.
55. Hudson RR, Bailey K, Skarecky D, Kwiakowski J, Ayala FJ. Evidence for positive selection in the superoxide dismutase (*Sod*) region of *Drosophila melanogaster*. Genetics. 1994;136:1329–40.
56. Sabeti PC, Reich DE, Higgins JM, Levine HZP, Richter DJ, Schaffner SF, et al. Detecting recent positive selection in the human genome from haplotype structure. Nature. 2002;419(6909):832–7.
57. Hanchard NA, Rockett KA, Spencer C, Coop G, Pinder M, Jallow M, et al. Screening for recently selected alleles by analysis of human haplotype similarity. Am J Hum Genet. 2006;78(1):153–9.
58. Sabeti PC, Varilly P, Fry B, Lohmueller J, Hostetter E, Cotsapas C, et al. Genome-wide detection and characterization of positive selection in human populations. Nature. 2007;449(7164):913–8.
59. Voight BF, Kudaravalli S, Wen X, Pritchard JK. A map of recent positive selection in the human genome. PLoS Biol. 2006;4(3):e72.
60. Wang ET, Kodama G, Baldi P, Moyzis RK. Global landscape of recent inferred Darwinian selection for Homo sapiens. Proc Natl Acad Sci U S A. 2006;103(1):135–40.
61. Pennings PS, Hermisson J. Soft Sweeps III: The signature of positive selection from recurrent mutation. PLoS Genet. 2006;2(12):e186.
62. Garud NR, Messer PW, Buzbas EO, Petrov DA. Recent selective sweeps in North American *Drosophila melanogaster* show signatures of soft sweeps. PLoS Genet. 2015;11(2):e1005004.
63. Field Y, Boyle EA, Telis N, Gao Z, Gaulton KJ, Golan D, et al. Detection of human adaptation during the past 2000 years. Science. 2016;354(6313):760–4.
64. Vitti JJ, Grossman SR, Sabeti PC. Detecting natural selection in genomic data. Annu Rev Genet. 2013;47:97–120.
65. Przeworski M. The signature of positive selection at randomly chosen loci. Genetics. 2002;160:1179–89.
66. Hermisson J, Pennings PS. Soft sweeps and beyond: understanding the patterns and probabilities of selection footprints under rapid adaptation. Methods Ecol Evol. 2017;8(6):700–16.
67. Pritchard JK, Pickrell JK, Coop G. The genetics of human adaptation: hard sweeps, soft sweeps, and polygenic adaptation. Curr Biol. 2010;20(4):R208–15.
68. Messer PW, Petrov DA. Population genomics of rapid adaptation by soft selective sweeps. Trends Ecol Evol. 2013;28(11):659–69.
69. Garud NR, Petrov DA. Elevated linkage disequilibrium and signatures of soft sweeps are common in *Drosophila melanogaster*. Genetics. 2016;203(2):863–80.
70. Jensen JD. On the unfounded enthusiasm for soft selective sweeps. Nat Commun. 2014;5:5281.
71. Berg JJ, Coop G. A population genetic signal of polygenic adaptation. PLoS Genet. 2014;10(8):e1004412.
72. Jain K, Stephan W. Rapid adaptation of a polygenic trait after a sudden environmental shift. Genetics. 2017;206(1):389–406.
73. Robertson A. A theory of limits in artificial selection. Philos Trans R Soc Lond B Biol Sci. 1960;153(951):16.
74. Kim Y, Nielsen R. Linkage disequilibrium as a signature of selective sweeps. Genetics. 2004;167(3):1513–24.
75. Jensen JD, Thornton KR, Bustamante CD, Aquadro CF. On the utility of linkage disequilibrium as a statistic for identifying targets of positive selection in nonequilibrium populations. Genetics. 2007;176(4):2371–9.
76. Messer PW, Neher RA. Estimating the strength of selective sweeps from deep population diversity data. Genetics. 2012;191(2):593–605.
77. Schrider DR, Kern AD. S/HIC: Robust identification of soft and hard sweeps using machine learning. PLoS Genet. 2016;12(3):e1005928.
78. Schrider DR, Kern AD. Soft sweeps are the dominant mode of adaptation in the human genome. Mol Biol Evol. 2017;34(8):1863–77.
79. Maynard Smith J, Haigh J. The hitch-hiking effect of a favourable gene. Genet Res. 1974;23:23–5.
80. Charlesworth B. Fundamental concepts in genetics: effective population size and patterns of molecular evolution and variation. Nat Rev Genet. 2009;10(3):195–205.
81. Braverman JM, Hudson RR, Kaplan NL, Langley CH, Stephan W. The hitchhiking effect on the site frequency spectrum of DNA polymorphisms. Genetics. 1995;140:783–96.
82. Charlesworth D, Charlesworth B, Morgan MT. The pattern of neutral molecular variation under the background selection model. Genetics. 1995;141:1619–32.
83. Hudson RR, Kaplan NL. Deleterious background selection with recombination. Genetics. 1995;141:1605–17.
84. Nordborg M, Charlesworth B, Charlesworth D. The effect of recombination on background selection. Genet Res. 1996;67:159–74.

Appendix B

Recombination in wild mice

B.1 Supplementary Material

Included here are the supplementary figures and tables for Chapter 2.

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

	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.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).

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

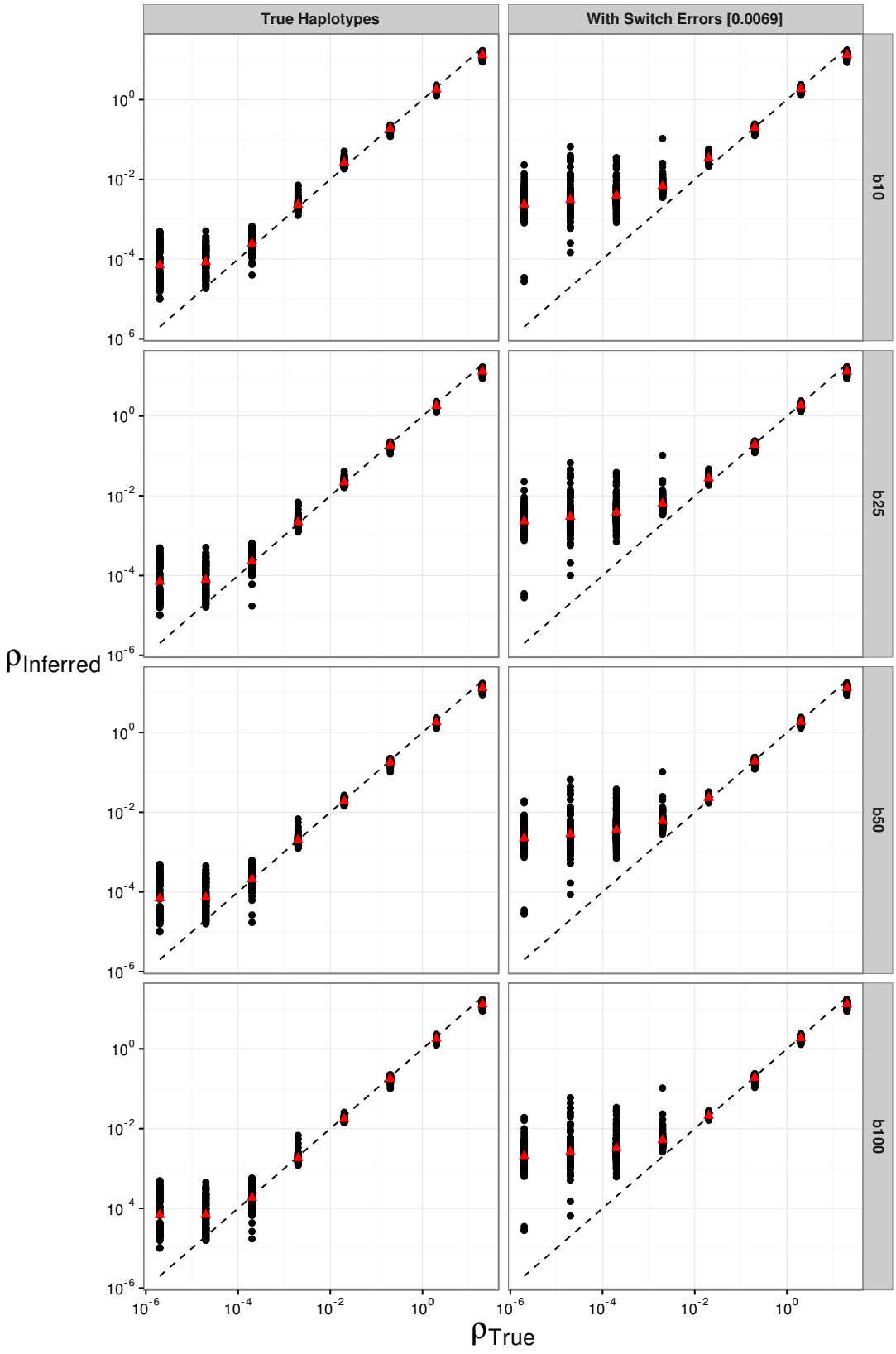
True Haplotypes
With Switch Errors [0.0069]
b10
b25
b50
b100
ρInferred
ρTrue
10^0
10^-2
10^-4
10^-6

Figure B.1

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

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.

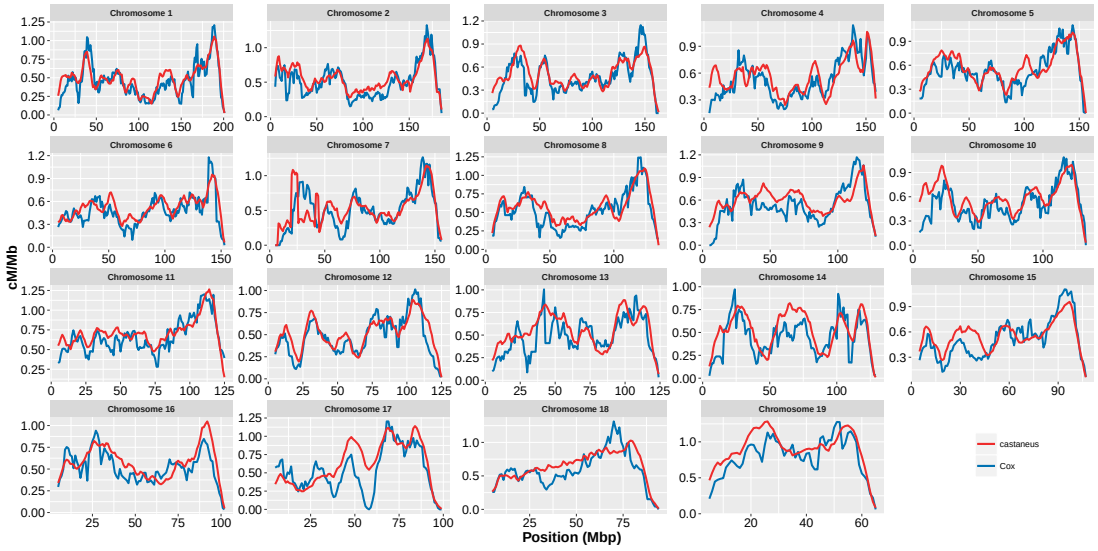

Figure B.2: The effect of switch errors on the mean recombination rate inferred using LDhelmet with a block penalty of 100. Each black point represents results for a window of 4000 SNPs, with 200 SNPs overlapping between adjacent windows, using sequences simulated in SLiM for a constant value of ρ/bp. 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

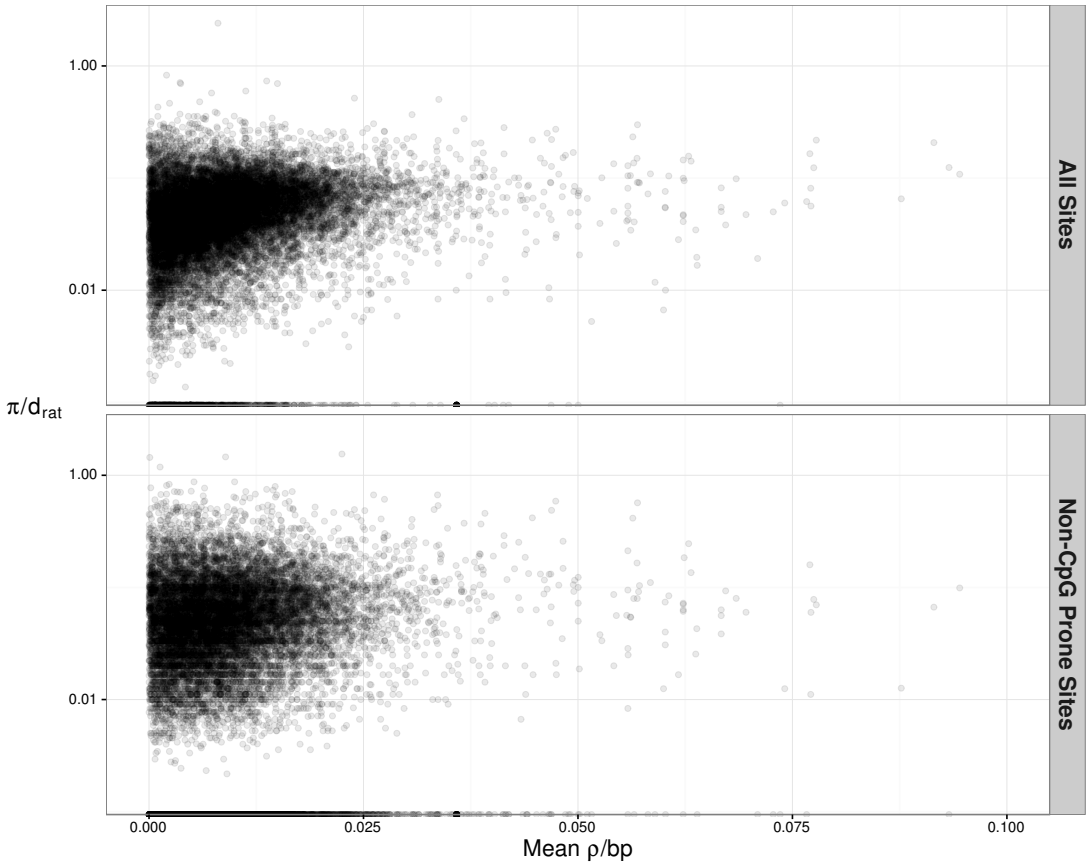

Figure B.3: The effect of switch errors on the mean recombination rate inferred using LDhelmet with a block penalty of 100. Each black point represents results for a window of 4000 SNPs, with 200 SNPs overlapping between adjacent windows, using sequences simulated in SLiM for a constant value of ρ/bp. 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

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

GENETICS | INVESTIGATION

The Recombination Landscape in Wild House Mice Inferred Using Population Genomic Data

Tom R. Booker,*,[1] Rob W. Ness,† and Peter D. Keightley*
*Institute of Evolutionary Biology, University of Edinburgh, EH9 3FL, United Kingdom and †Department of Biology, University of Toronto Mississauga, Ontario, L5L 1C6, Canada

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.

KEYWORDS *Mus musculus*; recombination; wild Mice; population genomics

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 population's 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, 2004; Auton and McVean 2007; Chan *et al.* 2012). Recombination rates estimated in this way reflect long-term variation in crossing-over in the population's 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).

doi: https://doi.org/10.1534/genetics.117.300063
Manuscript received February 27, 2017; accepted for publication July 19, 2017; published Early Online July 26, 2017.
Available freely online through the author-supported open access option.
Supplemental material is available online at www.genetics.org/lookup/suppl/doi:10.1534/genetics.117.300063/-/DC1.
[1]Corresponding author: Institute of Evolutionary Biology, University of Edinburgh, Ashworth Laboratories, Charlotte Auerbach Rd., EH9 3FL Edinburgh, UK. E-mail: t.r.booker@sms.ed.ac.uk

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.* 2016). The broad-scale recombination landscapes of the great apes are, however, strongly positively correlated (Stevison *et al.* 2011, 2016), 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 ~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) analyzed 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. Second, 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.

Materials and Methods

Polymorphism data for Mus musculus castaneus

We analyzed 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 $>20\times$ and the proportion of the genome with $>10\times$ coverage was $>80\%$ 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*.

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 pseudofemales where the true haplotypes are known (H12 + H28 = H40; H12 + H34 = H46; H28 + H34 = H62). We then jointly recalled variants in the seven female samples plus the three pseudofemales 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 pseudofemales. We filtered SNPs based on combinations of variant and genotype quality scores (QUAL and GQ, respectively) and on an individual's sequencing depth (DP) (Supplemental Material, 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 pseudofemale 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 three from the male samples). With these, we estimated the recombination rate landscape for *M. m. castaneus*.

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 2×10^{-6} to 2×10^{1} were simulated in SLiM v1.8 (Messer 2013). For each value of ρ, 0.5 Mbp of neutrally evolving sequence was simulated for populations of $N = 1000$ 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 10 Mbp of sequence for each value of ρ. 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 4400 SNPs with a 200 SNP overlap between windows. We analyzed 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 pseudofemale 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 pseudofemale

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.

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 analyzed 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 4400 SNPs, overlapping by 200 SNPs on either side were analyzed. 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 reran the LDhelmet analysis with a block penalty of 10. We analyzed all sites that passed the filters chosen using the pseudofemale 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).

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 analyzed 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 analyzed 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 $\rho = 4N_e r$. For comparison purposes, we converted these units to centimorgans per megabase using frequency-weighted means, as follows. LDhat and LDhelmet provide estimates of ρ (per kilobase pair and base pair, 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 pair's contribution to the total. By setting the total map length for each chromosome to that of Cox *et al.* (2009), we converted the cumulative ρ at each analyzed SNP position to centimorgan 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 centimorgans per megabase 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 megabase pair 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 centimorgans per megabase) in nonoverlapping windows of 1–20 Mbp for the *castaneus*, Cox and Brunschwig maps. For visual comparison of the *castaneus* and Cox maps, we plotted recombination rates in sliding windows of 10 Mbp, offset by 1 Mb.

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 analyzed 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.* (2015). We divided the genome into nonoverlapping windows of 2 kbp, and, using the maps generated with a block penalty of 10, classified as putative hotspots all windows where the recombination rate was at least 5× greater than the recombination rate in the surrounding 80 kbp. Recombination hotspots may be >2 kbp, so neighboring 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 intersubspecific 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 University of California Santa Cruz (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 analyzed by Smagulova *et al.* (2016), this procedure was repeated 1000 times. The maximum number of overlapping DSB and putative *castaneus* hotspots observed across all 1000 replicates was taken as an ∼0.1% significance threshold.

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 fourfold degenerate synonymous sites and current GC content for each gene. If a site was annotated as fourfold 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 fourfold 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 nonparametric 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).

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_recombination-maps.

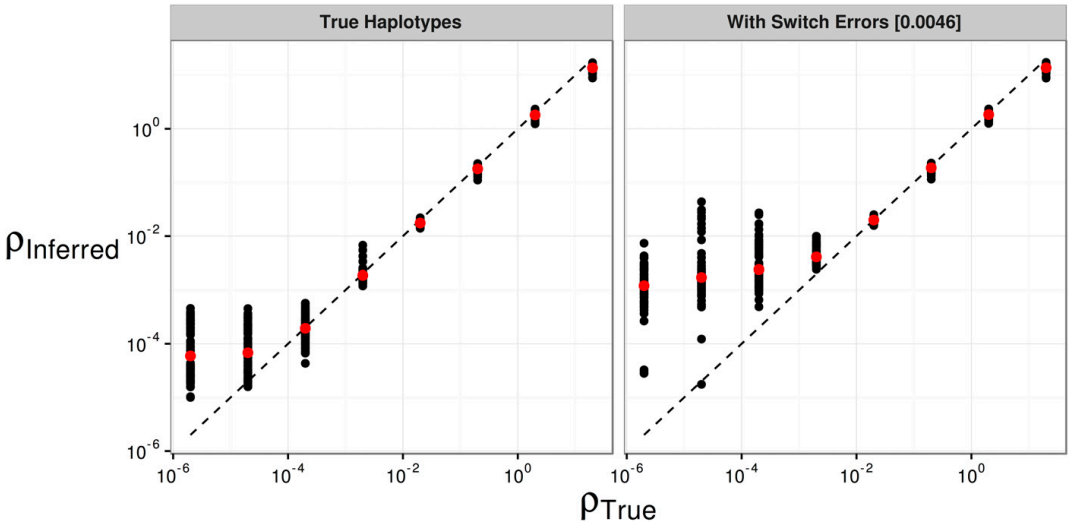

Figure 1 The effect of switch errors on the mean recombination rate inferred using LDhelmet with a block penalty of 100. Each black point represents results for a window of 4000 SNPs, with 200 SNPs overlapping between adjacent windows, using sequences simulated in SLiM for a constant value of ρ/bp. Red 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.

Results

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 pseudofemale 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.

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 ρ/bp $= 2 \times 10^{-4}$. 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 2×10^{-3} to 2×10^{2}. When the true ρ/bp was $<2 \times 10^{-3}$, however, LDhelmet overestimated the mean recombination rate for 0.5 Mbp 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 (Figure S2 and Table 1).

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 analyzed 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

Table 1 Summary of sex-averaged recombination rates estimated for the *M. m castaneus* autosomes compared with published rates

		castaneus		Brunschwig[b]	
Chromosome	Cox[a] cM/Mb	Freq. Weighted Mean	N_e Estimate	Freq. Weighted Mean	N_e Estimate
1	0.50	0.0079	395,000	0.000015	745
2	0.57	0.0088	386,000	0.000015	653
3	0.52	0.0083	400,000	0.000014	693
4	0.56	0.0091	408,000	0.000020	889
5	0.59	0.0090	382,000	0.000015	646
6	0.53	0.0089	421,000	0.000015	728
7	0.58	0.0100	429,000	0.000019	801
8	0.58	0.0094	404,000	0.000014	610
9	0.61	0.0096	394,000	0.000018	749
10	0.61	0.0096	392,000	0.000023	928
11	0.70	0.0102	365,000	0.000019	689
12	0.53	0.0089	420,000	0.000019	897
13	0.56	0.0095	426,000	0.000014	629
14	0.53	0.0084	395,000	0.000013	632
15	0.56	0.0083	371,000	0.000024	1080
16	0.59	0.0091	386,000	0.000017	721
17	0.65	0.0087	335,000	0.000052	2020
18	0.66	0.0098	371,000	0.000021	785
19	0.94	0.0122	323,000	0.000026	681
X	0.48	0.0026	137,000	—	—
Mean		0.0092		0.000020	

Rates for the *castaneus* and Brunschwig maps are presented in terms of $4N_er$/bp. Estimates of N_e were obtained by assuming the recombination rates from Cox *et al.* (2009).
[a] Cox *et al.* (2009)
[b] Brunschwig *et al.* (2012)

provides us with a baseline for which we have two *a priori* expectations. First, 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. Second, 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 centimorgans per megabase pair) 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.

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

We then compared intrachromosomal 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 ≥8 Mbp (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 analyzed 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 ∼500× 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 ∼100,000 and 500,000, respectively (Geraldes *et al.* 2008, 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).

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

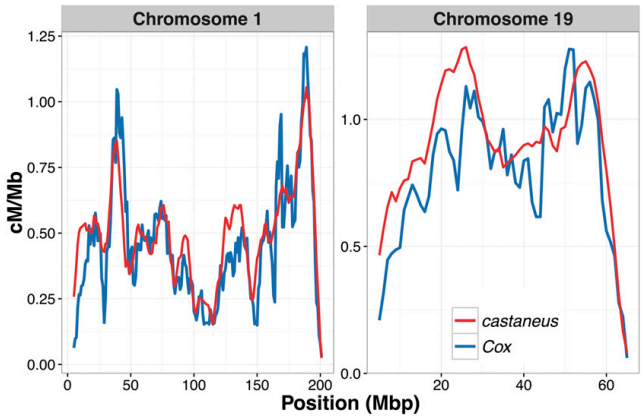

Figure 2 Comparison of sex-averaged recombination rates for chromosomes 1 and 19 of *M. musculus castaneus* inferred by LDhelmet (red) with rates estimated in the pedigree-based study of Cox *et al.* (2009) (blue). Recombination rates were scaled to units of centimorgans per megabase for the *castaneus* map by setting the total map length of each chromosome to the corresponding map length of Cox *et al.* (2009).

chromosome, there was an average of 15 hotspots per megabase pair. 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 Rockman's (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; <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 analyzed by Smagulova *et al.* (2016).

Correlation between recombination rate and properties of protein coding genes

There is evidence of pervasive natural selection acting in protein-coding genes and conserved noncoding elements of the murid genome (Halligan *et al.* 2010, 2011, 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 (Figure S6 and Table 2). 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 DSBs, 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 fourfold 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

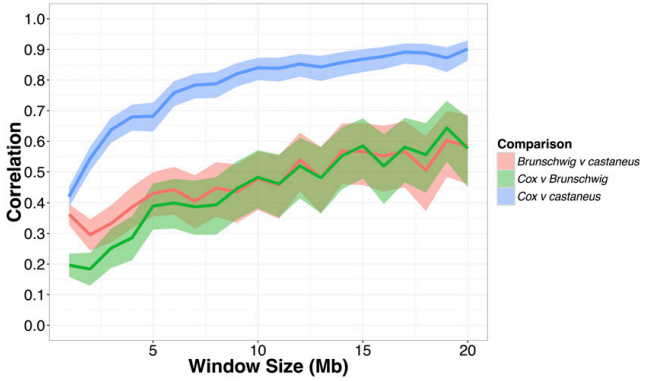

Figure 3 Pearson correlation coefficients between the recombination map inferred for *M. m. castaneus*, the Brunschwig *et al.* (2012) map and the Cox *et al.* (2009) map. Correlations were calculated in nonoverlapping windows of varying size across all autosomes. Confidence intervals (95%) are indicated by shading.

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.

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 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, ~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.* 2016). 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.* (2016) examined pairs of great ape species, and found that correlations between recombination maps (at the 1 Mbp 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 1 Mbp scale. Although this is only a single data point, it suggests that recombination

Table 2 Correlation coefficients between recombination rate and pairwise nucleotide diversity and divergence from the rat at fourfold degenerate sites for protein coding genes

	Correlation Coefficient	
	Non-CpG Prone Sites	All Sites
Nucleotide diversity (π)	0.090	0.20
Divergence from rat (d_{rat})	−0.038	0.062
Corrected diversity (π/d_{rat})	0.10	0.18

Nonparametric Kendall correlations were calculated for non-CpG prone sites and for all sites, regardless of base context. All coefficients shown are highly significant ($P < 10^{-10}$).

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.* (2016) 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. First, we used a much larger number of markers to resolve recombination rates than Brunschwig *et al.* (2012). Second, 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 >8 Mbp (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 2×10^{-3} ρ/bp (Figure 1 and 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 three-quarters 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 π/d_{rat} and recombination rate is 0.20 for all fourfold 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 and 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 nonsignificant for both *M. m. castaneus* and *M. m. domesticus* (Geraldes *et al.* 2011). This study analyzed 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).

Acknowledgments

We are grateful to Ben Jackson, Bettina Harr, Dan Halligan, Rory Craig, and two anonymous reviewers for comments on the manuscript. We thank Galina Petukhova and Kevin Brick for help with the double-strand break data from their 2016 study. T.B. is supported by a Biotechnology and Biological Sciences Research Council (BBSRC) East of Scotland BioScience Doctoral Training Partnership (EASTBIO) studentship. This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program (grant agreement No. 694212). R.W.N. was funded by the BBSRC (BB/L00237X/1).

Literature Cited

Auton, A., and G. McVean, 2007 Recombination rate estimation in the presence of hotspots. Genome Res. 17: 1219–1227.

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

Baudat, F., J. Buard, C. Grey, A. Fledel-Alon, C. Ober *et al.*, 2010 PRDM9 is a major determinant of meiotic recombination hotspots in humans and mice. Science 327: 836–840.

Baudat, F., Y. Imai, and B. de Massy, 2013 Meiotic recombination in mammals: localization and regulation. Nat. Rev. Genet. 14: 794–806.

Brick, K., F. Smagulova, P. Khil, R. D. Camerini-Otero, and G. V. Petukhova, 2012 Genetic recombination is directed away from functional genomic elements in mice. Nature 485: 642–645.

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

Cai, J. J., J. M. Macpherson, G. Sella, and D. A. Petrov, 2009 Pervasive hitchhiking at coding and regulatory sites in humans. PLoS Genet. 5: e1000336.

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

Cox, A., C. L. Ackert-Bicknell, B. L. Dumont, Y. Ding, J. T. Bell *et al.*, 2009 A new standard genetic map for the laboratory mouse. Genetics 182: 1335–1344.

Cutter, A. D., and B. A. Payseur, 2013 Genomic signatures of selection at linked sites: unifying the disparity among species. Nat. Rev. Genet. 14: 262–274.

Davies, B., E. Hatton, N. Altemose, J. G. Hussin, F. Pratto *et al.*, 2016 Re-engineering the zinc fingers of PRDM9 reverses hybrid sterility in mice. Nature 530: 171–176.

Delaneau, O., B. Howie, A. J. Cox, J. F. Zagury, and J. Marchini, 2013 Haplotype estimation using sequencing reads. Am. J. Hum. Genet. 93: 687–696.

Dumont, B. L., and B. A. Payseur, 2011 Genetic analysis of genomic-scale recombination rate evolution in house mice. PLoS Genet. 7: 11.

Dumont, B. L., M. A. White, B. Steffy, T. Wiltshire, and B. A. Payseur, 2011 Extensive recombination rate variation in the house mouse species complex inferred from genetic linkage maps. Genome Res. 21: 114–125.

Duret, L., and P. F. Arndt, 2008 The impact of recombination on nucleotide substitutions in the human genome. PLoS Genet. 4: e1000071.

Duret, L., and N. Galtier, 2009 Biased gene conversion and the evolution of mammalian genomic landscapes. Annu. Rev. Genomics Hum. Genet. 10: 285–311.

Geraldes, A., P. Basset, B. Gibson, K. L. Smith, B. Harr *et al.*, 2008 Inferring the history of speciation in house mice from autosomal, X-linked, Y-linked and mitochondrial genes. Mol. Ecol. 17: 5349–5363.

Geraldes, A., P. Basset, K. L. Smith, and M. W. Nachman, 2011 Higher differentiation among subspecies of the house mouse (*Mus musculus*) in genomic regions with low recombination. Mol. Ecol. 20: 4722–4736.

Grey, C., P. Barthes, G. Chauveau-Le Friec, F. Langa, F. Baudat *et al.*, 2011 Mouse PRDM9 DNA-binding specificity determines sites of histone h3 lysine 4 trimethylation for initiation of meiotic recombination. PLoS Biol. 9: e1001176.

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

Halligan, D. L., F. Oliver, J. Guthrie, K. C. Stemshorn, B. Harr *et al.*, 2011 Positive and negative selection in murine ultraconserved noncoding elements. Mol. Biol. Evol. 28: 2651–2660.

Halligan, D. L., A. Kousathanas, R. W. Ness, B. Harr, L. Eory *et al.*, 2013 Contributions of protein-coding and regulatory change to adaptive molecular evolution in murid rodents. PLoS Genet. 9: e1003995.

Hudson, R. R., 2001 Two-locus sampling distributions and their applications. Genetics 159: 12.

Jensen-Seaman, M. I., T. S. Furey, B. A. Payseur, Y. Lu, K. M. Roskin *et al.*, 2004 Comparative recombination rates in the rat, mouse and human genomes. Genome Res. 14: 528–538.

Johnston, S. E., C. Berenos, J. Slate, and J. M. Pemberton, 2016 Conserved genetic architecture underlying individual recombination rate variation in a wild population of soay sheep (*Ovis aries*). Genetics 203: 583–598.

Kaur, T., and M. V. Rockman, 2014 Crossover heterogeneity in the absence of hotspots in *Caenorhabditis elegans*. Genetics 196: 137–148.

Kono, H., M. Tamura, N. Osada, H. Suzuki, K. Abe *et al.*, 2014 PRDM9 polymorphism unveils mouse evolutionary tracks. DNA Res. 21: 315–326.

Kousathanas, A., D. L. Halligan, and P. D. Keightley, 2014 Faster-X adaptive protein evolution in house mice. Genetics 196: 1131–1143.

Lesecque, Y., S. Glemin, N. Lartillot, D. Mouchiroud, and L. Duret, 2014 The red queen model of recombination hotspots evolution in the light of archaic and modern human genomes. PLoS Genet. 10: e1004790.

Li, H., and R. Durbin, 2009 Fast and accurate short read alignment with burrows-wheeler transform. Bioinformatics 25: 1754–1760.

Li, H., B. Handsaker, A. Wysoker, T. Fennell, J. Ruan *et al.*, 2009 The sequence alignment/map format and samtools. Bioinformatics 25: 2078–2079.

Liu, E. Y., A. P. Morgan, E. J. Chesler, W. Wang, G. A. Churchill *et al.*, 2014 High-resolution sex-specific linkage maps of the mouse reveal polarized distribution of crossovers in male germline. Genetics 197: 91–106.

McVean, G., P. Awadalla, and P. Fearnhead, 2002 A coalescent-based method for detecting and estimating recombination from gene sequences. Genetics 160: 1231–1241.

McVean, G., S. R. Myers, S. Hunt, P. Deloukas, D. R. Bentley *et al.*, 2004 The fine-scale structure of recombination rate variation in the human genome. Science 304: 581–584.

Messer, P. W., 2013 SLiM: simulating evolution with selection and linkage. Genetics 194: 1037–1039.

Myers, S. R., R. Bowden, A. Tumian, R. E. Bontrop, C. Freeman *et al.*, 2010 Drive against hotspot motifs in primates implicates the PRDM9 gene in meiotic recombination. Science 327: 876–879.

Paigen, K., and P. Petkov, 2010 Mammalian recombination hot spots: properties, control and evolution. Nat. Rev. Genet. 11: 221–233.

Paigen, K., J. P. Szatkiewicz, K. Sawyer, N. Leahy, E. D. Parvanov *et al.*, 2008 The recombinational anatomy of a mouse chromosome. PLoS Genet. 4: e1000119.

Phung, T. N., C. D. Huber, and K. E. Lohmueller, 2016 Determining the effect of natural selection on linked neutral divergence across species. PLoS Genet. 12: e1006199.

Quinlan, A. R., and I. M. Hall, 2010 Bedtools: a flexible suite of utilities for comparing genomic features. Bioinformatics 26: 841–842.

R Core Team, 2016 *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna.

Schwartz, J. J., D. J. Roach, J. H. Thomas, and J. Shendure, 2014 Primate evolution of the recombination regulator PRDM9. Nat. Commun. 5: 4370.

Singhal, S., E. Leffler, K. Sannareddy, I. Turner, O. Venn *et al.*, 2015 Stable recombination hotspots in birds. Science 350: 6.

Smagulova, F., I. V. Gregoretti, K. Brick, P. Khil, R. D. Camerini-Otero *et al.*, 2011 Genome-wide analysis reveals novel molecular features of mouse recombination hotspots. Nature 472: 375–378.

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

Smukowski, C. S., and M. A. Noor, 2011 Recombination rate variation in closely related species. Heredity (Edinb) 107: 496–508.

Smukowski Heil, C. S., C. Ellison, M. Dubin, and M. A. Noor, 2015 Recombining without hotspots: a comprehensive evolutionary portrait of recombination in two closely related species of drosophila. Genome Biol. Evol. 7: 2829–2842.

Stevison, L. S., K. B. Hoehn, and M. A. Noor, 2011 Effects of inversions on within- and between-species recombination and divergence. Genome Biol. Evol. 3: 830–841.

Stevison, L. S., A. E. Woerner, J. M. Kidd, J. L. Kelley, K. R. Veeramah *et al.*, 2016 The time scale of recombination rate evolution in great apes. Mol. Biol. Evol. 33: 928–945

Wang, J. R., F. P. de Villena, and L. McMillan, 2012 Comparative analysis and visualization of multiple collinear genomes. BMC Bioinformatics 13: S13.

Wang, R. J., M. M. Gray, M. D. Parmenter, K. W. Broman, and B. A. Payseur, 2017 Recombination rate variation in mice from an isolated island. Mol. Ecol. 26: 457–470.

Winckler, W., S. R. Myers, D. J. Richter, R. C. Onofrio, G. J. McDonald *et al.*, 2005 Comparison of fine-scale recombination rates in humans and chimpanzees. Science 308: 107–111.

Yang, H., J. R. Wang, J. P. Didion, R. J. Buus, T. A. Bell *et al.*, 2011 Subspecific origin and haplotype diversity in the laboratory mouse. Nat. Genet. 43: 648–655.

Communicating editor: B. Payseur

Appendix C

Understanding patterns of genetic diversity

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):

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

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 Ne 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.