THE UNIVERSITY *of* EDINBURGH

The Fitness Effects of New Mutations and Adaptive Evolution in House Mice

Athanasios Kousathanas

Submitted for the degree of Doctor of Philosophy

University of Edinburgh

2013

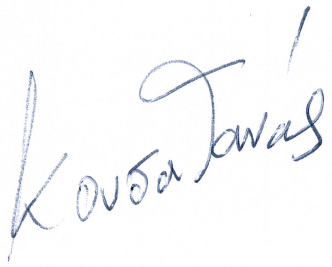
To my teachers Manolis Ladoukakis and Eleftherios Zouros

Declaration

I hereby declare that this thesis has been composed by myself, that the work contained herein is my own, and that no part of the thesis has been submitted for any other degree or professional qualification.

Athanasios Kousathanas,

Edinburgh, October 2013



Acknowledgements

Firstly, I would like to thank my supervisor Peter Keightley for his great help and mentorship, especially during the early stages of my PhD. I thank Peter for his patience going through countless versions of my manuscripts and for providing comments, a process which greatly improved my reasoning and writing skills over the years.

Many thanks go to Dan Halligan for helping me throughout my PhD with programming and data analysis, for very helpful discussions and comments on all of my manuscripts.

I thank Brian Charlesworth, Adam Eyre-Walker and Paul Sharp for criticism, discussions, and very helpful comments on my work.

I thank Matthew Hartfield, Rob Ness and Andrew Morgan, which were members of the Keightley group over the years, and also the members of the Genetics lab meeting, for helpful comments on my work and discussions.

I thank Richard Ennos, Trevor Bedford and Crispin Jordan for encouragement and for infusing me with inspiration and enthusiasm for science.

I thank my friends Georgios Koutsovoulos, Jose Campos, Roberta Bergero, Richard Perry and Sara Guirao Rico for emotional support and for being there to listen in my brightest and darkest moments.

I thank my family for encouragement and support.

I acknowledge that funding for this PhD was provided by a studentship awarded jointly by the Biotechnology and Biological Sciences Research Council (BBSRC) and the School of Biological Sciences-University of Edinburgh. I acknowledge funding for three months from a BBSRC grant awarded to Peter Keightley. I acknowledge funding from the Genetics Society UK to attend the 2012 meeting of the Society for Molecular Biology and Evolution in Dublin.

Publications

The following published papers have arisen from this thesis.

- Kousathanas A, Oliver F, Halligan DL, Keightley PD. 2011. Positive and negative selection on noncoding DNA close to protein-coding genes in wild house mice. *Molecular Biology and Evolution* **28**: 1183 –1191.
- Kousathanas A, Keightley PD. 2013. A comparison of models to infer the distribution of fitness effects of new mutations. *Genetics* **193**: 1197–1208.

The following paper has arisen from this thesis and has been submitted to *Genetics*.

- Kousathanas A, Halligan DL, Keightley PD. Faster-X adaptive protein evolution in house mice.

The author of this thesis has contributed to the following work which has been submitted to *PloS Genetics*.

- Halligan DL, Kousathanas A, Ness RW, Li H, Harr B, Eory L, Keane TM, Adams DJ, Keightley PD. Contributions of protein-coding and regulatory change to adaptive molecular evolution in murid rodents.

Table of Contents

Abstract of thesis

Knowledge of the distribution of fitness effects of new mutations (DFE) can enable us to quantify the amount of genetic change between species that is driven by natural selection and contributes to adaptive evolution. The primary focus of this thesis is the study of methods to infer the DFE and the study of adaptive evolution in the house mouse subspecies *Mus musculus castaneus*.

Firstly, I extended previous methodology to model the DFE based on polymorphism data. Methods that have previously been used to infer the DFE from polymorphism data have relied on the assumption of a unimodal distribution. I developed new models that can be used to fit DFEs of arbitrary complexity, and found that multimodality can be detected by these models given enough data. I used these new models to analyse polymorphism data from *Drosophila melanogaster* and *M. m. castaneus*, and found evidence for a unimodal DFE for *D. melanogaster* and a bimodal DFE for *M. m. castaneus*.

Secondly, I investigated the contribution of change in coding and non-coding DNA to evolutionary adaptation. I used a polymorphism dataset of ~80 loci from *M. m. castaneus* sequenced in 15 individuals to investigate selection in protein-coding genes and putatively regulatory DNA close to these genes. I found that, although protein-coding genes are much more selectively constrained than non-coding DNA, they experience similar rates of adaptive substitution. These results suggest that change in functional non-coding DNA sequences might be as important as

protein-coding genes to evolutionary adaptation.

Thirdly, I used whole genome data from 10 *M. m. castaneus* individuals to compare the rate of adaptive substitution in autosomal and X-linked genes. I found that, on average, X-linked genes have a 1.8 times faster rate of adaptive substitution than autosomal genes. I also found that faster-X evolution is more pronounced for male-specific genes. I used previously developed theory to show that these observations can be explained if new advantageous mutations are recessive, with an average dominance coefficient less than or equal to 0.25. These results can help to explain the long-studied phenomenon of the large effect of the X chromosome in speciation.

Chapter 1. General Introduction

In the general introduction I will firstly outline some basic principles of population genetics that are necessary for understanding the research presented in the thesis. Secondly, I will present the basic methodology to detect natural selection from comparison of DNA sequences. Thirdly, I will present the basic research subjects that have motivated the original research presented in the thesis.

1.1 Basic Principles.

Different forms of genes in a population are called alleles. Evolutionary forces change allele frequencies over time. These forces are mutation, gene flow, random genetic drift and natural selection. If no evolutionary forces act in a population, allele frequencies remain constant over time (Hardy-Weinberg principle). I briefly outline below the four evolutionary forces and their effect on allele frequency change and the maintenance of variation. I also describe some important concepts in population genetics such as the effective population size and the fixation probability of mutations.

Mutation. The total amount of hereditary information of an organism constitutes its genome and is stored in the deoxyribonucleic acid (DNA) as a sequence of four types of nucleotides: adenine (A), thymine (T), cytosine (C) and guanine (G). Adenine and guanine are purines, and thymine and cytosine are

pyrimidines.

Mutation is a heritable change in the sequence of DNA. Mutations can arise spontaneously, by errors of enzymes that process DNA (e.g. sequence repair or replication enzymes) or by physical or chemical mutagens (e.g. ultraviolet radiation). There are large-scale mutations that can affect whole chromosomes (e.g. inversions) and small scale mutations that affect small genomic segments (e.g. indels) or single nucleotides (i.e. point mutations). In this thesis, all references to mutations are for point mutations, unless otherwise stated. Point mutations cause the replacement of a single nucleotide base pair and can be either transitions (purine to purine or pyrimidine to pyrimidine), or transversions (purine to pyrimidine and *vice versa*). The rate at which mutations appear can vary between different regions of the genome and between species. For example, the mutation rate per generation at methylated CpG sites in mammals is ~10 times higher than non-CpG sites (Arndt et al. 2003) and humans have a ~10-fold higher mutation rate per generation than *Drosophila melanogaster* (Haag-Liautard et al. 2007; Keightley 2012).

Mutation generates new alleles of genes in populations and is the ultimate source of genetic variation. However, mutation is a very slow process in changing allele frequencies because the mutation rate per generation (μ) is very small for most organisms (Charlesworth and Charlesworth 2010). Assuming a locus with two alleles, A and a, with frequencies p and q, the forward mutation rate (A to a) is μ and the backward mutation rate (a to A) is v. In the absence of other evolutionary forces, and after a number of generations, an equilibrium is reached where $p=v/(\mu+v)$ and $q=\mu/(\mu+v)$.

Gene flow. Gene flow refers to the movement of alleles between populations. Gene flow occurs when individuals from one population migrate to breed with individuals of another population of the same species. In this thesis, I will not consider the effect of gene flow on allele frequency change since I will be studying evolution in single, non-subdivided populations.

Random genetic drift and effective population size. Random genetic drift is the process of change in allele frequency due to random sampling between generations in finite populations. Through genetic drift, the frequency of alleles fluctuates randomly over time, which can eventually lead to their loss or their fixation (i.e. obtain a frequency of 0 or 1, respectively). Therefore, genetic drift causes the passive loss of genetic variability over time.

Wright and Fisher modelled random genetic drift as binomial sampling between successive non-overlapping generations (Fisher 1930; Wright 1931). Assuming a locus with two alleles, *A* and *a*, with frequencies *p* and *q*, respectively, in a diploid population of size *N,* the probability that a sample of 2*N* gametes contains *j* alleles of type A can be calculated by the binomial distribution as:

$$P(j;2N,p)=\binom{2N}{j}p^{j}q^{2N-j} \tag{1.1}$$

One way to predict allele frequency change between generations that is caused by random genetic drift is to use equation 1.1 to build a transition probability matrix **M**, where **M** describes the probability of transitioning from any state *i* with

allele frequency $i/2N$ (i=0..2N) to any state j with allele frequency $j/2N$ (j=0..$2N$). By setting up a vector that describes the initial allele frequency distribution and by performing matrix multiplication we can obtain the allele frequency distribution at any generation. An alternative way to describe allele frequency change due to random genetic drift is to use the diffusion theory approximation (Fisher 1930; Ewens 1979; Kimura 1985).

The effective population size (N_e) is a central concept in population genetics and can be used to approximate the rate of random genetic drift. N_e is the number of breeding individuals in an idealised population that experiences the same amount of random genetic drift as the focal population (Wright 1931). N_e is usually smaller than the census size of a population (N). For example, if the population consists of many more females than males, N_e is closer to the number of males rather than the sum of males and females. This is because half of the alleles must come from either males or females. Moreover, N_e may be much smaller than N, if the population has experienced size changes in the past. For example, although the current human population size is ~7 billion, its N_e has been estimated to be ~10,000-20,000 (Yu et al. 2004). This is because human populations had historically much smaller size than the present and they may have also experienced frequent and severe bottlenecks. An estimate of the historical N_e for a population can be obtained from the pairwise nucleotide diversity (π) at loci that evolve neutrally and μ. Assuming that the population has not experienced any recent population size changes and that there is no population subdivision, we can calculate N_e by using the following equation:

$$N_e = \frac{\pi_{neutral}}{4\mu} \quad (1.2)$$

(Charlesworth 2009).

Natural selection. Individuals that carry different alleles can differ in their ability to survive and reproduce (i.e. they have different fitnesses). Natural selection is the process where alleles change in frequency due to the differential reproduction of their bearers. Natural selection is the only known evolutionary force which can lead to the adaptation of a population to its environment.

There are many types of natural selection. In this thesis I mainly study directional selection, which occurs when one homozygote for one of the alleles segregating in the population has the highest fitness. For a simple case of a locus with two alleles, the intensity of selection, or equivalently the selection coefficient, *s*, is equal to the difference in fitness between the homozygotes for the alternative alleles.

Directional selection leads either to the purging of deleterious alleles (purifying selection) or the fixation of advantageous alleles (positive selection) in the population. Directional selection causes the active loss of genetic variability over time. It should be noted, however, that selection does not always lead to loss of genetic variability. Balancing selection is a type of selection that actively maintains genetic variability and can manifest with several mechanisms. One such mechanism is heterozygote advantage, i.e. when individuals with heterozygous genotypes have higher fitness than individuals with homozygous genotypes.

Fixation probability of a new mutation, substitution and evolutionary

rate. Given sufficient time, a mutation will either be fixed or lost from the population. We can calculate the fixation probability (u) of a mutation given its initial frequency (p), the effective population size (N_e) and the strength of selection (s) as follows:

$$u=\frac{1-e^{-2N_e sp}}{1-e^{-2N_e s}} \tag{1.3}$$

(Fisher 1930; Kimura 1962). Note that equation 1.3 refers to weak selection (s<0.1).

We can consider some useful results that can be derived from equation 1.3. A new mutation has a frequency $1/2N$ in a diploid population. If s =0 (i.e. the mutation is neutral), its fate is entirely determined by random genetic drift, and u=$1/2N$. If s≠0, the fate of the mutation will be determined by drift and selection. The trajectory of mutations with $|s|\gg 1/N_e$ is deterministic: deleterious mutations never get fixed ($u\rightarrow 0$) and advantageous mutations have u=$2s$ (Haldane 1927; Kimura 1985). For mutations with $|s|\ll 1/N_e$ ('nearly neutral'), selection and drift both affect u. Therefore it is possible for mutations that are slightly deleterious to be fixed, especially in populations with small N_e.

Mutations that are eventually fixed are called substitutions and contribute to differences between populations. The substitution rate per nucleotide site per year λ is proportional to N, μ and u:

$$\lambda=2N\mu u \tag{1.4}$$

(Kimura 1985).

By using equation 1.4 and assuming that N=N_e, we can make simple predictions for λ. For neutral mutations (s=0), λ=μ. Therefore, λ in genomic regions that receive only

neutral mutations is independent of N_e. Moderately and strongly deleterious mutations ($s<-1/N_e$), do not contribute to λ. Slightly deleterious mutations (with $-1/N_e<s<0$) contribute to λ at a lower rate than neutral mutations. For slightly deleterious mutations, λ is faster in populations with small N_e and slower in populations with large N_e. For advantageous mutations, $\lambda=4N_e\mu s$, therefore λ is faster in populations with large N_e and slower in populations with small N_e. These results are very useful for generating expectations to test with real data from populations that differ in N_e and for determining the contribution of neutral, slightly deleterious and advantageous mutations in molecular evolution (see also section on selectionist-neutralist controversy).

Genetic linkage and selection. The fate of a mutation can be affected by the genetic background on which it arises. For example the spread of an advantageous mutation might be retarded if it is genetically linked to a deleterious mutation. The phenomenon of interference between genetically linked loci that are under selection is known as Hill-Robertson interference (HRI) (Hill and Robertson 1966). Another related phenomenon is genetic hitch-hiking where the change in frequency of an allele is affected by selection on linked alleles. Positively selected alleles on their way to fixation drag along neutral or even slightly deleterious alleles (selective sweep; Smith and Haigh 1974). Correspondingly, neutral alleles that are linked to deleterious alleles are removed from the population by selection (background selection; Charlesworth et al. 1993).

1.2 Detecting selection using DNA sequence data.

In this section I describe the basic principles of methods that can be used to detect the footprint of natural selection from the analysis of DNA sequence data. I focus on methods that were used to generate the original research presented in the thesis.

1.2.1 Comparison of between-species divergence.

The number of differences between homologous DNA sequences of two species (D) over the total number of sites compared (n_{sites}) can be used to calculate per site divergence(d):

$$d = \frac{D}{n_{sites}} \tag{1.5}$$

By calculating d using Equation 1.5 we ignore the possibility of multiple per-site substitutions, which is likely when the compared species are evolutionarily distant. We can take account of multiple substitutions by applying a correction to d as follows:

$$d = -\frac{3}{4}\ln\left(1-\frac{4D}{3n_{sites}}\right) \tag{1.6}$$

(Jukes and Cantor 1969).

There are other methods that can correct for more factors that can bias estimation of d, such as unequal frequencies of transition and transversion types of mutations (Graur and Li 2000).

To test for selection, we compare d for a region that we assume is evolving neutrally (the expected d; d_E) with the observed d in the focal region (d_O). If the ratio

d_O/d_E is significantly different from 1, then the focal region is likely to be under some form of selection. When $d_O/d_E<1$ then the focal region, or part of it, is under purifying selection. The fraction of sites in the focal region that are under purifying selection (the so-called selective constraint C) can be quantified as $1-d_O/d_E$. In contrast, when $d_O/d_E>1$, the focal region is likely to have been the target of positive selection.

Defining an appropriate 'neutral' class of sites to obtain d_E is critical, and there are two main caveats when doing so. Firstly, the neutral class must have the same mutation rate as the focal class. Secondly, we must have an *a priori* biological expectation that mutations that occur in the neutral class are functionally equivalent. Changes in synonymous sites of protein-coding genes do not result in changes in the amino-acid sequence of the protein and therefore are expected to evolve neutrally. For this reason, researchers who study the evolution of protein-coding genes usually compare non-synonymous divergence (d_N) to synonymous divergence (d_S) to test for selection. d_N is frequently smaller than d_S for most protein-coding genes and for a variety of species (Graur and Li 2000). Protein-coding genes with $d_N>d_S$ are rare. Such cases are mostly genes with reproductive (Swanson and Vacquier 2002) or immunity function (Tanaka and Nei 1989), which presumably evolve very rapidly due to sexual selection and host-parasite evolutionary arms races, respectively.

1.2.2 Comparison of nucleotide diversity.

Genetic variation is usually measured as nucleotide diversity at the DNA level. Given a sample of n alleles from a population, nucleotide diversity (π) can be quantified as:

$$\pi = 2\sum_{i}^{n}\sum_{j}^{i-1} x_i x_j \pi_{ij} \quad (1.7)$$

where x_i and x_j are the frequencies of the i^{th} and j^{th} sequences in the population and π_{ij} is the number of differences between the i^{th} and j^{th} sequences (Nei and Li 1979). Genomic regions that are under purifying selection are expected to display lower diversity than neutrally evolving regions (Kimura 1985). Therefore, comparison of nucleotide diversity of a focal genomic region with the diversity in a region assumed to evolve neutrally can allow the detection of selection. For protein-coding genes, diversity at nonsynonymous (π_N) and synonymous sites (π_S) is usually compared to infer selection. In a similar fashion to the comparison of divergence, $\pi_N/\pi_S < 1$ signifies the action of purifying selection on the gene. π_N/π_S can be a better statistic than d_N/d_S for detecting purifying selection and quantifying selective constraint when the gene under consideration is under both positive and negative selection. For example, if part of a gene is undergoing very rapid adaptive evolution and another part is under strong selective constraint, d_N/d_S might be close to 1, which suggests falsely that the gene is neutrally evolving. In contrast, π_N/π_S is robust to the action of positive selection because mutations that are under positive selection spend little time segregating and they do not contribute substantially to polymorphism (Nei 1987). However, π_N/π_S can be greater than 1 if some form of balancing selection is acting on the non-synonymous sites of the gene (Charlesworth 2006). Such cases are believed to be rare, for example for the major histocompatibility complex family of genes (MHC) in vertebrates (Piertney and Oliver 2005).

1.2.3 Testing for positive selection: The McDonald and Kreitman test.

The McDonald and Kreitman (MK) test can be used to test for non-neutral evolution of a locus and to detect positive selection. The MK test is a comparison of between species divergence and within species polymorphism at a focal class of sites (e.g. nonsynonymous sites of protein-coding genes) versus a class of sites that is assumed to be neutrally evolving (e.g. synonymous sites). To perform the test, we construct a 2X2 contingency table test with counts of polymorphic and divergent sites of the focal and neutral classes of sites (Table 1.1). In this table, P_N and P_S are the counts of nonsynonymous and synonymous sites, respectively, with intraspecific differences and D_N and D_S are the counts of nonsynonymous and synonymous sites, respectively, with interspecific differences.

Table 1.1. The 2X2 contingency table of the McDonald and Kreitman test (McDonald and Kreitman 1991). The null neutral hypothesis that is tested predicts that $P_N/P_S=D_N/D_S$..

Site class	Polymorphism	Divergence
Nonsynonymous (selected)	P_N	D_N
Synonymous (neutral)	P_S	D_S

A significant excess of divergence relative to polymorphism at the focal class (i.e. $D_N/D_S>P_N/P_S$) is interpreted as evidence for positive selection. The MK test can be

extended to estimate the proportion of substitutions that have been fixed by positive selection (α):

$$\alpha=1-\frac{D_S P_N}{D_N P_S} \quad (1.8)$$

(Fay et al. 2002; Smith and Eyre-Walker 2002).

As McDonald and Kreitman have noted in their original study, past population size changes can generate a significant MK test in the absence of selection. For example $\alpha>0$ can be generated if the past N_e was smaller than the present N_e (Eyre-Walker 2002). This works as follows: during a past period of small N_e, slightly deleterious nonsynonymous mutations become fixed in the population. These mutations do not segregate in the present population of larger N_e, because selection is more efficient at removing them. Therefore they contribute to D_N but not to P_N, and $\alpha>0$ even in the absence of adaptive substitutions. A strategy to minimise this problem is to choose to study populations that are not likely to have experienced severe and prolonged population size changes. Alternatively, one can use methods that take account of the estimated demographic history of the population when estimating α (Boyko et al. 2008; Eyre-Walker and Keightley 2009; Schneider et al. 2011).

1.2.4 Comparison of allele frequency spectra.

As discussed previously, new mutations introduce new alleles in populations. Let's assume that only two alleles can segregate at a site. A new allele introduced by a mutation (also called the derived allele) will be found initially at a low frequency in

the population, but its frequency might change due to drift or selection and displace the old allele that was previously fixed in the population (also called the ancestral allele). The relative proportion of derived alleles over different allele-frequency classes is called the allele frequency spectrum (AFS). If we know whether the segregating alleles in our sample are derived or ancestral we can infer the so-called unfolded AFS, otherwise we obtain the folded AFS. The inference of the derived/ancestral state of an allele is usually done by doing a sequence comparison to a closely related species (outgroup). Assuming that the compared outgroup sequence is not polymorphic and using simple parsimony, we infer that the ancestral allele is the one that is identical to the outgroup sequence. More sophisticated methodologies exist for inferring the derived/ancestral state of alleles and the unfolded spectrum (for example Hernandez et al. 2007).

The AFS obtained from a sample of individuals from the population can be used to detect selection. By using population genetics theory, we can predict the expected AFS under neutrality and compare it with the observed AFS. Selection can distort the AFS in a variety of ways. For example, negative selection produces an excess of rare variants and creates a more L-shaped AFS (Figure 1.1A). The signal of negative selection can be detected in both unfolded and folded AFS (Figure 1.1A). Positive selection produces an excess of high-frequency variants and a U-shaped AFS (Figure 1.1B). The signal of positive selection might be detected in the unfolded AFS, but is almost indistinguishable from neutrality in the folded AFS (Figure 1.1B).

Non-selective processes such as population size changes can distort the AFS in a similar way to selection. For example, population expansion and population

bottlenecks can produce an excess of rare variants similarly to negative selection (Figure 1.2A and Figure 1.2B respectively). Moreover, population contraction can produce a deficiency of rare variants similarly to positive selection (Figure 1.2A). To distinguish demographic effects from selection, we can compare the AFS for sites that we expect to evolve neutrally, such as synonymous sites, to the AFS from the focal sites.

A useful statistic to summarise the skew in the AFS is Tajima's *D* (Tajima 1989). The Tajima's *D* statistic is calculated by taking the difference between the observed pairwise diversity (π) and the expected diversity under Wright-Fisher equilibrium (Watterson's θ). A Tajima's *D* value of zero indicates no deviation from expectation. Negative values of Tajima's *D* indicate a positively skewed AFS, and are usually interpreted to be the result of negative selection or population expansion (Nielsen 2005; Charlesworth and Charlesworth 2010). Positive values of Tajima's *D* indicate an excess of intermediate frequency variants and are usually interpreted as suggesting balancing selection or population size contraction (Nielsen 2005; Charlesworth and Charlesworth 2010).

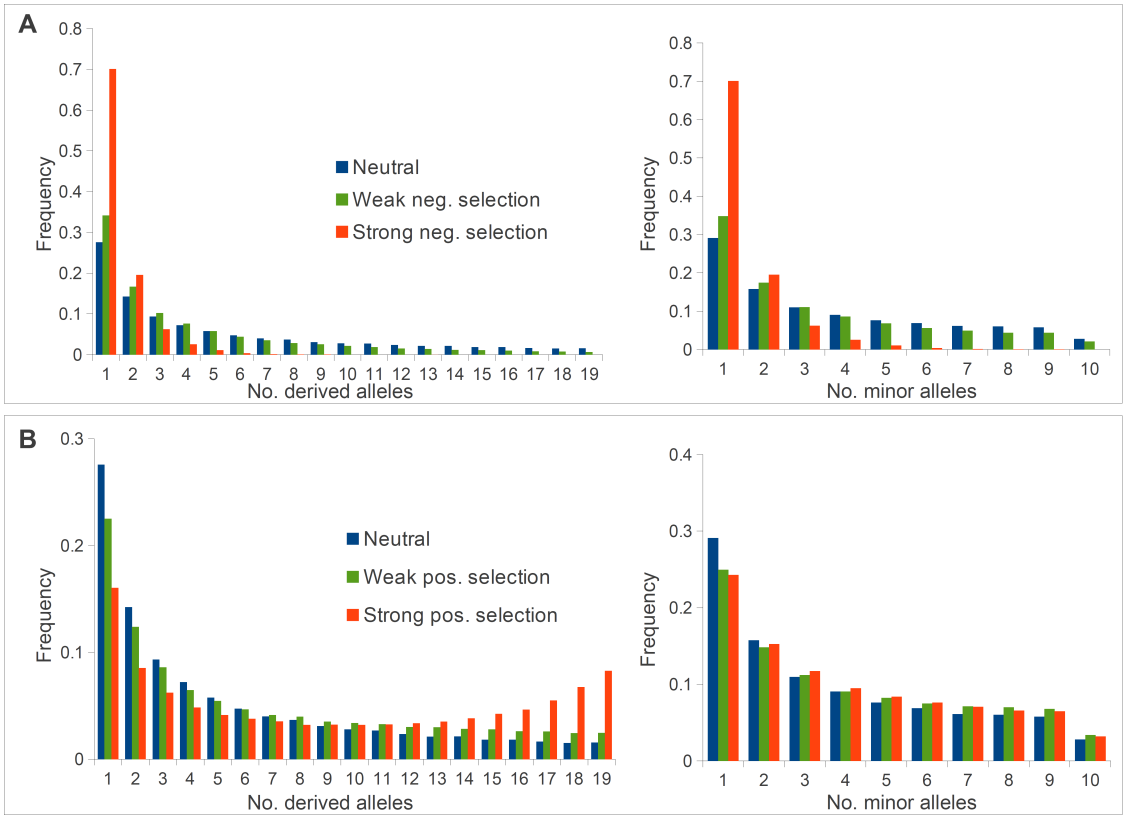


Figure 1.1. The expectation for the unfolded and folded allele frequency spectrum (left and right panels, respectively) when assuming (A) negative and (B) positive selection. For weak selection $N_e|s|=1$ and for strong selection $N_e|s|=10$.

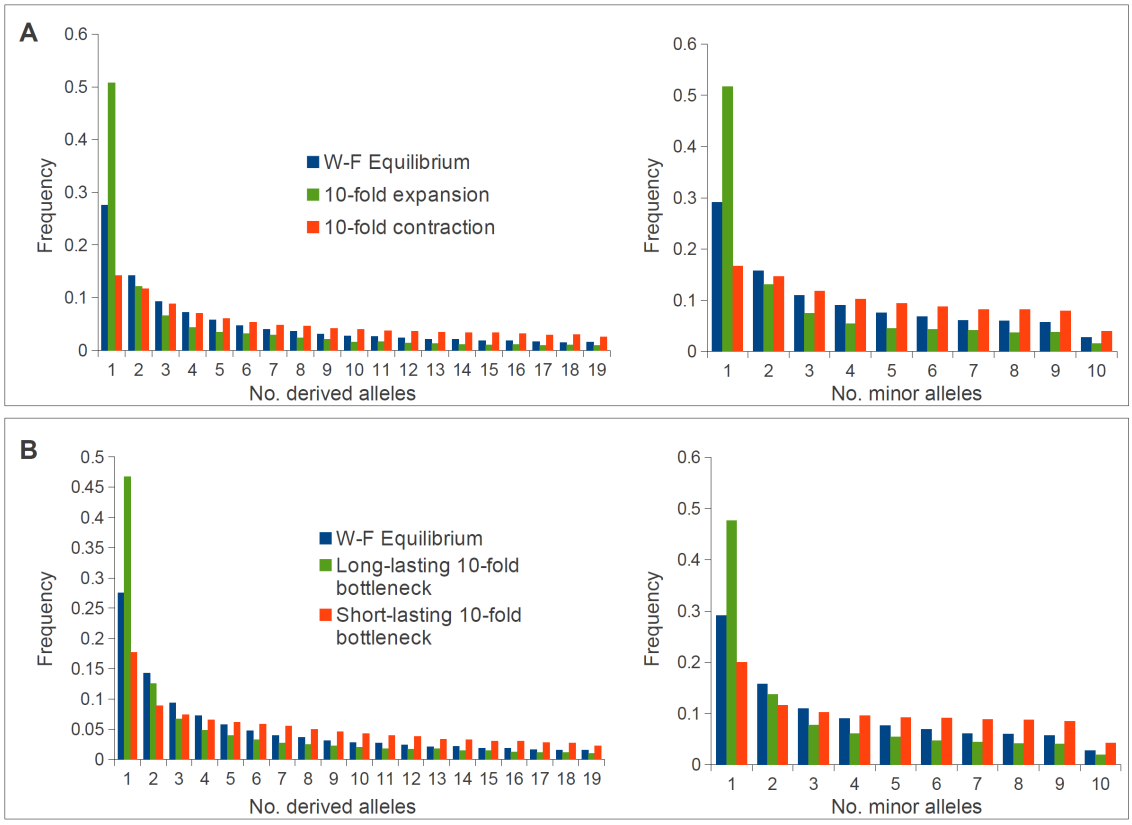


Figure 1.2. The expectation for the unfolded and folded allele frequency spectrum (left and right panels, respectively) under different demographic scenarios involving (A) a single population size change (B) two population size changes. W-F equilibrium refers to the expectation for the AFS for a population that is in Wright-Fisher equilibrium. Scenario A, involves a single population size change: either a 10-fold expansion or a 10-fold contraction 100 generations into the past. Scenario B involves a bottleneck: a population of an initial size N_1=100 is reduced to size N_2=10, 100 generations ago, and subsequently recovers to size 100 either 95 (short-lasting bottleneck) or 10 generations ago (short and long-lasting bottleneck, respectively).

1.3 State of research

I describe here the central research subjects that motivated the original research that is presented in the thesis. I begin by describing the debate on whether species differences at the molecular level are shaped by neutral or adaptive evolution. I continue by describing the research problem of inferring the distribution of fitness effects of new mutations. I then introduce the problem of locating which loci, protein-coding or non-coding, contribute mostly to adaptive evolution. Finally I conclude by presenting a central research question in genetics concerning the importance of the X chromosome in speciation.

1.3.1 Neutral versus adaptive evolution: the neutralist-selectionist controversy.

Origins of the controversy. In the middle of the 20th century, the neo-Darwinian synthesis provided for the first time an integrated framework to explain evolution (Dobzhansky 1937; Huxley 1942). Most evolutionary biologists during that time emphasised natural selection as the major force driving divergence between species (Mayr 1963; Simpson 1964). The prevailing view also held that variability in populations is maintained by selection and that very few segregating alleles are neutral (Mayr 1963). The first molecular data started to be collected in the 1960's and some observations appeared to be inconsistent with the established selectionist view. One such observation was that proteins appear to diverge at a constant rate per year (Zuckerkandl and Pauling 1965). Since most species live in different environments

and experience different selection pressures over time, a constant rate of molecular evolution per year is not expected if divergence between species is driven only by natural selection. Moreover, Haldane, one of the main workers of the neo-Darwinian synthesis, had suggested that the cost of selection can impose limits on the rate of molecular evolution (Haldane 1957). The observed rate of substitution in protein-coding genes seemed to exceed that limit (Kimura 1968). These observations led Kimura and Jukes and Cantor to independently suggest that most protein substitutions are neutral (Kimura 1968; King and Jukes 1969). Kimura later formalised this view as the neutral theory of molecular evolution (Kimura 1985).

According to the neutral theory, molecular divergence between species is mainly due to the fixation of neutral mutations, since deleterious mutations are very unlikely to be fixed and do not generally contribute much to sequence divergence (Kimura 1968). Moreover, according to the neutral theory, most segregating polymorphisms are neutral and constitute a phase of molecular evolution (Kimura and Ohta 1971). The neutral theory was challenged on several grounds as soon as it was proposed (Smith 1968; Richmond 1970) and was intensely debated for several years (Hey 1999). Although some of the original arguments of Kimura were shown to be wrong (most importantly the cost of selection argument; Smith 1968), the neutral theory proved to be very useful as a null hypothesis for tests for selection (Kreitman 1996). Moreover, the neutral theory was subsequently modified and expanded to model mutations that have small fitness effects, i.e. 'nearly neutral' mutations (Ohta 1973; Ohta 1992).

At the same time, alternative selectionist models were developed by Gillespie to

explain the molecular observations that the neutral theory purportedly explains (Gillespie 1991). Gillespie suggested that the molecular evolution of proteins is driven by bursts of adaptive evolution after frequent episodes of environmental change (Gillespie 1991).

Conceptual predictions of the rival theories (neutralist versus selectionist) can be compared and tested. The main distinguishing difference between the two theories concerns the contribution of adaptive substitutions to evolution. The neutral theory posits that adaptive substitutions are rare (Kimura 1985), whereas the selectionist theory posits that adaptive substitutions are common (Gillespie 1991). Another critical difference is the expectation regarding the effects of selection on linked loci (Hill-Robertson Interference and genetic hitchhiking) on the maintenance of genetic variability. The neutral theory expects these effects to have a negligible influence on patterns of genetic variability. In contrast, the selectionist theory expects these effects to be the main factor driving the maintenance of genetic variability. The central importance of linked selection on the selectionist theory is summarised by the concept of 'genetic draft' where levels of genetic variability are determined mainly by the rate of selective sweeps (Λ) and the recombination rate (r) rather than the mutation rate and N_e (Gillespie 1999; Gillespie 2000).

Current state of the neutralist-selectionist controversy and research questions. The application of neutrality tests (such as the McDonald-Kreitman (MK) test) have revealed departures from neutrality for many genes and for many species (Kreitman 1996). However, the occasional departure from neutrality is not inconsistent with the

neutral theory and recent research has focused on application of the MK test in genome-wide surveys to directly estimate the fraction of substitutions that have been fixed by positive selection (α). Estimates for α for amino-acid changes in protein-coding genes are consistently greater than 30% in *Drosophila melanogaster* (Begun et al. 2007; Shapiro et al. 2007; Mackay et al. 2012), but close to zero in humans (Boyko et al. 2008; Eyre-Walker and Keightley 2009). These observations appear to reject neutrality for *D. melanogaster* but not humans, which is hardly sufficient to resolve the neutralist-selectionist controversy. It has been suggested that humans might experience a slower rate of adaptive evolution than *Drosophila* because of their ~80 times smaller N_e (Eyre-Walker et al. 2006a; Eyre-Walker 2006; Eyre-Walker and Keightley 2009). This explanation rests on the assumption that a population with a small N_e has a smaller input of adaptive mutations and also experiences a stronger effect of genetic drift than a population with a large N_e. A current research subject is to investigate the relationship between α and N_e in species other than *D. melanogaster* and humans in order to elucidate the causes of the variation of α between species (Eyre-Walker 2006).

Another issue of contention among neutralists and selectionists is whether estimates of α using the MK approach can truly quantify adaptive substitution (Nei et al. 2010). Biological scenarios such as past demographic changes can generate genome-wide values of α that are significantly higher than 0 in the absence of adaptive substitutions (Eyre-Walker 2002). Consequently, the evidence from MK tests against the neutral theory has been challenged (Nei et al. 2010). A current research focus is to address weaknesses of the MK test and create a more solid

statistical foundation for its application (Eyre-Walker and Keightley 2009; Keightley and Eyre-Walker 2012).

1.3.2 The distribution of fitness effects of mutations (DFE).

The fitness effects of mutations are generally classified into three categories: neutral, deleterious and advantageous. Deleterious and advantageous mutations can have weak or strong effects on fitness. The distribution of fitness effects (DFE) describes the relative probability of sampling each type of effect.

Why knowing the DFE is important. There are several properties of the DFE that are important for both basic and applied genetics research questions. Firstly, by knowing the fraction of mutations that are deleterious (f_d) we can calculate the deleterious mutation rate per genome and per generation (U) (Kondrashov and Crow 1993; Eyre-Walker and Keightley 1999). Knowing U for a variety of species can help to understand the evolution and maintenance of sex (Kondrashov 1988) and to explain the evolution of haploid and diploid genetic systems (Kondrashov and Crow 1991). Secondly, by knowing the fraction of mutations that are neutral and nearly neutral, we can predict the contribution of advantageous mutations to molecular evolution (McDonald and Kreitman 1991; Eyre-Walker and Keightley 2009). Moreover, knowledge of the DFE is useful in many applied research subjects (Eyre-Walker and Keightley 2007), for example in medicine for calculating the health risk imposed by the accumulation of deleterious mutations (Crow 1997;

Eyre-Walker and Keightley 1999), in animal breeding for predicting long-term responses in selection experiments (Hill 1982), and in animal conservation for assessing the risk of population extinction in natural populations (Lande 1994).

Methods to quantify the DFE. To quantify the DFE, both experimental and population genetics approaches have been employed (Eyre-Walker and Keightley 2007). Experimental approaches usually involve the generation of mutation accumulation (MA) lines. These lines are derived from an ancestral genetically uniform line. The MA lines pass through many generations of accumulation of spontaneous or chemically-induced mutations. At the end of the experiment, the fitness of the MA lines is compared with the ancestral lines. Studies using the MA technique have been conducted in many species, including RNA viruses, bacteria, yeast, *Caenorhabditis elegans* and *Drosophila melanogaster* (Eyre-Walker and Keightley 2007). Experimental approaches can reveal mutations with large effects on fitness, but mutations with mild effects are largely undetected. For example an MA study in *C.elegans* estimated that only 1-4% of mutations had effects that were detectable in laboratory fitness assays (Davies et al. 1999).

Population genetic approaches can provide us with tools to quantify the DFE of mutations with weak or mild effects on fitness. The most basic approach is to compare the level of polymorphism or divergence of a genomic region to the neutral expectation. For example, the genomic distribution of d_N/d_S or π_N/π_S could be informative about the variation of selective pressures in the genome (Bustamante et al. 2005; Nielsen 2005). Alternatively, we can use the full distribution of allele

frequencies (AFV) for a set of genes or genomic regions. By using a model for the DFE (e.g. a gamma distribution), we can predict the expected AFV given a set of parameters. This model can then be fitted to AFVs from DNA sequence data to infer the DFE by using maximum likelihood or Bayesian approaches (Williamson et al. 2005; Eyre-Walker et al. 2006a; Keightley and Eyre-Walker 2007; Boyko et al. 2008).

What we currently know about the DFE and research directions. Early studies based on the MA technique have shown a large variation of the deleterious mutation rate per genome and per generation (U) across species, being lowest in bacteria (10^{-4}), intermediate in eukaryotes such as *C.elegans* (10^{-3}-10^{-2}) and *D. melanogaster* (10^{-2}-10^{-1}), and highest in RNA viruses (>1) (Halligan and Keightley 2009). In MA experiments the mutation rate (μ) and U are hard to tease apart, for example a high estimate of U could be due to a high μ or a high fraction of new mutations being deleterious. Therefore, an interpretation of the variation in U across organisms remained controversial until more recent technical progress allowed to directly estimate μ from the experiment and obtain more accurate estimates of U (Haag-Liautard et al. 2007). These new studies have suggested that for diploid sexual species, such as *D. melanogaster* and humans, U is likely to be higher than 1 (Halligan and Keightley 2009; Keightley 2012), a result that can provide an explanation for the evolution and the maintenance of sex. This is because at this limit (U=1) the advantages of sexual reproduction (e.g. more efficient purging of deleterious mutations from the population) become larger than its disadvantages (e.g.

two times lower number of breeding individuals).

Population genetic studies have shown that there is a large variation between species in the amount of new amino-acid mutations that fall into the class of effective neutrality, i.e. having an intensity of selection ($N_e s$) that is less than 1 (Table 1.1). This variation has been explained in terms of N_e, in accordance with neutral theory predictions: species with large N_e have a smaller fraction of nearly neutral mutations than species with large N_e (Table 1.1).

Regarding the shape of the DFE, results have been equivocal. Some MA studies have supported a highly leptokurtic distribution, with large variation of mutational effects, while others have supported a platykurtic distribution with most mutations having roughly similar effects (Eyre-Walker and Keightley 2007; Halligan and Keightley 2009). Results on the shape of the DFE from population genetic studies have been clearer, supporting a strongly leptokurtic DFE with a shape parameter that is less than 1 for several species (Eyre-Walker and Keightley 2007). It has been suggested that the discrepancy between MA studies and population genetic studies could be due to the DFE being complex and consisting of multiple modes (Davies et al. 1999; Eyre-Walker and Keightley 2007; Halligan and Keightley 2009). However, studies of complex DFE models that can take multimodal shapes are currently lacking (Keightley and Eyre-Walker 2010). More importantly, we currently don't know what biases in the estimation of important parameters of the DFE can arise when unimodal distributions are used to model multimodal DFEs.

Table 1.1. The proportion of amino-acid changing mutations assigned into three $N_e s$ categories for a variety of species. The estimates for N_e are from (Gossmann et al. 2012).

Species group	Species	Dataset	N_e	Nearly neutral ($N_e s$ 0-1)	Mildly deleterious ($N_e s$ 1-10)	Strongly deleterious ($N_e s$ >10)
Flowering plants	*C. grandiflora*	Slotte et al. 2010	641,262	0.07	0.07	0.86
	A. thaliana	Slotte et al. 2010	266,769	0.2	0.14	0.66
Flies	*D. melanogaster*	Shapiro et al. 2007	822,351	0.06	0.07	0.87
Mammals	House mouse	Halligan et al. 2010	573,567	0.1	0.11	0.79
	European rabbit	Carneiro et al. 2012	800,000	0.03	0.03	0.94
	Central chimpanzees	Hvilsom et al. 2012	110,000	0.2	0.05	0.75
	Humans	Boyko et al. 2008	20,974	0.22	0.13	0.65

1.3.3 Contribution of protein-coding versus regulatory change to adaptive evolution.

A central objective in evolutionary biology is to understand what types of genetic changes underlie phenotypic adaptations (Lewontin 1974). An enduring controversy is whether adaptive genetic change occurs more frequently within protein-coding genes or within non-coding DNA (Hoekstra and Coyne 2007; Stern and Orgogozo 2008). In this section I briefly outline the origin of the controversy and approaches that have been taken to resolve it.

Origin of the controversy. The functional role of non-coding DNA was first realised with the discovery of non-coding regulatory elements in enteric bacteria that can control the expression of genes in *cis* (Jacob and Monod 1961). Moreover, the subsequent finding that very few genetic differences between human and chimpanzees are within protein-coding genes suggested that the many morphological differences between them may be due to changes within regulatory non-coding DNA (King and Wilson 1975). These findings led to the idea that the basic functions of the cell (e.g. biochemical pathways) may have evolved early in the evolution of life, and are encoded in a basic tool-kit of structural genes (Jacob 1977), while more recent phenotypic adaptations may mostly concern the control of the spatio-temporal expression of structural genes, realised through changes in *cis*-regulatory non-coding DNA (Jacob 1977). Inspired by these ideas, researchers working on the evolution of development ('evo-devo') have developed the hypothesis that most morphological

adaptation is due to change in *cis*-regulatory elements rather than protein-coding genes (the *cis*-regulatory hypothesis; Carroll 2005a; Carroll 2005b; Stern and Orgogozo 2008). *Cis*-regulatory elements are posited to be less pleiotropically constrained than protein-coding genes, allowing changes to occur without disrupting their function, presumably by fine-tuning the expression of structural genes (Carroll 2005a; Carroll 2005b; Stern and Orgogozo 2008).

Other researchers have criticised many of the basic premises of the *cis*-regulatory hypothesis (Hoekstra and Coyne 2007). The notion that protein-coding genes do not differ greatly between humans and chimpanzees is not accurate: with nonsynonymous site divergence of 0.2% and assuming that there are 20,000 protein-coding genes, each with ~1,000 non-synonymous sites on average, we obtain a total of 40,000 amino-acid differences, which may be sufficient to explain the observed phenotypic differences between humans and chimpanzees. Moreover, many protein-coding genes perform regulatory roles (e.g. transcription factors). Therefore, even if phenotypic change is often realised through mutations that alter the regulation of gene expression, these mutations need not necessarily occur within non-coding DNA. Finally, weaker pleiotropic constraints do not necessarily allow faster adaptive evolution. If the selective coefficients for adaptive mutations in non-coding DNA are very small, the fixation of these mutations is more likely to be dominated by drift than selection, especially for species with small N_e.

Investigating the genetic basis of adaptation to resolve the controversy. A straightforward approach is to catalogue phenotypic adaptations in the wild and map

the genetic changes that underlie these adaptations (top-down approach). This approach usually involves a technique called association mapping, where phenotypic trait variation is correlated with genetic variation (Barrett and Hoekstra 2011). Studies using the association mapping have shown that genetic changes underlying phenotypic adaptations are frequently found within protein-coding genes (Hoekstra and Coyne 2007). However, it has been argued that this could be due to our better understanding of how phenotypes are affected by changes in protein-coding than changes in non-coding DNA (Stern and Orgogozo 2008). Moreover, studies using the top-down approach usually cannot resolve whether the cause of the studied adaptation is due to change within the coding region of the gene or within regulatory regions close to the gene.

A more indirect approach for investigating the genetic basis of adaptation is to detect the footprint of positive selection that has acted on different timescales through comparison of DNA sequences (bottom-up approach; Nielsen 2005). For example, loci that have been under recent positive selection show a distinctive pattern of neutral diversity and haplotype structure around the locus resulting from selective sweeps (Andolfatto 2001). Moreover, loci that have experienced frequent episodes of positive selection in the distant past display excess between-species divergence that is not accounted for by the neutral theory (McDonald and Kreitman 1991). Loci that display signatures of positive selection are frequently found to contain genetic variation that is associated with adaptive traits (Barrett and Hoekstra 2011). Therefore, bottom-up approaches can be complementary to top-down approaches in order to locate the locus of adaptation. The biggest challenge for

bottom-up approaches is to correctly take account of evolutionary forces that can produce similar footprints to positive selection (for example, population size changes).

Studies using the MK test on polymorphism and divergence data from *Drosophila* have revealed that the rate of adaptive substitution per year in protein-coding genes is approximately equal to that in non-coding regulatory regions that are upstream or downstream of genes (Andolfatto 2005; Begun et al. 2007; Haddrill et al. 2008). These results from *Drosophila* would suggest that change in *cis*-regulatory DNA is at least as important for adaptive evolution as protein-coding genes. Similar studies in humans have produced equivocal results. *Cis*-regulatory elements do not appear to experience a faster rate of adaptive evolution than protein-coding genes in humans (Eyre-Walker and Keightley 2009), although some elements that control expression of genes in brain development do display signals of positive selection (Torgerson et al. 2009). Consequently, the evidence from humans does not support the basic tenet of the *cis*-regulatory hypothesis that adaptive evolution for 'complex' organisms proceeds mostly through change in non-coding DNA. It has been argued that it might be harder to detect the signal of positive selection in humans due to their complex demographic history that presumably obscures tests of positive selection (Eyre-Walker and Keightley 2009). Expanding the study of positive selection in non-coding DNA of species other than *Drosophila* and humans could help to resolve the controversy and test the *cis*-regulatory hypothesis more comprehensively.

1.3.4 The two rules of speciation and their causes.

There are two patterns of observations in the field of speciation genetics that are found in almost all species with heteromorphic chromosomes: Haldane's rule and the large effect of the X chromosome in speciation (Coyne and Orr 1989). Researchers refer to these patterns as *rules* due to their universality (Coyne and Orr 1989). The investigation of these two rules of speciation has been at the forefront of speciation genetics research for more than a century (Coyne and Orr 2004). In this section, I briefly outline the rules and present evolutionary hypotheses that have been suggested to explain them.

Haldane's rule and its causes. When different species are crossed and one of the sexes in the hybrid offspring is inviable or sterile, this sex is the heterogametic sex (Haldane 1922) (males for XY systems and females for ZW systems). This pattern is known as the Haldane's rule and is observed in all taxa with heteromorphic sex chromosomes (Coyne and Orr 2004). Two hypotheses have been suggested to explain Haldane's rule; the dominance hypothesis and the faster-male hypothesis (Coyne and Orr 2004; Presgraves 2008).

According to the dominance hypothesis, Haldane's rule arises because X-linked mutations that cause inviability or sterility in species hybrids are more often recessive than dominant (Muller 1942; Turelli and Orr 1995). To understand this expectation, we need to consider how dominance of X-linked mutations can cause hybrid incompatibility problems in The effect of recessive mutations is made

manifest in XY males but masked in XX females. Therefore, X-linked recessive mutations can cause hybrid incompatibility problems only in males. In contrast, dominant mutations affect the phenotype in both males and females. Since the females have two copies of the X chromosome, X-linked dominant mutants are more likely to cause hybrid incompatibility problems in females. Experimental studies in *Drosophila* species have demonstrated that mutations that contribute to species incompatibilities are more often recessive than dominant (True et al. 1996; Masly and Presgraves 2007). These studies have given strong support to the dominance hypothesis as an explanation for Haldane's rule. Moreover, the central role of the X chromosome in the dominance hypothesis is consistent with observations that suggest that the X chromosome is a hotspot for speciation genes (Presgraves 2008).

According to the faster-male hypothesis, mutations which cause male sterility in hybrids accumulate faster than mutations that generally reduce fitness in hybrids (Wu et al. 1996). Genetic analysis has provided strong evidence that male-sterility mutations accumulate faster than other types of mutations that contribute to species incompatibilities (Hollocher and Wu 1996; Tao et al. 2003; Masly and Presgraves 2007). Rapid accumulation of male-sterility factors may be due to rapid adaptive evolution of male-specific genes caused by a higher intensity of sexual selection in males than females (or other causes; Wu et al. 1996). Growing molecular evidence from many species (including fruitflies, birds and humans) suggests that male-specific genes do indeed experience very rapid adaptive evolution (Ellegren and Parsch 2007). However, the fact that the faster-male hypothesis cannot explain Haldane's rule for species where the heterogametic sex is the female (ZW systems)

limits somewhat its scope (Coyne and Orr 2004).

The large effect of the X chromosome and its causes. Another phenomenon that is directly related to Haldane's rule is the disproportionate contribution of the X chromosome to species incompatibilities (large-X effect; Coyne and Orr 1989; Coyne and Orr 2004). For example in *Drosophila melanogaster*, the X chromosome harbors ~4 times more hybrid male sterility factors than other similarly sized chromosomes (Masly and Presgraves 2007). Moreover, the X chromosome displays a reduced amount of gene flow between populations of *Drosophila* (Kulathinal et al. 2009) and populations of house mice relative to other chromosomes (Payseur et al. 2004; Teeter et al. 2008). The evolutionary causes of the large-X effect are still unclear, but past work has narrowed down the list of possible explanations and these are outlined below.

Male-biased gene content of the X. The large-X effect may be due to a disproportionate number of male-biased genes on the X chromosome, compared with the autosomes (Presgraves 2008; Vicoso and Charlesworth 2009). However, gene expression studies in Drosophila have shown that the X chromosome is actually depauperate in genes with male-biased expression (Parisi et al. 2003). In mammals, the X chromosome is enriched for genes that are expressed early in spermatogenesis, but depauperate for genes that are expressed late in spermatogenesis, and are more likely to have male-specific function (Khil et al. 2004). Moreover, X-linked genes which are early-expressed in spermatogenesis in mice have been shown to be

female-biased (Zhang et al. 2010), which suggests that the X chromosome in mammals is demasculinised, as in *Drosophila*.

Faster-X evolution. The large-X effect could be a consequence of faster evolution of X-linked loci than autosomal loci. If new advantageous mutations are on average recessive, they are expected to fix faster on the X chromosome than the autosomes (Charlesworth et al. 1987). A direct test of this hypothesis is to compare the rate of substitution between autosomal and X-linked genes. In *Drosophila*, results have been equivocal. Depending on the selection of lineage to study divergence, and the method used (i.e. by comparing d_N/d_S ratios or performing the MK test) researchers have reached different conclusions, either supporting or rejecting faster-X evolution (Presgraves 2008; Vicoso and Charlesworth 2009; Mackay et al. 2012). In mammals, the evidence for faster-X evolution appears to be stronger (Vicoso and Charlesworth 2009; Hvilsom et al. 2012), although it might be limited to testis-specific genes (Khaitovich et al. 2005). Therefore, faster-X evolution could be contributing to the large-X effect, but it is likely to be limited to specific lineages or classes of genes.

Meiotic drive. Meiotic drive elements lead to segregation distortion (i.e. biased, non-Mendelian, transmission of alleles or chromosomes). "Selfish" meiotic drive elements can create fitness costs for their hosts which can lead to the evolution of genes that suppress the expression of the selfish elements. Evolutionary conflict between meiotic drive and suppressor genes can lead to their rapid divergence between species. Therefore, in hybrid crosses, drivers might be in a genetic

background that cannot sufficiently suppress their expression, which can, in turn, cause sterility or inviability of the hybrids. Theory predicts that sex-linked meiotic drive elements are more likely to invade a population than autosomal elements (Presgraves 2008). In accordance with theory, many more distorter elements have been found to be X-linked than autosomal in *Drosophila* (Jaenike 2001). Therefore, a higher concentration of meiotic drive elements on the X than the autosomes might contribute to the large-X effect (Presgraves 2008; Cocquet et al. 2012).

Special role of the X chromosome in spermatogenesis. The X chromosome is inactivated during the first meiotic stage of male spermatogenesis in mammals, a process known as meiotic sex chromosome inactivation or MSCI (Lifschytz and Lindsley 1972). It has been suggested that male spermatogenesis might be a process that is sensitive to perturbations of gene expression and that disruption of MSCI in hybrids might cause male sterility (Coyne and Orr 2004; Presgraves 2008). Laboratory crosses between *Mus musculus musculus* and *M. m. domesticus* have shown that MSCI is disrupted in sterile, but not in fertile, offspring (Good et al. 2008; Campbell et al. 2013). Therefore MSCI could contribute to the large-X effect, at least in mammals. Evidence exists that MSCI occurs in *Drosophila* (Hense et al. 2007) and birds (Schoenmakers et al. 2009), thus suggesting that disruption of MSCI in hybrids may be a universal explanation for the large-X effect. However, for both *Drosophila* and birds, the evidence for MSCI is highly disputed (Guioli et al. 2012; Mikhaylova and Nurminsky 2012)

1.4 Aims of the thesis.

This research thesis aims to address the problems outlined in the previous section and to inform the relevant controversies.

In Chapter 2, I aim to address the problem of inferring the distribution of fitness effects (DFE) of new mutations. The DFE is frequently modelled as a gamma distribution. However, increasing experimental and population genetics evidence is challenging the notion of a unimodal DFE. The evidence suggests that the DFE is likely to be complex and multimodal. In Chapter 2, I develop methodology to model complex DFEs. My aims for Chapter 2 are threefold: Firstly, to investigate whether a multimodal DFE can be detectable from the analysis of polymorphism data. Secondly, to uncover whether biases in the parameters of the DFE, such as the mean effect of a new mutation, can arise if a unimodal DFE is fitted to a DFE that is multimodal. Thirdly, I aim to apply the new methodology to infer the DFE by analysing polymorphism data from natural populations of *D. melanogaster* and *M. m. castaneus*.

In Chapter 3, I aim to inform the controversy on the locus of adaptation by quantifying the contribution of protein-coding and non-coding regulatory DNA to adaptive evolution in a population of house mice. House mice are a species of rodents that can be considered phenotypically as complex as humans. Therefore, according to the *cis*-regulatory hypothesis, we should expect a high contribution of non-coding regulatory DNA to adaptive evolution relative to the protein-coding genes. I study a population of house mice that has high levels of diversity and is

considered to be located in the ancestral range of the species. This population is unlikely to have experienced severe population size changes in the recent past, therefore creating ideal conditions for the application of tests for natural selection.

In Chapter 4, I analyse genome-wide polymorphism data for a population of house mice, to study the evolution of autosomal and X-linked genes. My main aim is to test the faster-X hypothesis that has been proposed to explain the large effect of the X chromosome in speciation. I test two main predictions of the faster-X hypothesis: Firstly, I test whether X-linked genes have a faster rate of adaptive evolution than autosomal genes; secondly, I test whether faster-X evolution is more intense in male-specific genes than female-specific genes.

Chapter 2. A Comparison of Models to Infer the Distribution of Fitness Effects of New Mutations.

The work presented in this chapter has been published as a research paper: Kousathanas A, Keightley PD. 2013. A comparison of models to infer the distribution of fitness effects of new mutations. *Genetics* **193**: 1197–1208.

I present the work as published with slight modifications. AK designed and performed the experiments, analysed the data and wrote the paper. PDK provided help with computer coding and comments on previous versions of the manuscript.

2.1 Summary

Knowing the distribution of fitness effects (DFE) of new mutations is important for several topics in evolutionary genetics. Existing computational methods to infer the DFE based on DNA polymorphism data have frequently assumed that the DFE can be approximated by a unimodal distribution, such as a lognormal or a gamma distribution. However, if the true DFE departs substantially from the assumed distribution (e.g. if the DFE is multimodal), this could lead to misleading inferences about its properties. We conducted simulations to test the performance of parametric

and non-parametric discretised distribution models to infer the properties of the DFE for cases in which the true DFE is unimodal, bimodal or multimodal. We found that lognormal and gamma distribution models can perform poorly in recovering the properties of the distribution if the true DFE is bimodal or multimodal, whereas discretised distribution models perform better. If there is a sufficient amount of data, the discretised models can detect a multimodal DFE and can accurately infer the mean effect and the average fixation probability of a new deleterious mutation. We fitted several models for the DFE of amino acid-changing mutations using whole-genome polymorphism data from *Drosophila melanogaster* and the house mouse subspecies *Mus musculus castaneus*. A lognormal DFE best explains the data for *D. melanogaster*, whereas we find evidence for a bimodal DFE in *M. m. castaneus*.

2.2 Introduction

New mutations generate genetic variation in the genome of every species. For example, it has been estimated that a newborn human has ~70 new mutations that originated in its parents' germlines (Keightley 2012). The fitness effects of new mutations can range from deleterious to neutral and to advantageous, and the relative frequencies of their effects is known as the distribution of fitness effects (DFE) of new mutations. Inferring the properties of the DFE is a long-standing goal of evolutionary genetics and is key to several important questions, including the evolution of sex and recombination, the prevalence of Muller's ratchet and the

constancy of the molecular clock (Charlesworth 1996; Eyre-Walker and Keightley 2007).

A number of methodologies have been developed to infer the DFE based on DNA sequence data (Sawyer et al. 2003; Nielsen and Yang 2003; Piganeau and Eyre-Walker 2003; Loewe et al. 2006; Eyre-Walker et al. 2006; Keightley and Eyre-Walker 2007; Boyko et al. 2008; Schneider et al. 2011; Wilson et al. 2011). All of these assume that there is a neutrally evolving class of sites, and contrast patterns of polymorphism and/or divergence from an outgroup with that of a tightly linked focal site class. Selection affecting the focal sites is expected to alter the pattern of polymorphism compared to that of the neutral class. A distribution of selection coefficients is then fitted to the data, and its properties inferred. The three most widely used methods are those developed by Eyre-Walker et al. (2006), Keightley and Eyre-Walker (2007) and Boyko et al. (2008). Keightley and Eyre-Walker (2007) use a Wright-Fisher transition-matrix approach (Ewens 1979), whereas Eyre-Walker et al. (2006) and Boyko et al. (2008) use a diffusion approximation (Sawyer and Hartl 1992; Williamson et al. 2005). All three methods have been reported to give similar results, but make slightly different assumptions. For example, they differ in the way in which they model demographic changes (e.g. population size changes). Eyre-Walker et al. (2006) use a heuristic, non-model based, approach, whereas the other two explicitly model some simple demographic scenarios. It is necessary to model demographic change, because this is known to alter patterns of polymorphism in ways that can resemble selection. Because these methods use allele-frequency

information (summarised as the site-frequency spectrum or SFS), they are expected to be sensitive to demographic change.

Several studies have employed the above methods to infer properties of the DFE of amino acid-changing mutations. In these analyses, a gamma distribution of fitness effects has often been assumed, since it is a flexible distribution with two parameters, the shape (*b*) and the scale (*a*). For example, for amino acid-changing mutations in *D. melanogaster*, the shape parameter has been estimated to be ~0.4 (implying a leptokurtic distribution), and most (>90%) of new mutations are inferred to be moderately to strongly deleterious, with effective strength of selection $N_e s$>10 (Keightley and Eyre-Walker 2007; Eyre-Walker and Keightley 2009). In humans, the DFE appears to be more even more leptokurtic than in *Drosophila* (i.e. the estimated shape parameter is ~0.2), and only ~60% of mutations appear to be moderately to strongly deleterious (Eyre-Walker et al. 2006b; Keightley and Eyre-Walker 2007; Boyko et al. 2008; Eyre-Walker and Keightley 2009). Differences between *Drosophila* and humans in the properties of the DFE have been attributed to a difference in their effective population size (N_e), the former being ~80 times larger (Eyre-Walker et al. 2002). An effect attributable to N_e has also been observed in several other species. For example, N_e in wild house mice is substantially larger than humans, but smaller than *Drosophila*, and ~70-80% of amino acid mutations are estimated to be moderately to strongly deleterious (Halligan et al. 2010; Kousathanas et al. 2011). *Capsella grandiflora* and *Arabidopsis thaliana* are two plant species with large and small N_e, respectively, and ~86% and ~66% of amino acid mutations

are estimated to be moderately to strongly deleterious, respectively (Foxe et al. 2008; Slotte et al. 2010). *Arabidopsis thaliana* and *Capsella grandiflora* also differ in their mating system (self-fertilising versus outcrossing), which could also contribute to the difference in the efficiency of natural selection between them.

Most of the above methods assume that the DFE can be approximated by a certain type of mathematical distribution, such as the gamma distribution. One would like, however, to have a more general approach to obtain information about the DFE without needing to assume an explicit distribution. Steps in this direction were taken by Keightley and Eyre-Walker (2010), who examined a model of multiple discrete selection coefficients rather than assuming a continuous distribution. However, Keightley and Eyre-Walker (2010) did not examine the performance of their models when the true distribution deviated from a gamma distribution. Boyko et al. (2008) also fitted several types of distributions and combinations of continuous distributions and discrete fixed effects when inferring the DFE for amino-acid changing mutations in humans. Wilson et al. (2011) recently developed a new method that assumes a series of discrete fixed selection coefficients, the density associated with each selection coefficient being estimated as a parameter. However, due to the complexity of the model, Wilson et al. (2011) needed to assume a constant population size.

Although several different types of parametric and non-parametric DFE models have been fitted to DNA polymorphism data, to our knowledge their performance in cases where the true DFE is bimodal or multimodal has not previously been investigated. In this study, we use simulations to examine cases

where the true DFE is unimodal, bimodal or multimodal. We analyse simulated data assuming six models for the DFE. The first two are parametric unimodal distributions: the lognormal and the gamma distribution. The third model is a parametric distribution that can be bimodal: the beta distribution. The fourth model is a discrete point mass distribution of selection coefficients, where the locations and the probability densities of each point mass (or 'spikes') are estimated parameters. We refer to this model as the spikes model, which is similar to the discretised model used by Keightley and Eyre-Walker (2010). The fifth model ('steps model') consists of multiple continuous, uniform distributions (or steps), the boundaries and probability densities of which are estimated parameters. The sixth model is a variant of the model used by Wilson at al. (2011), and assumes 6 fixed selection coefficients where only their probability densities are estimated parameters. We refer to this model as the 'fixed 6-spikes model'.

In this chapter, we use simulations to test the performance of the six models assuming various scenarios for the complexity of the true DFE. We further examine the performance of the six models for different allele sequencing efforts. We then test the robustness of the six models to the effects of population size changes and linkage. Finally, we fit the six models to protein polymorphism datasets from *D. melanogaster* and *M. m. castaneus,* each containing sequences of several thousand protein-coding genes.

2.3 Materials and Methods

Population genetic model and assumptions. In this study, we extended the methods developed by Keightley and Eyre-Walker (2007) to infer the distribution of fitness effects (DFE) of new mutations based on the allele frequency distribution of polymorphic nucleotide sites among individuals sampled from a population. This approach is based on Wright-Fisher population genetics theory, and makes a number of assumptions. We assumed that sites are unlinked, have the same mutation rate, and that polymorphic sites are biallelic. We assumed that there are two classes of sites in the genome, one 'neutral' and one 'selected'. The fates of new mutations in the neutral class are affected only by genetic drift. New mutations at selected sites are assumed to be unconditionally deleterious and to have additive effects on fitness. We defined the selection coefficient s as the fitness reduction experienced by the homozygote for the mutant allele compared to the homozygote for the wild-type allele. Therefore, the fitnesses of the wild-type, heterozygote and mutant homozygote are 1, 1-s/2 and 1-s, respectively.

Description of the modelled distributions of selection coefficients. New mutations affecting the selected class of sites are sampled from a probability distribution. We investigated six models for this probability distribution: the first is a lognormal distribution, which has two parameters: the mean or location (μ) and the standard deviation or scale (σ). The second is a gamma distribution, which has two parameters: the shape (b) and the scale (a). The third model is the beta distribution,

which has two shape parameters (k_1, k_2). The fourth model (spikes model) assumes m mutational effects classes (spikes), which are modelled as point masses. For each mutational effect class i (i = 1..m), the location s_i and the probability density (p_i) are estimated parameters, for a total of 2m-1 parameters. The fifth model (steps model) assumes m mutational effects classes, and each class i (i = 1..m) is modelled as a uniform distribution where the minimum and maximum values ($N_e s_{i-1}$ and $N_e s_i$ respectively) and the probability density (p_i) are estimated parameters. The minimum value of the first step is fixed to zero. We assumed that the start of each step is the end of the previous, that is, for step i, $N_e s_i = N_e s_{i-1}$, ensuring that there are no overlapping steps. The total number of parameters to be estimated is m for the minimum and maximum values values of the steps, plus m-1 for the probability density of each step, giving a total of 2m-1 parameters. For the sixth model (6-fixed spikes) I assumed 6 mutational effects classes (spikes), modelled as point masses arbitrarily fixed at $N_e s_1$=0, $N_e s_2$=1, $N_e s_3$=5, $N_e s_4$=10, $N_e s_5$=50, $N_e s_6 = N_e$. The probability densities of the fixed point masses were estimated parameters, for a total of 5 parameters.

Table 2.1. The selection models investigated in this study.

DFE Model	No. Parameters	Parameters
Lognormal	2	μ, σ (location, scale)
Gamma	2	a, b (scale, shape)
Beta	2	k_1, k_2 (shape 1, shape 2)
Spike	2m-1	For i (i = 1..m), $N_e s_i$ For i (i = 1..m-1), p_i
Step	2m-1	For i (i = 1..m), $N_e s_i$ For i (i = 1..m-1), p_i

6-fixed spikes	5	For i (i = 1..5), p_i

Demographic Model. Following Keightley and Eyre-Walker (2007), we also incorporated a simple demographic model of a step change from population size N_1 to population size N_2 at some time t in the past. N_1 is fixed at 100, the parameter t is estimated relative to N_2, and the parameter N_2 is estimated relative to N_1 (i.e. the magnitude of the size change is estimated). There may be little information to estimate the relative values of N_1 and N_2, so we also computed a weighted recent effective population size N_w:

$$N_w = \frac{N_1 w_1 + N_2 w_2}{w_1 + w_2} \tag{2.1}$$

where $w_1 = N_1\left(1 - \frac{1}{2N_2}\right)^t$ and $w_2 = N_2\left(1 - e^{-t/(2N_2)}\right)$ (Eyre-Walker and Keightley 2009). The weightings w_1 and w_2 are the expected contributions of neutral mutations to allele frequency variation t generations after the population size change. Note that N_w is only a scaling parameter to estimate $\overline{N_e s}$, and it is not an estimate of the effective population size (N_e).

We also incorporated a parameter f_0, which is the proportion of unmutated sites. Under selective neutrality and stationary equilibrium, 1-f_0 is proportional to the product of the mutation rate and the persistence time of a new mutation.

Generation of the expected allele-frequency vector and computation of

likelihood. We assumed that, at some point in the past, a population of size N_1 was at mutation-selection-drift equilibrium. This population then experienced a size change (either expansion or contraction) to size N_2, t generations from the present. Throughout this period, new mutations arise, which are neutral for the neutral class of sites, and deleterious with selection coefficients s sampled from a probability distribution $f(s)$ for the selected class. Following Keightley and Eyre-Walker (2007), we employed Wright-Fisher transition matrix methods to generate the expected allele frequency distribution at the present time for a set of parameter values f_0, t, N_2, and a given s value, and we stored it in a vector $\boldsymbol{v}(s)$. The lognormal, gamma, spike and step distributions can potentially have substantial parts of their density at $s>1$. We modelled the contribution of mutations for $s>1$ assuming that their frequency in the population goes down in proportion to the expectation at mutation-selection balance, following Keightley and Eyre-Walker (2007). The expected mean allele-frequency distribution $\boldsymbol{z}$ was obtained by integrating over the distribution of selection coefficients for all elements of $\boldsymbol{v}(s)$:

$$\boldsymbol{z} = \int_0^\infty \boldsymbol{v}(s)\, f(s|\Theta)\, ds \tag{2.2}$$

where Θ represents the parameters of the distribution of selection coefficients (e.g. a and b for the gamma distribution).

The numbers of derived alleles in a sample of n_T alleles constitute the SFSs, and were stored in vectors $\mathbf{q}(N)$ and $\mathbf{q}(S)$ for the selected and neutral sites, respectively. Numbers of alleles were binomial draws from a diploid population of size N_2. Since we did not distinguish between the derived and ancestral states, we

used only folded SFSs. We folded the SFS and the allele-frequency vector **z** as follows:

$$\boldsymbol{q}_i = \boldsymbol{q}_i + \boldsymbol{q}_{n_T - i} \quad \text{, for} \quad 0 \le i < n_T/2 \tag{2.3}$$

$$\boldsymbol{z}_i = \boldsymbol{z}_i + \boldsymbol{z}_{2N-i} \quad \text{, for} \quad 1 \le i \le 2N_2/2 \tag{2.4}$$

Under the assumption that numbers of derived alleles are binomially distributed, we computed the log likelihood of the observed allele frequency distributions (i.e. SFSs) for neutral and selected sites as:

$$logL = \sum_{i=0}^{n_T/2} \boldsymbol{q}_i \log\left(\sum_{j=0}^{N_2} \boldsymbol{z}_j \left(b(i|n_T, j/2N_2) + b(n_T - i|n_T, j/2N_2) \right) \right) \tag{2.5}$$

(Keightley and Eyre-Walker 2007), where $b(i|n, p)$ is the binomial probability for i derived alleles in a sample of n alleles with probability of occurrence p. We found the set of the parameter values that best fits the observed SFSs by maximizing the sum of the log-likelihoods calculated for the neutral and selected classes of sites.

Likelihood maximization. The parameters to be estimated are f_0, N_2, t, plus additional parameters, depending on the selection model implemented (Table 2.1). Maximization of the likelihood was done using a custom likelihood search algorithm for N_2, and the SIMPLEX algorithm (Nelder and Mead 1965) for the remaining parameters. The expected mean allele-frequency distribution $\boldsymbol{z}$ was precomputed for discrete sets of parameter combinations. As per Keightley and Eyre-Walker (2007), values of N_2 and t went from 2 to 1000, and from 1 to 5000, respectively, in steps increasing by 5% or 1, whichever was higher. We integrated over the distribution of

s, by using precomputed $\boldsymbol{z}$ vectors for 250 *s* values. Note that when using precomputed $\boldsymbol{z}$ vectors for discrete parameter values, it is likely in the estimation procedure to obtain exactly the same value for a certain parameter when analysing different datasets.

To increase the speed of the maximization procedure, we first estimated the demographic parameters N_2 and *t* and the parameter f_0 from the neutral SFS. We assumed the maximum likelihood (ML) estimates of N_2 and *t w*hen estimating the parameters from the selected SFS.

We generated starting values for the location parameters of the spikes and the steps by using a power series:

for spike or step *i* (*i*=1..*m*),

$$N_e s_i = N_e^{(\frac{i}{m} - r)} \tag{2.6}$$

where $N_e=N_w$ as calculated by equation 2.1 and *r* is a pseudorandom deviate from a normal distribution with a mean 0, and standard deviation 0.1. This power series was devised empirically, and has several desirable properties: the term $N_e^{i/m}$ places the spikes or steps at a reasonable distance from each other; the last spike or step is placed at N_e, therefore avoiding generating extremely large $N_e s$ values; the pseudorandom normal deviate *r* adds noise in the placement of the spikes/steps.

The starting values for the relative probability densities of the steps were set to 1/*m*. As the number of parameters increases, the possibility of multiple local maxima also increases. To ensure that the global maximum had been found, we performed 10 starts of the maximization algorithm for each run, each time using a

different seed for the pseudorandom number generator. We recorded the ML estimates that gave the highest likelihood in these runs.

Implementation of the model. Our simulations used a forward Wright-Fisher simulator to generate SFSs, and we then used ML to fit demographic and selection models and estimate the parameters. This was implemented in a recoded version of the C program *DFE-alpha* (Eyre-Walker and Keightley 2009). This version implements all of the models we describe, can be used to analyse SFS datasets in a similar way to *DFE-alpha*, and will be made available via the authors' website.

Simulations assuming a constant population size. We simulated SFS datasets assuming a diverse set of distributions of selection coefficients, including unimodal, bimodal and multimodal distributions. We performed forward simulations in which we assumed a constant population size (N_1=N_2=100). We used 10^6 neutral and 10^6 selected sites and sampled with replacement 64 alleles. We also compared simulations in which we sampled with replacement different numbers of alleles (8, 16, 32, 64, 128 and 256), while assuming a set number of sites (10^6). Parameter f_0 was set to 0.9. For each simulated dataset, we performed 100 replicate simulations.

Simulations assuming variable population size. We modelled population size changes as step changes from an initial population of size N_1=100 at stationary equilibrium. Time is expressed in units of N_1. We simulated two demographic

histories: a population expansion and a bottleneck. The simulated expansion was a step change to size N_2 (N_2/N_1=3.1), at time t_2/N_1=1. The simulated bottleneck was a reduction in population size N_2/N_1=0.72 at time t_2/N_1=1.1 and a subsequent expansion with a step change in size N_3/N_1=3.8 at time t_3/N_1=0.11. The parameters for the two simulated demographic scenarios were chosen to match the inferred histories of real populations. The simulated expansion matched that inferred for a population of wild mice (Halligan et al. 2010) and for the American population of humans with African ancestry (Boyko et al. 2008). The bottleneck scenario matched that inferred for the American population of humans with European ancestry (Boyko et al. 2008). For these simulations we assumed a gamma DFE with a=0.05 and b=0.5. For each simulated dataset we used 10^6 neutral and 10^6 selected sites, sampled 64 alleles and performed 20 replicate simulations.

Simulations with linkage. We used the C++ program *SLiM*, developed by Philip Messer and available at http://www.stanford.edu/ ~messer/software.html to perform simulations with linkage (Messer 2013). We simulated 1 Mbp long chromosomes. Each chromosome had 20 loci. Each locus consisted of 10 exons of length 100 bp each alternating with 1Kbp introns. The loci were at a distance of 40 Kbp from each other. We used exonic sites and the first 100 bp of introns as selected and neutral sites respectively. We simulated a population of size N=100 for 10N generations to reach stationary equilibrium, and sampled 64 chromosomes every 2N generations for 100N generations to obtain polymorphism data for a total of 10^6 selected and 10^6 neutral

sites. We assumed a mutation rate $4N_e\mu$=1% and simulated various levels of linkage between sites by assuming recombination rates ($4N_er$) varying between 10^{-5} to 1. We performed three types of simulations, varying the properties of the DFE for selected sites: Firstly, we assumed a bimodal DFE consisting of 2 spikes of selection coefficients (N_es_1=0, N_es_2=10, p_1=0.2), secondly we assumed a gamma DFE (a=0.05, b=0.5), and thirdly we assumed that 97% of sites were under negative selection (gamma DFE; a=0.05, b=0.5) and 3% were under positive selection (single spike DFE; N_es_1=10). We performed 20 replicate runs for each simulation type.

Evaluation of model performance. The performance of the models was assessed by their accuracy in inferring the mean effect ($\overline{N_e s}$), the average fixation probability of new deleterious and neutral mutations relative to the fixation probability of neutral mutations ($\bar{u}$) and the proportion of mutations falling into five N_es categories (0.0-0.1, 0.1-1.0, 1.0-10.0, 10.0-100.0, >100.0). $\overline{N_e s}$ and $\bar{u}$ are important quantities for several questions, including inferring the proportion of mutations fixed by positive selection and the rate of adaptive relative to neutral evolution (i.e. α and ω_α, respectively; (Eyre-Walker and Keightley 2009; Gossmann et al. 2010). $\overline{N_e s}$ was calculated by taking the arithmetic average of the selection coefficients over the range of s between 0 and 100 (i.e. the N_es range was between 0 and 10,000, for N_e=100). $\bar{u}$ was calculated by integrating over the DFE, as in Eyre-Walker and Keightley (2009):

$$\bar{u} = \int_0^\infty 2 N_e u(N_e, s) f(s|\Theta) ds \quad (2.7)$$

where u (N_e, s), is the fixation probability of a new deleterious mutation (Fisher 1930; Kimura 1957; Kimura 1962).

To assess the accuracy in recovering the properties (X) of the simulated distributions, we compared estimates (X_i) versus true values (X_{true}). For $\overline{N_e s}$ and $\bar{u}$, we calculated the relative error as:

$$rel.\ error(X) = \frac{X_i - X_{true}}{X_{true}} \quad (2.8)$$

We compared the goodness of fit between models by comparing their likelihoods and by comparing Akaike Information Criterion (*AIC*) scores. The *AIC* score penalizes parameter-rich models as follows:

$$AIC = 2k - 2\log(L) \quad (2.9)$$

where k is the number of parameters in the model, and L is the maximum likelihood for the estimated model. We considered an *AIC* difference greater than 2 as significant when comparing models. For the spike/step models we increased the number of fitted spike/steps until an improvement of less than 2 *AIC* units was obtained.

Analysis of *Drosophila* and house mouse datasets. We analysed polymorphism data for protein-coding genes of *Drosophila melanogaster* and *Mus musculus castaneus* using the six selection models described above. For these analyses we also fitted a simple demographic model of a step change in population size, similarly to the simulations assuming variable population size that were described above. The

initial population size (N_1) was assumed to be 100. The effective population size of real populations may be much larger than 100, but since the parameters of the demographic model are the magnitude of the size change (N_2/N_1) and the time of the size change event in generations scaled by N_1 (t_2/N_1), we do not expect the choice of N_1=100 to lead to biases in the estimation procedure, as shown previously by Keightley and Eyre-Walker (2007).

For *D. melanogaster*, we analysed a dataset of 17 alleles from individuals originating in East Africa (17 haploid Rwanda lines from the DPGP2 project; release version 2.0, http://www.dpgp.org/dpgp2/DPGP2.html; Pool et al. 2012). The dataset included polymorphism data for 8,367 autosomal genes. Regions with evidence of cosmopolitan (out-of-Africa) admixture were filtered-out. These regions had previously been identified by an identity-by-descent analysis by the DPGP team.

For *M. m. castaneus*, we used a dataset of 20 alleles from individuals sampled in NW India (Halligan et al 2010; Halligan et al. unpublished data). The dataset included data for 18,110 autosomal genes. CpG dinucleotides have substantially higher mutation rates in mammals (Arndt et al. 2003) and their frequencies differ between coding and noncoding DNA. Therefore for *M. m. castaneus*, we restricted the analysis to nonCpG-prone sites (sites not preceded by C or followed by G).

We quantified the DFE and calculated α and ω_a for non-synonymous 0-fold degenerate sites. Our method requires a neutrally evolving class of sites. We used synonymous 4-fold degenerate sites as the neutral class of sites. Extensive evidence for non-neutrality of synonymous sites exists for *D. melanogaster* (for example,

Lawrie et al. 2013). Violation of our assumption for neutrality of synonymous sites will therefore lead to an underestimation of the strength of selection at non-synonymous sites, but we do not expect biases when comparing the performance of the different selection models, since we used the same neutral class across these. To calculate α and ω_a we used the divergences at non-synonymous and synonymous sites between *D. melanogaster* and *D. yakuba* and between *M. m. castaneus* and rat, as follows:

$$\alpha = \frac{d_N - d_S \overline{u}}{d_N} \tag{2.10}$$

$$\omega_a = \frac{d_N - d_S \overline{u}}{d_S} \tag{2.11}$$

where d_N and d_S are the nucleotide divergences between the focal species and the outgroup at non-synonymous and synonymous sites, respectively.

2.4 Results

2.4.1. Simulations testing the performance of the models to infer unimodal and multi-modal DFEs

We simulated SFS datasets, choosing the parameters of the simulated distributions to create three biologically plausible scenarios for their complexity (i.e. unimodality, bimodality and multimodality; Figure 2.1). We then examined the performance of several models incorporating parametric or non-parametric distributions. We

considered several criteria for evaluating the performance of the tested models: the log-likelihood and *AIC* scores and the accuracy in estimating the mean effect of a new mutation ($\overline{N_e s}$) and the average fixation probability of a new mutation ($\bar{u}$). The accuracy in estimating $\overline{N_e s}$ and $\bar{u}$ was evaluated by calculating the relative error (equation 2.8). Δlog$L$, ΔAIC scores, rel.error ($\overline{N_e s}$) and rel.error ($\bar{u}$) for the tested models of each simulation set are shown in Table 2.2. We also examined the performance of the tested models in accurately inferring the proportion of mutations in five $N_e s$ ranges (Figure 2.2). We discuss in turn the results of each simulation set below. The estimates for the parameters of each of the six tested models for each simulation set (SIM1, SIM2, SIM3) are given in Appendix A2.1.

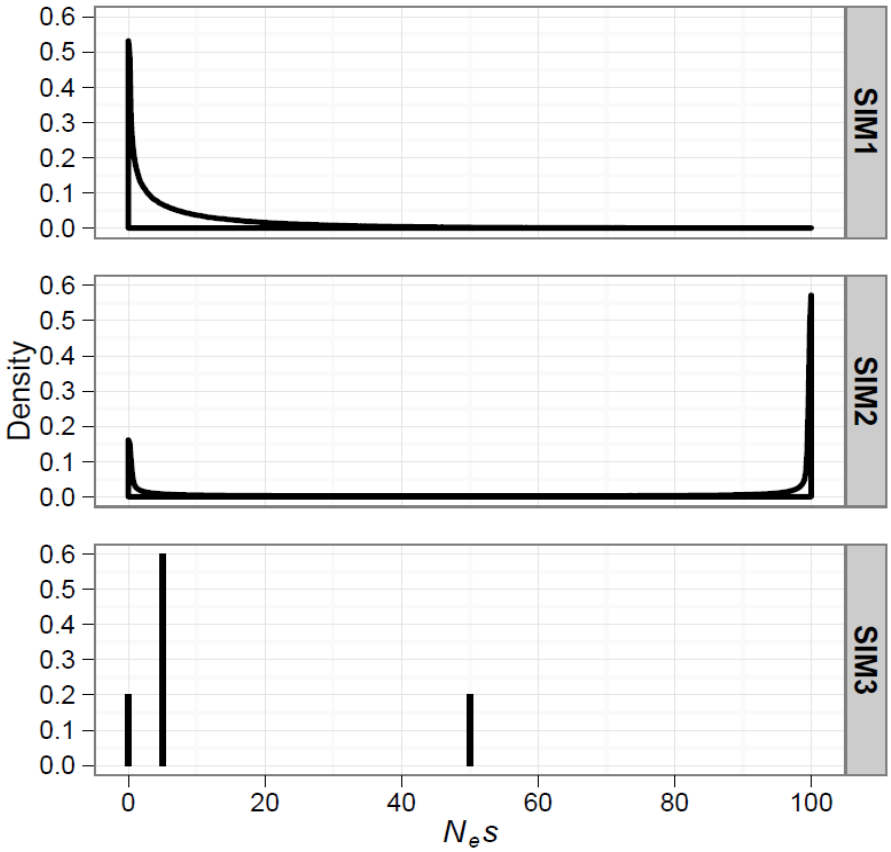


Figure 2.1. The simulated DFEs. For SIM1, we simulated a gamma DFE with scale $a = 0.05$ and shape $b = 0.5$. For SIM2, we simulated a beta DFE with shape parameters $k_1 = 0.2$ and $k_2 = 0.1$ rescaled to the $N_e s$ interval [0, 100]. For SIM3, the DFE was comprised of three selection coefficients, $N_e s_1 = 0$, $N_e s_2 = 5$, $N_e s_3 = 50$, with probability densities $p_1 = 0.2$, $p_2 = 0.6$, $p_3 = 0.2$.

Table 2.2. Statistics for the performance of tested models for each simulation set. The statistics reported are the mean log-likelihood and the mean *AIC* score difference from the highest scoring model and mean relative error in estimating $\overline{N_e s}$ and $\bar{u}$ across 100 replicates of each simulation set. The best-scoring model according to the *AIC* criterion is highlighted in light grey. 95% confidence intervals for rel.error ($\overline{N_e s}$) and rel.error ($\bar{u}$) are reported in brackets. For spikes and step models, the number of spikes/steps that best fitted the data according to the *AIC* criterion is shown in parentheses. Positive and negative values of %rel. error signify overestimation and underestimation of these parameters, respectively.

				% rel.error	
Simulation	**Model**	***Δ*log*L***	***ΔAIC***	$\overline{N_e s}$	$\bar{u}$
SIM1 (gamma)	Lognormal	-13.9	-27.8	94.3 [72.9, 116]	-14.4 [-16.8, -11.4]
	Gamma	-0.02	0	0.2 [-5.4, 5.7]	-1.47 [-3.96, 1.9]
	Beta	-0.3	-0.5	-5.7 [-10.6, -1.1]	-0.16 [-2.68, 3.30]
	Best spike (2)	-1.5	-4.9	-5 [-32, 464]	-14.7 [-32.4, 26.5]
	Best step (2)	0	-2	-11.8 [-18.4, -4.2]	-16.52 [-20.8, -10.6]
	6-fixed spikes	-0.6	-7.1	14.1 [4.4, 44.6]	5.51 [-17.7, 18.1]
SIM2 (bimodal beta)	Lognormal	-300	-597	979 [933, 1015]	-24.1 [-27.4, -21.2]
	Gamma	-46.4	-89.9	321 [279, 371]	-15.2 [-17.9, -12.7]
	Beta	-1.4	0	0.27	0.64

Simulation	Model	$\Delta \log L$	ΔAIC	% rel.error	
				$\overline{N_e s}$	$\bar{u}$
				[-1.74, 2.52]	[-2.96, 4.43]
	Best spike (3)	0	-3.1	-0.85 [-12.0, 173]	-12.9 [-29.1, 12.0]
	Best step (2)	-1.3	-1.8	18.3 [11.4, 25.7]	-9.31 [-16.3, -0.97]
	6-fixed spikes	-3.5	-10.2	2.26 [-7.24, 8.67]	-0.15 [-32.7, 18.2]
SIM3 (3-spike multimodal)	Lognormal	-29.5	-53	21.7 [7.85, 35.0]	-25.3 [-27.9, -22.1]
	Gamma	-6.9	-7.8	-26.1 [-30.1, -21.9]	-6.91 [-33.1, 2.96]
	Beta	-8.2	-10.4	-29.5 [-33.1, -25.9]	-26.3 [-28.8, -23.1]
	Best spike (3)	0	0	6.15 [-13.9, 107]	-3.48 [-47.6, 4.27]
	Best step (3)	-0.7	-1.3	20 [-45.4, 252]	-19.0 [-35.6, -13.2]
	6-fixed spikes	-0.6	-1.3	9.85 [-10.5, 43.8]	-13.6 [-40.0, 2.31]

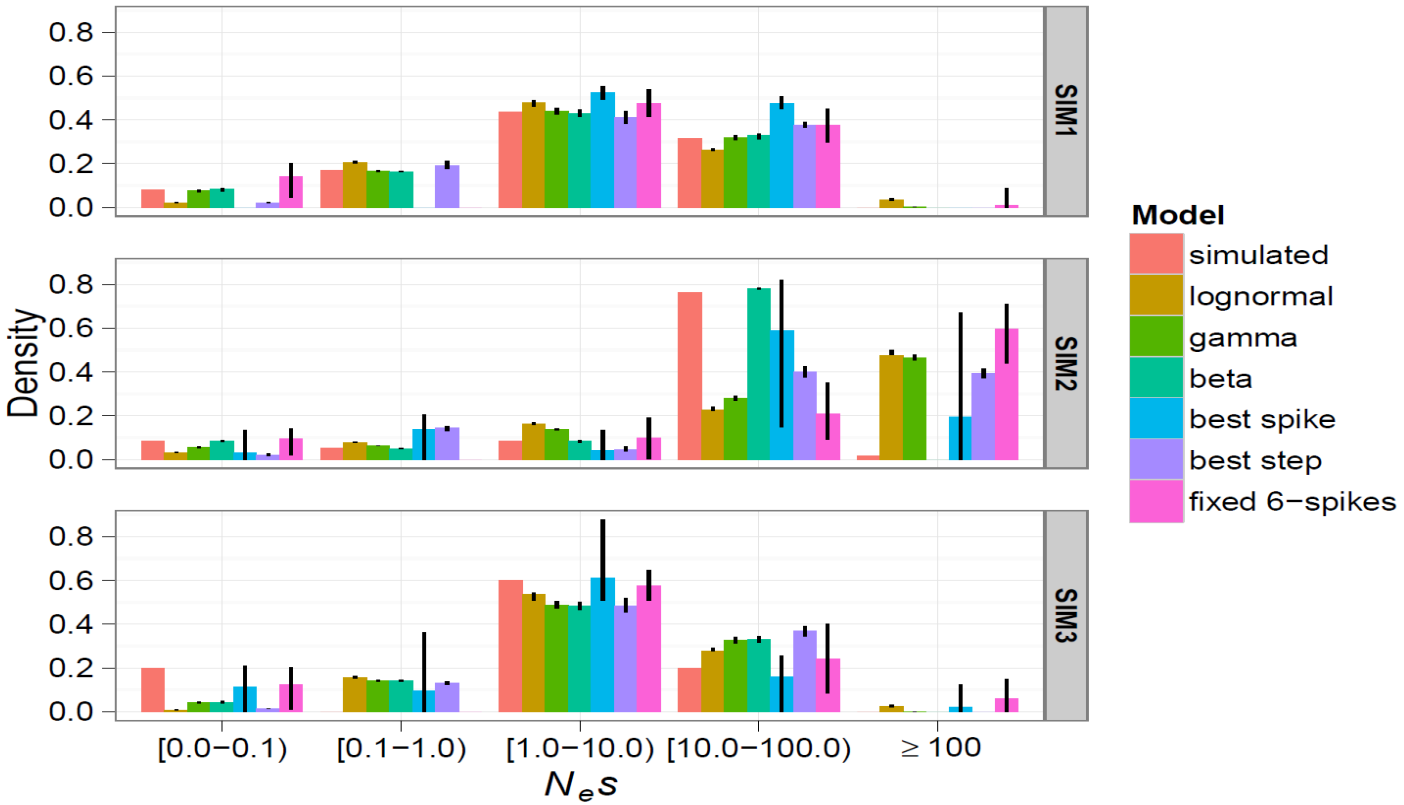


Figure 2.2. The mean estimated proportions of mutations in five $N_e s$ ranges for SIM1, SIM2 and SIM3, assuming a sequencing effort of 64 alleles and 10^6 neutral and selected sites. Error bars are the 5th and 95th percentiles of estimates over 100 simulation replicates.

A gamma distribution simulated (SIM1). To approximate a realistic scenario for protein-coding loci, where current information suggests a leptokurtic DFE and most sites under strong negative selection, we simulated a gamma DFE with scale $a = 0.05$ and shape $b = 0.5$ (SIM1; Figure 1).

As expected, the gamma model gave the best fit to the data, accurately estimating $\overline{N_e s}$ and $\bar{u}$ (SIM1; Table 2.2). The lognormal model performed more poorly, overestimating $\overline{N_e s}$ and underestimating $\bar{u}$, while the beta model gave a good fit (ΔAIC from the best-fitting model was -0.5) and accurately estimated $\overline{N_e s}$ and $\bar{u}$ (SIM1; Table 2.2). Based on their *AIC* scores, the best-fitting variable spike and variable steps models were the 2-spike and 2-step models, respectively (SIM1; Table 2.2), and these models fitted only slightly worse than the gamma model (SIM1; Table 2.2). However they did not recover $\overline{N_e s}$ and $\bar{u}$ as accurately as the gamma (SIM1; Table 2.2).

All models tested performed well in accurately recovering the proportions of mutations in the $N_e s$ ranges we examined (Figure 2.2). However, the lognormal and all the non-parametric models did not succeed in accurately assigning the proportions of mutations in the $N_e s$ ranges 0.0-0.1 and 0.1-1.0, presumably because there is little information to discriminate between these categories. In contrast, the gamma and beta models performed almost perfectly in assigning the proportions of mutations to these categories.

A bimodal beta distribution simulated (SIM2). We then investigated a beta distribution with shape parameters k_1=0.2 and k_2 = 0.1 rescaled to the N_es interval [0, 100] (SIM2; Figure 2.1). For this distribution, roughly 10% of selected sites are under weak negative selection (N_es < 1), another 10% are under moderately strong negative selection (N_es = 1-10) and the remaining 80% are under very strong negative selection (N_es>10). Such a bimodal distribution is intended to model protein-coding loci where amino-acid changing mutations are either neutral or strongly deleterious, with relatively few mutations of intermediate effect.

As expected, the beta model had the best *AIC* score (SIM2; Table 2.2), recovering $\overline{N_es}$ and $\bar{u}$ accurately (SIM2; Table 2.2). The unimodal lognormal and gamma models fitted the data very poorly (ΔAIC from beta = -597.2 for the lognormal and -89.9 for the gamma, SIM2; Table 2.2). $\overline{N_es}$ was grossly overestimated by the lognormal and gamma models (SIM2; Table 2.2). However, $\bar{u}$ was estimated relatively accurately by these models (SIM2; Table 2.2). The estimate for $\overline{N_es}$ can be heavily influenced by a long tail in the fitted distribution whereas $\bar{u}$ is mostly affected by effects in the N_es range 0-1. Therefore, the low accuracy of $\overline{N_es}$ and $\bar{u}$ estimates from the lognormal and gamma models presumably reflects a bad fit to the 'strong effects' part of the distribution (i.e. N_es>10), but there is a reasonably good fit to the 'nearly neutral effects' part of the distribution (i.e. 0<N_es<1). The best-fitting 3-spike and 2-step models and the fixed 6-spike model fitted almost as well as the beta distribution (SIM2; Table 2.2). These non-parametric

models accurately estimated $\overline{N_e s}$ and $\bar{u}$ (SIM2; Table 2.2). We observed that the lognormal, gamma and non-parametric models assigned substantial proportions of mutations into the $N_e s$>100 range (Figure 3), although the simulated distribution had a near-zero density in this range. Presumably, there is little information to precisely estimate the upper limit of the simulated distribution.

We also examined the performance of the models when varying the locations of the modes of a bimodal DFE. We investigated distributions with two classes of effects (2-spike): the first class of mutations was assumed to be neutral with $N_e s_1$ =0, and we varied the selection strength and probability density associated with the second class ($N_e s_2$ and p_2 respectively). We then fitted the gamma and the 3-step models to these distributions and compared their performance (Figure 2.3).

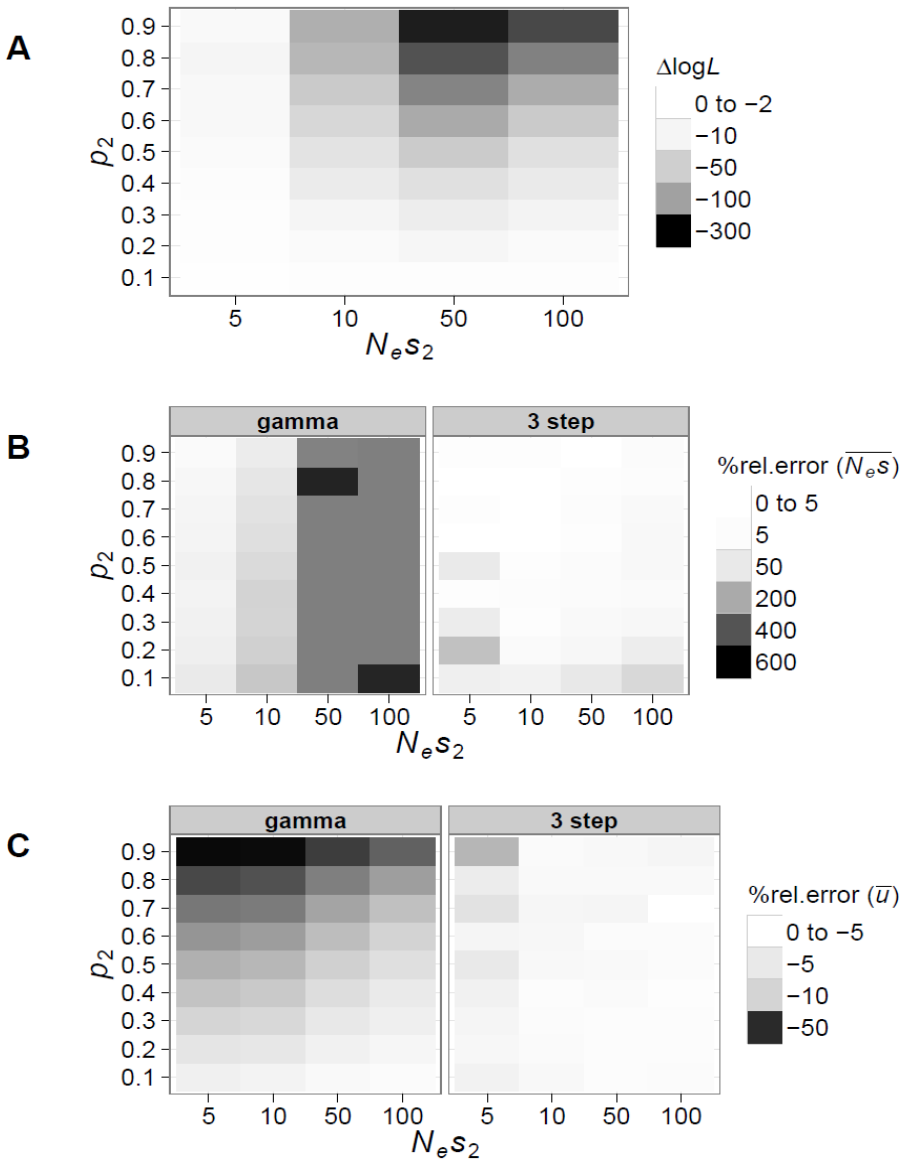


Figure 2.3. The performance of the gamma and 3-step models when fitted to bimodal DFEs. We simulated 2-spike DFEs with one spike fixed at $N_e s_1$ =0 and we varied the selection strength ($N_e s_2$) and probability density (p_2) of the second spike. (A) Δlog$L$ between the 3-step and gamma models fitted to the simulated DFEs as a function of $N_e s_2$ and p_2 . We also compared the % rel. error in estimating (B)

$\overline{N_e s}$ and (C) $\bar{u}$. Positive and negative values of % rel. error signify overestimation and underestimation of these parameters, respectively.

We found that for 2-spike distributions where $N_e s_2 \geq 10$ and $p_2 \geq 0.4$, the 3-step model significantly outperformed the gamma model (Figure 2.3A). We then examined the performance of the models in estimating $\overline{N_e s}$ and $\bar{u}$. We found that the gamma model overestimated $\overline{N_e s}$ when $N_e s_2 \geq 10$ and underestimated $\bar{u}$ for almost all parameter combinations of $N_e s_2$ and p_2 (Figure 2.3B and Figure 2.3C, respectively), while the 3-step model overestimated $\overline{N_e s}$ and underestimated $\bar{u}$ when $N_e s_2 < 10$ (Figure 2.3B and Figure 2.3C respectively).

A 3-spike multimodal distribution simulated (SIM3). To examine a case where the true DFE is more complex, we simulated a DFE comprising of three selection coefficients, $N_e s_1 = 0$, $N_e s_2 = 5$, $N_e s_3 = 50$, with probability densities p_1=0.2, p_2=0.6, p_3=0.2, respectively (SIM3; Figure 2.1). The choice of parameters was mainly based on generating 3 sufficiently distinct modes. As expected, a 3-spike model gave the best fit according to the *AIC* criterion (SIM3; Table 2.2). The other non-parametric models fitted almost equally well (*ΔAIC* was -1.3 for both the 3-steps model and the fixed 6-spikes model, SIM3; Table 2.2). The lognormal, gamma and beta models gave a poorer fit than the non-parametric models (*ΔAIC* was -53, -7.8 and -10.4 for the lognormal, gamma and beta models, respectively, SIM3; Table 2.2). However, we did not observe large differences in the accuracy of estimating $\overline{N_e s}$

and $\bar{u}$ between the models tested (SIM3; Table 2.2). The lognormal, best spike, best step and fixed 6-spike models slightly overestimated $\overline{N_e s}$, whereas the gamma and beta models slightly underestimated $\overline{N_e s}$ (SIM3; Table 2.2). All models tested slightly underestimated $\bar{u}$, the most accurate being the best fitting spike model (SIM3; Table 2.2).

The effect of increasing the allele sequencing effort. The primary goal of this section was to examine whether the general trends in the performance of the six models tested hold for different allele sequencing efforts. We compared the performance of the models for 8, 16, 32, 64, 128 and 256 alleles sequenced. For the gamma distribution (SIM1), increasing the sequencing effort led to more accurate estimates of $\overline{N_e s}$ but did not improve accuracy of estimating $\bar{u}$ (Figure 2.4A and Figure 2.4B, respectively). For the beta distribution (SIM2), increasing the allele sequencing effort increased the accuracy of estimating $\overline{N_e s}$ but the accuracy of estimating $\bar{u}$ did not increase for the spike, step and fixed 6-spikes models, and surprisingly decreased for the lognormal and gamma models (SIM2; Figure 2.4B). This decrease can be explained if we consider that the overall fit of the gamma and lognormal models improves as the number of alleles sequenced is increased, but the fit of the models to the $N_e s$ range 0-1 worsens (the good fit of the models to the $N_e s$ range 0-1 is crucial for an accurate estimate of $\bar{u}$). For the 3-spike multimodal distribution (SIM3), we observed that the parametric lognormal, gamma and beta

models showed no improvement in accuracy for estimating $\overline{N_e s}$ and $\bar{u}$ when increasing the number of alleles sequenced (SIM3; Figure 2.4A and Figure 2.4B, respectively). The spike, step and 6-fixed models at low sequencing efforts (8-32 alleles) had an inferior performance compared to the parametric models (SIM3; Figure 2.4). However, as the number of alleles sequenced was increased to 64 or greater the performance of these models became superior to the parametric models (SIM3; Figure 2.4).

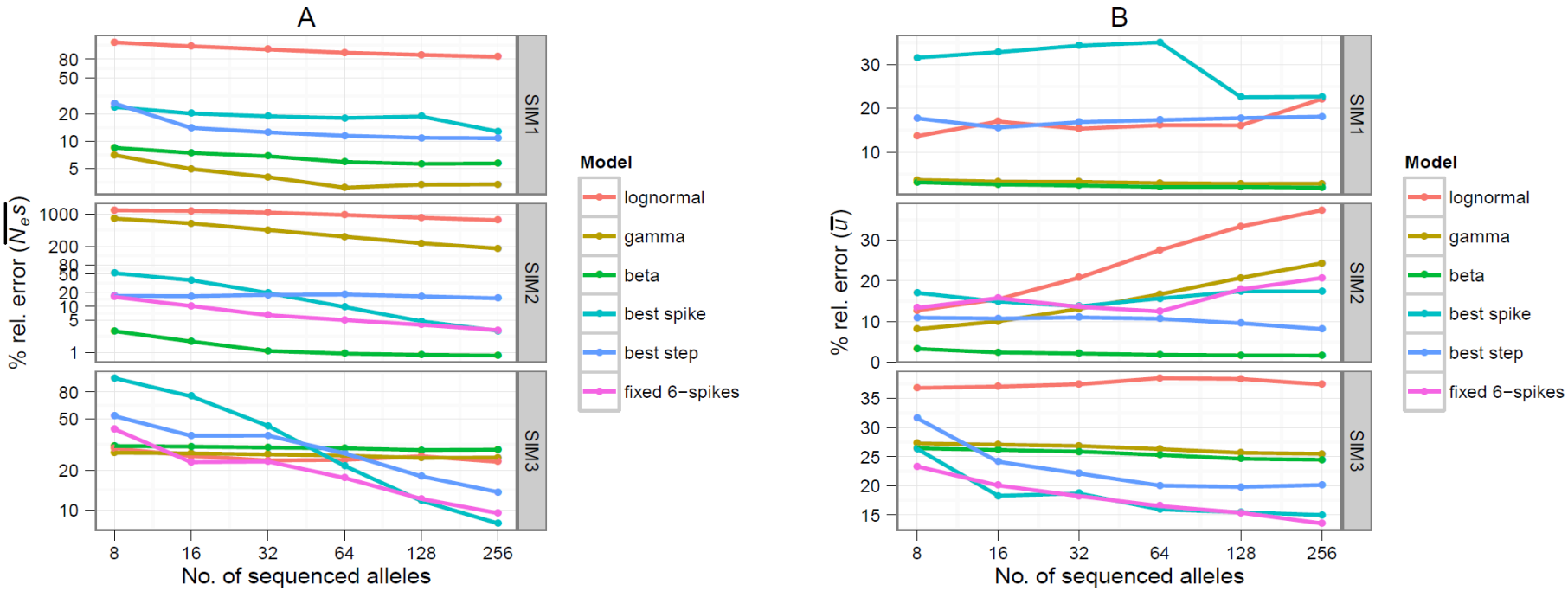


Figure 2.4. Mean estimates of % rel. error in estimating $(\overline{N_e s})$ (A), and $(\bar{u})$ (B) for the models tested when increasing the number of sequenced alleles for SIM1, SIM2 and SIM3. The y axis is log-scaled for panel A.

2.4.2. Simulations testing the robustness of the models to population size changes and linked selection.

The effect of incorporating a population size change. We then examined whether population size changes can affect the performance of the non-parametric relative to the parametric models by simulating two population histories: an expansion and a bottleneck. The expansion was a 3-fold step-change in population size. The bottleneck was a long-lasting 30% reduction in population size, followed by a short-lived 4-fold step expansion. For the selected sites, we assumed a gamma DFE with scale $a = 0.05$ and shape $b = 0.5$ (as for SIM1). Since our method can incorporate a model of a step change in population size, we fitted this model to the neutral data for both simulated histories.

For the expansion scenario, the demographic parameters of the step change were accurately estimated and the performance of the different selection models was similar to SIM1 (Table 2.3). For the bottleneck scenario, the 2-epoch demographic model appeared to mostly capture the second change in population size (Table 2.3). However, the non-parametric 2-spike and 2-step selection models fitted the data better than the parametric models (Table 2.3). Therefore a long-lasting bottleneck followed by rapid expansion can produce a signal in the data that is not fully accounted for by the fitted 2-step demographic scenario and can cause the spike and step models to overfit the data and produce spurious evidence for multimodality.

Other population histories such as a bottleneck followed by long-lasting recovery or expansion gave similar results to the 2-step expansion scenario (result not shown).

Table 2.3. Estimates of demographic parameters for the fitted step change in population size, goodness of fit and summary statistics for simulations assuming a population expansion and a bottleneck. A gamma DFE was assumed with a=0.05 and b=0.5. The statistics reported are the mean log-likelihood and the mean *AIC* score difference from the highest scoring model (Δlog$L$ and ΔAIC respectively), the mean estimate of the mean effect of a new mutation ($\overline{N_e s}$), and of the probability of fixation of a new mutation ($\bar{u}$). Only results for the best-fitting spike and step model according to the *AIC* criterion are shown. The 5th and 95th percentiles of estimates over 20 simulation replicates are shown in brackets.

Simulation	**Demography**		**Selection**				
	N_2/N_1	t/N_1	**Model**	Δ**logL**	ΔAIC	$\overline{N_e s}$	$\bar{u}$
Expansion	3.1 [3.1, 3.1]	0.97 [0.95, 1.0]	Lognormal	-5.1	-7.1	41 [36, 48]	0.13 [0.12, 0.13]
			Gamma	-1.6	0.0	17 [16, 19]	0.16 [0.15, 0.16]
			Beta	-1.9	-0.8	16 [15, 17]	0.16 [0.15, 0.17]
			Best spike (3)	0.0	-2.9	24 [14, 50]	0.12 [0.081, 0.21]
			Best step (2)	-2.3	-3.4	14 [13, 15]	0.12 [0.11, 0.13]
			6-fixed spikes	-1.1	-5.2	20 [17, 31]	0.15 [0.11, 0.20]
Bottleneck	5.3 [5.0, 6.0]	0.11 [0.10, 0.12]	Lognormal	-26.8	-51.7	24 [21, 28]	0.18 [0.17, 0.18]
			Gamma	-9.9	-17.9	12 [11, 13]	0.21 [0.20, 0.22]
			Beta	-8.2	-14.5	11 [11, 12]	0.21 [0.20, 0.22]
			Best spike (2)	-0.7	-1.5	8.4	0.17

Simulation	Demography		Selection				
	N_2/N_1	t/N_1	Model	ΔlogL	ΔAIC	$\overline{N_e s}$	$\bar{u}$
						[7.8, 8.8]	[0.14, 0.21]
			Best step (2)	0.0	0.0	8.8 [8.4, 9.3]	0.25 [0.20, 0.28]
			6-fixed spikes	-5.9	-15.8	13 [11, 16]	0.24 [0.20, 0.30]

Simulated values

Expansion scenario: N_2/N_1=3.1, t/N_1=1, $\overline{N_e s}$ = 17, $\bar{u}$ = 0.16

Bottleneck scenario: N_2/N_1=0.72, N_3/N_1=3.8, t_2/N_1=1.1, t_3/N_1=0.11, $\overline{N_e s}$ = 11, $\bar{u}$ = 0.20

The effect of linkage and selection. In our simulations we have assumed that sites are unlinked, but genomes of real organisms can exhibit various amounts of linkage. We performed simulations assuming a range of recombination rates between sites to examine how linkage can affect the performance of the 3-steps model in detecting a bimodal DFE. This performance is assessed by a significantly better fit of the 3-steps model than the gamma model.

Firstly, we investigated whether background selection alone could produce a spurious signature of a bimodal DFE by simulating a gamma DFE with a=0.05 and b=0.5. We observed a better fit of the 3-steps model than the gamma model for high levels of linkage (Figure 2.1A; upper panel). However, when we fitted a demographic model of a step change to the neutral sites, a procedure which has been suggested to control for the effects of linkage (Messer and Petrov 2013), the 3-steps and gamma models fitted the data equally well at all levels of linkage (Figure 2.1A; lower panel).

Secondly, we examined whether positive selection could produce a signature of a bimodal DFE. We simulated a gamma DFE with a=0.05 and b=0.5 for negatively selected mutations and a single spike for positively selected mutations with selection strength $N_e s_a$=10 and probability density p_a = 0.03, which is similar to what has been observed for protein-coding genes in *D. melanogaster* (Schneider et al. 2011). We observed results very similar to those we obtained by assuming only negative selection; Figure 2.1B). Therefore fitting a demographic model to the neutral sites appears to be sufficient for controlling the effects of linkage in

producing spurious evidence of a bimodal DFE.

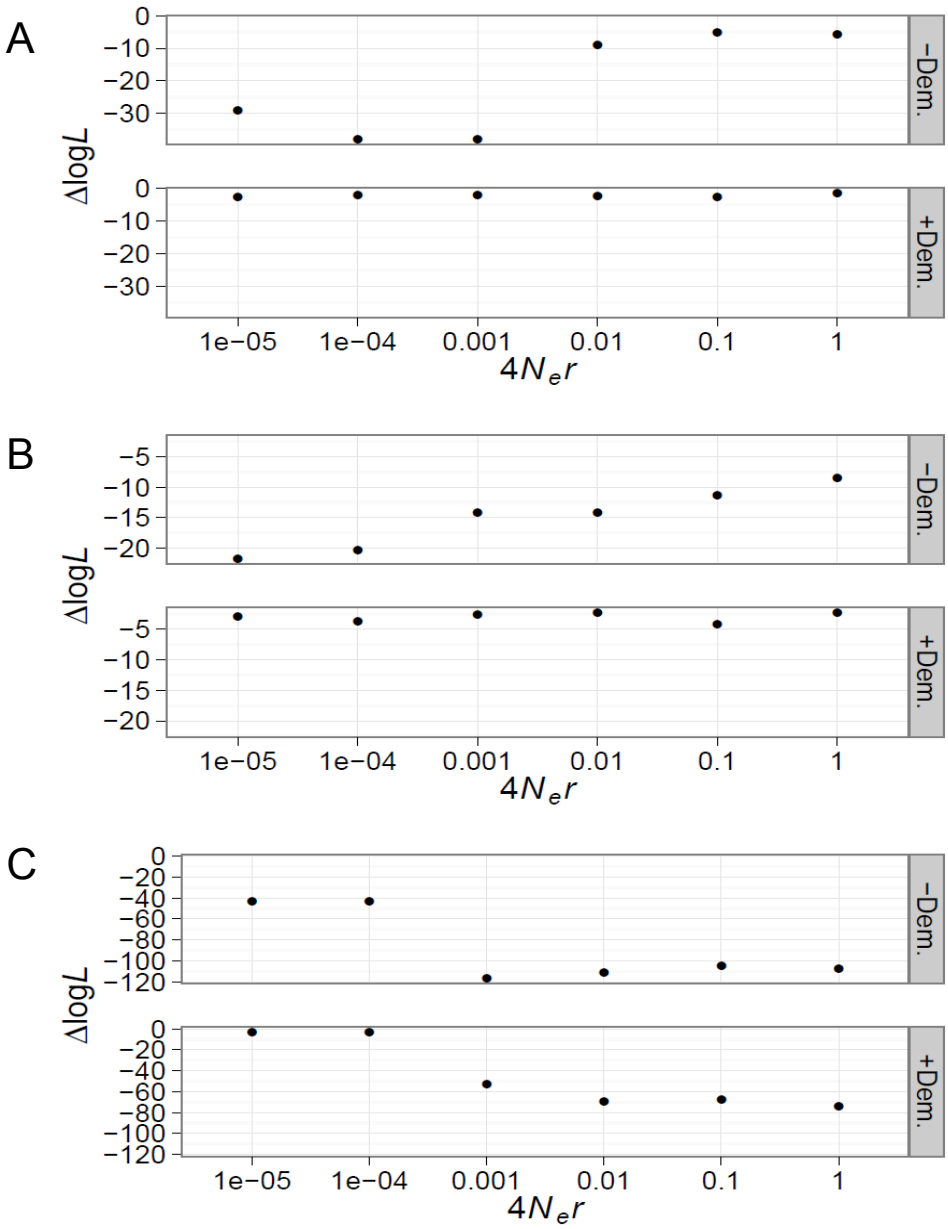


Figure 2.1. The effect of (A) background and (B) positive selection on producing spurious evidence for a bimodal DFE for various levels of linkage. (C) The effect of linkage on the power to detect a bimodal DFE. $\Delta\log L$ between the 3-step and gamma

model is shown for a range of recombination rates ($4N_e r$). Upper and lower panels contrast the results when fitting a demographic model to the neutral sites (the simulated population size is constant).

Thirdly, we investigated whether linkage could affect our power to detect a multimodal DFE with the non-parametric steps model. We simulated a bimodal 2-spike DFE with $N_e s_1 = 0$, $N_e s_2 = 10$ with probability densities p_1=0.2, p_2=0.8, respectively. We found that strong linkage can reduce the $\Delta \log L$ between 3-step and gamma models (Figure 2.1C; upper panel). The results were similar when we also fitted a demographic model of a step change to the neutral sites (Figure 2.1C; lower panel). Therefore, a true bimodal DFE would be harder to detect in genomic regions that exhibit strong linkage.

2.4.3 Analysis of protein polymorphism datasets from *D. melanogaster* and *M. m. castaneus.*

To account for demographic effects on our inferences of selection we fitted a step change in population size to synonymous sites. The step change model inferred a population expansion for both *D. melanogaster* and *M. m. castaneus* (Table 2.4) and fitted the data very well (Figure 2.1). Note that the size of the expansion (N_2/N_1) predicted for the two species is the same (2.79; Table 2.4). This is because the method employed searches a sparse parameter space consisting of only a few possible values for N_2/N_1, in order to make calculations less computationally

intensive (see also methods section).

Table 2.4. The demographic and selection parameter estimates obtained from the analysis of protein-coding loci in *D. melanogaster* and *M. m. castaneus*. The inferred parameters for the lognormal and beta model are given unscaled by $N_e=N_w$.

Species	Demography			Selection										f_0
	N_2/N_1	t/N_1	N_w	Model	$\mu/a/k_1$	$\sigma/b/k_2$	N_es_1	N_es_2	N_es_3	p_1	p_2	p_3	p_4	p_5	
D. melanogaster	2.79	0.11	110	Log-normal	-2.9	4.9	-	-	-	-	-	-	-	-	0.85
				Gamma	$1.6X10^{-4}$	0.33	-	-	-	-	-	-	-	-	0.85
				Beta	0.14	0.023	-	-	-	-	-	-	-	-	0.85
				Best spike (3)	-	-	$5.6X10^{-10}$	5.1	296	0.063	0.10	-	-	-	0.85
				Best step (2)	-	-	2.4	653	-	0.12	0.88	-	-	-	0.85
				6-fixed spikes	-	-	-	-	-	0.070	0.00	0.48	0.00	0.085	0.85
M. m. castaneus	2.79	1.48	182	Log-normal	-6.1	12	-	-	-	-	-	-	-	-	0.93
				Gamma	$2.1X10^{-7}$	0.12	-	-	-	-	-	-	-	-	0.93
				Beta	0.037	0.011	-	-	-	-	-	-	-	-	0.93
				Best spike (3)	-	-	$2.3X10^{-12}$	16.4	1056	0.19	0.12	-	-	-	0.93
				Best step (2)	-	-	$4.8X10^{-3}$	585	-	0.18	-	-	-	-	0.93
				6-fixed spikes	-	-	-	-	-	0.19	0.00	$5.3X10^{-3}$	0.025	0.00	0.93

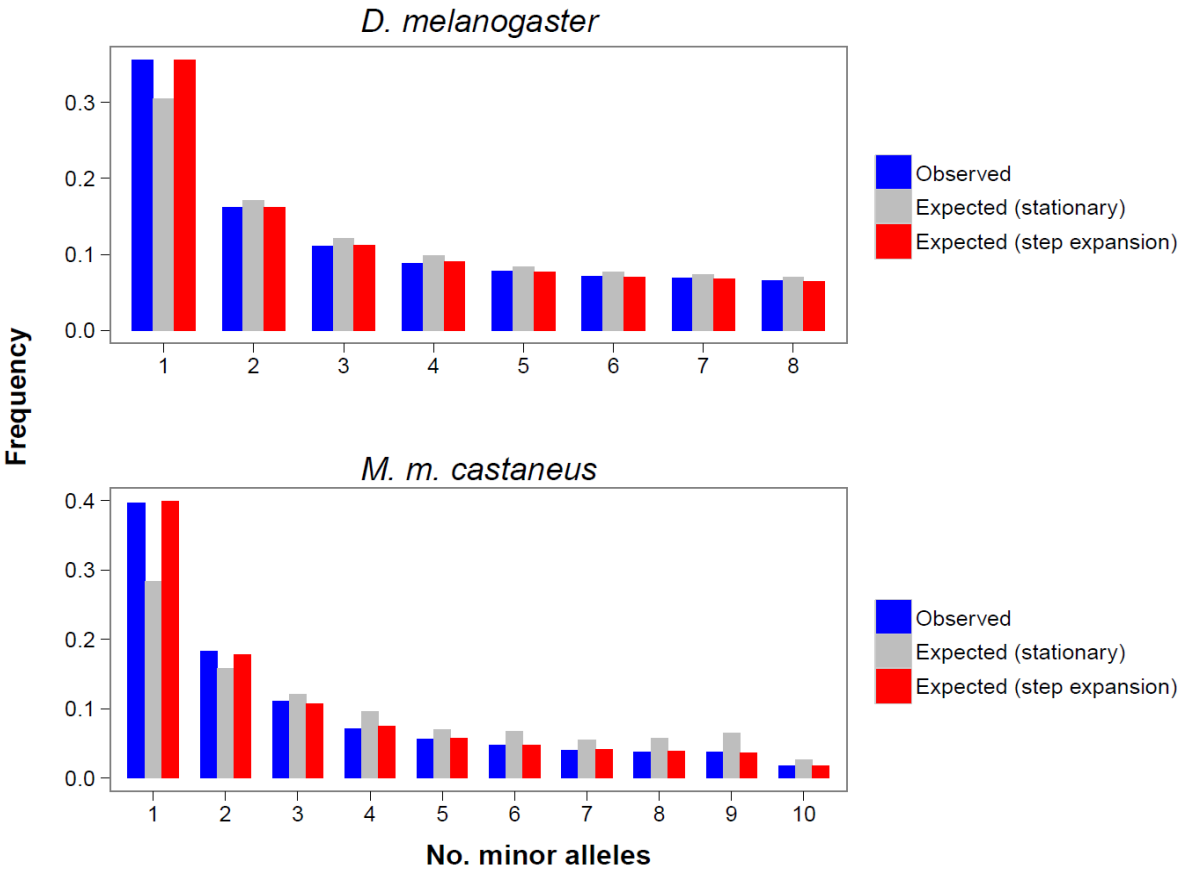


Figure 2.1. The observed site frequency spectrum and the expectation generated by assuming a stationary and the best-fitting expansion demographic models for *D. melanogaster* and *M. m. castaneus*. The expansion model was fitted to the synonymous site data.

We then fitted the lognormal, gamma, beta, variable spike, variable step and fixed 6-spike models to nonsynonymous sites to infer selection. For each dataset, we computed $\Delta\log L$, ΔAIC scores, the proportions of mutations falling into four $N_e s$ ranges (0-1, 1-10, 10-100, >100), $\overline{N_e s}$ and $\bar{u}$ (Table 2.5).

For *D. melanogaster*, we found that the best-fitting model according to the *AIC* criterion was the lognormal model, the gamma model having a slightly worse fit (ΔAIC from the lognormal was -5.1 units; Table 2.5). However, the estimated proportion of mutations in the $N_e s$ ranges examined, $\overline{N_e s}$ and $\bar{u}$ were very similar between these two models (Table 2.5). All models estimate that ~2-7% of new mutations are nearly neutral ($N_e s$ 0-1), a further ~4-20% are moderately to strongly deleterious ($N_e s$ 1-100), and ~80-90% are very strongly deleterious ($N_e s$ >100). The beta and 6-fixed spikes models gave a substantially poorer fit than the lognormal model (ΔAIC to lognormal was -187 units; Table 2.5). The main discernible difference was a ~10 times lower estimated $\overline{N_e s}$ for the beta and 6-fixed spikes models than the lognormal model. The beta and 6-spikes models do not allow selection strength $N_e s > N_e$ and their poor fit may be a consequence of a substantial proportion of mutational effects lying in that range.

Table 2.5. Log-likelihood and *AIC* score differences from the highest scoring model, the estimated proportion of mutations falling into four $N_e s$ ranges, the estimated mean effects of a new mutation ($\overline{N_e s}$), estimated mean probability of fixation of a new mutation ($\bar{u}$) and estimates of α and ω_a, obtained from the analysis of protein-coding loci in *D. melanogaster* and *M. m. castaneus*. The best-scoring model according to the *AIC* criterion is highlighted in light grey. Only results for the best-fitting spike and step models, based on the *AIC* criterion, are shown.

Species	**Model**	**Δlog*L***	**Δ*AIC***	$N_e s$				$\overline{N_e s}$	$\bar{u}$	**α**	**ω_a**
				[0-1)	**[1-10)**	**[10-100)**	**≥100**				
D. melanogaster	Lognormal	-0.8	0.0	0.044	0.064	0.11	0.78	1359.2	0.050	0.62	0.082
	Gamma	-3.3	-5.1	0.049	0.055	0.12	0.78	1624.1	0.054	0.59	0.079
	Beta	-94.2	-187.0	0.064	0.025	0.043	0.87	94.6	0.066	0.50	0.067
	Best spike (3)	0.0	-4.5	0.063	0.00	0.10	0.84	275.2	0.063	0.52	0.069
	Best step (2)	-3.2	-7.0	0.023	0.097	0.058	0.82	289.4	0.039	0.70	0.10
	6-fixed spikes	-72.3	-144.6	0.070	0.00	0.048	0.88	96.8	0.070	0.47	0.063
M. m. castaneus	Lognormal	-23.9	-41.8	0.17	0.052	0.061	0.72	1298.9	0.16	0.30	0.070
	Gamma	-21.2	-36.4	0.17	0.050	0.065	0.71	1840.1	0.16	0.29	0.069
	Beta	-4.4	-2.9	0.18	0.016	0.022	0.78	141.2	0.18	0.22	0.052
	Best spike (3)	0.0	0.0	0.19	0.00	0.12	0.69	755.4	0.19	0.20	0.047
	Best step (2)	-2.8	-1.6	0.18	0.0098	0.10	0.71	237.4	0.19	0.20	0.047
	6-fixed spikes	-2.9	-5.8	0.19	0.0053	0.02	0.78	142.6	0.19	0.20	0.046

For *M. m. castaneus*, the best-fitting model according to the *AIC* criterion was the 3-spike model (Table 2.5). The estimated parameter values were $N_e s_1$ =2.3X10^{-12}, $N_e s_2$ =16.4, $N_e s_3$ =1056, with probability densities p_1=0.19, p_2=0.12, p_3=0.69, respectively (Table S3). The 6-fixed spike, 2-step and beta models fitted only slightly worse than the 3-spike model, while the lognormal and gamma models had substantially worse fits (Table 2.5). The parameter estimates of the 3-spike model together with the good fit of the beta model support a bimodal DFE in *M. m. castaneus*. The DFE is inferred to have a peak at near neutrality ($N_e s$ 0-1) of density ~20%, and another peak at very strongly deleterious to lethal effects ($N_e s$ >100) with density ~70% (Table 2.5). Intermediate effects ($N_e s$ 1-100) are inferred to have a density of ~10% (Table 2.5).

The average fixation probability of a new deleterious mutation ($\bar{u}$) is an important quantity, since it can be used to estimate the fraction of adaptive substitutions between two species (Keightley and Eyre-Walker 2009). We calculated α and ω_a (equations 2.10 and 2.11) by using the estimated $\bar{u}$ for each model (Table 2.5). For *D. melanogaster*, we obtained values of α in the range 0.47-0.7 and ω_a 0.063-0.1 from the different models (Table 2.5). For *M. m. castaneus*, the lognormal and the gamma models gave slightly lower estimates for $\bar{u}$ and therefore higher estimates for α and ω_a (0.30 and 0.070, respectively; Table 3) than the best-fitting 3-spike model (0.20 and 0.047, respectively; Table 2.5).

2.5 Discussion

In this study, we have examined the performance of several models incorporating parametric and non-parametric distributions for inferring the properties of the DFE. Since the true DFE is of unknown complexity, and can have multiple modes, our purpose was to examine the performance of the different models when the true DFE was unimodal, bimodal or multimodal. We investigated parametric distributions, including the unimodal lognormal and gamma distributions, which are widely used to model the DFE, and the beta distribution, which can also take a bimodal shape. We also examined the performance of custom non-parametric models, including discretised distributions, where the selection coefficients are modelled as point masses, or uniform distributions, that are either variable or fixed.

For cases where the true DFE was a gamma distribution, a spike or step model with 2 or more classes performed almost as well as the gamma model. When the true DFE was a bimodal beta distribution, we found that the lognormal and gamma models fitted poorly, and produced inaccurate estimates of $\overline{N_e s}$, $\bar{u}$ and the density in several $N_e s$ ranges, most notably mutations with $N_e s$>100. When we simulated a more complex DFE, the biases affecting estimates of $\overline{N_e s}$ and $\bar{u}$ from the lognormal and gamma models were not as pronounced.

Accuracy in estimating $\overline{N_e s}$ and $\bar{u}$ seems to depend mostly on the density of the extreme tails of the DFE, irrespectively of its complexity. In our simulations, we frequently observed that a particular model could have a good overall fit, but

perform relatively poorly for parts of the DFE that are crucial for estimating $\overline{N_e s}$ or $\bar{u}$. For example, we consistently observed that $\bar{u}$ was not estimated with high accuracy if the models fitted were different from that simulated. Presumably, the SFS contains limited information about mutations with very small selective effects in the $N_e s$ range 0-1 implying that estimation of $\bar{u}$ strongly depends on the properties of the distribution assumed. Since $\bar{u}$ can be used for calculating the proportion of adaptive substitutions (α) and the rate of adaptive evolution (ω_a), underestimation of $\bar{u}$ would lead to overestimation of α and ω_a (and *vice versa*). When we examined a series of bimodal DFEs in which we varied the locations and densities of the two modes of the DFE, we observed substantial underestimation of $\bar{u}$ by the gamma model for cases where one mode of the DFE was at $N_e s$=0 with density <30% and the other mode was at a weakly to moderately deleterious effect with density >70%. Therefore, if the true DFE is bimodal, underestimation of $\bar{u}$ by the gamma model would be expected for genomic regions where most of the sites are under selection, such as protein-coding genes or conserved non-coding elements (CNEs), but not for genomic regions where most of the sites are evolving neutrally such as UTRs and introns.

We also applied the parametric and non-parametric models to infer the DFE for amino-acid-changing mutations in *D. melanogaster* and the house mouse *M. m. castaneus*, based on data from several thousand autosomal protein-coding genes. In *D. melanogaster*, we found that the lognormal model gave the best fit to the data, a result that is consistent with a previous study (Loewe and Charlesworth 2006). The

estimate for $\overline{N_e s}$ was 1360 by the best-fitting lognormal model. This estimate is similar to estimates obtained from a smaller dataset of Shapiro et al. (2007) analysed by Keightley and Eyre-Walker (2007). If we assume that the DFE for amino-acid changing mutations in *Drosophila* is lognormal, and that N_e is of the order 0.7×10^6 (Halligan et al. 2010), then the mean selection coefficient of new deleterious amino-acid changing mutations for *D. melanogaster* is of the order 2×10^{-3}. We also estimate that α and ω_α are 0.62 and 0.082, respectively. Reassuringly, the choice of the distribution to model the DFE does not strongly affect $\bar{u}$ and consequently α and ω_α. Regardless of the model assumed, α>0.47 *and* ω_α>0.063, supporting the presence of highly effective positive selection in *D. melanogaster*, as several other researchers have inferred (Sella et al. 2009).

In *M. m. castaneus*, we found that a 3-spike model gave the best fit to the SFS. The beta distribution also fitted almost as well as the 3-step model, while the lognormal and gamma models gave substantially poorer fits. These observations suggest that the DFE for new deleterious amino-acid changing mutations in *M. m. castaneus* is bimodal, with 20% of the distribution's density attributable to weakly deleterious mutations ($N_e s$ 0-1), and 70% to very strongly deleterious mutations ($N_e s$>100). We also obtained estimates for α and ω_α, of 0.20 and 0.046, respectively. We observed differences among the estimates of α and ω_α between different models, the lognormal and gamma models producing higher estimates than the best-fitting 3-spike and beta models. Underestimation of $\bar{u}$ by the gamma and lognormal models was observed in simulations where the true DFE was a bimodal beta of

similar properties to the inferred DFE for *M. m. castaneus*. It seems likely that fitting a lognormal or a gamma distribution to the DFE leads to overestimation of α and ω_α. Halligan et al. (2010), who fitted a gamma distribution to a small gene sample from *M. m. castaneus*, obtained larger estimates for α (α=0.37 for non-CpG-prone sites and using rat as outgroup) than those obtained in the present study.

There are some potential caveats to our study. Firstly, our models do not incorporate genetic linkage in the inference method. We investigated whether linkage and background or/and positive selection can affect inferences from the models tested, and found that under moderate linkage, spurious evidence for multimodality can be produced (assessed by a better fit of spike/step models to data than unimodal distributions). We can take account of the effects of linkage, however, by fitting a simple demographic model to the neutral class of sites (as is also suggested by Messer and Petrov 2013). Secondly, our 2-epoch demographic model is not sufficient for complex demographic histories, such as bottlenecks. Assuming a more realistic population history of a long-lasting bottleneck followed by a rapid expansion, we found that the spike/step models can overfit the data, producing spurious evidence for multimodality of the DFE. Therefore, when inferring the DFE using spike/step models it is necessary to fit a 3-epoch model to data from populations that have experienced bottlenecks. A 3-epoch model can be incorporated into the inference procedure of our method, but due to computational limitations it was not feasible to investigate its performance in simulations. However, a 3-epoch model fitted only slightly better to the folded synonymous SFS for *D. melanogaster* and *M. m.*

castaneus than a 2-epoch model ($\Delta \log L$ between the 2-epoch and 3-epoch model was 3 and 7, respectively; result not shown). Moreover, other demographic scenarios such as population subdivision have previously been shown not to produce biases in the estimates of the DFE. Note that the population samples that we examined are most likely from single, non-subdivided populations. The regions with evidence of admixture have been filtered-out for the *D. melanogaster* sample (see methods) and a previous study has shown no evidence for population structure for the *M. m. castaneus* sample (Halligan et al. 2010). Therefore, we do not expect a substantial effect of a complex demographic history on our inferences of selection in these populations. Thirdly, the fact that we infer a bimodal DFE for *M. m. castaneus* does not necessarily rule out a more complex DFE. It appears that there is limited information in the SFS, and our simulations indicate that at best 3 modes can be inferred, even for very large datasets. It is likely that the precise shape of the DFE cannot accurately be determined based on SFS data alone, as has been shown for the demographic history of a population (Myers et al. 2008).

In conclusion, we have shown that the DFE can be modelled reliably by non-parametric discretised models such as the spike and step models. The fit of these models is expected to be as good or better than parametric distributions, such as the gamma. They produce accurate estimates of the important parameters, notably $\overline{N_e s}$ and $\bar{u}$, and increasing the numbers of alleles sequenced will increase their performance. These models can also help in determining whether the DFE has multiple modes. We note that we have examined only one particular case of each

type of distribution (unimodal, bimodal, multimodal) and we do not consider the particular simulated examples as representatives of all possible unimodal, bimodal and multimodal distributions. However, our results are relevant in showing the limitations of fitting relatively inflexible distributions, such as the gamma distribution to the DFE, and illustrate the advantages of using a more general model such as the spike or step model to infer the DFE. Fitting the spike or the step model with different numbers of classes of mutational effects can be informative about the complexity of the DFE and identifying which $N_e s$ ranges we have little information on.

Chapter 3. Selection on genes and non-coding DNA in house mice

The work presented in this chapter has been published as a research paper:
Kousathanas A, Oliver F, Halligan DL, Keightley PD. 2011. Positive and negative selection on noncoding DNA close to protein-coding genes in wild house mice. *Molecular Biology and Evolution* **28**: 1183 –1191.

I present the work as published with small modifications. AK analysed the data and wrote the paper. FO conducted the sequencing experiments. DLH constructed the sequence alignments. DLH and PDK provided comments on previous versions of the manuscript.

3.1 Summary

During the past two decades, evidence has accumulated of adaptive evolution within protein-coding genes in a variety of species. However, with the exception of *Drosophila* and humans, little is known about the extent of adaptive evolution in non-coding DNA. Here, we study regions upstream and downstream of protein-coding genes in the house mouse *Mus musculus castaneus*, a species that has a much larger effective population size (N_e) than humans. We analyze polymorphism data for 78 genes from 15 wild-caught *M. m. castaneus* individuals, and divergence

from a closely related species, *Mus famulus*. We find high levels of nucleotide diversity and moderate levels of selective constraint in upstream and downstream regions compared to non-synonymous sites of protein-coding genes. From the polymorphism data, we estimate the distribution of fitness effects (DFE) of new mutations, and infer that most new mutations in upstream and downstream regions behave as effectively neutral and that only a small fraction are strongly negatively selected. We estimate the fraction of substitutions that have been driven to fixation by positive selection (α) and the ratio of adaptive to neutral divergence (ω_α). We find that α for upstream and downstream regions is much lower than α for non-synonymous sites. However, ω_α estimates are very similar for non-synonymous sites and upstream and downstream regions. We conclude that negative selection operating in upstream and downstream regions of *M. m. castaneus* is weak, and that the low values of α for upstream and downstream regions relative to non-synonymous sites are most likely due to the presence of a higher proportion of neutrally evolving sites and not due to lower absolute rates of adaptive substitution.

3.2 Introduction

In recent years, the search for evidence of adaptive evolution at the molecular level has been at the forefront of genetics research. A principal motivation has been to identify regions of the genome that have experienced adaptive evolution, since this might provide clues to their functional importance and may be informative about the features that make each species unique.

There have been a wealth of studies focusing on amino-acid changes in

protein-coding genes. Studies in *Drosophila*, employing variants of the McDonald-Kreitman (MK) test (McDonald and Kreitman 1991), suggest that a high proportion of amino acid substitutions are adaptive (α*;* the proportion of substitutions that have been fixed by positive selection is 50% or more) (Smith and Eyre-Walker 2002; Sawyer et al. 2003; Bierne and Eyre-Walker 2004; Welch 2006; Shapiro et al. 2007; Eyre-Walker and Keightley 2009), whereas in humans similar studies have produced low estimates of α (0-20%) (The Chimpanzee Sequencing and Analysis Consortium 2005; Zhang and Li 2005; Boyko et al. 2008; Eyre-Walker and Keightley 2009). These contrasting results between *Drosophila* and humans have been interpreted to be a consequence of different effective population sizes (N_e), i.e. the small N_e of the hominid lineage could have resulted in reduced efficacy of natural selection. Other evidence points to a positive relationship between α and recent N_e. For example, α for protein-coding genes has been estimated to be 50% or more in enteric bacteria, which have a large N_e (Charlesworth and Eyre-Walker 2006), close to zero in *Arabidopsis* (*A. lyrata* and *A. thaliana*), which have small N_e (Foxe et al. 2008), and about 40% in *Capsella grandiflora,* a species that is closely related to *A. thaliana* (divergence time ~10 MYA) and has a larger N_e (Slotte et al. 2010). A recent study of the house mouse *M. m. castaneus*, which has N_e comparable to *Drosophila*, produced a high estimate of α for protein-coding genes (~50%) (Halligan et al. 2010), again suggesting N_e as a determinant of the efficacy of positive selection across taxa. However, the possible relationship between α and recent N_e has been a controversial issue in the literature. A recent study of several species with varying N_e has found a positive correlation between α and recent N_e (Gossmann et al.

2012),while another found no significant correlation between these variables (Gayral et al. 2013).

Estimates of the frequency of adaptive nucleotide substitution in non-coding DNA are currently restricted to *Drosophila* and humans. In *Drosophila*, estimates of α for 5' and 3' untranslated regions (UTRs), are nearly as high as for protein-coding genes (50% or more; Kohn et al. 2004; Andolfatto 2005; Haddrill et al. 2008) and for introns and intergenic regions are relatively low (~10-20%; Andolfatto 2005; Haddrill et al. 2008). In humans, estimates of α for non-coding regions upstream and downstream of protein-coding genes are close to zero (Keightley, Lercher, et al. 2005; Eyre-Walker and Keightley 2009).

In this study, we investigate positive and negative selection operating on non-coding regions upstream and downstream of the protein-coding genes in a sample of the house mouse *M. m. castaneus* that were previously studied by Halligan et al. (2010). We study regions upstream and downstream of protein-coding genes, because they are known to be enriched for regulatory elements (Xie et al. 2005; Veyrieras et al. 2008), and are implicated in the control of transcription and translation (Gray and Wickens 1998; Shabalina and Spiridonov 2004). Previous studies in murids have shown that ~30% of sites in 5'- and 3'-UTRs and ~10% of sites that are within the first 3 to 5 kb upstream and downstream of the transcription start and stop codon, respectively, are subject to negative selection (Keightley, Lercher, et al. 2005; Gaffney and Keightley 2006). Here, we perform a more thorough investigation of negative selection operating in upstream and downstream regions, by estimating the full distribution of fitness effects of new mutations (DFE).

We then proceed to investigate positive selection by estimating α using a method that attempts to account for the presence of slightly deleterious mutations: the DFE is used to predict the expected divergence between two species caused by the fixation of neutral and slightly deleterious mutations, and is compared with the observed divergence (Eyre-Walker and Keightley 2009). The difference between the observed and expected divergence is used to estimate the amount of adaptive divergence and α. We also estimate ω_α, the rate of adaptive divergence relative to neutral divergence, which allows us to better compare rates of adaptive evolution between species, by controlling for the effects of N_e on the numbers of effectively neutral substitutions (Gossmann et al. 2010).

3.3 Materials and Methods

Sampling of mice. We analyzed 15 *M. m. castaneus* individuals sampled from 4 regions south of the Himalayas in the Himachal Pradesh state of India. Tests for population structure or admixture have been conducted in a previous study (Halligan et al. 2010) on this population sample using the program *Structure* (Pritchard et al. 2000). These tests had shown no significant evidence for population subdivision (Halligan et al. 2010). Therefore, we consider that our sample is from a single non-subdivided population.

We also generated sequence data from a *M. famulus* individual originating from India that was previously obtained from the Montpellier wild mice genetic repository (http://www.isem.cnrs.fr/spip.php?article4777). A more detailed description of the sampling of the mice can be found in Halligan et al. (2010).

Choice of genes. We analyzed 78 autosomal genes from *M. m. castaneus* whose human orthologs have also been sequenced as part of the Environmental Genome Project (EGP) (Livingston et al. 2004). The EGP dataset is enriched for genes that are involved in pathways for DNA repair, cell cycle control, drug metabolism, and apoptosis and is therefore non random (Livingston et al. 2004). The genes were chosen if there were African human polymorphism data available; this enabled us to more directly compare the results in humans with mice.

As part of this study, we successfully sequenced upstream and downstream regions for 49 and 51 genes respectively in 15 *M. m. castaneus* individuals and one *M. famulus* individual. We designed primers to amplify the upstream region of each gene, which lies approximately up to 500 bp upstream of the first codon of the first exon, as annotated in the reference mouse genome. Similarly, to amplify the downstream region of each gene, we designed primers that captured the region that lies approximately up to 500 bp downstream of the stop codon of the last exon in the reference mouse genome. We chose to sequence ~500 bp upstream and downstream of protein-coding genes, since evidence from studies of selective constraint, regulatory motifs and expression-QTLs suggest that there is a high density of functional elements in these regions (Xie et al. 2005; Gaffney and Keightley 2006; Veyrieras et al. 2008). Additionally, the interpretation of sequences further upstream and downstream from these regions was made more difficult by frequent indel variation, making calling of SNPs problematic. Details of the genes analyzed in this study and the upstream and downstream regions that were successfully sequenced are

given in Appendix A.3.1.

Note that we chose not to restrict our analyses to those genes for which upstream, downstream, exonic and intronic sequence data were all available because the smaller samples for intronic (65), upstream (49) and downstream (51) site classes are unbiased in relation to the larger dataset for exonic sequence (78). Moreover, for most genes we have both upstream and downstream sequence data (Appendix A.3.1). Additionally, in analyses where a putatively neutral (i.e. synonymous or intronic sites) and a selected class (i.e. non-synononymous, upstream or downstream) were required, we only analyzed genes for which both the neutral and the selected class were sequenced.

In this study, we have updated the dataset of Halligan et al. (2010) with new exonic and intronic sequence of the 78 genes. Instead of using the Halligan et al. (2010) dataset, we analyzed the updated exonic and intronic datasets, since we had ~20% new exonic sequence data and ~60% new intronic sequence data. The differences between the Halligan et al. (2010) dataset and the updated dataset used in this study are shown in Appendix A.3.2.

Sequencing. GoTaq DNA polymerase (Promega) was used in touchdown-style PCR reactions: an initial denaturation step of 95ºC for 15 minutes, followed by 28 cycles of 95ºC for 30 seconds, 62ºC for 45 seconds (reducing by 0.5ºC every cycle), 72ºC for 2 minutes, then 12 cycles of 95ºC for 30 seconds, 52 ºC for 45 seconds and 72ºC for 2 minutes, with a final extension step at 72ºC for 10 minutes. ExoSAP-IT (USB) was used for the purification of PCR products. If we obtained non-specific PCR

products, we designed new primers to try to increase the specificity. Sequencing was done using Big Dye Terminator Sequencing Kits (Applied Biosystems) on an ABI Prism 3730 DNA Analyzer, and both forward and reverse sequences were generated. CodonCode Aligner version 2.0.6 was used to analyze and detect variants (http://www.codoncode.com/aligner/). We used the *Phred* computer program, as employed in *CodonCode Aligner* to assess sequence quality. Sequences had an average *Phred* score of >60. All sequence traces were manually checked. Sites with a *Phred* score <30, which could be low quality sequence or heterozygotes, were manually checked but were not automatically excluded, in order to avoid excluding heterozygotes. Where sequence was found to be too low quality, due to multiple indels or repetitive regions, new amplicons were produced on either side of the difficult to sequence area. Such difficult to sequence areas were replaced by 'N's before further analysis. We used *CodonCode Aligner* to identify and analyze heterozygous indels and we checked very carefully SNPs that were not at Hardy-Weinberg equilibrium. Finally, we generated alignments of the 15 *M. m. castaneus*, the *M. famulus* individual and the *M. m. musculus* reference sequence using *CodonCode Aligner*, and checked all alignments by eye before further analysis.

Sequence processing. We obtained orthologous *Rattus norvegicus* sequences for each amplicon using a reciprocal-best-hits *BLAST* approach. To do this, we BLASTed the mouse reference sequence (mm9) for each amplicon, plus 200 bp of flanking DNA, against two different assemblies (labelled "standard" and "alternative") of the rat genome and searched for a reciprocal-best-hit. If we failed to

find a reciprocal-best-hit for the standard assembly, we searched the alternative assembly. Both assemblies were downloaded from UCSC genome browser; the standard was produced by the Baylor College of Medicine Human Genome Sequencing Center (BCM-HGSC) as part of the Rat Genome Sequencing Consortium, and the alternative was produced by Celera Genomics. If we failed to identify an ortholog via the reciprocal-best-hits approach, we checked the relevant section in the "multiz30way" whole genome sequence alignments of 30 vertebrates (http://genome.ucsc.edu/). We considered sequences to be orthologous if the sequence of interest was located entirely within a single unbroken alignment for mouse and rat. We realigned all alignments obtained by either method using *MAVID* (Bray and Pachter 2004) and then subsequently checked them all by eye. Any obviously mis-aligned sections identified when checking by eye were masked from any further analysis. Using this procedure, putatively orthologous rat sequences were obtained for at least part of every mouse amplicon.

We constructed alignments for each amplicon between the mouse reference (mm9) sequence, the sequences from all *M. m. castaneus* individuals, *M. famulus* and rat. We annotated sites according to the mouse reference genome into the following categories: 5', 3', intron or coding. Within the coding category, sites were categorised as 1st, 2nd or 3rd positions as well as the level of degeneracy in the genetic code (zero-fold-, two-fold-, or four-fold-degenerate). We excluded potential splice sites of introns (defined as the first 6bp or last 16bp of an intron) from any analysis. We also categorised sites on the basis of their CpG-prone status (defined as being preceded by a C or followed by a G in any species).

Summary statistics. We assume that segregating polymorphisms are biallelic. If there were more than two alleles segregating at a site we only consider the two most frequent alleles. We calculated two statistics for nucleotide diversity, π and Watterson's θ (θ_w):

Let the site frequency spectrum (SFS) of a class of sites be the vector $\mathbf{v}_i$, containing i ($0 \leq i < n$) segregating alleles in a sample of n alleles from the population. Then π and θ_w are calculated as follows:

$$\pi = 2 \frac{\sum_{i=1}^{n-1} i(n-i)\,\mathbf{v}_i}{n(n-1)} \tag{3.1}$$

$$\theta_w = \frac{\sum_{i=1}^{n-1} \mathbf{v}_i}{\sum_{i=1}^{n-1} \frac{1}{i}} \tag{3.2}$$

(Watterson 1975; Tajima 1983)

For our dataset, given that we sequenced both chromosomes of each of the 15 *M. m. castaneus* individuals, the sampled number of alleles (n) was 30 if the sequencing was successful for every individual. However, due to sequencing failures, our dataset did not contain 30 sequenced alleles for each site, so we calculated composite estimates of π and θ_w. We calculated π and θ_w for sites that had the same number of alleles sequenced (categories of coverage) and then calculated a weighted average across categories of coverage as per Halligan et al. (2010).

For a population at Wright-Fisher equilibrium, and assuming no selection, π and θ_w estimates are expected to be equal to one another. They are expected to differ, however, if there is a skew in the SFS towards low or high frequency alleles. The level of skew can be quantified by the Tajima's *D* statistic (Tajima 1989). However, to calculate *D*, there needs to be an equal number of alleles sequenced at each site. We therefore rejected any sites where we had fewer than 20 alleles sequenced. We then sampled without replacement 20 alleles from each of the remaining sites, such that the number of alleles sampled at each site was constant. We bootstrapped by gene with replacement 1000 times to perform statistical comparisons of *D* between different classes of sites or with zero.

Nucleotide divergence (*d*) between *M. m. castaneus* and both *M. famulus* and rat was calculated using the Kimura 2-parameter correction (Kimura 1980). We had multiple sequences for *M. m. castaneus*, so we computed an average divergence. We calculated evolutionary constraint *C* by comparing substitution rates at a putatively neutral and a selected site class. We use substitution rates at neutral sites to estimate expected numbers of subtitutions at selected sites. Expected (E_d) and observed (O_d) numbers of substitutions are compared, and constraint is calculated as $C_d = 1 - O_d/E_d$ (Eyre-Walker and Keightley 1999; Keightley and Gaffney 2003). We distinguish evolutionary constraint calculated using divergences and polymorphism (see below) by denoting them as C_d and C_p, respectively. We used synonymous sites or introns as the neutral class.

Non-synonymous and synonymous sites. We treated non-synonymous and

synonymous sites as in Li (1993) and Pamilo and Bianchi (1993): all zero-fold degenerate sites were treated as non-synonymous and four-fold degenerate sites as synonymous. At two-fold degenerate sites, transitions were considered synonymous and transversions non-synonymous. The 1st position of the codons for Arginine (AGA,AGG,CGA,CGG) was considered two-fold degenerate and the third position of the codons for Isoleucine (ATT, ATC, ATA) was considered four-fold degenerate. Unmutated two-fold degenerate sites were divided into non-synonymous and synonymous by considering the ratio of transitional and trasversional changes (t_s/t_v) as calculated at four-fold degenerate sites across all genes in a comparison of *M. m. castaneus* and *M. famulus*.

Distribution of fitness effects of new mutations. We employed a maximum likelihood (ML) approach described by Keightley and Eyre-Walker (2007) to infer the distribution of fitness effects of new mutations at non-synonymous sites of coding regions and in upstream and downstream regions as implemented in the program *DFE-alpha* (available online at: http://homepages.ed.ac.uk/eang33/).

DFE-alpha assumes two classes of sites, one neutral and one selected, and contrasts site frequency spectra (SFSs) of the two classes. Fitness effects of new mutations (s) are assumed to be zero in the neutral class, and unconditionally deleterious in the selected class, and are sampled from a gamma distribution with parameters a (scale) and b (shape). It has previously been shown that even in the presence of slightly advantageous mutations in the selected class of sites, the estimates of the parameters of the DFE of deleterious mutations will be unaffected

(Keightley and Eyre-Walker 2010). Additionally, if the true DFE is multimodal, we could misinfer its properties by fitting a gamma distribution which is unimodal (Kousathanas and Peter D. Keightley 2013). Multi-modal discretised distribution models have been previously developed (Chapter 2 in this thesis) that could potentially be used to account for this possibility. However, our dataset in this study was very small, which prohibited the use of these parameter-rich models (as demonstrated in Chapter 2, these models should be used for sequencing efforts of greater than 1 Mbp of sites and 8 sequenced alleles). Finally the *DFE-alpha* method assumes no linkage between sites, but this assumption is likely to be violated for natural populations. However, it has previously been shown that even in the presence of strong linkage the parameter estimates for the DFE will be accurate if a two-step size change is simultaneously fitted to the neutral class (Messer and Petrov 2013; Kousathanas and Keightley 2013).

DFE-alpha incorporates a simple demographic model: the population at an initial size N_1 experiences a step change to N_2, t generations in the past. Even though the demographic model implemented by *DFE-alpha* is simple, the estimates for the parameters of the DFE by *DFE-alpha* have previously been shown to be robust to more complex demographic histories such as bottlenecks and population subdivision (Eyre-Walker and Keightley 2009).

We use a constant N_1 of 100 so that the ratio N_2/N_1, i.e. the change in population size, is actually estimated. An additional parameter, f_0, estimates the proportion of unmutated sites. The parameter space of N_2/N_1, t/N_2, f_0, a and b is searched to find the values that maximize the likelihood of observing the neutral and

selected SFSs.

In order to account for variation in the number of alleles at each site, we generated SFSs for sites that had the same number of alleles sampled in both neutral and selected classes. We summed the log-likelihoods of each SFS to produce the overall log likelihood as per Halligan et al. (2010). We interpolated from the estimated parameters of the gamma distribution, the percentages of mutations that fall within four $N_e s$ ranges: 0 to 1, 1 to 10, 10 to 100, 100 to $+\infty$.

Estimating evolutionary constraint by using polymorphism data. Evolutionary constraint, calculated using divergences (C_d) as explained in the summary statistics section, will be biased downwards if some fraction of the observed divergence at the focal class is adaptive. We obtained a second estimate of evolutionary constraint, which is not subject to such biases, by using information from polymorphism data only (C_p). We first estimate the average fixation probability of new deleterious and neutral mutations relative to the fixation probability of neutral mutations ($\bar{u}$) at the focal class by integrating over the DFE, as in (Eyre-Walker and Keightley 2009):

$$\bar{u} = \int_0^\infty 2Nu(N, s) f(s|a, b)\, ds \tag{3.3}$$

where (u (N, s), is the fixation probability of a new deleterious mutation (N is assumed equal to N_e) .

C_p can then be calculated as:

$$C_d = 1 - \bar{u} \tag{3.4}$$

Quantifying adaptive evolution. To estimate the proportion of adaptive substitutions (α), approaches based on the McDonald-Kreitman test are frequently used (Eyre-Walker 2006). However, these approaches do not take into account slightly deleterious mutations, which contribute proportionally more to polymorphism than divergence and therefore can lead to underestimates of α. They also ignore demographic history, which can be problematic, since a population size change in the past could produce evolutionary signatures similar to selection. A recent extension of the McDonald-Kreitman test (*DFE-alpha*; Eyre-Walker and Keightley 2009) attempts to take into account both slightly deleterious mutations and population demography.

The nucleotide divergence of the neutral class (d_S) is assumed to be proportional to the mutation rate, and divergence due to deleterious mutations in the selected class is the product of the mutation rate and the average fixation probability of a new deleterious mutation ($\bar{u}$). We can estimate the expected divergence (d_{est}) in the selected class due to neutral and deleterious mutations as:

$$d_{est} = d_S \bar{u} \tag{3.5}$$

The difference between the observed (d_X) and estimated divergence (d_{est}), estimates the amount of adaptive divergence ($d_{adaptive}$) in the selected class (X). If we

scale $d_{adaptive}$ by d_X we obtain α, the fraction of adaptive substitutions in the selected class:

$$\alpha = \frac{d_{adaptive}}{d_X} \tag{3.6}$$

However, as noted by Gossmann et al. (2010), caution should be exercised when comparing estimates of α from different species or regions of the genome. Differences in the estimates of α could reflect differences in the contribution of slightly deleterious mutations to d_X rather than different rates of adaptive substitution. We can control for differences in the frequency of effectively neutral mutations in the selected class by computing ω_α, the ratio of $d_{adaptive}$ to d_S:

$$\omega_\alpha = \frac{d_{adaptive}}{d_S} \tag{3.7}$$

We also estimated α using a simple but frequently used method (Fay and Wu 2001; Smith and Eyre-Walker 2002):

$$\alpha_{FWW} = 1 - \frac{D_S P_X}{D_X P_S} \tag{3.8}$$

where D_X and D_S are counts of divergent sites between *M. m. castaneus* and an outgroup species for selected and neutral site classes respectively and P_X and P_S counts of polymorphic sites for selected and neutral site classes respectively.

Confidence intervals and standard error for all parameters were obtained by bootstrapping 1000 times by gene. *P* values, computed for comparisons between site classes or with zero, were obtained by two-tailed bootstrap tests.

Assumption of neutrality for synonymous sites and introns. To calculate selective constraint, the DFE, α and ω_α, we needed to use a neutrally evolving class of sites. We used two classes of sites as the neutral class: synonymous sites and introns. Current evidence from comparisons of the evolutionary rate of these site classes with ancestral repeats and pseudogenes in murids suggests that they experience overall very small selective constraints (Eory et al. 2010). Also note that we excluded potential splice sites of introns (defined as the first 6bp or last 16bp of an intron) from the analysis, since those have been documented to be under moderate selective constraint in mammals (Gaffney and Keightley 2006). Therefore we do not expect to substantially underestimate the strength of selection on non-synonymous sites and up/downstream non-coding regions by using synonymous sites or introns as the neutral standard. Moreover, since we used the same neutral classes for inferring selection on non-synonymous sites and up/downstream non-coding regions, we do not expect our estimates for the relative strength of selection between these classes of sites to be affected by a violation of the assumption of neutrality for synonymous sites or introns.

3.4 Results

3.4.1 Data and summary statistics

Description of data. Our dataset consists of sequences from 78 autosomal genes from a sample of 15 wild, unrelated *M. m castaneus* individuals sampled from NW

India. Part of the coding region of these genes and partial introns were sequenced in a previous study (Halligan et al. 2010). In this study, we focus on regions directly upstream and downstream of the coding region of these genes. We successfully amplified and sequenced ~500 bp upstream and downstream from the start and stop codon for a subset of 49 and 51 genes, respectively (Table 3.1). We have also updated the dataset of Halligan et al. (2010) by obtaining additional exonic and intronic sequence for the 78 genes and we compared our results from non-coding DNA to results from these new data. We successfully sequenced 20 alleles or more for ~90% of the sites (Table 3.1). We also sequenced the orthologous genes in a *M. famulus* individual, which we used together with the rat as an outgroup to estimate divergence, selective constraint, α and ω_α.

Table 3.1. Details of genes sequenced and percentages of sites sequenced for all 30 alleles and for at least 20 alleles.

Site class	No. genes	No. sites	Mean no. sites per gene [SD]	% sites sequenced	
				30 alleles	>20 alleles
Non-synonymous	78	34,532	443 [160]	60	95
Synonymous	78	13,056	167 [63]	60	94
Intron	65	43,672	672 [413]	45	88
Upstream	49	25,303	516 [132]	50	93
Downstream	51	26,622	522 [182]	57	91

Summary statistics. Nucleotide diversity, Tajima's *D*, divergence to *M. famulus* and rat and evolutionary constraint estimates are shown in Table 3.2. The upstream and downstream site classes show intermediate levels of nucleotide diversity (π= 0.56%

in both cases) and divergence to *M. famulus* (d= 2.49% for upstream and 2.38% for downstream) compared to non-synonymous sites (π= 0.15% and d= 0.82%), but are closer to the synonymous site estimates (π= 0.75% and d= 3.27%). Divergence to the rat is about five times higher than divergence to *M. famulus* for all site classes.

In contrast to non-synonymous sites, the upstream, downstream and intronic site classes do not show discernible differences in the shape of their SFSs compared to synonymous sites (Figure 3.1). Tajima's *D* estimates, which quantify the skew in the SFS, are significantly lower from zero for all cases examined, even for synonymous sites ($P<10^{-3}$ in all cases examined). A negative Tajima's *D* indicates an excess of rare variants, which can be caused by negative selection. However, population expansion, a prolonged population bottleneck or population subdivision can also produce a similar pattern. Different mutation rates between the compared regions could also alter the SFS. For example, CpG dinucleotides have higher mutation rates in mammals (Arndt et al. 2003) and their frequences differ between coding and non-coding DNA. However, if CpG-prone sites are excluded, we observe little changes in the SFS in all cases (Figure 3.1). We calculated evolutionary constraint C_d by comparing interspecific divergence between the focal site class and a putatively neutral site class. The estimate for C_d is moderately high for upstream and downstream site classes (24.1% and 28.4% respectively, Table 3.2) and significantly different from zero in both cases (P=0.004 for upstream; P=0.002 for downstream).

Table 3.2. Estimates of percentage diversity (%π, %θ_w) summed over all sites for *M. m. castaneus*, Tajima's *D*, percentage divergence (%*d*) to *M. famulus* and the rat and evolutionary constraint (C_d) calculated using synonymous sites as the neutral class.

Site class	**%π [SE]**	**%θ_w [SE]**	**Tajima's *D* [SE]**	**% *d* (*M. famulus*) [SE]**	**% *d* (rat) [SE]**	**% C_d (*M. famulus*) [SE]**
Non-synonymous	0.15 [0.02]	0.22 [0.02]	-0.93 [0.17]	0.82 [0.1]	3.69 [0.40]	74.5 [3.3]
Synonymous	0.75 [0.06]	0.93 [0.06]	-0.53 [0.12]	3.27 [0.21]	18.11 [0.56]	-
Intron	0.66 [0.04]	0.83 [0.04]	-0.75 [0.09]	2.90 [0.14]	15.61 [0.42]	12.1 [7.2]
Upstream	0.56 [0.05]	0.71 [0.06]	-0.78 [0.14]	2.49 [0.21]	12.2 [0.65]	24.1 [7.1]
Downstream	0.56 [0.06]	0.69 [0.06]	-0.59 [0.16]	2.38 [0.21]	11.78 [0.78]	28.4 [7.7]

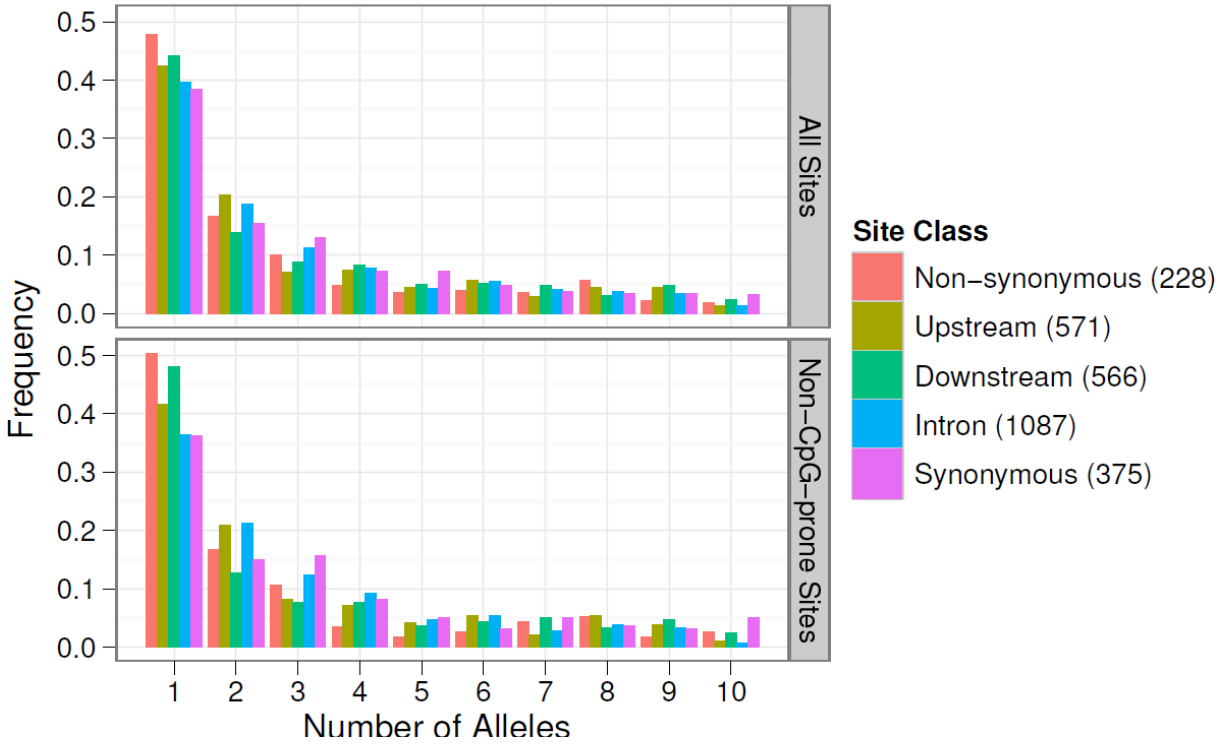


Figure 3.1. Plots of the site frequency spectra for non-synonymous, upstream, downstream, intron and synonymous site classes for all sites and for non-CpG-prone sites only. Numbers of polymorphic sites are given in parentheses.

3.4.2 The fitness effects of new mutations in genes and non-coding DNA

Inference of demographic history and the distribution of fitness effects of new mutations. We inferred the distribution of fitness effects of new mutations (DFE) along with demographic parameters using a maximum likelihood (ML) approach (Keightley and Eyre-Walker 2007). The demographic model was a step change in population size, and the selection model was a gamma distribution.

Firstly, we tested whether a model that incorporates demographic change plus selection (Model 3) fits the data significantly better than a model that assumes only demographic change (Model 1). The likelihood ratios for this comparison are highly significant in all cases examined ($-\Delta \log L$ values are reported in Table 3.3; $P< 10^{-2}$ in all cases, with d.f.=2). We also examined a model where we fitted only selection under constant population size (Model 2). We found that the fit of Model 2 to the data was significantly poorer than Model 3 in all cases examined ($-\Delta$logL values are reported in Table 3.3; $P< 10^{-9}$ in all cases, with d.f.=2). Therefore, Model 3 which incorporates both a gamma DFE and a step change in population size best explained the SFS data.

The ML estimates for the parameters of the demographic model indicated a population expansion ($N_2/N_1 \approx 3$; Table 3.4), which is consistent with the negative Tajima's *D* values for both synonymous and intronic sites (Table 3.2). The ML estimates for the mean strength of selection on new deleterious mutations ($N_e E(s)$) indicated much stronger negative selection on nonsynonymous sites than upstream or

downstream regions (Table 3.5), which is consistent with the results on selective constraint that were obtained from the analysis of interspecific divergence (Table 3.2).

Table 3.3. Likelihood-ratio tests contrasting models fitted to the SFS data when estimating the distribution of fitness effects. Models fitting only demography (M1) or only selection (M2) are contrasted with a model that fits both demography and selection (M3) to the data.

Site class		**-2*Δ*log*L***	
Neutral	**Selected**	**M1 vs. M3**	**M2 vs. M3**
Syn	Non-synonymous	312.4	47.2
	Upstream	12.4	50.5
	Downstream	13.2	43.7
Intron	Non-synonymous	412.6	120.3
	Upstream	10.6	116.8
	Downstream	12.1	126.0

M1: demography, no selection

M2: no demography, selection

M3: demography and selection

Table 3.4. Estimates of demographic parameters.

Site class		N_2/N_1 [95%CI]	t/N_2 [95%CI]
Neutral	**Selected**		
Synonymous	Non-synonymous	3.07 [2.31, 6]	0.27 [0.1, 0.91]
	Upstream	3.07 [2.31, 5.45]	0.27 [0.1, 0.77]
	Downstream	2.79 [2.1, → ∞]	0.21 [0.04, 1.32]
	Up+Down/stream	3.07 [2.31, 4.09]	0.39 [0.2, 0.95]
Intron	Non-synonymous	3.72 [2.79, → ∞]	0.68 [0.32, 1.29]
	Upstream	3.07 [2.31, → ∞]	0.51 [0.25, 1.22]
	Downstream	3.07 [2.31, → ∞]	0.49 [0.2, 1.36]
	Up+Down/stream	3.07 [2.31, 4.95]	0.23 [0.1, 0.49]

Table 3.5. Estimates of the mean strength of selection $N_e\mathrm{E}(s)$ and the shape (b) parameter of the gamma distribution.

Site class		$N_eE(s)$ [95%CI]	b [95%CI]
Neutral	**Selected**		
Synonymous	Non-synonymous	864 [75.3, 1.52X10^{10}]	0.24 [0.06, 0.45]
	Upstream	113 [0, 9.76X10^{3}]	0.05 [0.05, 0.49]
	Downstream	22.9 [0, 4.11X10^{3}]	0.08 [0.05, 96.4]
	Up+Down/stream	102 [0, 6.10X10^{3}]	0.05 [0.05, 0.36]
Intron	Non-synonymous	333 [34.4, 1.32X10^{7}]	0.27 [0.08, 0.54]
	Upstream	13.8 [0, 231]	0.05 [0.05, 0.48]
	Downstream	21.5 [0, 336]	0.05 [0.05, 0.44]
	Up+Down/stream	14.1 [0, 174]	0.05 [0.05, 0.15]

Due to the wide confidence intervals of our estimates for $N_eE(s)$ (Table 3.6) and the fact that $N_eE(s)$ can be strongly affected by extreme values of the gamma distribution, we cannot draw strong conclusions on the biological significance of our estimates for $N_eE(s)$. However, when interpolating the density of the distribution in four categories of selective effects (N_es): 0 to 1, 1 to 10, 10 to 100, 100 to $+\infty$ (Table 3.6) we obtained much more narrower confidence intervals. We also calculated evolutionary constraint C_p, a statistic that summarizes the DFE, and is the average probability of a new deleterious mutation to be lost.

For upstream and downstream sites, most new mutations fall into the effectively neutral category (N_es, 0 – 1) (69.8% and 67.5% respectively), which is in sharp constrast and significantly different ($P<0.05$ for both upstream and downstream in all comparisons) from the estimate for non-synonymous sites (15.4%). Although most new mutations in upstream and downstream sequences are effectively neutral, there is a substantial fraction (21.7% and 19.6% respectively, not significantly different from zero) of strongly selected mutations ($N_es>10$) in these regions. $\%C_p$ is moderately high for upstream (30.7%) and downstream (33.9%) site classes and is significantly different from zero only for the downstream site class ($P<0.05$).

Table 3.6. Estimates of percentages of mutations in four $N_e s$ ranges and evolutionary constraint estimated from polymorphism (C_p).

Site class		Percentage of mutations in $N_e s$ range [95%CI]				$\%C_p$ [95%CI]
Neutral	**Selected**	**0 - 1**	**1 - 10**	**10 - 100**	**100 - ∞**	
Syn	Non-synonymous	15.4 [9, 23.3]	11.4 [3.1, 18.6]	19.5 [3.6, 42.1]	53.7 [27.6, 71.8]	84.7 [75.9, 90.8]
	Upstream	69.8 [48.1, 100]	8.5 [0, 39.0]	9.4 [0, 20.3]	12.3 [0, 29.4]	30.7 [0, 53.3]
	Downstream	67.5 [29.1, 100]	13 [0, 69.2]	13.6 [0, 23.6]	6.0 [0, 28.0]	33.9 [7.9, 64]
	Up+Downstream	70.7 [52.6, 92.2]	8.6 [5.9, 31.8]	9.5 [0, 17.7]	11.2 [0, 27.4]	30.0 [10.4, 48.9]
Intron	Non-synonymous	15 [7.8, 24.3]	13.1 [4.4, 20.4]	24.1 [5.3, 47.5]	47.8 [21.8, 67.9]	83.0 [73.7, 89.6]
	Upstream	77.6 [56.6, 100]	9.3 [0, 35.2]	9.2 [0, 15.7]	3.8 [0, 15.3]	22.2 [0, 41.3]
	Downstream	75.9 [50.8, 100]	9.2 [0, 30.2]	9.5 [0, 19.7]	5.4 [0, 19.9]	33.9 [7.9, 64.0]
	Up+Downstream	77.1 [63.9, 100]	9.3 [0, 21.3]	9.3 [0, 14.5]	4.3 [0, 15.8]	23.1 [0, 35.9]

3.3.3 Adaptive evolution in genes and non-coding DNA

We then estimated the fraction of substitutions driven to fixation by positive selection (α) using an extension of the MK test (Eyre-Walker and Keightley 2009). This method uses neutral divergence between *M. m. castaneus* and an outgroup (either *M. famulus* or rat) along with the distribution of fitness effects, inferred from polymorphism data of *M. m. castaneus*, to estimate the expected divergence between *M. m. castaneus* and the outgroup. The difference between the observed and the

expected divergence estimates the adaptive divergence between *M. m. castaneus* and the outgroup. α is then calculated by scaling the adaptive divergence by the observed divergence (see detailed description in the methods).

Estimates of α for the non-synonymous, upstream and downstream site classes for *M. m. castaneus*, are presented in Table 3.7. We report moderately low estimates of α for upstream and downstream site classes (11.8% and 9.3% respectively), that are not significantly different from zero when using *M. famulus* as the outgroup, and synonymous sites as the neutral reference. By combining upstream and downstream sequences, we obtain a similar point estimate for α (9.3%) and a narrower confidence interval (-15.4, 36.3), which includes zero but excludes high estimates of α, and is not significantly different from the estimate for non-synonymous sites (P=0.078). The point estimates are very similar when using the rat as the outgroup or intronic sites are used as the neutral reference, but the confidence intervals are narrower for the latter case, since more data are included. The estimate for α for combined upstream and downstream sequences is significantly different from non-synonymous sites when using intronic sites as the neutral reference (P=0.014 when using *M. famulus* as the outgroup and P=0.018 when using the rat as the outgroup).

Table 3.7. The fraction of substitutions driven to fixation by positive selection (α) and the ratio of adaptive to neutral divergence (ω_α) estimated using *M. famulus* and the rat as outgroups.

Site class		% α [95%CI] outgroup		% ω_α [95%CI] outgroup	
Neutral	**Selected**	***M. famulus***	**Rat**	***M. famulus***	**Rat**
Synonymous	Non-synonymous	46.6 [18, 67.6]	43.5 [13.9, 64.2]	11.9 [4.4, 18.3]	9.1 [2.9, 13.7]
	Upstream	11.8 [-19.5, 43.8]	9.3 [-7.2, 53.9]	9.0 [-13.7, 35.5]	6.6 [-4.8, 37.7]
	Downstream	9.3 [-28, 64.3]	10.5 [-19.8, 54.7]	6.6 [-19.5, 48.4]	7.0 [-13, 36.2]
	Up+Downstream	9.3 [-15.4, 36.3]	6.7 [-10.7, 35.5]	5.2 [-19.7, 30.6]	4.6 [-7.1, 23.6]
Intron	Non-synonymous	54.0 [25.5, 74.7]	45.4 [13.7, 67.4]	15.3 [6.8, 22.6]	10.6 [3.5, 16.3]
	Upstream	12.6 [-8.3, 42.1]	9.3 [-8.3, 40.1]	10.8 [-6.7, 37.1]	7.3 [-6.2, 31.9]
	Downstream	9.6 [-16.5, 38.8]	5.3 [-20.7, 25]	7.8 [-13.3, 32.1]	4.1 [-17.1, 18.8]
	Up+Downstream	12.5 [-2.3, 27.5]	4.6 [-10.9, 17.6]	5.1 [-13.1, 22.4]	3.5 [-8.3, 13.8]

The different estimates of α between non-synonymous, upstream and downstream site classes might be due to differences in the rate of slightly deleterious, rather than adaptive substitutions (Gossmann et al. 2010). In order to take account of any differences in the slightly deleterious substitution rate between the selected site classes, we computed the ratio of adaptive divergence to neutral divergence (ω_α). The resulting ω_α estimates are very similar, and not significantly different, for the non-synonymous, upstream and downstream site classes (11.9%, 9%, 6.6%

respectively, Table 3.7). These results indicate that the lower estimates of α in the upstream and downstream site classes compared to non-synonymous sites, could be due to a higher porportion of the upstream and downstream sites evolving nearly neutrally, rather than a lower rate of adaptive substitution. The confidence intervals for ω_α are very wide when examining upstream and downstream site classes individually, but narrow down when we combine data from upstream and downstream site classes (the upper boundary for ω_α is never higher than 30.6%). Similarly with estimates for α, when using intronic sites as the neutral reference and the rat as the outgroup, we get narrower confidence intervals, since more data are included.

Finally, we used a simple, frequently used approach, to estimate α (Fay et al. 2001; Smith and Eyre-Walker 2002) in order to be able to make comparisons with studies that have not employed the Eyre-Walker and Keightley (2009) methodology. We controlled for slightly deleterious mutations by excluding low frequency polymorphisms (<10%) as suggested by (Fay et al. 2001). By using this method, we obtained zero or negative estimates of α for upstream and downstream site classes (Table 3.8), which roughly agree with the estimates we obtained using the Eyre-Walker and Keightley (2009) methodology.

Table 3.8. Percent estimates of the fraction of substitutions driven to fixation by positive selection using a simple extension of the McDonald-Kreitman test (Fay et al. 2001; Smith and Eyre-Walker 2002). We use *M. famulus* as outgroup.

Site class		% α_{FWW} [SE]	% α_{FWW}>10% [SE]
Neutral	**Selected**		
Synonymous	Non-synonymous	11.6 [10.8]	25.6 [13.9]
	Upstream	-4.1 [12.8]	1.3 [17]
	Downstream	-7.1 [15.1]	-6.4 [20.8]
	Up+Down/stream	-2.4 [11.9]	-1.4 [14.9]
Intron	Non-synonymous	9.5 [10.1]	31.5 [11.7]
	Upstream	2.3 [10.1]	9.8 [11.3]
	Downstream	0.8 [8.9]	-6.8 [14.5]
	Up+Down/stream	2.9 [6.9]	2.2 [7.8]

3.5 Discussion

In this study, we presented results suggesting that sites upstream and downstream of protein- coding regions in *M. m. castaneus* are, on average, under weak positive and negative selection. Several lines of evidence support this conclusion. Nucleotide diversity values in *M. m. castaneus* and divergence to *M. famulus* or rat in upstream and downstream regions are much higher than for non-synonymous sites, and slightly but significantly lower than synonymous sites. Evolutionary constraint is also significantly lower in upstream and downstream regions than for non-synonymous sites. Tajima's *D* estimates are not significantly different between site classes, except for the synonymous and non-synonymous sites comparison, which suggests either that all site classes investigated are under negative selection or

that a population expansion or bottleneck has occurred in the past in *M. m. castaneus*. Indeed, if we fit a simple demographic model of a step change in population size, we find evidence for population expansion of *M. m. castaneus*, which might explain the negative Tajima's *D* at synonymous sites. A population expansion might also explain the negative Tajima's *D* in upstream and downstream regions. However, a model of a demographic change plus negative selection fits the data significantly better than a model of demographic change with no selection or a model with selection only, in all cases examined. Therefore, we obtained statistically significant evidence for both a population expansion in *M. m. castaneus* in the past and negative selection acting on upstream and downstream regions. The DFE inferred for upstream and downstream regions implies that most new mutations have $N_e s$ values in the range of 0 to 1, but a small fraction are strongly deleterious. At non-synonymous sites, the pattern is reversed, and we infer that most new mutations are strongly deleterious. This result further supports the conclusion that upstream and downstream regions are, on average, under weak selective constraint compared to non-synonymous sites.

Our low point estimates of α for upstream and downstream regionssuggest that weak positive selection operates in these regions, compared with that acting on non-synonymous sites. The estimates for α and ω_α are not significanly different from zero. However, the 95% confidence intervals for combined upstream and downstream regions exclude α values that are higher than ~36%.

In sequencing diploid outbred individuals, regions lying in between heterozygous indels can be problematical for SNP calling. Our dataset contains less than 10% of such regions. If such regions are excluded, the estimates for constraint,

the parameters of the DFE and α are unchanged (result not shown). Another consideration about our dataset is that because the Environmental Genome Project (EGP) sample is not a random sample of genes (Livingston et al. 2004), we might have excluded genes that have high rates of adaptive evolution in upstream and downstream regions. For example, promoter regions of many neural and nutrition-related genes in humans have been found to be subject to positive selection (Haygood et al. 2007). However, a comparison of estimates of α for regions ~500 bp upstream and downstream of the start and stop codon of protein-coding genes in humans, obtained with the methodology employed in the present study and using the EGP and PGA (Akey et al. 2004) datasets, showed no significant differences between datasets (Eyre-Walker and Keightley 2009). Additionally, a comparison of estimates of α, obtained with the methodology employed in the present study and using the EGP, PGA and Boyko et al. (2008) datasets, has shown no significant differences for non-synonymous sites between datasets in humans (Halligan et al. 2010).

It has been suggested that regulatory non-coding regions might be more important for evolution than protein-coding genes in primates (King and Wilson 1975). However, studies that have used a simple extension of the McDonald-Kreitman test (Keightley et al. 2005) and the methodology employed in this study (Eyre-Walker and Keightley 2009) have estimated that α in upstream and downstream regions in humans is close to zero. Humans might have low rates of adaptive substitution in upstream and downstream regions because of their historically low N_e. However, in the current study we also obtain low estimates for α (~10%; not significantly different from zero) for upstream and downstream regions

in *M. m. castaneus*, a mammalian species with a N_e much larger than humans (Halligan et al. 2010). The low estimate of α in upstream and downstream regions in *M. m. castaneus* may be due to the sparse distribution of regulatory elements in the mammalian genome. Therefore, the upstream and downstream sequences we have focused on, could include a substantial amount of neutral sequence along with some functionally relevant elements. In order to control for differences between site classes in the contribution of slightly deleterious mutations to the observed divergence, we calculated the ratio of adaptive to neutral divergence (ω_α), and we obtained similar estimates for non-synonymous, upstream and downstream site classes (~5-10%; not significantly different from zero). Therefore, upstream and downstream regions of protein-coding genes in *M. m. castaneus* appear to have a similar absolute rate of adaptive substitution with non-synonymous sites. This finding implies that the difference in the estimate of α observed at non-synonymous sites between humans (~0-20%) and *M. m. castaneus* (~50%) might also be due to differences in the relative proportion of slightly deleterious mutations between the two species. More specifically, non-synonymous sites in humans might experience more nearly neutral substitutions than *M. m. castaneus* but have a similar rate of adaptive substitution as *M. m. castaneus*.

Finally, if non-coding regulatory elements are distributed over thousands of base pairs in the mammalian genome, then the net input of adaptive substitutions to regulatory regions of mammals could be higher than protein-coding genes. Eyre-Walker and Keightley's (2009) study in humans and our study in *M. m. castaneus* only examined ~500 bp upstream and downstream of the start and stop

codon, respectively, of a limited collection of protein-coding genes. We suggest that genome-wide studies of putative regulatory non-coding regions are needed in *M. m. castaneus* and humans, so that the role of regulatory regions to adaptation can be more confidently ascertained.

Chapter 4. Selection on autosomal and X-linked genes in house mice

The work in this chapter has been prepared as the following research paper (submitted in *Genetics*).

Kousathanas A, Halligan DL, Keightley PD. Faster-X adaptive protein evolution in house mice.

AK compiled and analysed the data and wrote the paper. DLH performed the SNP and genotype calling and wrote the text describing the details provided in Appendix A.4.1. DLH and PDK provided comments on previous versions of the manuscript.

4.1 Summary

The causes of the large effect of the X chromosome in reproductive isolation and speciation have long been debated. Charlesworth et al. (1987) demonstrated that X-linked loci are expected to have higher rates of adaptive evolution than autosomal loci if new mutations are on average recessive. Reproductive isolation should therefore evolve faster when contributing loci are located on the X chromosome (the faster-X hypothesis). In this study, we analysed genome-wide polymorphism data from the house mouse subspecies *Mus musculus castaneus* and divergence from *Mus famulus* and *Rattus norvegicus* to infer rates of adaptive evolution for autosomal and

X-linked protein-coding genes. We find significantly faster adaptive evolution for X-linked genes, particularly for those with male-specific expression, while autosomal and X-linked genes with female-specific expression evolve at similar rates. We also estimated rates of adaptive evolution for genes expressed during spermatogenesis, and found that X-linked genes that escape meiotic sex chromosome inactivation (MSCI) show rapid adaptive evolution. Our results suggest that faster-X adaptive evolution is either due to average recessivity of new advantageous mutations or to a special gene content of the X chromosome regulating male function and spermatogenesis. We discuss how our results can help to explain the large effect of the X chromosome in speciation.

4.2 Introduction

The X chromosome has a special role in speciation, harbouring a disproportionate number of loci contributing to reproductive isolation. This phenomenon, also known as the "large-X" effect (or large-Z for species where the female is the heterogametic sex), has been documented in several species of *Drosophila*, Lepidoptera, birds and mammals (Coyne and Orr 1989; Coyne 1992; Coyne and Orr 2004). Its causes are disputed, and several hypotheses have been proposed to explain it (Rice 1984; Charlesworth et al. 1987; Presgraves 2008: 200). One hypothesis rests on the fact that the X chromosome is found only in one copy in males, and therefore recessive mutations on the X are fully exposed to selection. If new advantageous mutations are partially recessive, X-linked loci are expected to have higher rates of adaptive evolution than autosomal loci, ("faster-X" hypothesis Charlesworth et al. 1987). If

true, faster-X evolution could partially or fully explain the large-X effect (Presgraves 2008).

The faster-X hypothesis has been highly influential, since it generated predictions that could be tested using genomic data. It also presented the intriguing possibility of estimating the dominance coefficient (h) of new advantageous mutations. Assuming an equal number of breeding females and males, that the distribution of fitness effects of new advantageous mutations does not differ between autosomes and the X, and that most adaptive substitutions are from new mutations rather than standing variation, then the ratio of the rates of adaptive evolution of X-linked loci over autosomal loci (R) is a function of h and the selective effects of new mutations in females (s_f) and males (s_m):

$$R \approx \frac{2h s_f + s_m}{2h(s_f + s_m)} \tag{4.1}$$

When $s_f = s_m$ this reduces to a simple function of h:

$$R \approx \frac{2h + 1}{4h} \tag{4.2}$$

(Charlesworth et al. 1987; Vicoso and Charlesworth 2006).

Several researchers set out to test the faster-X hypothesis, initially by comparing the rates of protein evolution (estimated using the ratio of divergence at non-synonymous sites to synonymous sites; d_N/d_S) between X-linked and autosomal genes (Betancourt et al. 2002; Counterman et al. 2004; Lu and Wu 2005; Musters et al. 2006; Mank et al. 2007; Mank et al. 2010). However, a higher d_N/d_S ratio for X-linked *versus* autosomal loci could be caused by reduced efficacy of negative selection on the X, due to its smaller effective population size (N_e) than the

autosomes. A more powerful way of testing for positive selection is using the McDonald and Kreitman test (McDonald and Kreitman 1991) and its derivatives, which contrast patterns of polymorphism and divergence at selected and neutral classes of sites, and can be used to estimate the rate of adaptive substitution (α). For *Drosophila*, some studies have found evidence for faster-X adaptive evolution (Begun et al. 2007; Baines et al. 2008a; Mackay et al. 2012), whereas others have not (Thornton et al. 2006; Connallon 2007). There have been studies comparing α between autosomal and X-linked genes in species other than *Drosophila*. A study that compared α for autosomes and X of two subspecies of the European rabbit found faster-X evolution for only one of the two species (Carneiro et al. 2012). Another recent study found strong evidence for faster-X adaptive evolution in Central chimpanzees (Hvilsom et al. 2012).

Apart from a faster overall rate of adaptive evolution of the X chromosome, additional predictions of the faster-X theory can be tested using genomic data. For example, equation 4.1 can be simplified to show that for mutations with $s_m>0$ and $s_f=0$ (e.g. mutations in male-specific genes) R becomes an inverse function of h:

$$R \approx \frac{1}{2h} \tag{4.3}$$

For mutations with $s_m=0$ and $s_f>0$ (e.g. mutations in female-specific genes), $R \approx 1$ (Charlesworth et al. 1987; Vicoso and Charlesworth 2006). Therefore, a more pronounced faster-X effect is expected for recessive mutations that are selected only in males, whereas no faster-X effect is expected for recessive mutations that are selected only in females (Charlesworth et al. 1987; Vicoso and Charlesworth 2006).

Mutations in genes with male-biased and female-biased expression are likely to have mostly effects on males and females, respectively. Therefore, a way to test for the prediction of the faster-X hypothesis regarding mutations with sex-specific effects is to compare the evolutionary rate of autosomal and X-linked genes with sex-biased or sex-specific expression. By following this rationale, Baines at al. (2008) found a stronger faster-X effect for genes with male-biased expression than unbiased or female-biased genes in *Drosophila* confirming the predictions of the faster-X hypothesis (Baines et al. 2008a).

Exposure of recessive mutations in males is not the only process that can create conditions for faster- or slower-X evolution. A different gene content of the X chromosome and the autosomes could underlie differences in their evolutionary rate. For example, the X chromosome might be enriched for classes of genes that evolve rapidly, such as genes that are narrowly expressed (i.e. expressed in a limited number of tissues;(Meisel et al. 2012a; Meisel et al. 2012b). Moreover, genes on the X chromosome experience global inactivation during spermatogenesis (a process known as meiotic sex chromosome inactivation or MSCI; Lifschytz and Lindsley 1972). Evidence for MSCI has been documented in *Drosophila*, birds and mammals (Hense et al. 2007; Turner 2007; Schoenmakers et al. 2009), and it has been suggested that MSCI could be a universal feature of species with heteromorphic chromosomes (Namekawa and Lee 2009). However, it is unknown whether MSCI can affect the evolutionary rate of X-linked genes.

House mice are one of the best studied species for the dynamics of speciation, and a large volume of evidence has been accumulated showing a large effect of the X

chromosome on hybrid incompatibilities
(Tucker et al. 1992: 92; Oka et al. 2004; Payseur et al. 2004; Storchová et al. 2004; Oka et al. 2007; Good et al. 2008; Teeter et al. 2008). Population genetic studies have shown reduced gene flow on X-linked than autosomal genes in the hybrid zone between *Mus musculus domesticus* and *Mus musculus musculus* in Europe (Tucker et al. 1992; Payseur et al. 2004; Teeter et al. 2008). Additionally, laboratory crosses between different *Mus musculus* strains have revealed that the X chromosome harbors a disproportionate number of genes that are associated with hybrid male sterility
(Oka et al. 2004; Storchová et al. 2004; Oka et al. 2007; Good et al. 2008).

In this study, we analyse genome-wide polymorphism data from *Mus musculus castaneus*, a subspecies of the *Mus musculus* species complex. Previous studies have estimated that up to 50% of nonsynonymous substitutions in protein-coding genes have been driven to fixation by positive selection in *M. m. castaneus* (Halligan et al. 2010; Kousathanas et al. 2011; Phifer-Rixey et al. 2012). However, these studies examined only small numbers of autosomal loci. Here, we contrast within-species polymorphism and between-species divergence for ~19,000 protein-coding genes and quantify the relative rates of adaptive protein evolution between the autosomes and the X chromosome. To test faster-X theory predictions, we investigate the evolution of genes that have biased expression in sex-specific tissues. We also examine the evolution of genes expressed at various stages of spermatogenesis.

4.3 Materials and Methods

Sampling of mice. We generated genomic sequence for 10 *Mus musculus castaneus* individuals collected in NW India (Baines and Harr 2007); 7 females and 3 males. The sampling strategy is detailed in a previous study (Halligan et al. 2010) and was aimed at sampling non-related individuals from a single population. Tests for population structure and admixture (using the program *STRUCTURE*; Pritchard et al. 2000) showed no evidence for hidden population substructure or admixture between differentiated subspecies in our population sample (Halligan et al. 2010). We also sequenced the genome of an individual *Mus famulus* obtained from the Montpellier wild mice genetic repository to use as an outgroup.

Genome sequencing and Illumina read mapping.

Illumina paired-end sequencing libraries were generated for each individual with fragment sizes 300-550 bp. Mapped sequence coverage was 21-42x (average 29x) per sampled animal. The libraries were run at a mixture of 76, 100 and 108bp read lengths on the Illumina GAIIx and HiSeq platforms. The program *SMALT* (http://www.sanger.ac.uk/resources/software/smalt/) using the parameters: -k 13 -s 6 was used to align the *M. m. castaneus* Illumina sequencing reads to the NCBIM37/mm9 unmasked reference genome. We also generated genomic sequence for *M. famulus* to be used as an outgroup. Since *M. famulus* sequence is diverged from the reference (NCBIM37/mm9), we used an iterative mapping procedure to improve alignment to the reference (more details are given in Appendix A.4.1).

SNP calling. We used the *SAMtools* package to call genotypes at each site (Li et al. 2009). This involves creating genotype likelihood files using *mpileup* and obtaining SNP calls for every site in the genome using an iterative Bayesian approach with *bcftools*. More details on the procedure to call SNPs are given in Appendix A.4.1. We excluded genotype calls that had no mapped reads or where there was significant evidence for departure from Hardy-Weinberg proportions (a cut-off of <0.0002 was used on the *P*-value of a X^2 based test obtained using *SAMtools*). For the X chromosome, SNP calls were made using females only, because *SAMtools* assumes diploids. Therefore, we had an allelic coverage of 20 for the autosomes, and 14 for the X chromosome.

Obtaining the sequences for protein-coding genes. We obtained gene coordinates from the Ensembl database version 62 (http://apr2011.archive.ensembl.org/index.html) for a total of 18,110 autosomal and 700 X-linked protein-coding genes with orthologues in both mouse and rat. For each gene, we obtained the coordinates for the canonical spliceform as annotated in the Ensembl database. We used these to obtain gene sequences for rat and to construct sequences for *M. m. castaneus* and *M. famulus* individuals based on their genotype calls. We then created separate alignments for each gene using MAFFT (Katoh et al. 2002) based on the translated amino-acid sequences and back-translated them to the DNA sequence to preserve the coding frame. We considered only 0-fold and 4-fold degenerate sites as nonsynonymous and synonymous, respectively.

The site-frequency spectrum (SFS) and summary statistics. We obtained the frequencies of the segregating alleles for each polymorphic site in our population sample by assuming that all sites are biallelic and excluding sites where more than two alleles were found in the population. We obtained the folded SFS by summing the sites over all possible minor allele frequencies. To summarise diversity, we calculated the average per site heterozygosity π (Tajima 1983). We quantified the relative skew of the SFS compared to what is expected at Wright-Fisher equilibrium and an infinite sites mutation model by calculating Tajima's *D* (Tajima 1989). Note that we bootstrapped by gene with replacement 1000 times to perform statistical comparisons of *D* between different classes of sites or with zero. We used *M. famulus* and the rat as outgroups to calculate between species nucleotide divergence. For the polymorphic sites in *M. m. castaneus*, we calculated the average divergence between the *M. m. castaneus* alleles at a site with the outgroup base, accounting for their frequencies. We applied a Jukes-Cantor multiple hits correction to the divergence estimates (Jukes and Cantor 1969). CpG dinucleotides have higher mutation rates in mammals, and their frequency is higher close to and within genes than non-coding DNA that is far away from genes (Arndt et al. 2003). For analyses, we excluded sites that were preceded by C or followed by a G, as suggested by a previous study (Gaffney and Keightley 2008), unless specifically noted.

Assumption of neutral evolution for synonymous sites. We used synonymous sites of protein-coding genes as the presumed neutral class for our analyses. As discussed

in the methods section of Chapter 3, current evidence suggests very small selective constraints in synonymous sites of murids (Eory et al. 2010) , therefore we do not expect substantial underestimation of the strength of selection in nonsynonymous sites of autosomal and X-linked genes. However, if the selection pressure is different on synonymous sites of autosomal and X-linked genes, it is possible that we will obtain artificial evidence for faster- or slower- X evolution as has been suggested previously for *D. melanogaster* (Campos et al. 2012). However, a previous study that examined patterns of codon-usage bias in autosomal and X-linked genes of rodents, found no evidence that codon-usage bias is due to selection for either autosomal or X-linked genes (Smith and Hurst 1999). Therefore we do not expect to misinfer the relative strength of selection on nonsynonymous sites of autosomal and X-linked genes due to a different strength of selection on synonymous sites of autosomal and X-linked genes.

Estimating the distribution of fitness effects of new deleterious mutations (DFE). To infer the DFE, we used a maximum likelihood (ML) method (*DFE-alpha*) that fits a selection and a demographic model to the SFSs of assumed selected and neutral classes of sites, respectively (Keightley and Eyre-Walker 2007). We used synonymous sites of protein-coding genes as the neutral class for our analyses and nonsynonymous sites as the selected class of sites.

Using *DFE-alpha* we firstly fitted a demographic model of a step change in population size in the past to the neutral SFS. It has previously been shown that bottlenecks or population subdivision do not greatly affect the accuracy of inference

of selection by *DFE-alpha* if a 2-epoch model is fitted to the neutral class (Keightley and Eyre-Walker 2010; Kousathanas and Keightley 2013). Nevertheless, we also fitted a three-epoch demographic model to the synonymous data to investigate whether our results are robust to a more complex demographic history and to investigate the possibility of a bottleneck in the studied population.

Using *DFE-alpha* we then fitted a gamma distribution to the selected SFS to infer the DFE of deleterious mutations. We assumed that new mutations in the selected class are unconditionally deleterious. In natural populations some fraction of new mutations might be advantageous, however it has previously been shown that these will not affect the estimates of the parameters of the DFE for deleterious mutations (Keightley and Eyre-Walker 2010). We also fitted multi-spike distributions to the nonsynonymous data to investigate whether our results are robust to a multimodal DFE. Moreover, *DFE-alpha* assumes that sites are unlinked, which could affect its inferences. However, it has previously been shown that the effect of linkage can be taken into account by fitting a 2-epoch demographic model to a neutral reference that is interdigitated with the selected sites (Kousathanas and Keightley 2013; Messer and Petrov 2013).

Measuring the rate of molecular adaptation. To infer the rate of adaptive divergence between two species, we use an extension of the McDonald-Kreitman (MK) test (McDonald and Kreitman 1991). The standard MK test compares the ratio of nonsynonymous to synonymous divergence (d_N/d_S) between two species with the ratio of nonsynonymous to synonymous polymorphism (p_N/p_S) within a species.

Because positively selected mutations are not expected to contribute substantially to polymorphism, an excess of d_N/d_S relative to p_N/p_S is interpreted to be the result of adaptive substitutions. The rate of molecular adaptation is usually quantified by calculating the proportion of substitutions that have been fixed by positive selection (α) as follows:

$$\alpha = \frac{d_N - d_S \frac{p_N}{p_S}}{d_N} \tag{4.4}$$

(Fay et al. 2001; Smith and Eyre-Walker 2002), where d_N is the observed nonsynonymous divergence between two species and $d_S(p_N/p_S)$ is the expected divergence from neutral and slightly deleterious mutations.

When comparing estimates of α between different classes of genes or between different species, differences in α can be due to a difference in the contribution of slightly deleterious mutations to d_N rather than a different rate of adaptive substitution. This can be controlled for by calculating the rate of adaptive relative to neutral substitution (ω_a):

$$\omega_a = \frac{d_N - d_S \frac{p_N}{p_S}}{d_S} \tag{4.5}$$

(Gossmann et al. 2010)(Gossmann et al. 2010)(Gossmann et al. 2010).

Given the inferred DFE from the polymorphism data, we can calculate the average fixation probability of new deleterious and neutral mutations relative to the fixation probability of neutral mutations ($\bar{u}$) by integrating over the DFE. We modified equations 4.4 and 4.5 and calculated α and ω_a as follows:

$$\alpha = \frac{d_N - d_S \overline{u}}{d_N} \tag{4.6}$$

$$\omega_a = \frac{d_N - d_S \overline{u}}{d_S} \tag{4.7}$$

(Eyre-Walker and Keightley 2009). There are two advantages of using fixation probabilities instead of polymorphic counts to infer α and ω_α. Firstly, slightly deleterious mutations that can contribute disproportionately to polymorphism in the selected class, leading to underestimation of α and ω_α, are explicitly modelled. Secondly, by using the framework to estimate the DFE as detailed above, the recent demographic history of the population can be taken into account. Demographic changes can produce a signal in the polymorphism data that can bias estimates of α and ω_α (Eyre-Walker 2002).

Keightley and Eyre-Walker (2012) showed that the estimates of α and ω_α can be biased if the divergence between the species compared is low relative to within species polymorphism (Keightley and Eyre-Walker 2012). We corrected the divergence estimates for the contribution of polymorphism by using their suggested approach (Keightley and Eyre-Walker 2012). Unless otherwise stated, our estimates of the d_N/d_S ratio, α and ω_α, are all corrected using that method.

We also used non-parametric estimators to calculate α, because they can be potentially more powerful when analysing small numbers of loci. We used the program *MKtest (Welch 2006)* to calculate α_{FWW} and α_{SEW} developed by (Fay et al. (2001) and (Smith and Eyre-Walker (2002), respectively. The first estimator (α_{FWW}) is calculated by summing the counts of divergent and polymorphic nonynonynymous

and synonymous sites (D_N, P_N, D_S, P_S, respectively) across genes and using the following equation:

$$\alpha_{FWW} = 1 - \frac{D_S P_N}{D_N P_S} \tag{4.8}$$

The α_{FWW} estimator has been shown to be biased if there is a correlation between selective constraint and diversity. The α_{SEW} estimator has been introduced by Smith and Eyre-Walker (2002) to control for this bias by averaging D_N, P_N, D_S, P_S across genes and using the following equation:

$$\alpha_{SEW} = 1 - \frac{\overline{D_S}}{\overline{D_N}} \overline{\left(\frac{P_N}{P_S + 1} \right)} \tag{4.9}$$

Estimates for α_{FWW} and α_{SEW} were not corrected for the contribution of polymorphism to divergence.

Statistical testing. Confidence intervals for parameter estimates were obtained by bootstrapping by gene 200 times, unless otherwise stated. To compare different classes of genes, we performed a non-parametric bootstrap test and unless otherwise stated, the 2-tailed *P* value is reported. Mann-Whitney *U* tests were performed using *R* (http://www.r-project.org/).

Analysis of gene expression. In order to define functional categories of genes, we analysed several gene expression datasets from microarray experiments. We used the

GNF gene expression atlas (Su et al. 2004) to define male- and female-specific genes. This dataset contains measurements of gene expression for several thousand mouse and human genes in a large number of tissues (61 in mice). For this dataset, we defined a gene as expressed in a tissue when its expression value was higher than the median (=140.5) for the whole microarray experiment following the authors' suggestions (Su et al. 2004). We defined a gene as specifically expressed in a tissue when the expression of that gene in the focal tissue was 2-fold higher than the median expression of the gene over all tissues, excluding the focal tissue. Male-specific genes were defined as those that specifically expressed in testis or prostate, whereas female-specific ones were defined as those with expression specifically in ovary or uterus. To calculate the expression breadth (τ) of each gene in our dataset, we used the following formula:

$$\tau = \frac{\sum_{i=1}^{N} \frac{1 - \log T_i}{\log T_{max}}}{N - 1} \tag{4.10}$$

where N is the number of tissues examined, T_i is the expression value in each tissue and T_{max} is the maximum expression over all tissues (Liao et al. 2006).

To define genes expressed at different stages of spermatogenesis, we used the dataset of (Namekawa et al. 2006). This contains gene expression measured in four types of germ cells, corresponding to different stages of spermatogenesis. These are A and B spermatogonia, pachytene spermatocytes and round spermatids. A and B spermatogonia correspond to the early pre-meiotic stage of spermatogenesis (stage 1), pachytene spermatocytes respresent the stage where meiotic sex inactivation of the X chromosome (MSCI) occurs (stage 2), and round spermatids are mature

postmeiotic cells (stage 3). The dataset contains gene expression levels for each cell type computed from microarray signal intensities. Expression values had been scaled to a trimmed mean signal intensity of 125 for each microarray chip. There were two replicates per cell type and we averaged the expression levels of the replicates. Genes that had a signal intensity < 100 at all stages of spermatogenesis were considered as not expressed during spermatogenesis (following suggestions of Namekawa et al. 2006). A gene was considered as expressed during a stage if its expression value was higher than 125. We defined three groups of genes based on their expression during stage 1 and stage 3: group A for genes that are expressed in stage 1, and not in stage 3, group B for genes that are expressed during both stages 1 and 3, and group C for genes that are non-expressed in stage 1 and expressed in stage 3. These groups of genes correspond roughly to the groups defined by Namekawa et al. 2006.

4.4 Results

4.4.1 Diversity and divergence for autosomal and X-linked loci.

We analysed polymorphism within *M. m. castaneus* and divergence from *M. famulus* and the rat for nonsynonymous and synonymous sites in a total of 18,110 autosomal and 700 X-linked protein-coding loci. We examined results for all sites and non-CpG-prone sites separately.

The pairwise nucleotide diversity at synonymous sites (π_S) is substantially lower for X-linked than for autosomal loci (P<0.01; Table 4.1). X-linked

synonymous site divergence (d_S) from *M. famulus* and rat is also significantly lower than that of the autosomes ($P<0.01$ for all comparisons; Table 4.1). If we assume that synonymous sites evolve neutrally, then π_S is proportional to the product of the effective population size (N_e) and the mutation rate (μ) (Charlesworth 2009). Therefore, the lower π_S for X-linked loci relative to the autosomes could be either attributed to a lower N_e or μ. After controlling for a difference in μ between the chromosomes by dividing π_S with d_S from *M. famulus*, we obtained a diversity ratio (X/A) equal to 0.58, which should reflect the ratio of effective population sizes between the X and the autosomes (i.e. N_{eX}/N_{eA}). The X chromosome is found in two copies in females and one copy in males, and as a result, this ratio is expected to be 0.75 under neutrality and if the male and female sex ratio and reproductive success are equal. The observed N_{eX}/N_{eA} is signicantly lower than this expectation ($P<0.01$ for all comparisons including just that for non-CpG-prone sites and using *M. famulus* or rat to calculate divergence). The observed reduction in X-linked diversity could be explained by a bottleneck (Wall et al. 2002; Pool and Nielsen 2007), by unequal variance in reproductive success between males and females (Charlesworth 2001) or by a larger effect of selective sweeps or background selection in eliminating X-linked synonymous diversity (Hutter et al. 2007).

Table 4.1. Number of sites and summary statistics for nonsynonymous and synonymous sites of autosomal and X-linked loci. Statistics are given for all sites and non-CpG-prone sites. 95% confidence intervals are given in brackets.

Cpg-prone status	Site class	Chr.	No. sites (Mbp)	% π	% d (*M. famulus*)	% d (rat)	Tajima's D
All sites	Nonsyn	A	17.7	0.136 [0.133, 0.139]	0.590 [0.579, 0.601]	3.64 [3.58, 3.70]	-0.906 [-0.928, -0.884]
		X	0.634	0.057 [0.05, 0.063]	0.594 [0.537, 0.658]	3.91 [3.53, 4.39]	-0.829 [-0.954, -0.704]
	Syn	A	4.33	0.849 [0.839, 0.857]	3.53 [3.51, 3.56]	18.2 [18.1, 18.3]	-0.609 [-0.623, -0.595]
		X	0.147	0.372 [0.348, 0.397]	2.70 [2.60, 2.81]	15.0 [14.3, 15.7]	-0.636 [-0.731, -0.542]
Non-CpG-prone	Nonsyn	A	9.59	0.125 [0.122, 0.128]	0.575 [0.565, 0.586]	3.45 [3.39, 3.51]	-0.817 [-0.843, -0.790]
		X	0.370	0.0536 [0.0470, 0.0607]	0.592 [0.533, 0.656]	3.72 [3.36, 4.16]	-0.833 [-0.983, -0.674]
	Syn	A	1.49	0.631 [0.622, 0.641]	2.73 [2.71, 2.76]	14.8 [14.8, 14.9]	-0.643 [-0.665, -0.619]
		X	0.0495	0.306 [0.276, 0.335]	2.34 [2.21, 2.49]	13.7 [13.2, 14.3]	-0.726 [-0.898, -0.562]

We then obtained the folded or minor allele site-frequency spectrum (SFS) for each site class by summing the minor allele frequency over all sites per class (Figure 4.1). We also generated the expected SFS for a population at equilibrium under a neutral Wright-Fisher model for comparison (Figure 4.1). We observed a deviation from the equilibrium expectation for autosomal and X-linked genes, for both synonymous and non-synonymous sites (Figure 4.1), consistent with negative Tajima's *D* values for all site classes (Table 4.1). If we assume that synonymous sites are selectively neutral, their negative *D* values either suggest a population expansion, or, an effect of Hill-Robertson interference from nearby sites under selection (Hill and Robertson 1966).

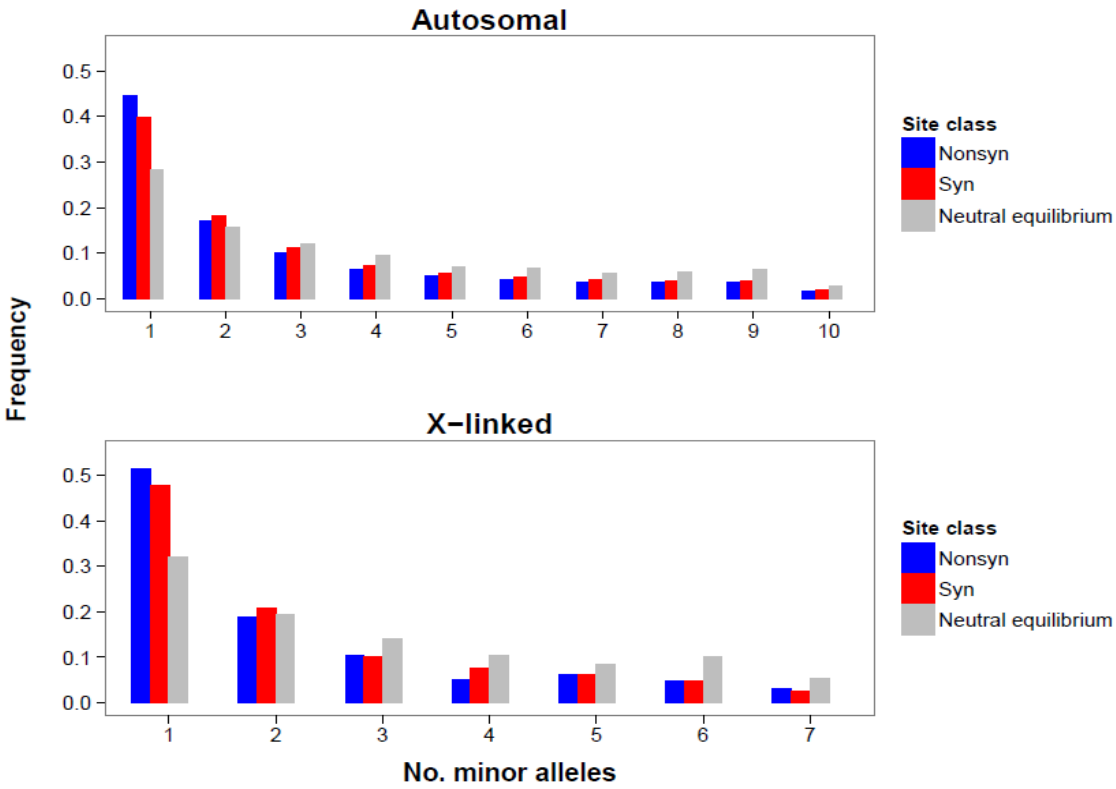


Figure 4.1. The site frequency spectrum for nonsynonymous and synonymous site classes for autosomal and X-linked genes. In grey we show the expected SFS for an equilibrium population under a neutral Wright-Fisher model of evolution. The SFSs are for non-CpG prone sites.

4.4.2 The fitness effects of new mutations in autosomal and X-linked loci

Model fitting to infer demography and selection. We estimated the distribution of fitness effects of new deleterious mutations (DFE) for autosomal and X-linked loci by applying a maximum likelihood approach (*DFE-alpha*) that fits a demographic and a

selection model to the SFSs from neutral and selected classes of sites (Keightley and Eyre-Walker 2007). We used synonymous sites to infer effects of population size changes and nonsynonymous sites to infer selection. A 2-epoch demographic model gave a good fit to autosomal and X-linked synonymous data (Figure 4.2), and a 3-epoch model produced only a marginally better fit (Table 4.2 and Figure 4.2).

We then fitted several types of models to the nonsynonymous data to infer the DFE (Table 4.2 and Figure 4.3). A model with 3 discrete selection coefficients had a better fit than the gamma distribution to the autosomal data (Table 4.2). However, the 3-spike model did not fit substantially better than the gamma distribution to the X-linked data (Table 4.2). For consistency of the analysis of autosomal and X-linked loci we used the gamma model to infer the DFE, while controlling for the population history inferred from the 2-epoch model. Note that we do investigate below the effect of fitting different models of the DFE and different demographic models on our inferences.

The DFE for autosomal and X-linked loci. The inferred parameters for the best-fitting 2-epoch and gamma distribution models are given in Table 4.3. The 2-epoch model gave evidence of an expansion for both autosomal and X-linked loci (Table 4.3). Even though our analysis included several thousand genes, the mean strength of selection for deleterious mutations (N_eE(s)) was very imprecisely estimated for both autosomal and X-linked loci (indicated by the very wide confidence intervals; Table 4.3). The point estimate for N_eE(s) for the X chromosome is very high ($3.73X10^8$). Even if we assume an N_e of 10^6 for *M. m. castaneus,* the E(s) value would still be higher than 100. This high value for E(s) should not be considered realistic, but rather an artifact of the

method to estimate the DFE. As discussed in the previous chapters, when the inferred gamma distribution is highly leptokurtic there is a disproportionately large contribution of mutations with strong effects to the mean, and since the method allows for s>1, E(*s*) could also be inferred to be much higher than 1. These excessively large estimates for E(*s*) do not have biological significance.

The shape parameter (*b*) was estimated with more precision than N_eE(*s*) and indicated a strongly leptokurtic DFE for both autosomal and X-linked loci (Table 4.3). We did not observe significant differences in the parameters of the DFE between autosomal and X-linked loci.

Table 4.2. Goodness of fit of demographic and selection models. The demographic models were fitted to the synonymous sites and the selection models to nonsynonymous sites for autosomal and X-linked genes separately. The log-likelihood difference ($\Delta \log L$) and the corrected akaike information criterion difference (ΔAIC) from the best fitted model is reported. The spike and step models consist of discrete selection coefficients that are fitted to the selected data. To infer the DFE with these models, we incremented the number of spikes/steps until the improvement of fitting additional spikes/steps is less than 2 *AIC* units. In parentheses we report the number of spikes/steps of the best-fitting spike model.

Sites	Chr.	Model	$\Delta \log L$	ΔAIC
Synonymous	A	Stationary	-1,727	-3,446
		2-epoch	-7	-10.1
		3-epoch	0	0
	X	Stationary	-36.7	-68.8
		2-epoch	-0.2	0
		3-epoch	0	-3.5
Nonsynonymous	A	No selection	-24,133.4	-48,256.8
		Gamma	-21.3	-36.5
		Spike (3)	0	0
		Step (2)	-2.8	-1.6
	X	No selection	-407.4	-810.4
		Gamma	-0.2	0
		Spike (3)	0	-5.5
		Step (2)	0	-1.5

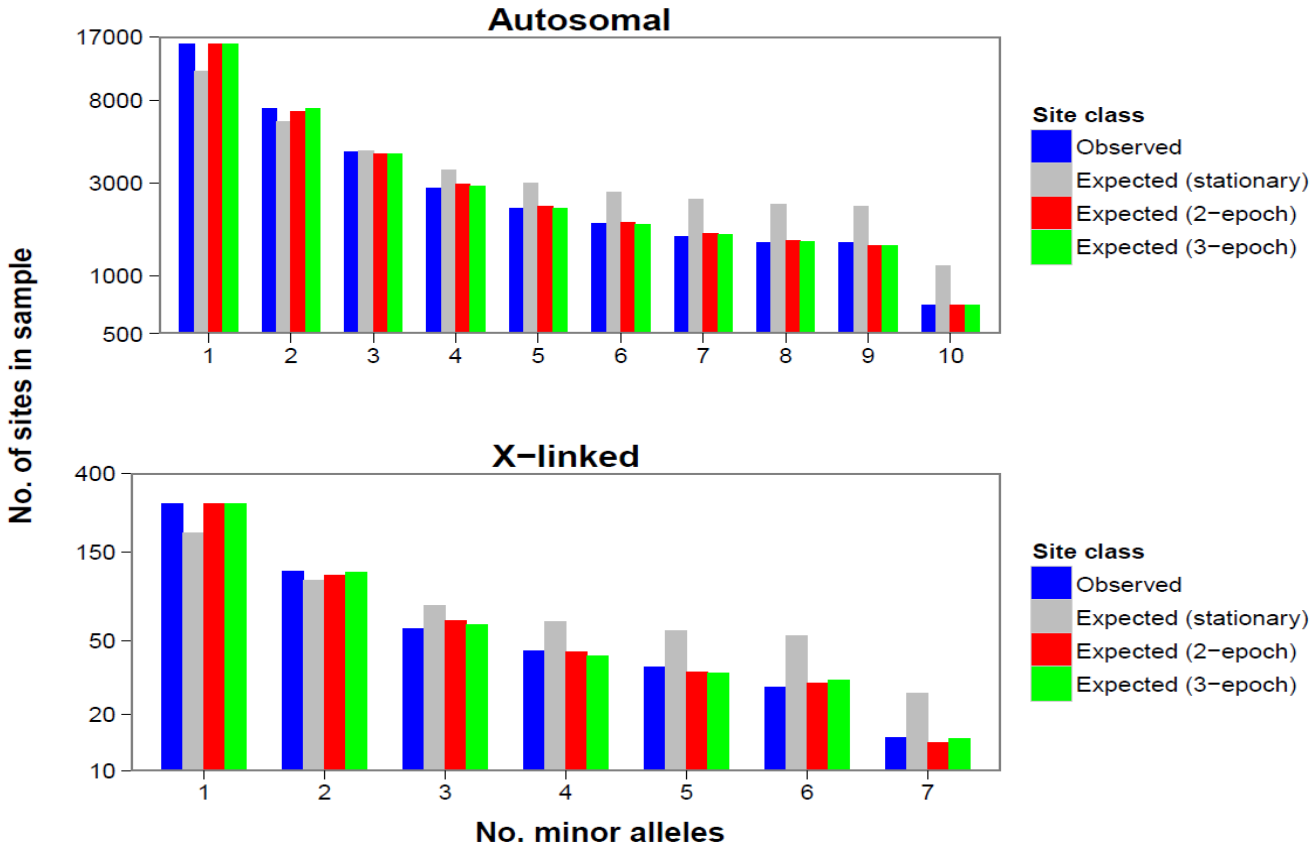


Figure 4.2. The observed synonymous site frequency spectrum and the expectation generated by assuming a stationary population size and two demographic models.The SFSs are for non-CpG prone sites.

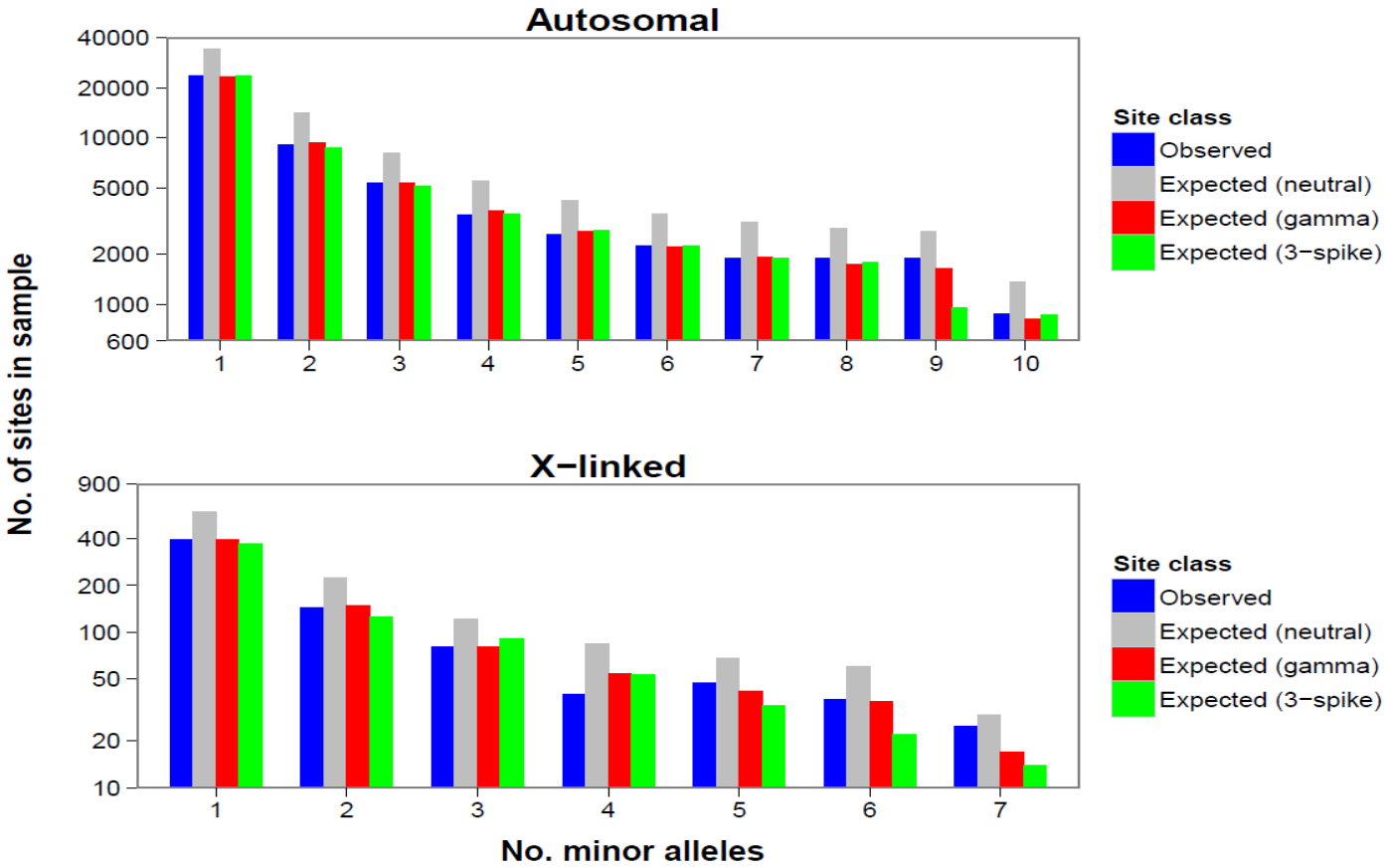


Figure 4.3. The observed nonsynonymous site frequency spectrum and the expectation generated by assuming no selection, a gamma distribution and a distribution consisting of 3 discrete selection coefficients. The SFSs are for non-CpG prone sites.

Table 4.3. Estimates and 95% confidence intervals for parameters of the 2-epoch demographic model and the gamma DFE for autosomal and X-linked loci. The 2-epoch model parameters are the magnitude of a population size change (N_2/N_1) and the time in generations since the size change (t_2/N_1), and are scaled by N_1 which is the initial size of the population. The parameter estimates for the gamma DFE are the mean strength of selection (N_eE(s)) and the shape (b) of the distribution.

Chr.	Demography (2-epoch)		Selection (gamma DFE)	
	N_2/N_1	t_2/N_1	N_eE(s)	b
A	2.79 [2.79, 3.07]	1.47 [1.3, 2.09]	$9.47X10^5$ [$2.74X10^5$, $3.36X10^7$]	0.11 [0.088, 0.13]
X	4.09 [3.07, 7.26]	2.19 [0.85, 6.14]	$3.73X10^8$ [$1.23X10^4$, $1.87X10^{15}$]	0.085 [0.05, 0.21]

We also compared the proportion of mutations assigned to four $N_e s$ ranges, and found that they did not differ between X-linked and autosomal loci (Figure 4.4). These results suggest that the efficcacy of purifying selection acting on nonsynonymous mutations is similar in X-linked and autosomal genes on average. This is unexpected, given that we infer that the X chromosome N_e is smaller than ¾ that of the autosomal genes and therefore experiences a stronger effect of drift. One possible explanation is that new deleterious mutations are on average recessive and therefore are removed more efficiently from the X than the autosomes. Presumably these two processes (smaller N_e of the X chromosome and recessivity of new deleterious mutations) cancel out to some extent.

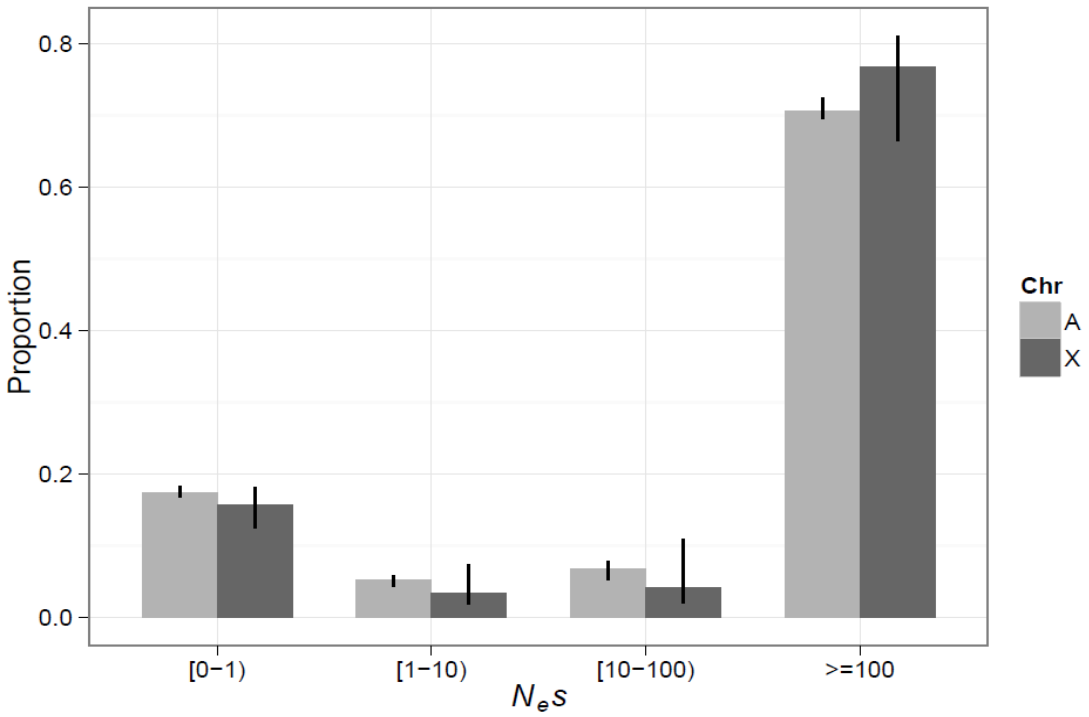


Figure 4.4. The distribution of fitness effects of new nonsynonymous mutations binned into four classes of effects for autosomal and X-linked genes. The estimates are for non-CpG- prone sites. 95% confidence intervals were generated by bootstrapping by gene.

4.4.3 Adaptive evolution in autosomal and X-linked loci

The inferred parameters of the DFE can be used together with the divergence between two species to infer the proportion of adaptive substitutions and the rate of adaptive relative to the rate of neutral substitution (α and ω_a respectively, calculated using equations 4.6 and 4.7). We considered significantly different ω_a values between two compared classes of genes as indicating different rates of adaptive evolution, but we also computed and compared d_N/d_S and α because these are more widely used than ω_a and can thus be compared with other studies. We compared estimates of d_N/d_S, α and ω_a for nonsynonymous sites of autosomal and X-linked genes (Table 4.4). X-linked loci have significantly higher d_N/d_S, α and ω_a than autosomal loci ($P<0.05$ for all parameter comparisons between autosomes and the X, using either *M. famulus* or rat as the outgroup; Table 4.4). These results provide strong support for faster X adaptive protein evolution. Faster X evolution was also inferred when using a 3-epoch model to infer the demographic history or with a DFE model consisting of three point masses (Table 4.5), or when using non-parametric estimators of α (panel 'All' of Figure 4.9). Note that the results in Table 4.4 were produced by correcting for the contribution of divergence to polymorphism (as per Keightley and Eyre-Walker 2012), but we did not use this correction to produce the results in Table 4.5, because it was not implemented in *DFE-alpha* for some of the selection models that were used to produce the results in Table 4.5. Since the estimates in Table 4.5 for α and ω_a were produced by using the rat as an outgroup, we expect a minimal effect of not applying the correction to the estimates.

Table 4.4. Estimates of d_N/d_S, α and ω_a for autosomal and X-linked genes using *M. famulus* and rat as outgroups. The estimates are for non-CpG prone sites. 95% confidence intervals are given in brackets.

Outgroup	Chr.	d_N/d_S	α	ω_a
M. famulus	A	0.238 [0.235, 0.255]	0.316 [0.289, 0.365]	0.0753 [0.0695, 0.0923]
	X	0.287 [0.257, 0.330]	0.480 [0.363, 0.646]	0.138 [0.0983, 0.201]
Rat	A	0.241 [0.238,0.246]	0.322 [0.284,0.347]	0.0774 [0.0687,0.0843]
	X	0.278 [0.252,0.300]	0.463 [0.356,0.617]	0.129 [0.0956,0.172]

Table 4.5. Point estimates for $\frac{\overline{u_{del}}}{u_{neu}}$, α and ω_α when fitting different combinations of demographic and selection models to autosomal and X-linked data. The models used in the present study are highlighted in light grey. We used the rat as an outgroup to calculate α and ω_a. The divergences from *M. m castaneus* were not corrected for the contribution of polymorphism to divergence because this correction is not implemented yet for some combinations of models. This is not expected to affect our inferences for α and ω_a, because the rat is distantly related to *M. m. castaneus* (~18% synonymous divergence). The models used in the present study are highlighted in light grey.

Chr.	Demographic model	Selection model	$\frac{\overline{u_{del}}}{u_{neu}}$	α (rat)	ω_a (rat)
A	2-epoch	Gamma	0.17	0.28	0.065
		Spike (3)	0.19	0.20	0.046
	3-epoch	Gamma	0.16	0.29	0.068
X	2-epoch	Gamma	0.15	0.45	0.12
		Spike (3)	0.17	0.38	0.10
	3-epoch	Gamma	0.15	0.45	0.12

4.4.4 Evolution of male and female-specific genes.

The faster-X effect is expected to be most pronounced if selection acts on males only, whereas equal rates of adaptive evolution are expected if selection acts on females only (Charlesworth et al. 1987; Vicoso and Charlesworth 2006). To investigate this prediction, we compared rates of adaptive evolution for genes with sex-specific expression. We used gene expression data for several tissues of mice from the Atlas of Gene Expression (Su et al. 2004) to define male-specific genes as those that had specific expression in testis or prostate and female-specific genes as those that had specific expression in ovary or uterus. We found significantly faster-X adaptive evolution for male but not for female-specific genes (Figure 4.5). Estimates for d_N/d_S, α and ω_a were not significantly different between the autosomes and the X chromosome for genes that do not show male or female-specific expression (Figure 4.5). This could be due to lack of power to detect a significant faster-X effect for these genes. We obtained similar results for male and female-specific genes using non-parametric estimators of α, although genes lacking sex-specific expression showed significantly faster-X adaptive evolution with this method (panel 'Sex-specific', Figure 4.9).

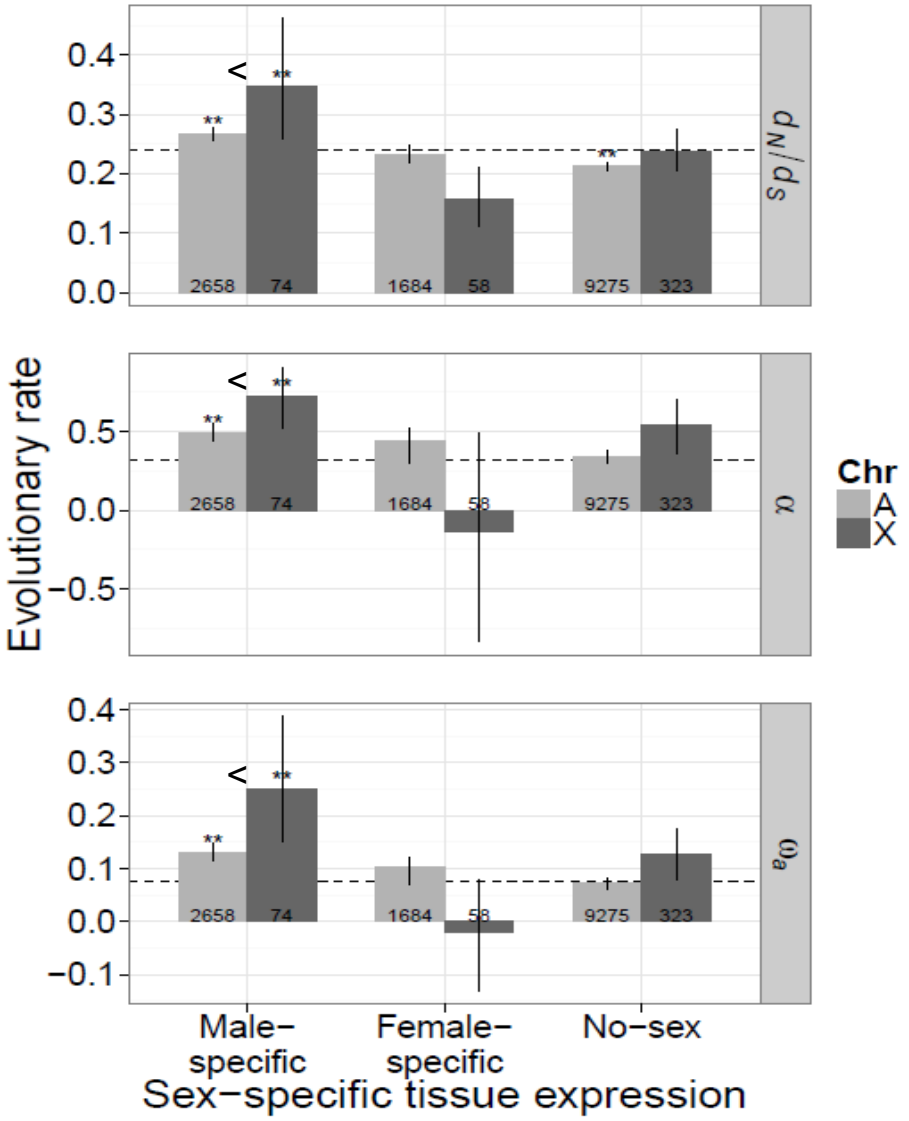


Figure 4.5. Molecular evolution of genes that have male- or female-specific expression and non-sex-specific expression. Estimates for d_N/d_S, α and ω_a were calculated using *M. famulus* as the outgroup. Error bars are 95% confidence intervals (CIs) obtained by bootstrapping by gene. Two-tailed bootstrap tests were performed to compare d_N/d_S, α and ω_a estimates with the autosomal average (indicated by the dashed line), and between autosomal and X-linked genes of each class. Stars indicate significance for the comparison to the autosomal average (* P<0.05, ** P<0.01). Signs indicate significance for the comparisons between autosomal and X-linked genes (one sign; P<0.05). The numbers in the boxes indicate the number of genes analysed within each class.

Tissue specificity of male and female-specific genes. Previous studies have shown that narrowly expressed genes have higher d_N/d_S values than widely expressed genes (Liao et al. 2006), and that the X chromosome is enriched for genes with narrow expression (Meisel et al. 2012a). To investigate whether a difference in tissue specificity between autosomal and X-linked genes could affect our results, we calculated the breadth of expression (τ) using equation 4.10. Small values of τ correspond to broad expression, whereas large values correspond to narrow expression. We found that τ is not significantly different between autosomal and X-linked genes that have male- or female-specific expression (Mann-Whitney *U* test $P>0.05$; Figure 4.6). This is expected, because we investigated genes with expression specific to male and female reproductive tissues, which are enriched in narrowly expressed genes. X-linked genes that were not male- or female-specific had a significantly higher τ than autosomal genes (Mann-Whitney *U* test $P<0.01$; Figure 4.6). Therefore, a narrower breadth of expression of X-linked than autosomal genes might partially account for faster-X evolution of genes that are non male- or female-specific.

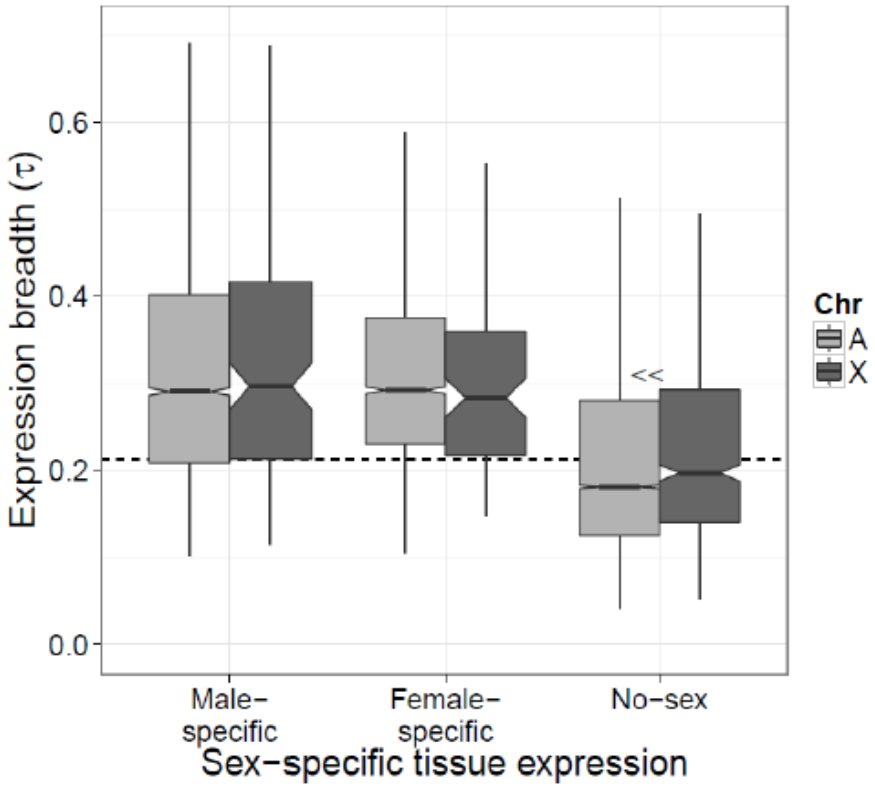


Figure 4.6. Breadth of expression of genes that have male- or female-specific expression and non-sex-specific expression. Boxes indicate 25th and 75th percentiles of the distribution of τ and whiskers are approximate 95% CIs. Solid line within boxes indicates the median τ and notches are approximate 95% CIs for the median. Dashed line indicates the genomic average τ. A Mann-Whitney *U* test was performed to compare median τ between autosomal and X-linked genes of each class. Signs indicate significance for the comparisons between autosomal and X-linked genes (one sign; $P<0.05$, two signs; $P<0.01$).

4.4.5 Evolution of genes expressed during spermatogenesis.

During spermatogenesis, each diploid spermatogonium cell undergoes two rounds of meiosis to give 4 haploid spermatids. X-linked recessive mutations are exposed to selection only early in spermatogenesis (pre-meiotically), but X-linked and autosomal mutations are exposed to selection late in spermatogenesis (post-meiotically). Therefore, we expect faster-X evolution only for genes that are expressed early in spermatogenesis. Moreover, during the first meiosis in spermatogenesis (meiosis I), X-linked genes experience global suppression of their expression (meiotic sex chromosome inactivation; MSCI (Lifschytz and Lindsley 1972). However, a few X-linked genes escape MSCI and are expressed post-meiotically (Namekawa et al. 2006). We obtained gene expression data for male germ cells at different spermatogenetic stages (dataset of Namekawa et al. 2006) to investigate the evolutionary rate of genes that have different expression patterns during spermatogenesis.

We defined three groups of genes: genes that are expressed pre-meiotically and suppressed post-meiotically (group A; Figure 4.7), genes that are expressed both pre-meiotically and post-meiotically (group B; Figure 4.7), and genes that are suppressed pre-meiotically and expressed post-meiotically (group C; Figure 4.7).

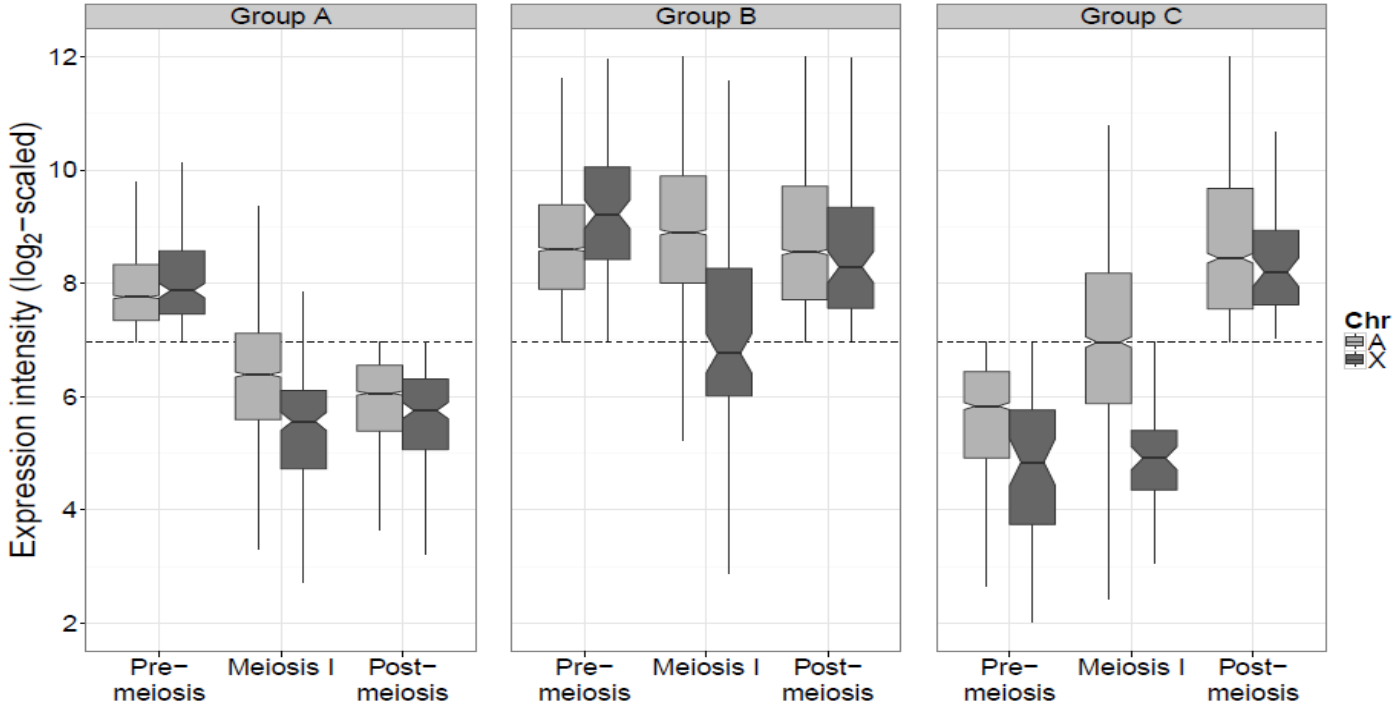


Figure **4.7.** Expression pattern (A) of three groups of genes expressed during spermatogenesis. Group A are genes which are expressed exclusively pre-meiotically, group B are genes which are expressed during pre-meiosis and post-meiosis, and group C are genes which are expressed exclusively post-meiotically. Boxes indicate 25th and 75th percentiles of $\log_2$ expression intensity and whiskers are approximate 95% confidence intervals (CIs). Solid line within boxes indicates median expression intensity and notches are approximate 95% CIs for the median. The dashed line indicates the expression intensity threshold that was used to define a gene as being expressed.

We then calculated d_N/d_S, α and ω_a for autosomal and X-linked genes within each group of spermatogenesis-expressed genes. Genes expressed only pre-meiotically (group A) had similar d_N/d_S, α and ω_a for the X chromosome and the autosomes (Figure 4.8). This is unexpected, because new advantageous X-linked recessive mutations in group A genes are exposed to selection in males. However, genes with exclusively early expression in spermatogenesis have been reported to be female-biased in expression (Zhang et al. 2010) and these are expected to evolve at similar rates for the autosomes and the X chromosome (Charlesworth et al. 1987; Vicoso and Charlesworth 2009). Therefore, the observation that group A X-linked and autosomal genes evolve at a similar rate does not necessarily contradict faster-X theory. Genes expressed both pre- and post-meiotically (group B) had significantly higher d_N/d_S, α and ω_a for the X chromosome than the autosomes (Figure 4.8). For group B, X-linked genes have a different expression profile from autosomal genes, because MSCI affects only the X chromosome (Figure 4.7). Therefore, X-linked and autosomal genes within group B might not be comparable. However, X-linked genes in group B have significantly higher α and ω_a than the autosomal average (Figure 4.8). The rapid adaptive evolution of X-linked group B genes might be related to their escape from MSCI. Genes expressed exclusively post-meiotically (group C) had similar d_N/d_S, α and ω_a for the X chromosome and the autosomes (Figure 4.8). This is expected, since cells are haploid late in spermatogenesis on both the autosomes and the X chromosome and recessive mutations are therefore exposed to selection.

Non-parametric estimators of α produced equivocal results (panel 'Sperm'; Figure 4.9). The Fay et al. (2001) estimator (α_{FWW}) showed similar results to

DFE-alpha, but the Smith and Eyre-Walker (2002) estimator (α_{SEW}) showed significantly higher α for X-linked than autosomal genes for all classes of genes examined (panel 'Sperm'; Figure 4.9). However, both estimators showed α to be significantly higher from zero for X-linked genes of group B and autosomal genes of group C, which is consistent with our findings when using *DFE-alpha*. All methods consistently show rapid adaptive evolution for X-linked genes that escape MSCI and also autosomal genes that are post-meiotically expressed.

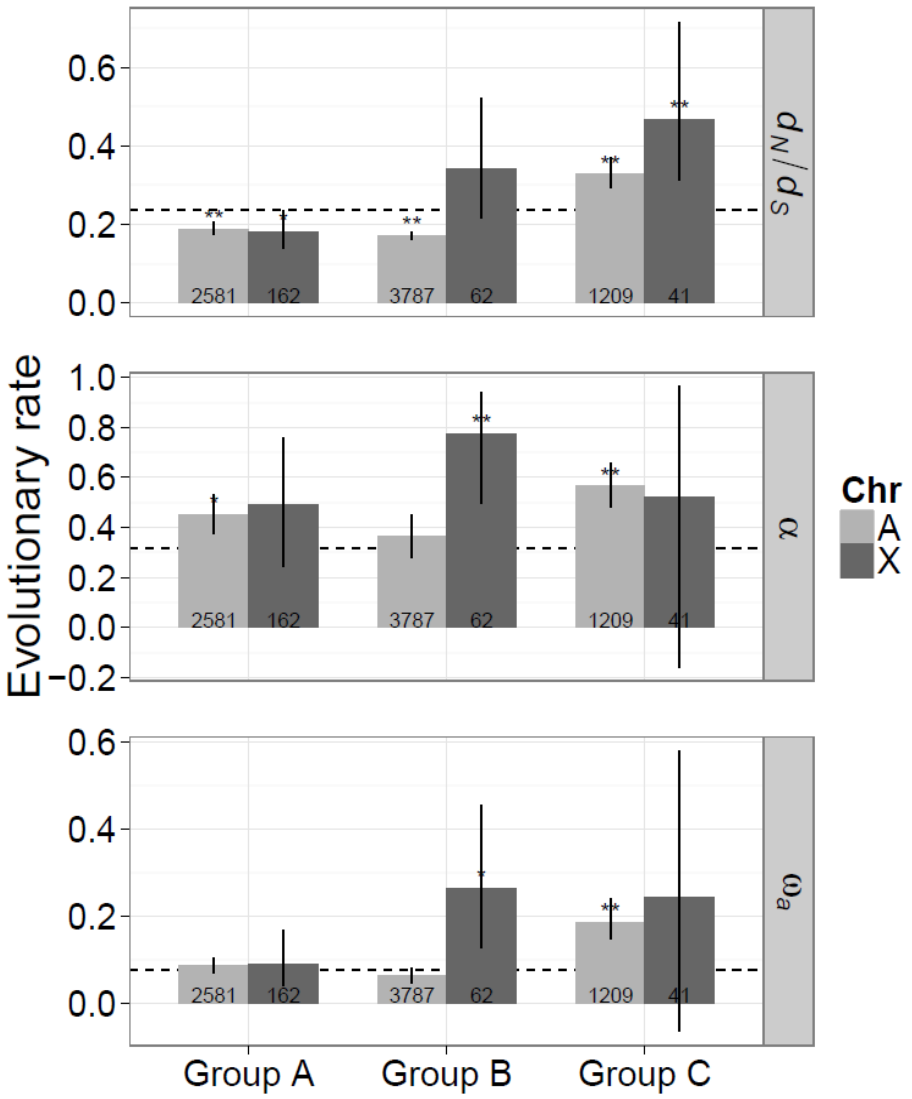


Figure 4.8. Estimates for d_N/d_S, α and ω_a for group A, group B and group C. d_N/d_S, α and ω_a were calculated by using *M. famulus* as the outgroup. Error bars are 95% CIs obtained by bootstrapping by gene. Two-tailed bootstrap tests were performed to compare d_N/d_S, α and ω_a estimates with the autosomal average (indicated by the dashed line), and between autosomal and X-linked genes of each class. Stars indicate significance for the comparison to the autosomal average (* P<0.05, ** P<0.01). The numbers in the boxes indicate the number of genes analysed within each class.

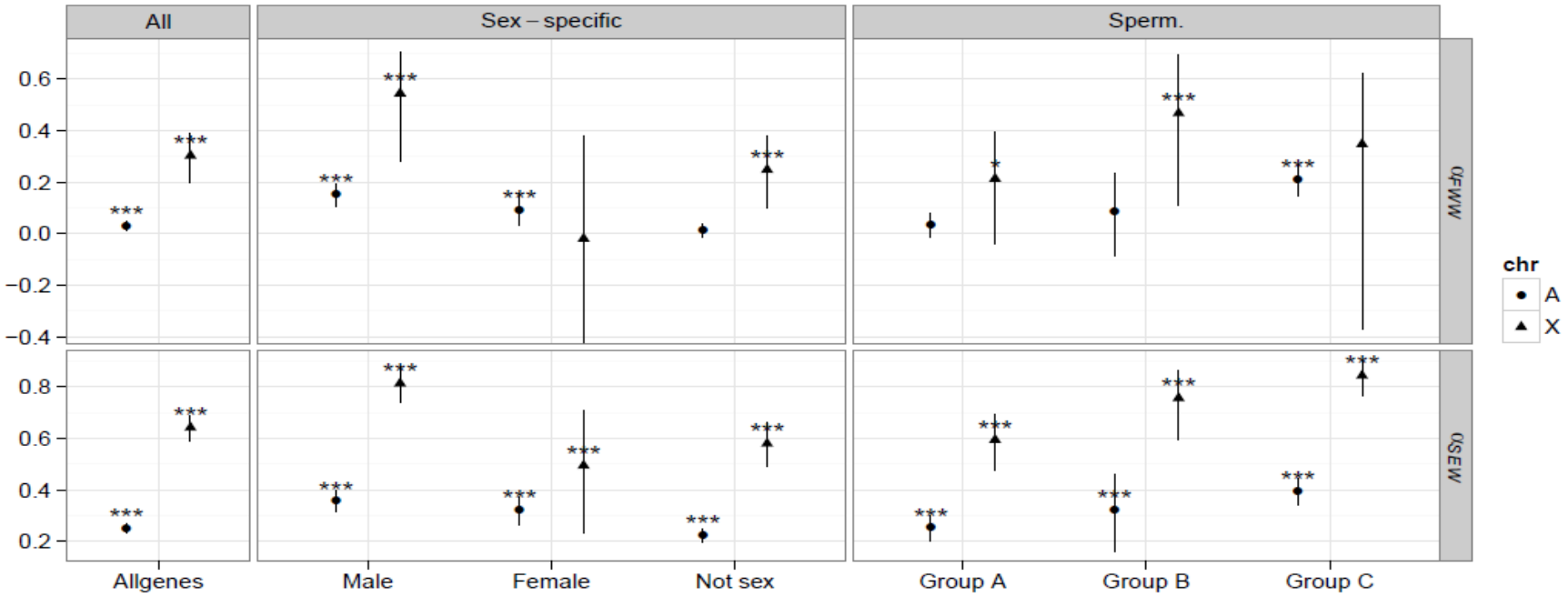


Figure 4.9. Estimates for α using non-parametric estimators. We used the Fay et al. (2001) and Smith and Eyre Walker (2002) estimators (α_{FWW} and α_{SEW}, respectively; Fay et al. 2001; Smith and Eyre-Walker 2002**).** Autosomal and X-linked genes are compared for all genes (All), genes with sex-specific expression (Sex-specific) and genes with a different expression pattern during spermatogenesis (Sperm.). Estimates were obtained by using *M. famulus* as the ougroup. Error bars are 95% confidence intervals obtained by bootstrapping by gene 10,000 times. Stars indicate significance for α>0 (* P<0.05, ** P<0.01, ***P<0.001).

4.5 Discussion

Our main finding is that the X chromosome has a higher rate of adaptive protein evolution than the autosomes. X-linked genes with male-specific expression evolve particularly rapidly, whereas X-linked and autosomal genes with female-specific expression evolve at similar rates. These observations can be explained if new advantageous mutations are on average recessive (Charlesworth et al. 1987). We can use our estimate of the ratio of the rate of adaptive substitution of X-linked to autosomal loci ($R=\omega_{aX}/\omega_{aA}$) to predict the average dominance coefficient (h) for new, advantageous mutations. Assuming that $N_{eX}/N_{eA}=0.75$, and that reproductive success of males and females are equal, we can use equations 4.2 and 4.3 to predict h (Table 4.6). The predicted value of h is ~0.2 (Table 4.6).

Table 4.6. Quantitative prediction for the dominance coefficient (h) based on our estimated ratio of adaptive substitution for X-linked over autosomal loci ($R=\omega_{aX}/\omega_{aA}$). We assumed that $N_{eX}/N_{eA}=0.75$, and that reproductive success of males and females are equal. We calculated R using *M. famulus* as the outgroup. 95% confidence intervals are given in brackets.

Gene class	Expected R	Estimated R	Predicted h
All genes	$R \approx \frac{2h+1}{4h}$	1.8 [1.2, 2.6]	0.19 [0.12, 0.36]
Not sex-specific		1.6 [0.78, 2.5]	0.24 [0.13, 0.89]
Male-specific	$R \approx \frac{1}{2h}$	2.4 [1.3, 4.3]	0.21 [0.12, 0.38]

Connallon et al. (2012) showed that genetic architecture underlying bouts of adaptive substitution can influence the assumptions of the theoretical predictions of Charlesworth, Coyne and Barton (1987) and that a contribution of standing variation to adaptive substitution can dampen the predicted relationship between R and h (Connallon et al. 2012). According to Connallon et al. (2012) results, our estimated R would be roughly consistent with an average h of ~0.25, assuming that N_{eX}/N_{eA}=0.75, that a large number of genes (>>1) contribute to individual bouts of adaptation and that most adaptive substitutions involve new mutations (Connallon et al. 2012). Moreover, Vicoso and Charlesworth (2009) showed that the relationship between R and h is sensitive to N_{eX}/N_{eA} and that recessivity of new advantageous mutations would be expected even for R=1, if N_{eX}/N_{eA}<0.75 (Vicoso and Charlesworth 2009). Given that we estimated N_{eX}/N_{eA} < 0.75, we expect new advantageous mutations to be recessive, on average, with h<0.25.

The causes of reduced diversity on the X chromosome. The X chromosome is expected to have 0.75 of the autosomal diversity, if the population is at equilibrium, and males and females have equal reproductive success. We observed a X/A diversity ratio significantly lower than this expectation (0.58). This could be explained by unequal variance in reproductive success between males and females (Charlesworth 2001), by population size reductions or bottlenecks (Wall et al. 2002; Pool and Nielsen 2007), or by a stronger effect of selective sweeps or background selection in eliminating X-linked synonymous diversity (Hutter et al. 2007).

If females have a larger variance in reproductive success than males, then the

female N_e is expected to be smaller than the male value. This could produce a X/A diversity ratio smaller than 0.75. Comparisons of autosomal and mitochondrial diversity in populations of *Mus musculus* have shown evidence for a larger female than male N_e (Baines and Harr 2007), which would predict a X/A diversity ratio higher than 0.75 (i.e. contrary to our observation). Unequal variance in reproductive success between males and females is therefore an unlikely explanation for the observed X/A diversity ratio.

Population size reductions or bottlenecks have been shown to reduce the X/A diversity ratio (Wall et al. 2002; Pool and Nielsen 2007). A 2-step demographic model gave a good fit to the autosomal and X-linked synonymous SFSs, providing evidence for a population expansion, which has been shown to increase the X/A diversity ratio (Pool and Nielsen 2007). However, given that synonymous polymorphism can also be affected by selection on linked sites, and this could bias demographic inference, we do not consider our inference of an expansion compelling, and cannot definitively exclude the possibility of a bottleneck in the history of *M. m. castaneus*, which could at least partially explain the reduced diversity on the X chromosome.

Finally, the X chromosome may experience a stronger effect of selective sweeps or background selection than the autosomes, reducing neutral diversity linked to selected loci. A stronger effect of selective sweeps on neutral diversity is plausible, given that we find a higher rate of adaptive evolution on the X than the autosomes.

Genes expressed during spermatogenesis. Previous studies have revealed evidence

for rapid evolution of genes expressed during spermatogenesis (Torgerson et al. 2002; Torgerson and Singh 2006), particularly those expressed post-meiotically (Good and Nachman 2005). We found that X-linked genes that are expressed during spermatogenesis and escape meotic sex chromosome inactivation (MSCI) experience rapid adaptive evolution. A previous study found that most X-linked genes that escape MSCI in mice are additions to the X chromosome less than 50 MYA (Zhang et al. 2010). Rapid adaptive evolution of these genes is presumably due to their young age (Zhang et al. 2010), or a consequence of an evolutionary arms race with selfish genetic elements that control expression during spermatogenesis (Presgraves 2008).

Explanations for the large-X. Faster-X evolution due to average recessivity of advantageous mutations can explain the large effect of the X chromosome in speciation. This is because loci that contribute to hybrid incompatibilities will evolve faster when located on the X chromosome than the autosomes (Presgraves 2008). Future studies should focus on documenting precisely the excess of X-linked relative to autosomal loci that cause hybrid incompatibility in mice, and investigating to whether a 1.8 times faster rate of adaptive substitutions on the X can explain that excess.

Another explanation for the large-X effect that is compatible with our data is related to the regulation of genes expressed during spermatogenesis (Presgraves 2008). Spermatogenesis might be inherently sensitive to perturbations, that are likely to occur in hybrids. For example, a recent study in *M. musculus* found a strong

association between X-linked hybrid male sterility and disruption of MSCI (Campbell et al. 2013). The phenomenon of MSCI may be universal to species with heteromorphic chromosomes (Namekawa and Lee 2009) and it has been suggested that MSCI evolved as a defense mechanism against selfish genetic elements such as sex-ratio distorters (Meiklejohn and Tao 2010). Recurrent bouts of invasion of selfish genetic elements could trigger an evolutionary arms race with the host to suppress their expression. Therefore, the faster-X evolution that we observe could be a consequence of a high concentration of genes on the X chromosome that evolve rapidly due to genetic conflict. Indeed, sex-ratio distorters have been shown to be more likely to invade a population when located on the sex chromosomes than the autosomes (Frank 1991; Hurst and Pomiankowski 1991). Male-specific genes, and genes that are expressed post-meiotically in spermatogenesis, might be especially likely to be involved in these arms races, which could explain the pronounced faster-X evolution of those genes. Future studies should focus on extensive mapping of selfish genetic elements such as sex-ratio distorters and further dissecting their potential evolutionary link with MSCI.

Chapter 5. General discussion

In this thesis, I have developed methods to quantify the effects of new mutations on fitness and investigated natural selection at the molecular level in a population of house mice (*Mus musculus castaneus*). In this chapter, I describe a summary of the findings for each research chapter in the thesis. I then make some comments about the novelty of the findings that have been presented. I go on to discuss limitations of the methods. Lastly, I present the general implications of the work for various subjects of evolutionary genetics research and suggest future directions of research.

5.1 Summary of findings

In Chapter 2, it was shown that fitting unimodal distributions, such as the lognormal and gamma distributions, to infer a distribution of fitness effects (DFE) that is multimodal will lead to misinference of biologically important properties of the DFE, such as the mean effect ($\overline{N_e s}$) and the mean fixation probability of a new mutation ($\bar{u}$). To address this issue, I modelled the DFE either as series of discrete selective effects (spikes-model) or continuous uniform distributions of selective effects (steps-model). I showed with simulations that the spikes and steps models fit better to multi-modal DFEs than the lognormal and gamma distributions. Moreover, I showed that the spikes and steps models perform well in accurately inferring $\overline{N_e s}$ and $\bar{u}$ when the true DFE is unimodal or multimodal. An analysis of large polymorphism datasets from natural populations of *Drosophila melanogaster* and *M.*

m. castaneus to study the DFE of nonsynonymous mutations revealed evidence for a lognormal DFE in *D. melanogaster* and a bimodal DFE in *M. m. castaneus*.

In Chapter 3, I quantified the strength of negative selection and the rate of adaptive substitution for nonsynonymous sites of protein-coding genes and non-coding DNA upstream and downstream of genes by analysing polymorphism data from a population of *M. m. castaneus*. I found that non-coding DNA is much less selectively constrained than nonsynonymous sites. Moreover, I found that the fraction of substitutions that were driven to fixation by positive selection (α) is much smaller in non-coding DNA than in nonsynonymous sites. However, I found that the rate of adaptive relative to neutral substitution (ω_α) is rather similar between nonsynonymous sites and non-coding DNA. I concluded that the higher α for nonsynonymous sites than non-coding DNA is only due to higher constraint on non-synonymous sites rather than a higher rate of adaptive substitution than non-coding DNA.

In Chapter 4, I quantified the rate of adaptive substitution in nonsynonymous sites of the majority of protein-coding genes in *M. m. castaneus* in order to test the predictions of a hypothesis that has been suggested to explain Haldane's rule and the large effect of the X chromosome in speciation (the so-called faster-X hypothesis). I firstly compared the rate of adaptive substitution between autosomal and X-linked genes, and found that X-linked genes experience a 1.8 times faster rate of adaptive substitution than autosomal genes. Moreover, I found that X-linked male-specific genes have a 2.4 times faster rate of adaptive substitution than autosomal male-specific genes, but female-specific X-linked and autosomal genes evolve at

similar rates. These observations were found to be consistent with the predictions of the faster-X hypothesis. I also analysed spermatogenesis-expressed genes and found that genes that are expressed in the early pre-meiotic stage of spermatogenesis do not display faster-X evolution. Finally, I found that autosomal genes that are expressed in the post-meiotic stage of spermatogenesis and X-linked genes that escape meiotic sex chromosome inactivation (MSCI) experience a rapid rate of adaptive substitution. I concluded that the observed faster-X evolution could be either due to average recessivity of new advantageous mutations or more intense genetic conflict on the X chromosome than the autosomes.

5.2 Novelty of findings

Previous studies have modelled the DFE as discrete selection coefficients (Keightley and Eyre-Walker 2010; Wilson et al. 2011) or mixtures of distributions (Boyko et al. 2008). However, the discrete models were used *ad-hoc* in those studies, without justifying their use or demonstrating whether they are better in inferring the DFE than other widely used distributions, such as the gamma. In Chapter 2, I demonstrated for the first time that the signal of a multi-modal DFE does exist in the site-frequency spectrum (SFS) and can be detected by using the spikes/steps models. I investigated the biases that can arise in inferring important parameters of the DFE when fitting unimodal distributions to multi-modal DFEs and demonstrate under which conditions the spikes/steps models are better in inferring the DFE than unimodal distributions. I consolidated these findings by demonstrating that they are robust to complex population histories and linked selection. Finally, I inferred the

DFE for nonsynonymous changes in *D. melanogaster* and *M. m. castaneus* by using the newly developed spikes/steps models to analyse newly generated, genome-wide, polymorphism datasets.

There have been many studies investigating selective constraint on protein-coding genes and non-coding DNA in mammals (Keightley, Kryukov, et al. 2005; Keightley, Lercher, et al. 2005; Gaffney and Keightley 2006). The work presented in Chapter 3 is the first study to quantify the DFE and positive selection on non-coding DNA in a mammalian species other than human. The work on the DFE in non-coding DNA revealed similar results to previous studies. However, as I discuss below, the findings from the study of positive selection on non-coding DNA are surprising and have serious implications on the central question of the locus of adaptation.

The work presented in Chapter 4 is the most comprehensive test of the faster-X hypothesis that has ever been conducted. This work is the first to test multiple predictions of the faster-X hypothesis by using polymorphism data for the majority of autosomal and X-linked genes. Moreover, previous tests of the faster-X hypothesis compared d_N/d_S ratios (Betancourt et al. 2002; Mank et al. 2010) or α (Baines et al. 2008a; Hvilsom et al. 2012; Mackay et al. 2012) between autosomal and X-linked genes, but both of these parameters have serious limitations. A high d_N/d_S can be produced by either strong positive selection or weak selective constraint. Moreover, differences in α between classes of sites or genes can only indicate a difference in selective constraint rather than a difference in the efficiency of positive selection. In Chapter 4, I estimate ω_a, which robust to these limitations (Gossmann et

al. 2010).

5.3 Limitations

Linkage. The methods that I used to infer the DFE and the rate of adaptive substitution make several assumptions. The assumption of unlinked sites is perhaps the most likely to be violated for real datasets. In Chapter 2, I showed with simulation that even moderate levels of linkage can cause the spike/step models to overfit the data and produce spurious evidence of multimodality. I also demonstrated that fitting a demographic model to neutral sites that are interdigitated with the selected sites can control for the effects of linkage. However, the use of a demographic model to control for the effects of linkage, although it appears to work, is hardly sufficient for several reasons. Firstly, since this correction procedure does not explicitly model genetic linkage, it is uncertain for what types of DFEs the correction actually works. I investigated only a small parameter space, in respect to the properties of the simulated DFE, and it is reasonable to imagine that for certain DFEs it might fail to work, producing yet unknown biases. Secondly, the demographic history that is inferred by using this method is very likely to be wrong. For example, an apparent population size change, such as an expansion, might only be an artifact of the effect of linked selection. By assuming that the parameters of the demographic model are nuisance parameters, the scope of the analysis gets limited somewhat.

Population size changes. Population size changes can produce artificial evidence for

positive selection and estimates of $\alpha>0$, in the absence of adaptive substitutions. The method used to infer α does correct for recent demographic changes, but there are at least two reasons to suspect that this may not be sufficient, and that our estimates of α may be biased. Firstly, as was outlined above, linked selection can produce artificial evidence for population size changes. It is unknown whether the inference of selection would be robust to the combined effects of a complex demographic history and linked selection. Secondly, ancient population size changes do not leave a signature in the polymorphism data, but they can potentially affect divergence between two species. If *M. m. castaneus* had experienced prolonged bottlenecks in the distant past, these might have caused slightly deleterious mutations to become fixed. Those mutations will not segregate in the present, much larger, N_e of *M. m. castaneus* and thus produce $\alpha>0$.

Assumption of neutral evolution for synonymous sites. In most of my analyses I assumed that synonymous sites are evolving neutrally. It is possible that this assumption is violated. A process known as codon usage bias can create conditions for selection at synonymous sites. Codon usage bias occurs when alternative codons for an amino acid occur in different frequencies. This phenomenon has been documented for several species (Hershberg and Petrov 2008), including mice (Eyre-Walker 1991). Codon-usage bias can result from mutational biases or selection (Hershberg and Petrov 2008). A different selection pressure on codon-usage bias for autosomal and X-linked genes has been documented in *Drosophila* and it has been suggested that this can lead to artificial evidence for faster-X evolution (Campos et

al. 2012). However, there is no evidence for a significant role of selection on shaping codon-usage bias in mice (Eyre-Walker 1991; Urrutia and Hurst 2001; Yang and Nielsen 2008).

Codon-usage bias is not the only process that create non-neutrality of synonymous variants. For example, synonymous mutations can affect splicing and/or stability of mRNA trancsripts (Chamary et al. 2006). If synonymous sites experience selective constraints, our estimates of the strength of selection on nonsynonymous sites and non-coding DNA will be underestimates. However, our conclusions on the relative rate of adaptive substitution between nonsynonymous sites and non-coding DNA or between autosomal and X-linked genes will be unaffected.

Definition of *cis*-regulatory DNA. In Chapter 3, I quantified the rate of adaptive substitution in non-coding DNA that was ~500 bp upstream and downstream of protein-coding genes. Based on their location close to the genes, these regions are likely to contain *cis*-regulatory elements, such as promoters of transcription. However, I do not provide direct evidence, such as motifs, for the existence of *cis*-regulatory elements in these regions. Therefore, I might have underestimated the level of selective constraint and the rate of adaptive substitution on *cis*-regulatory DNA.

Estimates of α by different studies. The point estimates for α for nonsynonymous sites of autosomal protein-coding genes in *M. m. castaneus* varied substantially between studies. In Chapter 3, where a small sample of genes was used, α was

estimated to be ~0.4-0.5, depending on the outgroup used and the class of sites that was used as the neutral standard. In Chapter 4, where a much larger number of genes was used, α was estimated to be lower (~0.3; non-significantly lower than the Chapter 3 estimate). The higher estimate for α from Chapter 3 is likely due to the biased sample of genes that was used to perform the analysis. The gene sample from Chapter 3 is enriched in genes that are orthologous to human genes that are associated with human genetic diseases whose susceptibility is influenced by environmental challenge (Livingston et al. 2004). Therefore, the α estimate from Chapter 4 which was produced by analysis of a very large and unbiased sample of genes is likely to be more accurate than the estimate from Chapter 3. However, in Chapter 2, when I used discretised distributions to model the DFE, I obtained an even lower estimate for α than the estimate from Chapter 4 (0.2 versus 0.3, respectively), even though I used exactly the same data in the two analyses. Since the discretised distribution models were shown to fit the nonsynonymous data better than the gamma distribution model, I conclude that the most likely estimate for α for nonsynonymous sites of autosomal protein-coding genes in *M. m. castaneus* and when using the rat as the outgroup is the one from Chapter 2, i.e. 0.2.

5.4 General implications of findings and future directions.

The results presented in the thesis are informative for many subjects of evolutionary genetics research and I discuss here their general implications and suggest future directions of research.

Selectionist-neutralist controversy. The study of natural selection in *M. m. castaneus* revealed substantial evidence of positive selection on protein-coding genes (Chapters 3 and 4). In Chapter 3, I analysed a small number of protein-coding genes and in Chapter 4, I expanded the analysis to a large proportion of the protein-coding genes in the mouse genome. The proportion of nonsynonymous substitutions that were fixed by positive selection (α) was consistently found to be significantly greater from zero in both studies, even though the estimate for alpha varied substantially between the two studies (0.5 versus 0.3 in Chapters 3 and 4, respectively)This result is in direct contrast to the prediction of the neutral theory for only a minor, insignificant, contribution of adaptive substitutions to molecular evolution (Kimura 1985). For the past decade, evidence has continuously accumulated against the neutral theory (Fay 2011). Continued use of the neutral theory may prove to be an impediment to progress in evolutionary genetics research (Hahn 2008). Future work should focus on developing a new theory of molecular evolution that can explain observations such as those presented in Chapters 3 and 4 and that will replace the neutral theory as the 'null' hypothesis.

The distribution of fitness effects of new mutations. The DFE is frequently modelled as a unimodal distribution such as the gamma or lognormal distributions (Eyre-Walker and Keightley 2007). In Chapter 2, it was shown that fitting unimodal distributions to multimodal DFEs can lead to significant biases in the estimates of important parameters of the DFE, such as $\overline{N_e s}$ and $\overline{u}$. Therefore, estimates of these

parameters from previous work that modelled the DFE as a unimodal distribution might be biased. Since the true shape of the DFE is unknown, it is desirable to model the DFE without assuming a particular shape (Keightley and Eyre-Walker 2010). I took steps in this direction by developing the spikes and steps models. The spikes and steps models generally had as good or better performance than the gamma/lognormal models when fitted to unimodal or multi-modal DFEs. These results suggest that the gamma and lognormal distributions should perhaps be abandoned as models of the DFE and the spikes or steps models be used in all cases instead. Future studies should focus on improving the more realistic steps model. For example, instead of modelling each step as a uniform distribution, one could allow for more complex distributions. This could be especially useful for modelling the part of the DFE that lies in the $N_e s$ range of 0 to 1. The shape of the DFE within this range is critical for estimating $\bar{u}$.

The evidence for a bimodal DFE for nonsynonymous mutations in *M. m. castaneus* suggests that the DFE can indeed be complex, justifying the use of spikes/steps models to infer it. Future studies should use the spike/steps models to infer the DFE for a greater variety of species and investigate whether biological parameters (e.g. N_e) can predict the level of complexity of the DFE.

Finally, the simulation work in Chapter 2 revealed that the SFS contains only limited information on the DFE. Therefore, by using only SFS data we can only gain a very rough, low-resolution, picture of the DFE, irrespective of the model that is used to infer it. Future work should focus on developing methods that use more types of data to infer the DFE. For example, the pattern of sequence diversity, linkage

disequilibrium and the haplotype structure in the vicinity of a gene might be informative on the strength of negative and positive selection that acts on the gene.

The locus of adaptation: Protein-coding genes or non-coding DNA? The finding in Chapter 3 that negative selection is significantly stronger on nonsynonymous sites than non-coding DNA supports the notion that *cis*-regulatory DNA experiences less pleiotropic constraints than protein-coding genes, and is potentially more amenable to change. However, the finding that positive selection is weak on non-coding DNA and only 5-10% of non-coding substitutions are adaptive (Chapter 3) suggests that non-coding change is dominated by neutral substitutions. Interestingly, the rate of adaptive relative to neutral substitution (ω_a) is similar for nonsynonymous sites and non-coding DNA (Chapter 3), which suggests that non-coding DNA receives at least as many adaptive substitutions per base pair and per unit of time as nonsynonymous sites. Therefore, the results from Chapter 3 appear to give equal support to the two alternative hypotheses on the locus of adaptation.

A step forward in resolving the debate on the locus of adaptation might be to do a genome-wide study and analyse polymorphism data for most of the protein-coding genes and most of the DNA sequence that has *cis*-regulatory function in the genome. The answer to the debate will ultimately depend on the total input of adaptive substitutions in protein-coding genes versus non-coding DNA. Moreover, since both weakly and strongly selected mutations contribute to adaptive substitutions, it will be of interest to calculate the mean selective effect of mutations that are under positive selection for protein-coding genes and non-coding DNA. A

class of sites might have a high input of adaptive substitutions, but the effect of the mutations might be very weak, thus resulting in only small changes in fitness.

The two rules of speciation and their causes. Haldane's rule and the large effect of the X chromosome in speciation are phenomena that have been studied intensively by geneticists in order to uncover their evolutionary causes (Coyne and Orr 2004). Many hypotheses have been suggested to explain these phenomena (outlined in Chapter 1). The study of natural selection at the molecular level can allow tests of predictions of these hypotheses. The observation from Chapter 4 that X-linked genes experience a faster rate of adaptive substitution than autosomal genes in *M. m. castaneus* is consistent with the faster-X hypothesis (Charlesworth et al. 1987). The magnitude of faster-X evolution that is inferred in Chapter 4 can be explained if we assume that most adaptive substitutions come from new mutations rather than standing variation, and that advantageous mutations are partially recessive with an average dominance coefficient that is less than 0.25. However, analysis of genes that are expressed during different stages of spermatogenesis produced equivocal support to the faster-X hypothesis. In particular, autosomal and X-linked genes that are expressed in the pre-meiotic stage of spermatogenesis were found to have a similar rate of adaptive substitution. If new advantageous mutations were on average recessive, the pre-meiotically expressed genes would be expected to display the strongest evidence for faster-X evolution. Previous studies have reported that genes that are pre-meiotically expressed in spermatogenesis are female-biased in expression and therefore potentially under stronger selection in females (Zhang et al.

2010) and such genes are not expected to display faster-X evolution (Charlesworth et al. 1987). Therefore the evidence from the analysis of spermatogenesis-expressed genes neither favours nor contradicts the faster-X hypothesis.

Increasing evidence suggests that adaptation from standing variation is more common than previously thought (Karasov et al. 2010; Garud et al. 2013), which would suggest that partial dominance of new advantageous mutations might not be sufficient to explain faster-X evolution (Connallon et al. 2012) and consequently might not be capable of explaining the two rules of speciation. Therefore, alternative hypotheses should be considered to explain the results from Chapter 4. It has been suggested that intragenomic conflict over the transmission of sex chromosomes has shaped their special features and evolution (Meiklejohn and Tao 2010). Such conflict can arise from selfish genetic elements such as segregation distorters (also known as meiotic drive elements) (Presgraves 2008; Presgraves 2010). Segregation distorters are predicted to accumulate faster on sex chromosomes than autosomes (Frank 1991; Hurst and Pomiankowski 1991) and evidence for a disproportionate number of such elements on the X chromosome has been found in *Drosophila* (Jaenike 2001). Therefore, an explanation that involves preferential accumulation of segregation distorters and other selfish genetic elements on the X chromosome, which evolves rapidly due to genetic conflict, might be sufficient to explain the observed faster-X evolution. However, extensive mapping work for these elements is currently lacking (Meiklejohn and Tao 2010). Whether intragenomic conflict over the transmission of sex chromosomes can also explain the two rules of speciation is highly contested (Coyne and Orr 2004; Meiklejohn and Tao 2010; Presgraves 2010).

Future work on the subject should focus on quantifying the rate of adaptive evolution on autosomes and the X chromosome for more species. Future studies should go beyond MK-test approaches and attempt to estimate the type (hard or soft) and frequency of selective sweeps on the X chromosome. In this way, a critical assumption of the faster-X hypothesis, that is the higher frequency of hard than soft sweeps, will be more thoroughly tested. Finally, work on sex chromosome evolution should also focus on thoroughly documenting and mapping segregation distorter elements in order to assess their importance in shaping the special features of the sex chromosomes.

Bibliography

Akey JM, Eberle MA, Rieder MJ, Carlson CS, Shriver MD, Nickerson DA, Kruglyak L. 2004. Population History and Natural Selection Shape Patterns of Genetic Variation in 132 Genes. PLoS Biol 2:e286.

Andolfatto P. 2001. Adaptive hitchhiking effects on genome variability. Curr. Opin. Genet. Dev 11:635–641.

Andolfatto P. 2005. Adaptive evolution of non-coding DNA in Drosophila. Nature 437:1149–1152.

Arndt PF, Petrov DA, Hwa T. 2003. Distinct changes of genomic biases in nucleotide substitution at the time of mammalian radiation. Mol Biol Evol 20:1887–1896.

Baines JF, Harr B. 2007. Reduced X-Linked Diversity in Derived Populations of House Mice. Genetics 175:1911–1921.

Baines JF, Sawyer SA, Hartl DL, Parsch J. 2008. Effects of X-Linkage and Sex-Biased Gene Expression on the Rate of Adaptive Protein Evolution in Drosophila. Mol Biol Evol 25:1639–1650.

Barrett RDH, Hoekstra HE. 2011. Molecular spandrels: tests of adaptation at the genetic level. Nat Rev Genet 12:767–780.

Begun DJ, Holloway AK, Stevens K, et al. 2007. Population Genomics: Whole-Genome Analysis of Polymorphism and Divergence in Drosophila simulans. PLoS Biol 5:e310.

Betancourt AJ, Presgraves DC, Swanson WJ. 2002. A Test for Faster X Evolution in Drosophila. Mol Biol Evol 19:1816–1819.

Bierne N, Eyre-Walker A. 2004. The Genomic Rate of Adaptive Amino Acid Substitution in Drosophila. Mol Biol Evol 21:1350–1360.

Boyko AR, Williamson SH, Indap AR, et al. 2008. Assessing the evolutionary impact of amino acid mutations in the human genome. PLoS Genet 4:e1000083.

Bray N, Pachter L. 2004. MAVID: Constrained Ancestral Alignment of Multiple Sequences. Genome Research 14:693 –699.

Bustamante CD, Fledel-Alon A, Williamson S, et al. 2005. Natural selection on

protein-coding genes in the human genome. Nature 437:1153–1157.

Campbell P, Good JM, Nachman MW. 2013b. Meiotic Sex Chromosome Inactivation Is Disrupted in Sterile Hybrid Male House Mice. Genetics 193:819–828.

Campos JL, Zeng K, Parker DJ, Charlesworth B, Haddrill PR. 2013. Codon Usage Bias and Effective Population Sizes on the X Chromosome versus the Autosomes in Drosophila melanogaster. Mol Biol Evol 30:811–823.

Carneiro M, Albert FW, Melo-Ferreira J, Galtier N, Gayral P, Blanco-Aguiar JA, Villafuerte R, Nachman MW, Ferrand N. 2012. Evidence for widespread positive and purifying selection across the European rabbit (Oryctolagus cuniculus) genome. Mol. Biol. Evol. 29:1837–1849.

Carroll SB. 2005a. Evolution at Two Levels: On Genes and Form. PLoS Biol 3:e245.

Carroll SB. 2005b. Endless forms most beautiful: the new science of evo devo and the making of the animal kingdom. WW Norton & Company

Chamary JV, Parmley JL, Hurst LD. 2006. Hearing silence: non-neutral evolution at synonymous sites in mammals. Nat Rev Genet 7:98–108.

Charlesworth B. 1996. The good fairy godmother of evolutionary genetics. Current Biology 6:220.

Charlesworth B. 2001. The effect of life-history and mode of inheritance on neutral genetic variability. Genet. Res. 77:153–166.

Charlesworth B. 2009. Effective population size and patterns of molecular evolution and variation. Nat Rev Genet 10:195–205.

Charlesworth B, Charlesworth D. 2010. Elements of evolutionary genetics. Available from: http://www.publish.csiro.au/nid/20/pid/6182.htm

Charlesworth B, Coyne JA, Barton NH. 1987. The Relative Rates of Evolution of Sex Chromosomes and Autosomes. The American Naturalist 130:113–146.

Charlesworth B, Morgan MT, Charlesworth D. 1993. The Effect of Deleterious Mutations on Neutral Molecular Variation. Genetics 134:1289–1303.

Charlesworth D. 2006. Balancing Selection and Its Effects on Sequences in Nearby Genome Regions. PLoS Genet 2:e64.

Charlesworth J, Eyre-Walker A. 2006. The Rate of Adaptive Evolution in Enteric Bacteria. Mol Biol Evol 23:1348–1356.

Cocquet J, Ellis PJI, Mahadevaiah SK, Affara NA, Vaiman D, Burgoyne PS. 2012. A

Genetic Basis for a Postmeiotic X Versus Y Chromosome Intragenomic Conflict in the Mouse. PLoS Genet 8:e1002900.

Connallon T. 2007. Adaptive Protein Evolution of X-linked and Autosomal Genes in Drosophila: Implications for Faster-X Hypotheses. Mol Biol Evol 24:2566–2572.

Connallon T, Singh ND, Clark AG. 2012. Impact of Genetic Architecture on the Relative Rates of X versus Autosomal Adaptive Substitution. Mol Biol Evol 29:1933–1942.

Counterman BA, Ortíz-Barrientos D, Noor MAF. 2004. Using comparative genomic data to test for fast-X evolution. Evolution 58:656–660.

Coyne JA. 1992. Genetics and speciation. Nature 355:511–515.

Coyne JA, Orr HA. 1989. Two rules of speciation. Speciation and its Consequences:180–207.

Coyne JA, Orr HA. 2004. Speciation. Sinauer Associates Sunderland, MA

Crow JF. 1997. The high spontaneous mutation rate: Is it a health risk? PNAS 94:8380–8386.

Davies EK, Peters AD, Keightley PD. 1999. High Frequency of Cryptic Deleterious Mutations in Caenorhabditis elegans. Science 285:1748–1751.

Dobzhansky TG. 1937. Genetics and the Origin of Species. Columbia University Press

Ellegren H, Parsch J. 2007. The evolution of sex-biased genes and sex-biased gene expression. Nat Rev Genet 8:689–698.

Eory L, Halligan DL, Keightley PD. 2010. Distributions of selectively constrained sites and deleterious mutation rates in the hominid and murid genomes. Mol. Biol. Evol 27:177–192.

Ewens WJ. 1979. Mathematical population genetics. Springer-Verlag.

Eyre-Walker A. 2002. Changing Effective Population Size and the McDonald-Kreitman Test. Genetics 162:2017–2024.

Eyre-Walker A. 2006. The genomic rate of adaptive evolution. Trends Ecol. Evol. (Amst.) 21:569–575.

Eyre-Walker AC. 1991. An analysis of codon usage in mammals: selection or mutation bias? J. Mol. Evol. 33:442–449.

Eyre-Walker A, Keightley PD. 1999. High genomic deleterious mutation rates in hominids. Nature 397:344–347.

Eyre-Walker A, Keightley PD. 2007. The distribution of fitness effects of new mutations. Nat. Rev. Genet 8:610–618.

Eyre-Walker A, Keightley PD. 2009. Estimating the Rate of Adaptive Molecular Evolution in the Presence of Slightly Deleterious Mutations and Population Size Change. Mol Biol Evol 26:2097–2108.

Eyre-Walker A, Keightley PD, Smith NGC, Gaffney D. 2002. Quantifying the slightly deleterious mutation model of molecular evolution. Mol Biol Evol 19:2142 –2149.

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

Fay JC. 2011. Weighing the evidence for adaptation at the molecular level. Trends in Genetics 27:343–349.

Fay JC, Wu C-I. 2001. The neutral theory in the genomic era. Current Opinion in Genetics & Development 11:642–646.

Fay JC, Wyckoff GJ, Wu C-I. 2001. Positive and Negative Selection on the Human Genome. Genetics 158:1227–1234.

Fay JC, Wyckoff GJ, Wu C-I. 2002. Testing the neutral theory of molecular evolution with genomic data from Drosophila. Nature 415:1024–1026.

Fisher RA. 1930. The genetical theory of natural selection. Oxford, England: Clarendon Press

Foxe JP, Dar V-N, Zheng H, Nordborg M, Gaut BS, Wright SI. 2008. Selection on Amino Acid Substitutions in Arabidopsis. Mol Biol Evol 25:1375–1383.

Frank SA. 1991. Divergence of Meiotic Drive-Suppression Systems as an Explanation for Sex- Biased Hybrid Sterility and Inviability. Evolution 45:262–267.

Gaffney DJ, Keightley PD. 2006. Genomic selective constraints in murid noncoding DNA. PLoS Genet 2:e204.

Gaffney DJ, Keightley PD. 2008. Effect of the assignment of ancestral CpG state on the estimation of nucleotide substitution rates in mammals. BMC Evol. Biol 8:265.

Garud NR, Messer PW, Buzbas EO, Petrov DA. 2013. Soft selective sweeps are the

primary mode of recent adaptation in Drosophila melanogaster. arXiv:1303.0906 [Internet]. Available from: http://arxiv.org/abs/1303.0906

Gayral P, Melo-Ferreira J, Glémin S, et al. 2013. Reference-Free Population Genomics from Next-Generation Transcriptome Data and the Vertebrate–Invertebrate Gap. PLoS Genet 9:e1003457.

Gillespie JH. 1991. The causes of molecular evolution. Oxford University Press Available from: http://books.google.co.uk/books?hl=en&lr=&id=257cnXAoREwC&oi=fnd&pg=PR7&dq=gillespie+1991&ots=KIq382X-YK&sig=Pn_rUqEjNFtLyBtiIIoGUhuDvw0

Gillespie JH. 1999. The Role of Population Size in Molecular Evolution. Theoretical Population Biology 55:145–156.

Gillespie JH. 2000. Genetic Drift in an Infinite Population: The Pseudohitchhiking Model. Genetics 155:909–919.

Good JM, Dean MD, Nachman MW. 2008. A Complex Genetic Basis to X-Linked Hybrid Male Sterility Between Two Species of House Mice. Genetics 179:2213–2228.

Good JM, Nachman MW. 2005. Rates of Protein Evolution Are Positively Correlated with Developmental Timing of Expression During Mouse Spermatogenesis. Mol Biol Evol 22:1044–1052.

Gossmann TI, Keightley PD, Eyre-Walker A. 2012. The Effect of Variation in the Effective Population Size on the Rate of Adaptive Molecular Evolution in Eukaryotes. Genome Biol Evol 4:658–667.

Gossmann TI, Song B-H, Windsor AJ, Mitchell-Olds T, Dixon CJ, Kapralov MV, Filatov DA, Eyre-Walker A. 2010. Genome wide analyses reveal little evidence for adaptive evolution in many plantspecies. Mol Biol Evol 27:1822–1832.

Graur D, Li W-H. 2000. Fundamentals of molecular evolution. Sinauer Associates Sunderland, MA Available from: http://www.lavoisier.fr/livre/notice.asp?ouvrage=1727647

Gray NK, Wickens M. 1998. Control of Translation Initiation in Animals. Annu. Rev. Cell. Dev. Biol. 14:399–458.

Guioli S, Lovell-Badge R, Turner JMA. 2012. Error-Prone ZW Pairing and No Evidence for Meiotic Sex Chromosome Inactivation in the Chicken Germ Line. PLoS Genet 8:e1002560.

Haag-Liautard C, Dorris M, Maside X, Macaskill S, Halligan DL, Houle D, Charlesworth B, Keightley PD. 2007. Direct estimation of per nucleotide and genomic deleterious mutation rates in Drosophila. Nature 445:82–85.

Haddrill PR, Bachtrog D, Andolfatto P. 2008. Positive and Negative Selection on Noncoding DNA in Drosophila simulans. Mol Biol Evol 25:1825–1834.

Hahn MW. 2008. Toward a selection theory of molecular evolution. Evolution 62:255–265.

Haldane JBS. 1922. Sex ratio and unisexual sterility in hybrid animals. Journ. of Gen. 12:101–109.

Haldane JBS. 1927. A mathematical theory of natural and artificial selection, part V: selection and mutation. In: Mathematical Proceedings of the Cambridge Philosophical Society. Vol. 23. p. 838–844. Available from: http://journals.cambridge.org/abstract_S0305004100015644

Haldane JBS. 1957. The cost of natural selection. Journal of Genetics 55:511–524.

Halligan DL, Keightley PD. 2009. Spontaneous mutation accumulation studies in evolutionary genetics. Annual Review of Ecology, Evolution, and Systematics 40:151–172.

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

Haygood R, Fedrigo O, Hanson B, Yokoyama K-D, Wray GA. 2007. Promoter regions of many neural- and nutrition-related genes have experienced positive selection during human evolution. Nat Genet 39:1140–1144.

Hense W, Baines JF, Parsch J. 2007. X Chromosome Inactivation during Drosophila Spermatogenesis. PLoS Biol 5:e273.

Hernandez RD, Williamson SH, Bustamante CD. 2007. Context Dependence, Ancestral Misidentification, and Spurious Signatures of Natural Selection. Mol Biol Evol 24:1792–1800.

Hershberg R, Petrov DA. 2008. Selection on Codon Bias. Annual Review of Genetics 42:287–299.

Hey J. 1999. The neutralist, the fly and the selectionist. Trends in ecology & evolution 14:35–38.

Hill WG. 1982. Rates of change in quantitative traits from fixation of new mutations. PNAS 79:142–145.

Hill WG, Robertson A. 1966. The Effect of Linkage on Limits to Artificial Selection. Genetics Research 8:269–294.

Hoekstra HE, Coyne JA. 2007. The locus of evolution: evo devo and the genetics of adaptation. Evolution 61:995–1016.

Hollocher H, Wu C-I. 1996. The Genetics of Reproductive Isolation in the Drosophila simulans Clade: X vs. Autosomal Effects and Male vs. Female Effects. Genetics 143:1243–1255.

Hurst LD, Pomiankowski A. 1991. Causes of sex ratio bias may account for unisexual sterility in hybrids: a new explanation of Haldane's rule and related phenomena. Genetics 128:841–858.

Hutter S, Li H, Beisswanger S, Lorenzo DD, Stephan W. 2007. Distinctly Different Sex Ratios in African and European Populations of Drosophila melanogaster Inferred From Chromosomewide Single Nucleotide Polymorphism Data. Genetics 177:469–480.

Huxley J. 1942. Evolution. The Modern Synthesis. Evolution. The Modern Synthesis. [Internet]. Available from: http://www.cabdirect.org/abstracts/19432202794.html

Hvilsom C, Qian Y, Bataillon T, et al. 2012. Extensive X-linked adaptive evolution in central chimpanzees. PNAS 109:2054–2059.

Jacob F. 1977. Evolution and tinkering. Science 196:1161–1166.

Jacob F, Monod J. 1961. Genetic regulatory mechanisms in the synthesis of proteins. Journal of molecular biology 3:318–356.

Jaenike J. 2001. Sex Chromosome Meiotic Drive. Annual Review of Ecology and Systematics 32:25–49.

Jukes T, Cantor C. 1969. {Evolution of protein molecules}. In: Munro M, editor. Mammalian protein metabolism. Vol. III. Academic Press. p. 21–132.

Karasov T, Messer PW, Petrov DA. 2010. Evidence that Adaptation in Drosophila Is Not Limited by Mutation at Single Sites. PLoS Genet 6:e1000924.

Katoh K, Misawa K, Kuma K, Miyata T. 2002. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucl. Acids Res. 30:3059–3066.

Keightley PD. 2012. Rates and fitness consequences of new mutations in humans. Genetics 190:295–304.

Keightley PD, Eyre-Walker A. 2007. Joint inference of the distribution of fitness effects of deleterious mutations and population demography based on nucleotide polymorphism frequencies. Genetics 177:2251–2261.

Keightley PD, Eyre-Walker A. 2010. What can we learn about the distribution of fitness effects of new mutations from DNA sequence data? Philosophical Transactions of the Royal Society B: Biological Sciences 365:1187–1193.

Keightley PD, Gaffney DJ. 2003. Functional constraints and frequency of deleterious mutations in noncoding DNA of rodents. Proceedings of the National Academy of Sciences of the United States of America 100:13402 –13406.

Keightley PD, Kryukov G, Sunyaev S, Halligan DL, Gaffney DJ. 2005. Evolutionary constraints in conserved nongenic sequences of mammals. Genome Res 15:1373–1378.

Keightley PD, Lercher MJ, Eyre-Walker A. 2005. Evidence for Widespread Degradation of Gene Control Regions in Hominid Genomes. PLoS Biol 3:e42.

Keightley P, Eyre-Walker A. 2012. Estimating the Rate of Adaptive Molecular Evolution When the Evolutionary Divergence Between Species is Small. Journal of Molecular Evolution 74:61–68.

Khaitovich P, Hellmann I, Enard W, Nowick K, Leinweber M, Franz H, Weiss G, Lachmann M, Pääbo S. 2005. Parallel Patterns of Evolution in the Genomes and Transcriptomes of Humans and Chimpanzees. Science 309:1850–1854.

Khil PP, Smirnova NA, Romanienko PJ, Camerini-Otero RD. 2004. The mouse X chromosome is enriched for sex-biased genes not subject to selection by meiotic sex chromosome inactivation. Nat Genet 36:642–646.

Kimura M. 1957. Some problems of stochastic processes in genetics. The Annals of Mathematical Statistics:882–901.

Kimura M. 1962. On the probability of fixation of mutant genes in a population. Genetics 47:713–719.

Kimura M. 1968. Evolutionary Rate at the Molecular Level. Nature 217:624–626.

Kimura M. 1980. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. Journal of Molecular Evolution 16:111–120.

Kimura M. 1985. The neutral theory of molecular evolution. Cambridge Univ Pr

Kimura M, Ohta T. 1971. Protein polymorphism as a phase of molecular evolution. Nature 229:467–469.

King JL, Jukes TH. 1969. Non-Darwinian evolution. Science 164:788–798.

King MC, Wilson AC. 1975. Evolution at two levels in humans and chimpanzees. Science 188:107–116.

Kohn MH, Fang S, Wu C-I. 2004. Inference of Positive and Negative Selection on the 5' Regulatory Regions of Drosophila Genes. Mol Biol Evol 21:374–383.

Kondrashov AS. 1988. Deleterious mutations and the evolution of sexual reproduction. Nature 336:435–440.

Kondrashov AS, Crow JF. 1991. Haploidy or diploidy: which is better? Nature 351:314–315.

Kondrashov AS, Crow JF. 1993. A molecular approach to estimating the human deleterious mutation rate. Hum. Mutat. 2:229–234.

Kousathanas A, Keightley Peter D. 2013. A comparison of models to infer the distribution of fitness effects of new mutations. Genetics 193:1197–1208.

Kousathanas A, Oliver F, Halligan DL, Keightley PD. 2011. Positive and negative selection on noncoding DNA close to protein-coding genes in wild house mice. Mol Biol Evol 28:1183 –1191.

Kreitman M. 1996. The neutral theory is dead. Long live the neutral theory. BioEssays 18:678–683.

Kulathinal RJ, Stevison LS, Noor MAF. 2009. The Genomics of Speciation in Drosophila: Diversity, Divergence, and Introgression Estimated Using Low-Coverage Genome Sequencing. PLoS Genet 5:e1000550.

Lande R. 1994. Risk of Population Extinction from Fixation of New Deleterious Mutations. Evolution 48:1460–1469.

Lawrie DS, Messer PW, Hershberg R, Petrov DA. 2013. Strong Purifying Selection at Synonymous Sites in D. melanogaster. PLoS Genet 9:e1003527.

Lewontin RC. 1974. The genetic basis of evolutionary change. Columbia University Press New York Available from: http://dannyreviews.com/h/The_Genetic_Basis_of_Evolutionary_Change.html

Liao B-Y, Scott NM, Zhang J. 2006. Impacts of Gene Essentiality, Expression Pattern, and Gene Compactness on the Evolutionary Rate of Mammalian

Proteins. Mol Biol Evol 23:2072–2080.

Lifschytz E, Lindsley DL. 1972. The Role of X-Chromosome Inactivation during Spermatogenesis. PNAS 69:182–186.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R. 2009. The Sequence Alignment/Map format and SAMtools. Bioinformatics 25:2078–2079.

Livingston RJ, von Niederhausern A, Jegga AG, et al. 2004. Pattern of Sequence Variation Across 213 Environmental Response Genes. Genome Research 14:1821–1831.

Li WH. 1993. Unbiased estimation of the rates of synonymous and nonsynonymous substitution. Journal of Molecular Evolution 36:96–99.

Loewe L, Charlesworth B. 2006. Inferring the distribution of mutational effects on fitness in Drosophila. Biol. Lett. 2:426–430.

Loewe L, Charlesworth B, Bartolomé C, Nöel V. 2006. Estimating selection on nonsynonymous mutations. Genetics 172:1079–1092.

Lu J, Wu C-I. 2005. Weak selection revealed by the whole-genome comparison of the X chromosome and autosomes of human and chimpanzee. PNAS 102:4063–4067.

Mackay TFC, Richards S, Stone EA, et al. 2012. The Drosophila melanogaster Genetic Reference Panel. Nature 482:173–178.

Mank JE, Axelsson E, Ellegren H. 2007. Fast-X on the Z: Rapid evolution of sex-linked genes in birds. Genome Res. 17:618–624.

Mank JE, Vicoso B, Berlin S, Charlesworth B. 2010. Effective Population Size and the Faster-x Effect: Empirical Results and Their Interpretation. Evolution 64:663–674.

Masly JP, Presgraves DC. 2007. High-Resolution Genome-Wide Dissection of the Two Rules of Speciation in Drosophila. PLoS Biol 5:e243.

Mayr E. 1963. Animal species and evolution. Animal species and their evolution. [Internet]. Available from: http://www.cabdirect.org/abstracts/19640100703.html

McDonald JH, Kreitman M. 1991. Adaptive protein evolution at the Adh locus in Drosophila. Nature 351:652–654.

Meiklejohn CD, Tao Y. 2010. Genetic conflict and sex chromosome evolution.

Trends in Ecology & Evolution 25:215–223.

Meisel RP, Malone JH, Clark AG. 2012a. Disentangling the relationship between sex-biased gene expression and X-linkage. Genome Res. 22:1255–1265.

Meisel RP, Malone JH, Clark AG. 2012b. Faster-X evolution of gene expression in Drosophila. PLoS Genet. 8:e1003013.

Messer PW, Petrov DA. 2013. Frequent adaptation and the McDonald–Kreitman test. PNAS 110:8615–8620.

Mikhaylova LM, Nurminsky DI. 2012. No severe and global X chromosome inactivation in meiotic male germline of Drosophila. BMC Biology 10:50.

Muller HJ. 1942. Isolating mechanisms, evolution and temperature. In: Biol. Symp. Vol. 6. p. 71–125.

Musters H, Huntley MA, Singh RS. 2006. A Genomic Comparison of Faster-Sex, Faster-X, and Faster-Male Evolution Between Drosophila melanogaster and Drosophila pseudoobscura. Journal of Molecular Evolution 62:693–700.

Myers S, Fefferman C, Patterson N. 2008. Can one learn history from the allelic spectrum? Theoretical Population Biology 73:342–348.

Namekawa SH, Lee JT. 2009. XY and ZW: Is Meiotic Sex Chromosome Inactivation the Rule in Evolution? PLoS Genet 5:e1000493.

Namekawa SH, Park PJ, Zhang L-F, Shima JE, McCarrey JR, Griswold MD, Lee JT. 2006. Postmeiotic Sex Chromatin in the Male Germline of Mice. Current Biology 16:660–667.

Nei M. 1987. Molecular evolutionary genetics. Columbia University Press

Nei M, Li WH. 1979. Mathematical model for studying genetic variation in terms of restriction endonucleases. Proc Natl Acad Sci U S A 76:5269–5273.

Nei M, Suzuki Y, Nozawa M. 2010. The Neutral Theory of Molecular Evolution in the Genomic Era. Annual Review of Genomics and Human Genetics 11.

Nelder JA, Mead R. 1965. A Simplex method for function minimization. The Computer Journal 7:308–313.

Nielsen R. 2005. Molecular Signatures of Natural Selection. Annu. Rev. Genet. 39:197–218.

Nielsen R, Yang Z. 2003. Estimating the distribution of selection coefficients from phylogenetic data with applications to mitochondrial and viral DNA. Mol

Biol Evol 20:1231–1239.

Ohta T. 1973. Slightly Deleterious Mutant Substitutions in Evolution. Nature 246:96–98.

Ohta T. 1992. The nearly neutral theory of molecular evolution. Annual Review of Ecology and Systematics 23:263–286.

Oka A, Aoto T, Totsuka Y, et al. 2007. Disruption of Genetic Interaction Between Two Autosomal Regions and the X Chromosome Causes Reproductive Isolation Between Mouse Strains Derived From Different Subspecies. Genetics 175:185–197.

Oka A, Mita A, Sakurai-Yamatani N, Yamamoto H, Takagi N, Takano-Shimizu T, Toshimori K, Moriwaki K, Shiroishi T. 2004. Hybrid Breakdown Caused by Substitution of the X Chromosome Between Two Mouse Subspecies. Genetics 166:913–924.

Pamilo P, Bianchi N. 1993. Evolution of the Zfx and Zfy genes: rates and interdependence between the genes. Mol Biol Evol 10:271–281.

Parisi M, Nuttall R, Naiman D, Bouffard G, Malley J, Andrews J, Eastman S, Oliver B. 2003. Paucity of Genes on the Drosophila X Chromosome Showing Male-Biased Expression. Science 299:697–700.

Payseur BA, Krenz JG, Nachman MW. 2004. Differential patterns of introgression across the X chromosome in a hybrid zone between two species of house mice. Evolution 58:2064–2078.

Phifer-Rixey M, Bonhomme F, Boursot P, Churchill GA, Piálek J, Tucker PK, Nachman MW. 2012. Adaptive Evolution and Effective Population Size in Wild House Mice. Mol Biol Evol 29:2949–2955.

Piertney SB, Oliver MK. 2005. The evolutionary ecology of the major histocompatibility complex. Heredity 96:7–21.

Piganeau G, Eyre-Walker A. 2003. Estimating the distribution of fitness effects from DNA sequence data: implications for the molecular clock. PNAS 100:10335–10340.

Pool JE, Corbett-Detig RB, Sugino RP, et al. 2012. Population Genomics of Sub-Saharan Drosophila melanogaster: African Diversity and Non-African Admixture. PLoS Genet 8:e1003080.

Pool JE, Nielsen R. 2007. Population size changes reshape genomic patterns of diversity. Evolution 61:3001–3006.

Presgraves DC. 2008. Sex chromosomes and speciation in Drosophila. Trends in Genetics 24:336–343.

Presgraves DC. 2010. The molecular evolutionary basis of species formation. Nat Rev Genet 11:175–180.

Pritchard JK, Stephens M, Donnelly P. 2000. Inference of Population Structure Using Multilocus Genotype Data. Genetics 155:945–959.

Rice WR. 1984. Sex Chromosomes and the Evolution of Sexual Dimorphism. Evolution 38:735–742.

Richmond RC. 1970. Non-Darwinian Evolution: A Critique. Nature 225:1025–1028.

Sawyer S, Hartl DL. 1992. Population genetics of polymorphism and divergence. Genetics 132:1161–1176.

Sawyer S, Kulathinal R, Bustamante C, Hartl D. 2003. Bayesian analysis suggests that most aminoacid replacements in Drosophila are driven by positive selection. Journal of Molecular Evolution 57:S154–S164.

Schneider A, Charlesworth B, Eyre-Walker A, Keightley PD. 2011. A Method for inferring the rate of occurrence and fitness effects of advantageous mutations. Genetics 189:1427–1437.

Schoenmakers S, Wassenaar E, Hoogerbrugge JW, Laven JSE, Grootegoed JA, Baarends WM. 2009. Female Meiotic Sex Chromosome Inactivation in Chicken. PLoS Genet 5:e1000466.

Sella G, Petrov DA, Przeworski M, Andolfatto P. 2009. Pervasive natural selection in the Drosophila genome? PLoS Genet 5:e1000495.

Shabalina SA, Spiridonov NA. 2004. The mammalian transcriptome and the function of non-coding DNA sequences. Genome biology 5.

Shapiro JA, Huang W, Zhang C, et al. 2007. Adaptive genic evolution in the Drosophila genomes. Proceedings of the National Academy of Sciences 104:2271–2276.

Simpson GG. 1964. Organisms and molecules in evolution. Science 146:1535–1538.

Slotte T, Foxe JP, Hazzouri KM, Wright SI. 2010. Genome-wide evidence for efficient positive and purifying selection in Capsella grandiflora, a plant species with a large effective population size. Mol Biol Evol 27:1813 –1821.

Smith JM. 1968. “ Haldane’s dilemma” and the rate of evolution. Nature 219:1114.

Smith JM, Haigh J. 1974. The Hitch-Hiking Effect of a Favourable Gene. Genetics Research 23:23–35.

Smith NGC, Hurst LD. 1999. The Causes of Synonymous Rate Variation in the Rodent Genome: Can Substitution Rates Be Used to Estimate the Sex Bias in Mutation Rate? Genetics 152:661–673.

Smith NG., Eyre-Walker A. 2002. Adaptive protein evolution in Drosophila. Nature 415:1022–1024.

Stern DL, Orgogozo V. 2008. The Loci of Evolution: How Predictable is Genetic Evolution? Evolution 62:2155–2177.

Storchová R, Gregorová S, Buckiová D, Kyselová V, Divina P, Forejt J. 2004. Genetic analysis of X-linked hybrid sterility in the house mouse. Mamm. Genome 15:515–524.

Su AI, Wiltshire T, Batalov S, et al. 2004. A gene atlas of the mouse and human protein-encoding transcriptomes. PNAS 101:6062–6067.

Swanson WJ, Vacquier VD. 2002. The rapid evolution of reproductive proteins. Nat Rev Genet 3:137–144.

Tajima F. 1983. Evolutionary Relationship of DNA Sequences in Finite Populations. Genetics 105:437–460.

Tajima F. 1989. Statistical Method for Testing the Neutral Mutation Hypothesis by DNA Polymorphism. Genetics 123:585–595.

Tanaka T, Nei M. 1989. Positive darwinian selection observed at the variable-region genes of immunoglobulins. Mol Biol Evol 6:447–459.

Tao Y, Chen S, Hartl DL, Laurie CC. 2003. Genetic Dissection of Hybrid Incompatibilities Between Drosophila simulans and D. mauritiana. I. Differential Accumulation of Hybrid Male Sterility Effects on the X and Autosomes. Genetics 164:1383–1398.

Teeter KC, Payseur BA, Harris LW, et al. 2008. Genome-wide patterns of gene flow across a house mouse hybrid zone. Genome Res. 18:67–76.

The Chimpanzee Sequencing and Analysis Consortium. 2005. Initial sequence of the chimpanzee genome and comparison with the human genome. Nature 437:69–87.

Thornton K, Bachtrog D, Andolfatto P. 2006. X chromosomes and autosomes evolve at similar rates in Drosophila: No evidence for faster-X protein evolution.

Genome Research 16:498 –504.

Torgerson DG, Boyko AR, Hernandez RD, et al. 2009. Evolutionary Processes Acting on Candidate cis-Regulatory Regions in Humans Inferred from Patterns of Polymorphism and Divergence. PLoS Genet 5:e1000592.

Torgerson DG, Kulathinal RJ, Singh RS. 2002. Mammalian Sperm Proteins Are Rapidly Evolving: Evidence of Positive Selection in Functionally Diverse Genes. Mol Biol Evol 19:1973–1980.

Torgerson DG, Singh RS. 2006. Enhanced adaptive evolution of sperm-expressed genes on the mammalian X chromosome. Heredity 96:39–44.

True JR, Weir BS, Laurie CC. 1996. A Genome-Wide Survey of Hybrid Incompatibility Factors by the Introgression of Marked Segments of Drosophila mauritiana Chromosomes into Drosophila simulans. Genetics 142:819–837.

Tucker PK, Sage RD, Warner J, Wilson AC, Eicher EM. 1992. Abrupt Cline for Sex Chromosomes in a Hybrid Zone between Two Species of Mice. Evolution 46:1146–1163.

Turelli M, Orr HA. 1995. The dominance theory of Haldane's rule. Genetics 140:389–402.

Turner JMA. 2007. Meiotic sex chromosome inactivation. Development 134:1823–1831.

Urrutia AO, Hurst LD. 2001. Codon Usage Bias Covaries With Expression Breadth and the Rate of Synonymous Evolution in Humans, but This Is Not Evidence for Selection. Genetics 159:1191–1199.

Veyrieras J-B, Kudaravalli S, Kim SY, Dermitzakis ET, Gilad Y, Stephens M, Pritchard JK. 2008. High-Resolution Mapping of Expression-QTLs Yields Insight into Human Gene Regulation. PLoS Genet 4:e1000214.

Vicoso B, Charlesworth B. 2006. Evolution on the X chromosome: unusual patterns and processes. Nat Rev Genet 7:645–653.

Vicoso B, Charlesworth B. 2009. Effective Population Size and the Faster-x Effect: An Extended Model. Evolution 63:2413–2426.

Wall JD, Andolfatto P, Przeworski M. 2002. Testing models of selection and demography in Drosophila simulans. Genetics 162:203.

Watterson GA. 1975. On the number of segregating sites in genetical models without

recombination. Theoretical Population Biology 7:256–276.

Welch JJ. 2006. Estimating the Genomewide Rate of Adaptive Protein Evolution in Drosophila. Genetics 173:821–837.

Williamson SH, Hernandez R, Fledel-Alon A, Zhu L, Nielsen R, Bustamante CD. 2005. Simultaneous inference of selection and population growth from patterns of variation in the human genome. PNAS 102:7882–7887.

Wilson DJ, Hernandez RD, Andolfatto P, Przeworski M. 2011. A Population genetics-phylogenetics approach to inferring natural selection in coding sequences. PLoS Genet 7:e1002395.

Wright S. 1931. Evolution in Mendelian populations. Genetics 16:97.

Wu C-I, Johnson NA, Palopoli MF. 1996. Haldane's rule and its legacy: Why are there so many sterile males? Trends in Ecology & Evolution 11:281–284.

Xie X, Lu J, Kulbokas EJ, Golub TR, Mootha V, Lindblad-Toh K, Lander ES, Kellis M. 2005. Systematic discovery of regulatory motifs in human promoters and 3[prime] UTRs by comparison of several mammals. Nature 434:338–345.

Yang Z, Nielsen R. 2008. Mutation-Selection Models of Codon Substitution and Their Use to Estimate Selective Strengths on Codon Usage. Mol Biol Evol 25:568–579.

Yu N, Jensen-Seaman MI, Chemnick L, Ryder O, Li W-H. 2004. Nucleotide Diversity in Gorillas. Genetics 166:1375–1383.

Zhang L, Li W-H. 2005. Human SNPs Reveal No Evidence of Frequent Positive Selection. Mol Biol Evol 22:2504–2507.

Zhang YE, Vibranovski MD, Landback P, Marais GAB, Long M. 2010. Chromosomal Redistribution of Male-Biased Genes in Mammalian Evolution with Two Bursts of Gene Gain on the X Chromosome. PLoS Biol 8:e1000494.

Zuckerkandl E, Pauling L. 1965. Evolutionary divergence and convergence in proteins. Evolving genes and proteins 97:97–166.

Appendix

The appendix comprises of two parts (A and B).

In part A, I give supplementary information for Chapters 2, 3 and 4.

In part B, I provide reprints of the published papers that have arisen from this thesis.

Appendix A

A.2 Supplementary material for Chapter 2

A.2.1 Table. The median estimates and the 5th and 95th percentiles for the parameters of each of the tested models for each simulation set over 100 replicates (SIM1, SIM2, SIM3). The inferred parameters for the lognormal and beta model are given unscaled by N_e=100.

Model	Parameter	SIM1	SIM2	SIM3
Lognormal	μ	-3.3 [3.2, 3.3]	-0.10 [2.1X10^{-3}, 0.2]	-3.1 [3.6, 3.1]
	σ	1.8 [1.7, 1.9]	3.7 [3.6, 3.85]	1.6 [1.5, 1.6]
Gamma	a	0.051 [0.046, 0.056]	0.0012 [0.0010, 0.0013]	0.069 [0.0062, 0.0076]
	b	0.51 [0.49, 0.53]	0.33 [0.32, 0.34]	0.66 [0.63, 0.69]
Beta	k_1	0.49 [0.46, 0.51]	0.20 [0.19, 0.22]	0.63 [0.60, 0.67]
	k_2	4.7 [4.3, 5.2]	0.10 [0.09, 0.11]	6.3 [5.7, 7.0]
Best spike	$N_e s_1$	1.2 [1.1, 1.3]	0.49 [0, -0.8]	4.9X10^{-8} [1.5X10^{-12}, 1.1]
	$N_e s_2$	16 [14, 18]	13 [2.9, 47]	5.3 [4.5, 8.5]
	$N_e s_3$	-	92 [76, 710]	58 [34, 496]
	p_1	0.52 [0.49, 0.55]	0.18 [0.13, 0.2]	0.21 [0.19, 0.41]
	p_2	-	0.13 [0.05, 0.58]	0.57 [0.45, 0.64]
Best step	$N_e s_1$	2.2 [1.8, 2.6]	0.86 [0.6, 1.2]	1.4X10^{-3} [1.7X10^{-11}, 1.8]
	$N_e s_2$	29 [24.8, 33.6]	189 [178, 202]	10 [7.6, 15]
	$N_e s_3$	-	-	157 [58, 6X10^{6}]
	p_1	0.47	0.17	0.11

Model	Parameter	SIM1	SIM2	SIM3
		[0.41, 0.52]	[0.15, 0.18]	[0.090, 0.28]
	p_2	-	-	0.73 [0.58, 0.75]
6-fixed spikes	p_1	0.14 [0.04, 0.2]	0.10 [0.02, 0.14]	0.13 [0.011, 0.20]
	p_2	0.20 [0.11, 0.38]	0.070 [0, 0.19]	0.12 [$7.1X10^{-11}$, 0.34]
	p_3	0.27 [0.09, 0.38]	$2.2X10^{-8}$ [0, 0.11]	0.44 [0.22, 0.61]
	p_4	0.23 [0.12, 0.38]	0.070 [0, 0.16]	0.13 [$1.9X10^{-9}$, 0.28]
	p_5	0.14 [0.01, 0.18]	0.13 [0, 0.34]	0.14 [0, 0.2]

A.3 Supplementary material for Chapter 3

A.3.1 Table. Genes analyzed in Chapter 3 and details on bp that were successfully sequenced for exonic, intronic, upstream and downstream site classes.

Code	Gene name	Exonic	Intronic	5p (UTR)	3p (UTR)	5p (intergenic)	3p (intergenic)	total 5p	total 3p
101	Gucy1a3	646	-	-	-	-	-	-	-
102	Pla2g4a	545	1477	687	438	-	163	687	601
103	Umps	779	721	22	647	553	-	575	647
104	Egfr	456	896	280	-	331	685	611	685
106	Orc3l	432	1165	33	593	390	-	423	593
107	Cela1	499	1625	27	-	532	-	559	-
108	Cyp2c39	403	483	49	335	415	229	464	564
109	Nf1	756	663	-	573	541	-	541	573
110	Xpc	886	299	89	517	556	-	645	517
111	Angptl7	698	795	173	573	422	-	595	573
112	Prkd1	689	768	-	550	-	-	-	550
113	mmp21	767	774	7	136	620	283	627	419
114	mad2l2	338	780	305	283	273	356	578	639
115	prdx1	471	1610	545	468	-	-	545	468
116	bace1	758	705	499	624	75	-	574	624
117	Cyp2b10	512	825	21	375	538	212	559	587
118	cdc20	724	792	222	62	203	497	425	559
119	fgfr2	504	1224	394	580	-	-	394	580
120	Mmp15	835	696	532	493	71	-	603	493
121	Cyp17a1	679	793	157	160	324	404	481	564
122	Adm	568	704	609	535	19	-	628	535
124	Tert	829	1066	29	28	547	457	576	485
125	Cyp4b1	540	787	18	-	619	-	637	-
126	Rrm2b	429	470	83	1451	279	-	362	1451
127	prdx2	513	852	400	228	-	245	400	473
130	mt3	109	10	-	28	-	-	-	28
131	Mms19	668	833	23	505	448	-	471	505

132	mpo	1111	1479	423	460	-	-	423	460
133	Epx	699	1130	122	610	373	-	495	610
134	tgfb1i1	713	281	-	347	-	235	-	582
135	mapk11	452	772	-	370	-	-	-	370
137	Cyp1a1	981	781	537	525	-	-	537	525
139	stk25	758	843	400	393	166	-	566	393
140	csk	779	1186	554	539	-	-	554	539
141	por	440	821	312	368	-	155	312	523
143	mmp16	557	1004	104	344	196	-	300	344
144	rev1	857	1408	451	301	-	307	451	608
145	adh1	590	664	-	132	382	417	382	549
147	adh4	492	105	-	-	-	-	-	-
148	adh5	500	549	-	386	-	167	-	553
149	itgal	803	624	121	191	406	285	527	476
150	tpo	681	1370	501	464	-	85	501	549
151	Capn3	479	412	73	414	586	169	659	583
152	ddb1	709	601	81	570	564	-	645	570
153	pdlim1	476	25	314	346	526	278	840	624
154	tjp1	1013	345	408	684	263	-	671	684
155	ppib	135	323	97	-	579	-	676	-
157	cdc42bpb	710	470	-	501	-	-	-	501
158	aoc2	588	393	30	-	465	-	495	-
159	aoc3	972	321	334	-	295	-	629	-
160	ckb	969	639	380	207	188	272	568	479
162	Nos2	614	878	411	76	-	400	411	476
163	fmo3	702	70	-	-	-	-	-	-
164	tdp1	592	551	18	180	-	25	18	205
165	gab1	661	479	464	449	-	-	464	449
166	abp1	1212	812	400	-	-	-	400	-
167	poln	719	45	-	231	-	116	-	347
168	fgf21	585	564	-	-	-	-	-	-
170	blm	1255	454	-	-	-	356	-	356
174	hspa5	595	-	-	239	-	-	-	239
175	msr1	360	-	-	-	-	-	-	-
176	pax3	469	20	-	-	-	-	-	-
177	abl1	221	-	-	-	-	-	-	-

178	recql4	262	42	-	-	-	-	-	-
180	casp8	621	-	-	-	-	-	-	-
181	osr1	396	883	434	-	-	-	434	-
182	rxra	443	18	-	-	-	-	-	-
189	dclre1b	512	-	-	-	-	-	-	-
190	cdc37	279	31	-	-	-	-	-	-
191	neil3	566	10	-	-	-	-	-	-
192	foxm1	384	-	-	-	-	-	-	-
193	fmo5	291	-	-	-	-	-	-	-
194	scara3	435	-	-	-	-	-	-	-
196	sprr3	499	-	-	-	-	-	-	-
197	sepp1	836	456	385	315	-	-	385	315
198	birc2	467	-	-	-	-	-	-	-
199	fgf5	564	-	-	-	-	-	-	-
200	chrna4	551	-	-	-	-	-	-	-

A.3.2 Table. A comparison of the dataset analysed in Chapter 3 and the Halligan et al. (2010) dataset.

Code	Gene name	Exonic bp (Halligan et al 2010)	Exonic bp (this study)	Intronic bp (Halligan et al. 2010)	Intronic bp (this study)
101	Gucy1a3	0	646	0	0
102	Pla2g4a	460	545	871	1477
103	Umps	670	779	721	721
104	Egfr	381	456	896	896
106	Orc3l	305	432	929	1165
107	Cela1	369	499	922	1625
108	Cyp2c39	187	403	483	483
109	Nf1	699	756	663	663
110	Xpc	675	886	0	299
111	Angptl7	461	698	795	795
112	Prkd1	546	689	768	768
113	mmp21	531	767	774	774
114	mad2l2	292	338	744	780
115	prdx1	281	471	1021	1610
116	bace1	537	758	705	705

117	Cyp2b10	310	512	159	825
118	cdc20	505	724	792	792
119	fgfr2	300	504	998	1224
120	Mmp15	569	835	696	696
121	Cyp17a1	455	679	793	793
122	Adm	354	568	291	704
124	Tert	637	829	681	1066
125	Cyp4b1	483	541	787	787
126	Rrm2b	382	429	470	470
127	prdx2	244	513	280	852
130	mt3	109	109	10	10
131	Mms19	554	668	751	833
132	mpo	832	1111	499	1479
133	Epx	508	699	764	1130
134	tgfb1i1	389	713	137	281
135	mapk11	452	452	772	772
137	Cyp1a1	295	981	274	781
139	stk25	689	758	625	843
140	csk	581	779	610	1186
141	por	212	440	326	821
143	mmp16	303	557	367	1004
144	rev1	812	857	313	1408
145	adh1	566	590	536	664
147	adh4	492	492	105	105
148	adh5	480	500	222	549
149	itgal	617	803	557	624
150	tpo	549	681	708	1370
151	Capn3	407	479	347	412
152	ddb1	608	709	601	601
153	pdlim1	344	476	25	25
154	tjp1	998	1013	304	345
155	ppib	114	135	323	323
157	cdc42bpb	598	710	48	470
158	aoc2	225	588	223	393
159	aoc3	928	972	321	321
160	ckb	787	969	405	639

162	Nos2	532	614	32	878
163	fmo3	702	702	70	70
164	tdp1	561	592	73	551
165	gab1	548	661	0	479
166	abp1	1083	1212	0	812
167	poln	638	719	31	45
168	fgf21	585	585	127	564
170	blm	1255	1255	0	454
174	hspa5	475	595	0	0
175	msr1	360	360	0	0
176	pax3	469	469	20	20
177	abl1	221	221		0
178	recql4	262	262	42	42
180	casp8	621	621	0	0
181	osr1	396	396	0	883
182	rxra	443	443	18	18
189	dclre1b	512	512	0	0
190	cdc37	279	279	31	31
191	neil3	566	566	10	10
192	foxm1	384	384	100	0
193	fmo5	291	291	0	0
194	scara3	435	435	0	0
196	sprr3	499	499	0	0
197	sepp1	735	836	0	456
198	birc2	467	467	0	0
199	fgf5	564	564	0	0
200	chrna4	551	551	0	0

A.4 Supplementary material for Chapter 4

Declaration: The Supplementary text below was not written by the author of this thesis. It is provided only for clarification of details regarding the sequencing techniques that were employed to generate the data analysed in this chapter.

A.4.1 Supplementary text. Details on Sequencing, SNP and genotype calling

Estimation of the accuracy of Illumina sequencing.

We attempted to check the accuracy of the Illumina SNP calls by comparing Illumina-based genotype calls with those made using traditional Sanger sequencing technology of the same individuals from a previous study (Halligan et al. 2010). We attempted to reduce the error rate within the Sanger based genotype calls by choosing regions of the genome for which we had Sanger sequence chromatograms in both directions, from coding sequence data only (where alignments are less error prone) and where both the forward and reverse sequence chromatograms were clear and had little background noise.

We were able to make genotype comparisons at a total of 16,249 sites, and were able to compare a total of 99,459 individual genotype calls. The majority of these sites were called as invariant by both methods, only 244 sites being called as variant by either method. At these sites, we observed a total of 55 discrepant SNP calls (over a total of 33 sites), however for 34 SNP calls (covering 20 sites) the error

could confidently be assigned to the Sanger technology and for 19 SNP calls (11 sites) the error could tentatively be assigned to the Sanger technology. Assignment of the error to Sanger sequencing in these cases resulted from several observations. Firstly, in nine of these cases, we could identify strong evidence for the genotype called by Illumina in the Sanger chromatograms. In some cases, this was due to an incorrect heterozygous genotype code being used (e.g. R, implying A/G instead of Y implying C/T) when calling the Sanger genotypes, though it is worth noting that these errors would not affect the inferred site frequency spectrum. In other cases, where Sanger called a genotype as homozygous and Illumina called the genotype as heterozygous, it was clear upon re-inspection that two peaks were evident in the sequence chromatogram corresponding to the two bases called by Illumina. Secondly, in six cases, the Sanger and Illumina genotypes matched for all individuals, but genotypes for two individuals were swapped. The most parsimonious explanation for this would be an error in labelling tubes during preparation for Sanger sequencing, since this was only observed in two of the 80 amplicons. Again, an error of this type would not affect the inferred site frequency spectrum. Thirdly, for the remaining 19 cases, we could confidently assign the error to cases of single allele amplification when carrying out Sanger sequencing. In all of these cases, heterozygous individuals called by Illumina were homozygous when called by Sanger, the Sanger amplicons showed no heterozygosity throughout their length, and we could confidently identify a heterozygous position within one of the Sanger primer sites from the Illumina sequences of the individuals that were discrepant.

Furthermore, in all of these cases the Illumina read depth was not abnormally high, which would be predicted if reads from paralogs were aligned to the same region.

For 19 discrepant SNP calls that could be tentatively assigned as Sanger errors, 13 were from a single Sanger amplicon. The SNP calls throughout this amplicon were consistent with three individuals being swapped. The remaining six SNP calls tentatively assigned as Sanger errors were comprised of five homozygous genotype calls in Illumina, but heterozygous calls in Sanger and one heterozygous call in Illumina but homozygous in Sanger. In the first instance it is possible that background noise in sequence traces caused an incorrect genotype call in the Sanger technology. In the second instance the discrepancy could be due to a recent duplication combined with mapping of Illumina reads from a duplicate region to the same genomic section, or alternatively, single allele amplification of the Sanger amplicon.

In only one case could we confidently assign the error to Illumina sequencing (and in this case the reported genotype quality from SAMtools had an exceptionally low value of 3) . In one other case we could tentatively assign the error to Illumina sequencing. The results indicate that for this dataset our Illumina sequencing is much more accurate than the Sanger sequence data for the same regions, and furthermore, that the Illumina sequencing based error rate is low. Accepting that we have two Illumina errors, the error rate = 2/99,459 = 0.002% per genotype call or 2/16,249 = 0.012% per site.

SNP calling.

We created bcf (genotype likelihood) files for each chromosome from the individual BAM files using 'samtools mpileup' with options -D -S -g -m 2 -F 0.0005 -P ILLUMINA (Li et al. 2009). We then used 'bcftools view' with options -A -g to obtain SNP calls for every site in the genome. bcftools allows the specification of a prior allele frequency spectrum (AFS), which can improve genotype calls at each site. We obtained a suitable prior AFS for the genome using an iterative approach (see http://samtools.sourceforge.net/mpileup.shtml). We used bcftools to estimate a posterior allele frequency spectrum (AFS) for all sites on chromosome 1, then used the AFS as a prior (using option -P) for a second call to bcftools, and iterated until the prior and posterior converged. The final posterior AFS was then used as a prior to obtain genotype calls for the whole genome, which were used to obtain site frequency spectra for specific genomic regions. We called all genotypes using an approximate *M. m. castaneus* reference sequence, which is identical to the NCBIM37/mm9 reference sequence, but with all SNPs at a frequency of >0.5 replaced with the major allele observed amongst the *M. m. castaneus* individuals. This reduced the number of SNP calls representing fixed differences between the mouse reference and the *M. m. castaneus* sequences and reduced the number of triallelic SNP calls (which can arise if a variant in *M. m. castaneus* also has a fixed difference to the reference).

Construction of the M. famulus genome sequence.

M. famulus is divergent from *M. m. castaneus* and the *M. m. musculus* reference sequence and assembly of its genome sequence is therefore worth some special consideration. Specifically, divergent regions in the genome will reduce the efficacy or accuracy of the final sequence because reads with too many differences to the reference cannot be mapped properly to the reference genome. To mitigate this effect, we used an 'iterative mapping' approach where successive rounds of read mapping are conducted and after each iteration a new genome sequence is generated for use in the next iteration. In effect, we are converting the original reference genome to a *M. famulus* reference genome by changing the divergent sites to match those from *M. famulus*. Therefore, regions of high divergence where reads cannot be aligned initially may eventually be assembled as divergent sites are eliminated from the reference.

We aligned each of the lanes of data to the reference genome independently using BWA v0.5.9. We used SAMtools v0.1.16 mpileup to call variant SNPs and converted variant positions in the reference to match the all high quality homozygous variant calls (genotype quality, GQ > 40). With this approach we ignored all shared polymorphisms that are heterozygous in our *M. famulus* sample and more importantly we also ignore potential indel divergence. We discarded indels and SNPs neighbouring indels to avoid converting regions of the genome where read mapping has erroneously generated indels due to repeats and to retain the same position indices of genomic features as the reference genome. The new reference was then used to repeat this process. The most improvement in terms of positions covered and

reads mapped occurred in the first and second iteration, after which the gains made with successive iterations plateaued. We carried out a total of five iterations over which the number of reads mapped increased from 72.6% to 84.0% and the median coverage improved from improved from 23x to 25x. After five iterations we called the final genotype of the *M. famulus* genome using the same methods described for the *M. m. castaneus*.

End of supplementary text

Appendix B

Positive and Negative Selection on Noncoding DNA Close to Protein-Coding Genes in Wild House Mice

Athanasios Kousathanas,*[,1] Fiona Oliver,[1] Daniel L. Halligan,[1] and Peter D. Keightley[1]

[1]Institute of Evolutionary Biology, School of Biological Sciences, University of Edinburgh, Edinburgh, United Kingdom

***Corresponding author**: E-mail: a.kousathanas@sms.ed.ac.uk.*

Associate editor: Michael Nachman

Abstract

During the past two decades, evidence has accumulated of adaptive evolution within protein-coding genes in a variety of species. However, with the exception of *Drosophila* and humans, little is known about the extent of adaptive evolution in noncoding DNA. Here, we study regions upstream and downstream of protein-coding genes in the house mouse *Mus musculus castaneus*, a species that has a much larger effective population size (N_e) than humans. We analyze polymorphism data for 78 genes from 15 wild-caught *M. m. castaneus* individuals and divergence to a closely related species, *Mus famulus*. We find high levels of nucleotide diversity and moderate levels of selective constraint in upstream and downstream regions compared with nonsynonymous sites of protein-coding genes. From the polymorphism data, we estimate the distribution of fitness effects (DFE) of new mutations and infer that most new mutations in upstream and downstream regions behave as effectively neutral and that only a small fraction is strongly negatively selected. We also estimate the fraction of substitutions that have been driven to fixation by positive selection (α) and the ratio of adaptive to neutral divergence (ω_α). We find that α for upstream and downstream regions (~10%) is much lower than α for nonsynonymous sites (~50%). However, ω_α estimates are very similar for nonsynonymous sites (~10%) and upstream and downstream regions (~5%). We conclude that negative selection operating in upstream and downstream regions of *M. m. castaneus* is weak and that the low values of α for upstream and downstream regions relative to nonsynonymous sites are most likely due to the presence of a higher proportion of neutrally evolving sites and not due to lower absolute rates of adaptive substitution.

Key words: adaptive evolution, effective population size, wild mice, selective constraint, upstream and downstream regions.

Introduction

In recent years, the search for evidence of adaptive evolution at the molecular level has been at the forefront of genetics research. A principal motivation has been to identify regions of the genome that have experienced adaptive evolution because this might provide clues to their functional importance and may be informative about the features that make each species unique. There have been a wealth of studies focusing on protein-coding genes. Studies in *Drosophila*, employing variants of the McDonald–Kreitman (MK) test (McDonald and Kreitman 1991), suggest a high proportion of adaptive amino acid substitutions (α) (50% or more; Smith and Eyre-Walker 2002; Sawyer et al. 2003; Bierne and Eyre-Walker 2004; Welch 2006; Shapiro et al. 2007; Eyre-Walker and Keightley 2009), whereas in humans similar studies have produced low estimates of α (0–20%; The Chimpanzee Sequencing and Analysis Consortium 2005; Zhang and Li 2005; Boyko et al. 2008; Eyre-Walker and Keightley 2009). These contrasting results between *Drosophila* and humans have been interpreted to be a consequence of different effective population sizes (N_e), that is, the small N_e of the hominid lineage could have resulted in a reduced efficacy of natural selection. Other evidence points to a positive relationship between α and recent N_e. For example, α for protein-coding genes has been estimated to be 50% or more in enteric bacteria, which have a large N_e (Charlesworth and Eyre-Walker 2006), close to zero in *Arabidopsis* (*A. lyrata* and *A. thaliana*), which has small N_e (Foxe et al. 2008), and about 40% in *Capsella grandiflora*, a species closely related to *A. thaliana* with a larger N_e (Slotte et al. 2010). A recent study of the house mouse *Mus musculus castaneus*, which has N_e comparable with *Drosophila*, also produced a high estimate of α for protein-coding genes (~50%; Halligan et al. 2010), again suggesting that N_e is a strong determinant of the efficacy of positive selection on amino acid-changing mutations across taxa.

Estimates of the frequency of adaptive nucleotide substitution in noncoding DNA are currently restricted to *Drosophila* and humans. In *Drosophila*, estimates of α for 5′ and 3′ untranslated regions (UTRs) are nearly as high as for protein-coding genes (i.e., 50% or more; Kohn et al. 2004; Andolfatto 2005; Haddrill et al. 2008) and for introns and intergenic regions are relatively low (~10–20%; Andolfatto 2005; Haddrill et al. 2008). In humans, estimates of α for noncoding regions upstream and downstream of protein-coding genes are close to zero (Keightley et al. 2005; Eyre-Walker and Keightley 2009).

In this study, we investigate positive and negative selection operating on noncoding regions upstream and downstream of the protein-coding genes in a sample of the house mouse *M. m. castaneus* that were previously studied by Halligan et al. (2010). We study regions upstream and

downstream of protein-coding genes that are known to be enriched for regulatory elements (Xie et al. 2005; Veyrieras et al. 2008) and are implicated in the control of transcription and translation (Gray and Wickens 1998; Shabalina and Spiridonov 2004). Previous studies in murids have shown that ~30% of sites in 5′- and 3′-UTRs and ~10% of sites that are within the first 3–5 kb upstream and downstream of the transcription start and stop codon, respectively, are subject to negative selection (Keightley et al. 2005; Gaffney and Keightley 2006). Here, we perform a more thorough investigation of negative selection operating in upstream and downstream regions by estimating the full distribution of fitness effects (DFE) of new mutations. We then proceed to investigate positive selection by estimating α using a method that attempts to account for the presence of slightly deleterious mutations: The DFE is used to predict the expected divergence between two species caused by the fixation of neutral and slightly deleterious mutations, and this is compared with the observed divergence (Eyre-Walker and Keightley 2009). The difference between the observed and expected divergence is used to estimate the amount of adaptive divergence and α. We also estimate ω_a, the rate of adaptive divergence relative to neutral divergence, which allows us to better compare rates of adaptive evolution between species, by controlling for the effects of N_e on the numbers of effectively neutral substitutions (Gossmann et al. 2010).

Materials and Methods

Sampling of Mice

We analyzed 15 *M. m. castaneus* individuals sampled from four regions south of the Himalayas in the Himachal Pradesh state of India. We also generated sequence data from a *Mus famulus* individual originating from India that was previously obtained from the Montpellier wild mice genetic repository (http://www.isem.cnrs.fr/spip.php?article4777). A more detailed description of the sampling of the mice can be found in Halligan et al. (2010).

Choice of Genes

We analyzed 78 autosomal genes from *M. m. castaneus* whose human orthologs have also been sequenced as part of the Environmental Genome Project (EGP) (Livingston et al. 2004). The EGP data set is enriched for genes that are involved in pathways for DNA repair, cell cycle control, drug metabolism, and apoptosis and therefore non random (Livingston et al. 2004). The genes were chosen if there were African human polymorphism data available; this enabled us to more directly compare the results in humans with mice.

As part of this study, we successfully sequenced upstream and downstream regions for 49 and 51 genes, respectively, in 15 *M. m. castaneus* individuals and one *M. famulus* individual. We designed primers to amplify the upstream region of each gene, which lies approximately up to 500 bp upstream of the first codon of the first exon as annotated in the reference mouse genome. Similarly, to amplify the downstream region of each gene, we designed primers that captured the region that lies approximately up to 500 bp downstream of the stop codon of the last exon in the reference mouse genome. We chose to sequence ~500 bp upstream and downstream of protein-coding genes because evidence from studies of selective constraint, regulatory motifs, and expression-quantitative trait loci suggest that there is high density of functional elements in these regions (Xie et al. 2005; Gaffney and Keightley 2006; Veyrieras et al. 2008). Additionally, the interpretation of sequences further upstream and downstream from these regions was made more difficult by frequent indel variation, making calling of single nucleotide polymorphisms (SNPs) problematic. In supplementary table S1, Supplementary Material online, we give details of the genes analyzed in this study and the upstream and downstream regions that we successfully sequenced. We chose not to restrict our analyses to those genes for which upstream, downstream, exonic, and intronic sequence data were available because the smaller samples for intronic (65), upstream (49), and downstream (51) site classes are unbiased in relation to the larger data set for exonic sequence (78). The set of genes for which we have upstream and downstream sequence data overlap in almost all cases. Additionally, in analyses where a putatively neutral (i.e., synonymous or intronic sites) and a selected class (i.e., nonsynonymous, upstream, or downstream) were required, we only analyzed genes for which both the neutral and the selected class were sequenced. We have also updated the data set of Halligan et al. (2010) with new exonic and intronic sequence of the 78 genes. We reanalyzed the exonic and intronic data sets because we had ~20% new exonic sequence data and ~60% new intronic sequence data. Supplementary table S2, Supplementary Material online, shows in detail the differences between the Halligan et al. (2010) data set and the updated data set used in this study.

Sequencing

GoTaq DNA polymerase (Promega) was used in touchdown-style polymerase chain reactions (PCRs): an initial denaturation step of 95 °C for 15 min, followed by 28 cycles of 95 °C for 30 s, 62 °C for 45 s (reducing by 0.5 °C every cycle), 72 °C for 2 min, then 12 cycles of 95 °C for 30 s, 52 °C for 45 s, and 72 °C for 2 min, with a final extension step at 72 °C for 10 min. ExoSAP-IT (USB) was used for the purification of PCR products. If we obtained nonspecific PCR products, we designed new primers to try to increase the specificity. Sequencing was done using Big Dye Terminator Sequencing Kits (Applied Biosystems, Foster City, CA) on an ABI Prism 3730 DNA Analyzer, and both forward and reverse sequences were generated. CodonCode Aligner version 2.0.6 was used to analyze and detect variants (http://www.codoncode.com/aligner/). We used Phred computer program as employed in CodonCode Aligner to assess sequence quality. Sequences had an average Phred score of >60. All sequence traces were manually checked. Sites with a Phred score <30, which could be low quality sequence or heterozygotes, were manually checked but were not automatically excluded in order to avoid excluding

heterozygotes. Where sequence was found to be too low quality, due to multiple indels or repetitive regions, new amplicons were produced on either side of the difficult to sequence area. Such difficult to sequence areas were replaced by 'N's before further analysis. Finally, we generated alignments between the 15 *M. m. castaneus*, the *M. famulus* individual, and the *M. m. musculus* reference sequence using CodonCode Aligner, and all alignments were checked by eye before further analysis.

Dealing with Heterozygous Indels

Heterozygous indels can make SNP calls from sequence traces problematic, and we took several steps to ensure that SNP calls were accurate. First, we obtained sequences from both strands for every amplicon. This allowed us to accurately call SNPs up to a heterozygous indel without having to interpret the sequence traces after the occurrence of heterozygous indel. Second, for many of the genomic regions that we sequenced, we designed multiple overlapping amplicons to cover the region to ensure that SNP calls on either side of heterozygous indels were accurate. Third, we used CodonCode Aligner to call SNPs. This software has specialized algorithms that are able to call SNPs, following heterozygous indels, by interpreting the out-of-sync signals from both strands in the sequence traces. Fourth, we also manually checked all sequence traces and SNP calls by eye, including manually disentangling out-of-sync signals in sequence traces where heterozygous indels were encountered. Fifth, we also very carefully checked SNPs at sites where the genotype frequencies of the individuals were not at Hardy–Weinberg equilibrium (because this signal could be indicative of heterozygotes being miscalled as homozygotes).

Sequence Processing

We obtained orthologous *Rattus norvegicus* sequences for each amplicon using a reciprocal best-hits Blast approach. To do this, we Blasted the mouse reference sequence (mm9) for each amplicon, plus 200 bp of flanking DNA, against two different assemblies (labeled "standard" and "alternative") of the rat genome and searched for a reciprocal best hit. If we failed to find a reciprocal best hit for the standard assembly, we searched the alternative assembly. Both assemblies were downloaded from the University of California–Santa Cruz (UCSC) genome browser, the standard was produced by the Baylor College of Medicine Human Genome Sequencing Center (BCM-HGSC) as part of the Rat Genome Sequencing Consortium, and the alternative was produced by Celera Genomics. If we failed to identify an ortholog via the reciprocal best-hits approach, we checked the relevant section in the "multiz30way" whole-genome sequence alignments of 30 vertebrates (http://genome.ucsc.edu/). We considered sequences to be orthologous if the sequence of interest was located entirely within a single unbroken alignment for mouse and rat. We realigned all alignments obtained by either method using MAVID (Bray and Pachter 2004) and then subsequently checked them all by eye. Any obviously misaligned sections identified when checking by eye were masked from any further analysis. Using this procedure, putatively orthologous rat sequences were obtained for at least part of every mouse amplicon.

We constructed alignments for each amplicon between the mouse reference (mm9) sequence, the sequences from all *M. m. castaneus* individuals, *M. famulus*, and rat. We annotated sites according to the mouse reference genome into the following categories: 5', 3', intron or coding. Within the coding category, sites were categorized as first, second, or third positions as well as the level of degeneracy in the genetic code (zero-fold-, 2-fold-, or 4-fold degenerate). We excluded potential splice sites of introns (defined as the first 6 bp or last 16 bp of an intron) from any analysis. We also categorized sites on the basis of their CpG-prone status (defined as being preceded by a C or followed by a G in any species).

Summary Statistics

We assume that segregating polymorphisms are biallelic. If there were more than two alleles segregating at a site, we only consider the two most frequent alleles. We calculated two statistics for nucleotide diversity, π and Watterson's θ (θ_w):

Let the site frequency spectrum (SFS) of a class of sites be the vector ν_i, containing i ($0 \leq i < n$) segregating alleles in a sample of n alleles from the population. Then, π and θ_w are calculated as follows (Watterson 1975; Tajima 1983):

$$\pi = 2\frac{\sum_{i=1}^{n-1} i(n-i)\nu_i}{n(n-1)}, \quad \theta_w = \frac{\sum_{i=1}^{n-1} \nu_i}{\sum_{i=1}^{n-1} \frac{1}{i}}.$$

For our data set, given that we sequenced both chromosomes of each of the 15 *M. m. castaneus* individuals, the sampled number of alleles (n) was 30 if the sequencing was successful for every individual. However, due to sequencing failures, our data set did not contain 30 sequenced alleles for each site, so we calculated composite estimates of π and θ_w. We calculated π and θ_w for sites that had the same number of alleles sequenced (categories of coverage) and then summed the estimates across categories of coverage as per Halligan et al. (2010).

For a population at Fisher–Wright equilibrium, and assuming no selection, π and θ_w estimates are expected to be equal to one another. They are expected to differ, however, if there is a skew in the SFS toward low- or high-frequency alleles. The level of skew can be quantified by the Tajima's D statistic (Tajima 1989). However, to calculate D, there needs to be an equal number of alleles sequenced at each site. We therefore rejected any sites where we had fewer than 20 alleles sequenced. We then sampled without replacement 20 alleles from each of the remaining sites such that the number of alleles sampled at each site was constant.

Nucleotide divergence (d) between *M. m. castaneus* and both *M. famulus* and rat was calculated using the Kimura 2-parameter correction (Kimura 1980). We had multiple sequences for *M. m. castaneus*, so we computed an average

divergence. We calculated evolutionary constraint C_d by comparing substitution rates at a putatively neutral and a selected site class. We use substitution rates at neutral sites to estimate expected numbers of substitutions at selected sites. Expected (E_d) and observed (O_d) numbers of substitutions are compared, and constraint is calculated as $C_d = 1 - O_d/E_d$ (Eyre-Walker and Keightley 1999; Keightley and Gaffney 2003). We used synonymous sites as the neutral class.

Nonsynonymous and Synonymous Sites

We treated nonsynonymous and synonymous sites as in Li (1993) and Pamilo and Bianchi (1993): All zero-fold degenerate sites were treated as nonsynonymous and 4-fold degenerate sites as synonymous. At 2-fold degenerate sites, transitions were considered synonymous and transversions nonsynonymous. Unmutated 2-fold degenerate sites were divided into nonsynonymous and synonymous by considering the ratio of transitional and transversional changes (t_s/t_v) as calculated at 4-fold degenerate sites across all genes in a comparison of *M. m. castaneus* and *M. famulus*.

DFE of New Mutations

We employed a maximum likelihood (ML) approach described by Keightley and Eyre-Walker (2007) to infer the DFE of new mutations at nonsynonymous sites of coding regions and in upstream and downstream regions as implemented in the program DFE-alpha (available online at http://homepages.ed.ac.uk/eang33/). The method assumes two classes of sites, one neutral and one selected, and contrasts SFSs of the two classes. Fitness effects of new mutations (s) are assumed to be zero in the neutral class and unconditionally deleterious in the selected class and are sampled from a gamma distribution with parameters a (scale) and b (shape). The method also incorporates a simple demographic model: The population at an initial size N_1 experiences a step change to N_2, t generations in the past. We use a constant N_1 of 100 so that the ratio N_2/N_1, that is, the change in population size, is actually estimated. An additional parameter, f_0, estimates the proportion of unmutated sites. The parameter space of N_2/N_1, t/N_2, f_0, a, and b is searched to find the values that maximize the likelihood of observing the neutral and selected SFSs. In order to account for variation in the number of alleles at each site, we generated SFSs for sites that had the same number of alleles sampled in both neutral and selected classes. We summed the log-likelihoods of each SFS to produce the overall log-likelihood as per Halligan et al. (2010). We also explored the likelihood surface for the demographic and selection parameters to check for any irregularities such as local maxima (see supplementary figs. S1 and S2, Supplementary Material online). The likelihood surfaces were plotted using *R* (Ihaka and Gentleman 1996). We interpolated from the estimated parameters of the gamma distribution, the percentages of mutations that fall within four $N_e s$ ranges: 0–1, 1–10, 10–100, and >100.

Estimating Evolutionary Constraint by Using Polymorphism Data

Evolutionary constraint, calculated using divergences (C_d), will be biased downward if some fraction of the observed divergence at the focal class is adaptive. We obtained a second estimate of evolutionary constraint, which is not subject to such biases, by using information from polymorphism data only (C_p). We first estimate the average fixation probability of a new, negatively selected mutation ($\bar{u}$) at the focal class by integrating over the DFE as in Eyre-Walker and Keightley (2009):

$$\bar{u} = \int_0^\infty 2Nu(N,s)f\langle s|a,b\rangle \, \mathrm{d}s, \quad (1)$$

where $u(N, s)$ is the fixation probability of a new negatively selected mutation (N is assumed equal to N_e).

C_p can then be calculated as follows:

$$C_p = 1 - \bar{u}. \quad (2)$$

Adaptive Evolution

To estimate the proportion of adaptive substitutions (α), approaches based on the MK test are frequently used (Eyre-Walker 2006). However, these approaches do not take into account slightly deleterious mutations, which contribute proportionally more to polymorphism than divergence and therefore can lead to underestimates of α. They also ignore demographic history, which can be problematic, because a population size change in the past could produce evolutionary signatures similar to selection. A recent extension of the MK test (Eyre-Walker and Keightley 2009) attempts to take into account both slightly deleterious mutations and population demography.

The method assumes a neutral and a selected class of sites. The DFE in the selected class is first estimated, simultaneously accounting for demography as in Keightley and Eyre-Walker (2007). The nucleotide divergence of the neutral class (d_S) is assumed to be proportional to the mutation rate and divergence due to deleterious mutations in the selected class is the product of the mutation rate and the average fixation probability of a new deleterious mutation ($\bar{u}$). We can estimate the expected divergence (d_{est}) in the selected class due to neutral and deleterious mutations as follows:

$$d_{est} = d_S \bar{u}. \quad (3)$$

The difference between the observed (d_X) and estimated divergence (d_{est}) estimates the amount of adaptive divergence ($d_{adaptive} = d_X - d_{est}$) in the selected class (X). If we scale $d_{adaptive}$ by d_X we obtain α, the fraction of adaptive substitutions in the selected class:

$$\alpha = \frac{d_{adaptive}}{d_X}. \quad (4)$$

However, as noted by Gossmann et al. (2010), caution should be exercised when comparing estimates of α from

Selection on Noncoding DNA in Wild Mice · doi:10.1093/molbev/msq299 MBE

Table 1. Details of Genes Sequenced and Percentages of Sites Sequenced for All 30 Alleles and for at Least 20 Alleles.

				% Sites Sequenced	
Site Class	Number of Genes	Number of Sites	Mean Number Sites Per Gene (SD)	30 Alleles	>20 Alleles
Nonsynonymous	78	34,532	443 (160)	60	95
Synonymous	78	13,056	167 (63)	60	94
Intron	65	43,672	672 (413)	45	88
Upstream	49	25,303	516 (132)	50	93
Downstream	51	26,622	522 (182)	57	91

different species. Differences in the estimates of α could reflect differences in the contribution of slightly deleterious mutations to d_X rather than different rates of adaptive substitution. We can control for differences in the frequency of effectively neutral mutations in the selected class by computing ω_α, the ratio of $d_{adaptive}$ to d_S:

$$\omega_\alpha = \frac{d_{adaptive}}{d_S} \quad (5)$$

We also estimated α using a simple but frequently used method (Fay and Wu 2001;):

$$\alpha_{FWW} = 1 - \frac{D_S P_X}{D_X P_S}, \quad (6)$$

where D_X and D_S are counts of divergent sites between *M. m. castaneus* and an outgroup species for selected and neutral site classes, respectively, and P_X and P_S counts of polymorphic sites for selected and neutral site classes, respectively.

Confidence intervals and standard error for all parameters were obtained by bootstrapping 1,000 times by gene. *P* values, computed for comparisons between site classes or with zero, were obtained by two-tailed bootstrap tests.

Results

Data

Our data set consists of sequences from 78 autosomal genes from a sample of 15 wild, unrelated *M. m castaneus* individuals sampled from NW India. Part of the coding region of these genes and partial introns were sequenced in a previous study (Halligan et al. 2010). In this study, we focus on regions directly upstream and downstream of the coding region of these genes. We successfully amplified and sequenced ~500 bp upstream and downstream from the start and stop codon for a subset of 49 and 51 of the 78 genes, respectively (table 1). We have also updated the data set of Halligan et al. (2010) by obtaining additional exonic and intronic sequence for the 78 genes, and we compared our results from noncoding DNA with results from these new data. We successfully sequenced 20 alleles or more for ~90% of the sites (table 1). We also sequenced the orthologous genes in a *M. famulus* individual, which we used together with the rat as an outgroup to estimate divergence, constraint, α and ω_α.

Summary Statistics

Nucleotide diversity, Tajima's *D*, divergence to *M. famulus* and rat and evolutionary constraint estimates are shown in table 2. The upstream and downstream site classes show intermediate levels of nucleotide diversity (π = 0.56% in both cases) and divergence to *M. famulus* (d = 2.49% for upstream and 2.38% for downstream) compared with nonsynonymous sites (π = 0.15% and d = 0.82%) but are closer to the synonymous site estimates (π = 0.75% and d = 3.27%). Divergence to the rat is about five times higher than divergence to *M. famulus* for all site classes.

In contrast to nonsynonymous sites, the upstream, downstream, and intronic site classes do not show discernible differences in the shape of their SFSs compared with synonymous sites (fig. 1). Tajima's *D* estimates, which quantify the skew in the SFS, are significantly lower than zero for all cases examined even for synonymous sites ($P < 10^{-3}$ in all cases examined). A negative Tajima's *D* indicates an excess of rare variants, which can be caused by negative selection. However, population expansion, a prolonged population bottleneck or population subdivision can also produce a similar pattern. Different mutation rates between the compared regions could also alter the SFS. For example, CpG dinucleotides have higher mutation rates in mammals (Arndt et al. 2003) and their frequencies differ between coding and noncoding DNA. However, if CpG-prone sites are excluded, we observe little changes in the SFS in all cases (fig. 1).

We calculate evolutionary constraint C_d by comparing inter specific divergence between the focal site class and a putatively neutral site class. The estimate for C_d is moderately high for upstream and downstream site classes (24.1% and 28.4%, respectively, table 2) and significantly different than zero in both cases (P = 0.004 for upstream; P = 0.002 for downstream).

DFE of New Mutations

We inferred the DFE of new mutations along with demographic parameters using a ML-based approach (Keightley and Eyre-Walker 2007). We first tested whether a model that incorporates demographic change plus selection (Model 3) fits the data significantly better than a model that assumes only demographic change (Model 1). The likelihood ratios for this comparison were highly significant in all cases examined ($-\Delta\log L$ values are reported in table 3; $P < 10^{-2}$ in all cases, with d.f. = 2). We also examined a model where we fitted only selection under constant population size (Model 2). We found that the fit of Model 2 to the data was significantly poorer than Model 3 in all cases examined ($-\Delta\log L$ values are reported in table 3; $P < 10^{-9}$ in all cases, with d.f. = 2).

Table 2. Estimates of Percentage Diversity (% π, % θ_w) Summed over All Sites for *M. m. castaneus*, Tajima's *D*, Percentage Divergence (% *d*) to *Mus famulus*, and the Rat and Evolutionary Constraint (C_d) Calculated using Synonymous Sites as the Neutral Class.

Site Class	% π (SE)	% θ_w (SE)	Tajima's *D* (SE)	% *d* (*M. famulus*) (SE)	% *d* (rat) (SE)	% C_d (*M. famulus*) (SE)
Nonsynonymous	0.15 (0.02)	0.22 (0.02)	−0.93 (0.17)	0.82 (0.10)	3.69 (0.40)	74.5 (3.3)
Synonymous	0.75 (0.06)	0.93 (0.06)	−0.53 (0.12)	3.27 (0.21)	18.11 (0.56)	—
Intron	0.66 (0.04)	0.83 (0.04)	−0.75 (0.09)	2.90 (0.14)	15.61 (0.42)	12.1 (7.2)
Upstream	0.56 (0.05)	0.71 (0.06)	−0.78 (0.14)	2.49 (0.21)	12.20 (0.65)	24.1 (7.1)
Downstream	0.56 (0.06)	0.69 (0.06)	−0.59 (0.16)	2.38 (0.21)	11.78 (0.78)	28.4 (7.7)

Using the inferred parameters of the DFE (see supplementary tables S3 and S4, Supplementary Material online, for estimates of the demographic parameters, the mean selective effect ($N_e s$), and the shape parameter of the fitted gamma distribution), we can estimate the proportion of mutations falling into each of four categories of selective effects ($N_e s$): 0–1, 1–10, 10–100, and >100 (table 4). For upstream and downstream sites, most new mutations fall into the effectively neutral category ($N_e s$, 0–1) (69.8% and 67.5%, respectively), which is in sharp contrast and significantly different ($P < 0.05$ for both upstream and downstream in all comparisons) from the estimate for nonsynonymous sites (15.4%). Although most new mutations in upstream and downstream sequences are effectively neutral, there is a substantial fraction (21.7% and 19.6%, respectively, not significantly different from zero) of strongly selected mutations ($N_e s > 10$) in these regions.

We also calculate evolutionary constraint C_p, a statistic that summarizes the DFE, and is the average probability of a new deleterious mutation to be lost. $\%C_p$ is moderately high for upstream (30.7%) and downstream (33.9%) site classes and is significantly different from zero only for the downstream site class ($P < 0.05$).

Adaptive Evolution

We then estimated the fraction of substitutions driven to fixation by positive selection (α) using an extension of the MK test (Eyre-Walker and Keightley 2009). This method uses neutral divergence between *M. m. castaneus* and an outgroup (either *M. famulus* or rat) along with the DFE, inferred from polymorphism data of *M. m. castaneus* to infer the expected divergence between *M. m. castaneus* and the outgroup. The difference between the observed and the expected divergence estimates the adaptive divergence between *M. m. castaneus* and the outgroup. α is then calculated by scaling the adaptive divergence by the observed divergence (see detailed description in the Materials and Methods). Estimates of α for the nonsynonymous, upstream, and downstream site classes for *M. m. castaneus* are presented in table 5. We report moderately low estimates of α for upstream and downstream site classes (11.8% and 9.3%, respectively), which are not significantly different from zero when using *M. famulus* as the outgroup and synonymous sites as the neutral reference. By combining upstream and downstream sequences, we obtain a similar point estimate for α (9.3%) and a narrower confidence interval (−15.4/36.3), which includes zero but excludes high estimates of α and is not significantly different from the estimate for nonsynonymous sites ($P = 0.078$). The point estimates are very similar when using the rat as the outgroup or intronic sites are used as the neutral reference, but the confidence intervals are narrower for the latter case because more data are included. The estimate for α for combined upstream and downstream sequences is significantly different from nonsynonymous sites when using intronic sites as the neutral reference ($P = 0.014$ when

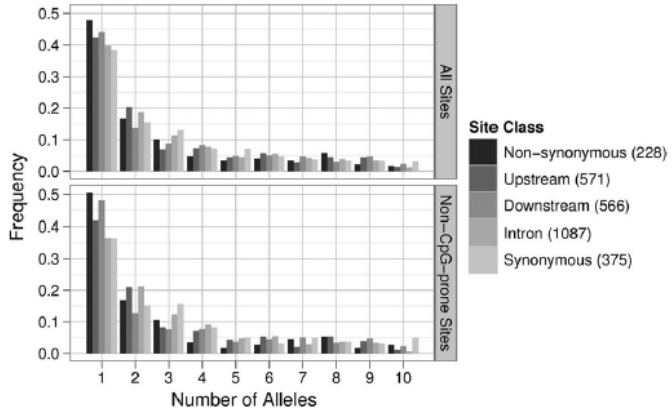


Fig. 1 Plots of the SFSs for nonsynonymous, upstream, downstream, intron, and synonymous site classes for all sites and for non-CpG-prone sites only. Numbers of polymorphic sites are given in parentheses and refer to the "All Sites" plot.

Table 3. Likelihood-Ratio Tests Contrasting Models Fitted to the SFS Data when Estimating the DFE. Models Fitting Only Demography (M1) or Only Selection (M2) Are Contrasted with a Model That Fits Both Demography and Selection (M3) to the Data.

Site Class		$-2\Delta\log L$	
Neutral	Selected	M1 versus M3	M2 versus M3
Synonymous	Nonsynonymous	312.4	47.2
	Upstream	12.4	50.5
	Downstream	13.2	43.7
Intron	Nonsynonymous	412.6	120.3
	Upstream	10.6	116.8
	Downstream	12.1	126.0

NOTE.—M1, demography, no selection; M2, no demography, selection; M3, demography and selection

using *M. famulus* as the outgroup and $P = 0.018$ when using the rat as the outgroup).

The different estimates of α between nonsynonymous, upstream, and downstream site classes might be due to differences in the rate of slightly deleterious rather than adaptive substitution (Gossmann et al. 2010). In order to account for any differences in the slightly deleterious substitution rate between the selected site classes, we computed the ratio of adaptive divergence to neutral divergence (ω_α). The resulting ω_α estimates are very similar and not significantly different for the nonsynonymous, upstream, and downstream site classes (11.9%, 9%, 6.6%, respectively, table 5). These results indicate that the lower estimates of α in the upstream and downstream site classes compared with nonsynonymous sites could be due to a higher proportion of the upstream and downstream sites evolving close to neutrally rather than a lower absolute rate of adaptive substitution. The confidence intervals for ω_α are very wide when examining upstream and downstream site classes individually but narrow down when we combine data from upstream and downstream site classes (the upper boundary for ω_α is never higher than 30.6%). Similarly, with estimates for α, when using intronic sites as the neutral reference and the rat as the outgroup, we get narrower confidence intervals because more data are included.

Finally, we use a simple frequently used approach to estimate α (Fay et al. 2001) in order to be able to make comparisons with studies that have not employed the Eyre-Walker and Keightley (2009) methodology. We control for slightly deleterious mutations by excluding low-frequency polymorphisms (<10%) as suggested by (Fay et al. 2001). By using this method, we obtained zero or negative estimates of α for upstream and downstream site classes (supplementary table S5, Supplementary Material online) that roughly agree with the estimates we obtained using the Eyre-Walker and Keightley (2009) methodology.

Discussion

In this study, we present results suggesting that sites upstream and downstream of protein-coding regions in *M. m. castaneus* are, on average, under weak positive and negative selection. Several lines of evidence support this conclusion. Nucleotide diversity in *M. m. castaneus* and divergence to *M. famulus* or rat in upstream and downstream regions are much higher than for nonsynonymous sites and slightly but significantly lower than synonymous sites. Evolutionary constraint is also significantly lower in upstream and downstream regions than for nonsynonymous sites. Tajima's *D* estimates are not significantly different between site classes, except for the synonymous and nonsynonymous sites comparison, which suggests either that all site classes investigated are under negative selection or that a population expansion or bottleneck has occurred in the past in *M. m. castaneus*. Indeed, if we fit a simple demographic model of a step change in population size, we find evidence for population expansion of *M. m. castaneus*, which might explain the negative Tajima's *D* at synonymous sites. A population expansion might also explain the negative Tajima's *D* in upstream and downstream regions. However, a model of a demographic change plus negative selection fits the data significantly better than a model of demographic change with no selection or a model with selection only in all cases examined. Therefore, we obtained statistically significant evidence for both a population expansion in *M. m. castaneus* in the past and negative selection acting on upstream and downstream regions. The DFE inferred for upstream and downstream regions implies that most new mutations have $N_e s$ values in range of 0–1, but a small fraction is strongly negatively selected. At nonsynonymous sites the pattern is reversed because most new mutations are strongly deleterious. This result further supports the conclusion that upstream and downstream regions are, on average, under weak selective constraint compared with nonsynonymous sites.

Table 4. Estimates of Percentages of Mutations in Four $N_e S$ Ranges and Evolutionary Constraint Estimated from Polymorphism (C_p).

Site Class		Percentage of Mutations in $N_e s$ Range [95% CI]				
Neutral	Selected	0–1	1–10	10–100	>100	% C_p [95% CI]
Synonymous	Nonsynonymous	15.4 [9/23.3]	11.4 [3.1/18.6]	19.5 [3.6/42.1]	53.7 [27.6/71.8]	84.7 [75.9/90.8]
	Upstream	69.8 [48.1/100]	8.5 [0/39.0]	9.4 [0/20.3]	12.3 [0/29.4]	30.7 [0/53.3]
	Downstream	67.5 [29.1/100]	13.0 [0/69.2]	13.6 [0/23.6]	6.0 [0/28.0]	33.9 [7.9/64]
	Upstream + Downstream	70.7 [52.6/92.2]	8.6 [5.9/31.8]	9.5 [0/17.7]	11.2 [0/27.4]	30.0 [10.4/48.9]
Intron	Nonsynonymous	15.0 [7.8/24.3]	13.1 [4.4/20.4]	24.1 [5.3/47.5]	47.8 [21.8/67.9]	83.0 [73.7/89.6]
	Upstream	77.6 [56.6/100]	9.3 [0/35.2]	9.2 [0/15.7]	3.8 [0/15.3]	22.2 [0/41.3]
	Downstream	75.9 [50.8/100]	9.2 [0/30.2]	9.5 [0/19.7]	5.4 [0/19.9]	33.9 [7.9/64.0]
	Upstream + Downstream	77.1 [63.9/100]	9.3 [0/21.3]	9.3 [0/14.5]	4.3 [0/15.8]	23.1 [0/35.9]

Kousathanas et al. · doi:10.1093/molbev/msq299 MBE

Table 5. The Fraction of Substitutions Driven to Fixation by Positive Selection (α) and the Ratio of Adaptive to Neutral Divergence (ω_α) Estimated Using *Mus famulus* and the Rat as Outgroups.

Site Class		% α [95% CI] Outgroup		% ω_α [95% CI] Outgroup	
Neutral	Selected	*M. famulus*	Rat	*M. famulus*	Rat
Synonymous	Nonsynonymous	46.6 [18.0/67.6]	43.5 [13.9/64.2]	11.9 [4.4/18.3]	9.1 [2.9/13.7]
	Upstream	11.8 [−19.5/43.8]	9.3 [−7.2/53.9]	9.0 [−13.7/35.5]	6.6 [−4.8/37.7]
	Downstream	9.3 [−28.0/64.3]	10.5 [−19.8/54.7]	6.6 [−19.5/48.4]	7.0 [−13/36.2]
	Upstream + Downstream	9.3 [−15.4/36.3]	6.7 [−10.7/35.5]	5.2 [−19.7/30.6]	4.6 [−7.1/23.6]
Intron	Nonsynonymous	54.0 [25.5/74.7]	45.4 [13.7/67.4]	15.3 [6.8/22.6]	10.6 [3.5/16.3]
	Upstream	12.6 [−8.3/42.1]	9.3 [−8.3/40.1]	10.8 [−6.7/37.1]	7.3 [−6.2/31.9]
	Downstream	9.6 [−16.5/38.8]	5.3 [−20.7/25]	7.8 [−13.3/32.1]	4.1 [−17.1/18.8]
	Upstream + Downstream	12.5 [−2.3/27.5]	4.6 [−10.9/17.6]	5.1 [−13.1/22.4]	3.5 [−8.3/13.8]

Our low point estimates of *a* for upstream and downstream regions are suggestive of weak positive selection operating in these regions compared with nonsynonymous sites. Although the confidence intervals for the estimates are very wide when we examine upstream and downstream regions individually, the confidence intervals for combined upstream and downstream regions exclude rates of adaptive substitution higher than ~36% on upstream and downstream regions of *M. m. castaneus*.

In sequencing diploid outbred individuals, regions between heterozygous indels can be problematical for SNP calling. Our data set contains <10% of such regions. If such regions are excluded, we obtain almost identical estimates for constraint, the parameters of the DFE and α (results not shown). Another consideration about our data set is that because the EGP sample is not random (Livingston et al. 2004), we might have excluded genes that have high rates of adaptive evolution in upstream and downstream regions. For example, promoter regions of many neural and nutrition-related genes in humans have been found to be subject to positive selection (Haygood et al. 2007). However, a comparison of estimates of α for regions ~500 bp upstream and downstream of the start and stop codon of protein-coding genes in humans, obtained with the methodology employed in the present study and using the EGP and PGA (Program for Genomic Applications; Akey et al. 2004) data sets, showed nonsignificant differences between data sets (Eyre-Walker and Keightley 2009). Additionally, a comparison of estimates of α, obtained with the methodology employed in the present study and using the EGP, PGA, and Boyko et al. (2008) data sets, has shown non significant differences between data sets for nonsynonymous sites in humans (Halligan et al. 2010).

It has been suggested that regulatory noncoding regions might be more important for evolution than protein-coding genes in primates (King and Wilson 1975). However, studies that have used a simple extension of the MK test (Keightley et al. 2005) and the methodology employed in this study (Eyre-Walker and Keightley 2009) have estimated that α in upstream and downstream regions in humans is close to zero. Humans might have low rates of adaptive substitution in upstream and downstream regions because of their historically low N_e. However, in the current study, we also obtain low estimates for α for upstream and downstream regions in *M. m. castaneus*, a mammalian species with a N_e much larger than humans (Halligan et al. 2010). The low estimate of α in upstream and downstream regions in *M. m. castaneus* may be due to the sparse distribution of regulatory elements in the mammalian genome. Therefore, the upstream and downstream sequences we have focused on could include a substantial amount of neutral sequence along with some functionally relevant elements.

In order to control for differences between site classes in the contribution of slightly deleterious mutations to the observed divergence, we calculated the ratio of adaptive to neutral divergence (ω_α), and we obtained similar estimates for nonsynonymous, upstream, and downstream site classes (~5–10%). Therefore, upstream and downstream regions of protein-coding genes in *M. m. castaneus* appear to have a similar absolute rate of adaptive substitution with nonsynonymous sites. This finding implies that the difference in the estimate of α observed at nonsynonymous sites between humans (~0–20%) and *M. m. castaneus* (~50%) might also be due to differences in the relative proportion of slightly deleterious mutations between the two species.

Finally, if noncoding regulatory elements are distributed over many thousands of base pairs in the mammalian genome, then the net input of adaptive substitutions to regulatory regions of mammals could be higher than protein-coding genes. Eyre-Walker and Keightley's (2009) study in humans and our study in *M. m. castaneus* only examined ~500 bp upstream and downstream of the start and stop codon, respectively, of a limited collection of protein-coding genes. We suggest that genome-wide studies of putative regulatory noncoding regions are needed in *M. m. castaneus* and humans so that the role of regulatory regions to adaptation can be more confidently ascertained.

Supplementary Material

Supplementary figures S1 and S2 and supplementary tables S1–S5 are available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

We thank Bettina Harr for providing the *M. m. castaneus* samples. We would also like to thank Bettina Harr, Adam Eyre-Walker, Paul Sharp, and three anonymous referees for providing insightful comments on the manuscript. We

acknowledge funding from grants from the Biotechnology and Biological Sciences Research Council (BBSRC) and the Wellcome Trust. A.K. is funded by a BBSRC postgraduate studentship.

References

Akey JM, Eberle MA, Rieder MJ, Carlson CS, Shriver MD, Nickerson DA, Kruglyak L. 2004. Population history and natural selection shape patterns of genetic variation in 132 genes. *PLoS Biol.* 2:e286.

Andolfatto P. 2005. Adaptive evolution of non-coding DNA in Drosophila. *Nature* 437:1149–1152.

Arndt PF, Petrov DA, Hwa T. 2003. Distinct changes of genomic biases in nucleotide substitution at the time of mammalian radiation. *Mol Biol Evol.* 20:1887–1896.

Bierne N, Eyre-Walker A. 2004. The genomic rate of adaptive amino acid substitution in Drosophila. *Mol Biol Evol.* 21:1350–1360.

Boyko AR, Williamson SH, Indap AR, et al. 2008. Assessing the evolutionary impact of amino acid mutations in the human genome. *PLoS Genet.* 4:e1000083.

Bray N, Pachter L. 2004. MAVID: constrained ancestral alignment of multiple sequences. *Genome Res.* 14:693–699.

Charlesworth J, Eyre-Walker A. 2006. The rate of adaptive evolution in enteric bacteria. *Mol Biol Evol.* 23:1348–1356.

Eyre-Walker A. 2006. The genomic rate of adaptive evolution. *Trends Ecol. Evol (Amst).* 21:569–575.

Eyre-Walker A, Keightley PD. 1999. High genomic deleterious mutation rates in hominids. *Nature* 397:344–347.

Eyre-Walker A, Keightley PD. 2009. Estimating the rate of adaptive molecular evolution in the presence of slightly deleterious mutations and population size change. *Mol Biol Evol.* 26:2097–2108.

Fay JC, Wu C. 2001. The neutral theory in the genomic era. *Curr Opin Genet Dev.* 11:642–646.

Fay JC, Wyckoff GJ, Wu C. 2001. Positive and negative selection on the human genome. *Genetics* 158:1227–1234.

Foxe JP, Dar V, Zheng H, Nordborg M, Gaut BS, Wright SI. 2008. Selection on amino acid substitutions in Arabidopsis. *Mol Biol Evol.* 25:1375–1383.

Gaffney DJ, Keightley PD. 2006. Genomic selective constraints in murid noncoding DNA. *PLoS Genet.* 2:e204.

Gossmann TI, Song B, Windsor AJ, Mitchell-Olds T, Dixon CJ, Kapralov MV, Filatov DA, Eyre-Walker A. 2010. Genome wide analyses reveal little evidence for adaptive evolution in many plant species. *Mol Biol Evol.* 27:1822–1832.

Gray NK, Wickens M. 1998. Control of translation initiation in animals. *Annu Rev Cell Dev Biol.* 14:399–458.

Haddrill PR, Bachtrog D, Andolfatto P. 2008. Positive and negative selection on noncoding DNA in Drosophila simulans. *Mol Biol Evol.* 25:1825–1834.

Halligan DL, Oliver F, Eyre-Walker A, Harr B, Keightley PD. 2010. Evidence for pervasive adaptive protein evolution in wild mice. *PLoS Genet.* 6:e1000825.

Haygood R, Fedrigo O, Hanson B, Yokoyama K, Wray GA. 2007. Promoter regions of many neural- and nutrition-related genes have experienced positive selection during human evolution. *Nat Genet.* 39:1140–1144.

Ihaka R, Gentleman R. 1996. R: a language for data analysis and graphics. *J Comput Graph Stat.* 5:299–314.

Keightley PD, Eyre-Walker A. 2007. Joint inference of the distribution of fitness effects of deleterious mutations and population demography based on nucleotide polymorphism frequencies. *Genetics* 177:2251–2261.

Keightley PD, Gaffney DJ. 2003. Functional constraints and frequency of deleterious mutations in noncoding DNA of rodents. *Proc Natl Acad Sci U S A.* 100:13402–13406.

Keightley PD, Lercher MJ, Eyre-Walker A. 2005. Evidence for widespread degradation of gene control regions in hominid genomes. *PLoS Biol.* 3:e42.

Kimura M. 1980. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *J Mol Evol.* 16:111–120.

King MC, Wilson AC. 1975. Evolution at two levels in humans and chimpanzees. *Science* 188:107–116.

Kohn MH, Fang S, Wu C. 2004. Inference of positive and negative selection on the 5′ regulatory regions of Drosophila genes. *Mol Biol Evol.* 21:374–383.

Li WH. 1993. Unbiased estimation of the rates of synonymous and nonsynonymous substitution. *J Mol Evol.* 36:96–99.

Livingston RJ, von Niederhausern A, Jegga AG, Crawford DC, Carlson CS, Rieder MJ, Gowrisankar S, Aronow BJ, Weiss RB, Nickerson DA. 2004. Pattern of sequence variation across 213 environmental response genes. *Genome Res.* 14:1821–1831.

McDonald JH, Kreitman M. 1991. Adaptive protein evolution at the Adh locus in Drosophila. *Nature* 351:652–654.

Pamilo P, Bianchi N. 1993. Evolution of the Zfx and Zfy genes: rates and interdependence between the genes. *Mol Biol Evol.* 10:271–281.

Sawyer S, Kulathinal R, Bustamante C, Hartl D. 2003. Bayesian analysis suggests that most amino acid replacements in Drosophila are driven by positive selection. *J Mol Evol.* 57:S154–S164.

Shabalina SA, Spiridonov NA. 2004. The mammalian transcriptome and the function of non-coding DNA sequences. *Genome Biol.* 5:105.

Shapiro JA, Huang W, Zhang C, et al. (12 co-authors). 2007. Adaptive genic evolution in the Drosophila genomes. *Proc Natl Acad Sci.* 104:2271–2276.

Slotte T, Foxe JP, Hazzouri KM, Wright SI. 2010. Genome-wide evidence for efficient positive and purifying selection in Capsella grandiflora, a plant species with a large effective population size. *Mol Biol Evol.* 27:1813–1821.

Smith NG, Eyre-Walker A. 2002. Adaptive protein evolution in Drosophila. *Nature* 415:1022–1024.

Tajima F. 1983. Evolutionary relationship of DNA sequences in finite populations. *Genetics* 105:437–460.

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

The Chimpanzee Sequencing and Analysis Consortium. 2005. Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature* 437:69–87.

Veyrieras J, Kudaravalli S, Kim SY, Dermitzakis ET, Gilad Y, Stephens M, Pritchard JK. 2008. High-resolution mapping of expression-QTLs yields insight into human gene regulation. *PLoS Genet.* 4:e1000214.

Watterson GA. 1975. On the number of segregating sites in genetical models without recombination. *Theor Popul Biol.* 7:256–276.

Welch JJ. 2006. Estimating the genomewide rate of adaptive protein evolution in Drosophila. *Genetics* 173:821–837.

Xie X, Lu J, Kulbokas EJ, Golub TR, Mootha V, Lindblad-Toh K, Lander ES, Kellis M. 2005. Systematic discovery of regulatory motifs in human promoters and 3[prime] UTRs by comparison of several mammals. *Nature* 434:338–345.

Zhang L, Li W. 2005. Human SNPs reveal no evidence of frequent positive selection. *Mol Biol Evol.* 22:2504–2507.

INVESTIGATION

A Comparison of Models to Infer the Distribution of Fitness Effects of New Mutations

Athanasios Kousathanas[1] and Peter D. Keightley
Institute of Evolutionary Biology, School of Biological Sciences, University of Edinburgh, Edinburgh EH9 3JT, United Kingdom

ABSTRACT Knowing the distribution of fitness effects (DFE) of new mutations is important for several topics in evolutionary genetics. Existing computational methods with which to infer the DFE based on DNA polymorphism data have frequently assumed that the DFE can be approximated by a unimodal distribution, such as a lognormal or a gamma distribution. However, if the true DFE departs substantially from the assumed distribution (*e.g.*, if the DFE is multimodal), this could lead to misleading inferences about its properties. We conducted simulations to test the performance of parametric and nonparametric discretized distribution models to infer the properties of the DFE for cases in which the true DFE is unimodal, bimodal, or multimodal. We found that lognormal and gamma distribution models can perform poorly in recovering the properties of the distribution if the true DFE is bimodal or multimodal, whereas discretized distribution models perform better. If there is a sufficient amount of data, the discretized models can detect a multimodal DFE and can accurately infer the mean effect and the average fixation probability of a new deleterious mutation. We fitted several models for the DFE of amino acid-changing mutations using whole-genome polymorphism data from *Drosophila melanogaster* and the house mouse subspecies *Mus musculus castaneus*. A lognormal DFE best explains the data for *D. melanogaster*, whereas we find evidence for a bimodal DFE in *M. m. castaneus*.

NEW mutations generate genetic variation in the genome of every species. For example, it has been estimated that a newborn human has ~70 new mutations that originated in its parents' germlines (Keightley 2012). The fitness effects of new mutations can range from deleterious to neutral and to advantageous, and the relative frequencies of their effects is known as the distribution of fitness effects (DFE) of new mutations. Inferring the properties of the DFE is a long-standing goal of evolutionary genetics and is key to several important questions, including the evolution of sex and recombination, the prevalence of Muller's ratchet, and the constancy of the molecular clock (Charlesworth 1996; Eyre-Walker and Keightley 2007).

A number of methodologies have been developed to infer the DFE based on DNA sequence data (Sawyer *et al.* 2003; Nielsen and Yang 2003; Piganeau and Eyre-Walker 2003; Loewe *et al.* 2006; Eyre-Walker *et al.* 2006; Keightley and Eyre-Walker 2007; Boyko *et al.* 2008; Schneider *et al.* 2011; Wilson *et al.* 2011). All of these assume that there is a neutrally evolving class of sites and contrast patterns of polymorphism and/or divergence from an outgroup with that of a tightly linked focal site class. Selection affecting the focal sites is expected to alter the pattern of polymorphism compared to that of the neutral class. A distribution of selection coefficients is then fitted to the data and its properties inferred. The three most widely used methods are those developed by Eyre-Walker *et al.* (2006), Keightley and Eyre-Walker (2007), and Boyko *et al.* (2008). Keightley and Eyre-Walker (2007) use a Wright–Fisher transition-matrix approach (Ewens 1979), whereas Eyre-Walker *et al.* (2006) and Boyko *et al.* (2008) use a diffusion approximation (Sawyer and Hartl 1992; Williamson *et al.* 2005). All three methods have been reported to give similar results, but make slightly different assumptions. For example, they differ in the way in which they model demographic changes (*e.g.*, population size changes). Eyre-Walker *et al.* (2006) use a heuristic approach, whereas the other two explicitly model some simple demographic scenarios. It is necessary to model demographic change, because this is known to alter patterns of polymorphism in ways that can resemble selection. Because these methods use allele-frequency information (summarized as the

doi: 10.1534/genetics.112.148023
Manuscript received November 26, 2012; accepted for publication January 12, 2013
Supporting information is available online at http://www.genetics.org/lookup/suppl/doi:10.1534/genetics.112.148023/-/DC1.
[1]Corresponding author: West Mains Rd., Edinburgh EH9 3JT, Scotland. E-mail: a.kousathanas@sms.ed.ac.uk

site-frequency spectrum or SFS), they are expected to be sensitive to demographic change.

Several studies have employed the above methods to infer properties of the DFE of amino acid-changing mutations. In these analyses, a gamma distribution of fitness effects has often been assumed, since it is a flexible distribution with two parameters, the shape (b) and the scale (a). For example, for amino acid-changing mutations in *Drosophila melanogaster*, the shape parameter has been estimated to be ~0.4 (implying a leptokurtic distribution), and most (>90%) new mutations are inferred to be moderately to strongly deleterious, with effective strength of selection $N_e s > 10$ (Keightley and Eyre-Walker 2007; Eyre-Walker and Keightley 2009). In humans, the DFE appears to be more even more leptokurtic than in *Drosophila* (*i.e.*, the estimated shape parameter is ~0.2), and only ~60% of mutations appear to be moderately to strongly deleterious (Eyre-Walker *et al.* 2006; Keightley and Eyre-Walker 2007; Boyko *et al.* 2008; Eyre-Walker and Keightley 2009). Differences between *Drosophila* and humans in the properties of the DFE have been attributed to a difference in their effective population size (N_e), the former being at least 2 orders of magnitude larger (Eyre-Walker *et al.* 2002). An effect attributable to N_e has also been observed in several other species. For example, N_e in wild house mice is substantially larger than humans but smaller than *Drosophila*, and ~70–80% of amino acid mutations are estimated to be moderately to strongly deleterious (Halligan *et al.* 2010; Kousathanas *et al.* 2011). *Capsella grandiflora* and *Aribidopsis thaliana* are two plant species with large and small N_e, respectively, and ~86% and ~66% of amino acid mutations are estimated to be moderately to strongly deleterious, respectively (Foxe *et al.* 2008; Slotte *et al.* 2010).

Most of the above methods assume that the DFE can be approximated by a certain type of mathematical distribution, such as the gamma distribution. One would like, however, to have a more general approach to obtain information about the DFE without needing to assume an explicit distribution. Steps in this direction were taken by Keightley and Eyre-Walker (2010), who examined a model of multiple discrete selection coefficients rather than assuming a continuous distribution. However, Keightley and Eyre-Walker (2010) did not examine the performance of their models when the true distribution deviated from a gamma distribution. Boyko *et al.* (2008) also fitted several types of distributions and combinations of continuous distributions and discrete fixed effects when inferring the DFE for amino acid-changing mutations in humans. Wilson *et al.* (2011) recently developed a new method that assumes a series of discrete fixed selection coefficients, the density associated with each selection coefficient estimated as a parameter. However, due to the complexity of the model, Wilson *et al.* (2011) needed to assume constant population size.

Although several different types of parametric and nonparametric DFE models have been fitted to DNA polymorphism data, to our knowledge their performance in cases where the true DFE is bimodal or multimodal has not previously been investigated. In this study, we use simulations to examine cases in which the true DFE is unimodal, bimodal, or multimodal. We analyze simulated data assuming six models for the DFE. The first two are parametric unimodal distributions: the lognormal and the gamma distribution. The third model is a parametric distribution that can be bimodal: the beta distribution. The fourth model is a discrete point mass distribution of selection coefficients where the locations and the probability densities of each point mass (or "spikes") are estimated parameters. We refer to this model as the spikes model, which is similar to the discretized model used by Keightley and Eyre-Walker (2010). The fifth model ("steps" model') consists of multiple continuous, uniform distributions (or steps), the boundaries and probability densities of which are estimated parameters. The sixth model is a variant of the model used by Wilson at al. (2011) and assumes six fixed selection coefficients where only their probability densities are estimated parameters. We refer to this model as the "fixed six-spikes" model. We use simulations to test the performance of the six models assuming various scenarios for the complexity of the true DFE. We go on to fit the six models to protein polymorphism data sets from *D. melanogaster* and *Mus musculus castaneus*, each containing sequences of several thousand protein-coding genes.

Materials and Methods

Population genetic model and assumptions

In this study, we extend the methods developed by Keightley and Eyre-Walker (2007) to infer the DFE of new mutations based on the allele frequency distribution of polymorphic nucleotide sites among individuals sampled from a population. This approach is based on Wright–Fisher population genetics theory and makes a number of assumptions. We assume that sites are unlinked and have the same mutation rate and that polymorphic *sites are biallelic*. We assume that there are two classes of sites in the genome, one "neutral" and one "selected." The fates of new mutations in the neutral class are affected only by genetic drift. New mutations at selected sites are assumed to be unconditionally deleterious and to have additive effects on fitness. We define the selection coefficient s as the fitness reduction experienced by the homozygote for the mutant allele compared to the homozygote for the wild-type allele. Therefore, the fitnesses of the wild-type, heterozygote, and mutant homozygote are 1, $1 - s/2$ and $1 - s$, respectively.

Description of the modeled distributions of selection coefficients

New mutations affecting the selected class of sites are sampled from a probability distribution. We investigated six models for this probability distribution: the first is a lognormal distribution, which has two parameters: the mean or location (μ) and the standard deviation or scale

(σ). The second is a gamma distribution, which has two parameters: the shape (b) and the scale (a). The third model is the beta distribution, which has two shape parameters (k_1, k_2). The fourth model (spikes model) assumes m mutational effects classes (spikes), which are modeled as point masses. For each mutational effect class i ($i = 1...m$), the location s_i and the probability density (p_i) are estimated parameters, for a total of $2m - 1$ parameters. The fifth model (steps model) assumes m mutational effects classes, and each class i ($i = 1...m$) is modeled as a uniform distribution where the minimum and maximum values ($N_e s_{i-1}$ and $N_e s_i$, respectively) and the probability density (p_i) are estimated parameters. The minimum value of the first step is fixed to zero. We assume that the start of each step is the end of the previous, that is, for step i, $N_e s_i = N_e s_{i-1}$, ensuring that there are no overlapping steps. The total number of parameters to be estimated is m for the minimum and maximum values values of the steps plus $m - 1$ for the probability density of each step, giving a total of $2m - 1$ parameters. The sixth model (fixed six-spikes) assumes six mutational effects classes (spikes), modeled as point masses arbitrarily fixed at $N_e s_1 = 0$, $N_e s_2 = 1$, $N_e s_3 = 5$, $N_e s_4 = 10$, $N_e s_5 = 50$, $N_e s_6 = N_e$. The probability densities of the fixed point masses are estimated parameters, for a total of five parameters.

Demographic model

Following Keightley and Eyre-Walker (2007), we also incorporate a simple demographic model of a step change from population size N_1 to population size N_2 at some time t in the past. N_1 is fixed at 100, the parameter t is estimated relative to N_2, and the parameter N_2 is estimated relative to N_1 (*i.e.*, the magnitude of the size change is estimated). There may be little information with which to estimate the relative values of N_1 and N_2 so we also compute a weighted recent effective population size N_w,

$$N_w = \frac{N_1 w_1 + N_2 w_2}{w_1 + w_2}, \tag{1}$$

where $w_1 = N_1(1-1/2N_2)^t$ and $w_2 = N_2(1-e^{-t/(2N_2)})$ (Eyre-Walker and Keightley (2009)). We also incorporate a parameter f_0, which is the proportion of unmutated sites. Under selective neutrality and stationary equilibrium, $1 - f_0$ is proportional to the product of the mutation rate and the persistence time of a new mutation.

Generation of the expected allele-frequency vector and computation of likelihood

We assume that at some point in the past, a population of size N_1 was at mutation–selection–drift equilibrium. This population then experienced a size change (either expansion or contraction) to size N_2 t generations from the present. Throughout this period, new mutations arise, which are neutral for the neutral class of sites and deleterious with selection coefficients s sampled from a probability distribution $f(s)$ for the selected class. Following Keightley and Eyre-Walker (2007), we employ Wright–Fisher transition matrix methods to generate the expected allele frequency distribution at the present time for a set of parameter values f_0, t, N_2, and a given s value, and we store it in vector $\mathbf{v}(s)$. The lognormal, gamma, spike, and step distributions can potentially have substantial parts of their density at $s > 1$. We modeled the contribution of mutations for $s > 1$ assuming that their frequency in the population goes down in proportion to the expectation at mutation–selection balance, following Keightley and Eyre-Walker (2007). The expected mean allele-frequency distribution $\mathbf{z}$ is obtained by integrating over the distribution of selection coefficients for all elements of $\mathbf{v}(s)$,

$$z = \int_0^\infty v(s) f(s|\Theta)\, ds, \tag{2}$$

where Θ represents the parameters of the distribution of selection coefficients (*e.g.*, a and b for the gamma distribution).

The numbers of derived alleles in a sample of n_T alleles constitute the SFSs and are stored in vectors $\mathbf{q}(N)$ and $\mathbf{q}(S)$ for the selected and neutral sites, respectively. Numbers of alleles are binomial draws from a diploid population of size N_2. Since we do not distinguish between the derived and ancestral states, we use only folded SFSs. We fold the SFS and the allele-frequency vector $\mathbf{z}$ as follows:

$$q_i = q_i + q_{n_T - i}, \quad \text{for } 0 \le i < n_T/2 \tag{3}$$
$$z_i = z_i + z_{2N-i}, \quad \text{for } 1 \le i \le 2N_2/2 \tag{4}$$

Under the assumption that numbers of derived alleles are binomially distributed, we compute the log likelihood of the observed allele frequency distributions (*i.e.*, SFSs) for neutral and selected sites as

$$\log L = \sum_{i=0}^{n_T/2} q_i \log\left(\sum_{j=0}^{N_2} z_j (b\langle i|n_T, j/2N_2\rangle + b\langle n_T - i|n_T, j/2N_2\rangle)\right) \tag{5}$$

(Keightley and Eyre-Walker 2007), where $b\langle i|n, p\rangle$ is the binomial probability for i derived alleles in a sample of n alleles with probability of occurrence p. We find the set of the parameter values that best fits the observed SFSs by maximizing the sum of the log likelihoods calculated for the neutral and selected classes of sites.

Likelihood maximization

The parameters to be estimated are f_0, N_2, t, plus additional parameters, depending on the selection model implemented (Table 1). Maximization of the likelihood was done using a custom likelihood search algorithm for N_2, and the SIMPLEX algorithm (Nelder and Mead 1965) for the remaining parameters. To increase the speed of the maximization procedure, we first estimated the demographic parameters N_2 and t and the parameter f_0 from the neutral SFS. We assumed the maximum likelihood (ML) estimates of N_2 and t when estimating the parameters from the selected SFS.

Table 1 The selection models investigated in this study

DFE Model	No. Parameters	Parameters
Lognormal	2	μ, σ (location, scale)
Gamma	2	a, b (scale, shape)
Beta	2	k_1, k_2 (shape 1, shape 2)
Spike	$2m - 1$	For i ($i = 1..m$), $N_e s_i$
		For i ($i = 1..m - 1$), p_i
Step	$2m - 1$	For i ($i = 1..m$) $N_e s_i$
		For i ($i = 1..m - 1$), p_i
Six-fixed spikes	5	For i ($i = 1...5$), p_i

We generated starting values for the location parameters of the spikes and the steps by using a power series,

$$\text{for spike or step } i(i = 1 \ldots m), \quad N_e s_i = N_e^{(i/m-r)}, \qquad (6)$$

where $N_e = N_w$ as calculated by Equation 1 and r is a pseudorandom deviate from a normal distribution with a mean 0 and standard deviation 0.1. This power series was devised empirically and has several desirable properties: the term $N_e{}^{i/m}$ places the spikes or steps at a reasonable distance from each other; the last spike or step is placed at N_e, therefore avoiding generating extremely large $N_e s$ values; the pseudorandom normal deviate r adds noise in the placement of the spikes/steps.

The starting values for the relative probability densities of the steps were set to $1/m$. As the number of parameters increases, the possibility of multiple local maxima also increases. To ensure that the global maximum had been found, we performed 10 starts of the maximization algorithm for each run, each time using a different seed for the pseudorandom number generator. We recorded the ML estimates that gave the highest likelihood in these runs.

Implementation of the model

Our simulations used a forward Wright–Fisher simulator to generate SFSs and we then used ML to fit demographic and selection models and estimate the parameters. This was implemented in a recoded version of the C program DFE-alpha (Eyre-Walker and Keightley 2009). This version implements all of the models we describe, can be used to analyze SFS data sets in a similar way to DFE-alpha, and will be made available via the authors' website.

Simulations assuming a constant population size

We simulated SFS data sets assuming a diverse set of distributions of selection coefficients, including unimodal, bimodal, and multimodal distributions. We performed simulations in which we assumed a constant population size ($N_1 = N_2 = 100$). We used 10^6 neutral and 10^6 selected sites and sampled 64 alleles. Parameter f_0 was set to 0.9. We also compared simulations in which we assumed different numbers of sequenced alleles (8, 16, 32, 64, 128, and 256), while assuming a set number of sites (10^6). For each simulated data set, we performed 100 replicate simulations.

Simulations assuming variable population size

We modeled population size changes as step changes from an initial population of size $N_1 = 100$ at stationary equilibrium. Time is expressed in units of N_1. We simulated two demographic histories: a population expansion and a bottleneck. The simulated expansion was a step change to size N_2 ($N_2/N_1 = 3.1$), at time $t_2/N_1 = 1$. The simulated bottleneck was a reduction in population size $N_2/N_1 = 0.72$ at time $t_2/N_1 = 1.1$ and a subsequent expansion with a step change in size $N_3/N_1 = 3.8$ at time $t_3/N_1 = 0.11$. The parameters for the two simulated demographic scenarios were chosen to match the inferred histories of real populations. The simulated expansion matches that inferred for a population of wild mice (Halligan *et al.* 2010) and for the American population of humans with African ancestry (Boyko *et al.* 2008). The bottleneck scenario matches that inferred for the American population of humans with European ancestry (Boyko *et al.* 2008). For these simulations we assumed a gamma DFE with $a = 0.05$ and $b = 0.5$. For each simulated data set we used 10^6 neutral and 10^6 selected sites, sampled 64 alleles, and performed 20 replicate simulations.

Simulations with linkage

We used C++ program *SLiM*, developed by Philip Messer and available at http://www.stanford.edu/~messer/software.html to perform simulations with linkage (Messer 2013). We simulated 1-Mbp-long chromosomes. Each chromosome had 20 loci. Each locus consisted of 10 exons of length 100 bp each alternating with 1-kbp introns. The loci were at a distance of 40 kbp from each other. We used exonic sites and the first 100 bp of introns as selected and neutral sites respectively. We simulated a population of size $N = 100$ for $10N$ generations to reach stationary equilibrium and sampled 64 chromosomes every $2N$ generations for $100N$ generations to obtain polymorphism data for a total of 10^6 selected and 10^6 neutral sites. We assumed a mutation rate $4N_e\mu = 1\%$ and simulated various levels of linkage between sites by assuming recombination rates ($4N_e r$) varying between 10^{-5} and 1. We performed three types of simulations, varying the properties of the DFE for selected sites: First, we assumed a gamma DFE ($a = 0.05$, $b = 0.5$), second we assumed that 97% of sites were under negative selection (gamma DFE; $a = 0.05$, $b = 0.5$) and 3% were under positive selection (single spike DFE; $N_e s_1 = 10$), and third we assumed a bimodal DFE consisting of two spikes of selection coefficients ($N_e s_1 = 0$, $N_e s_2 = 10$, $p1 = 0.2$). We performed 20 replicate runs for each simulation type.

Evaluation of model performance

We are interested in knowing how well the mean effect ($\overline{N_e s}$), the mean fixation probability of a new deleterious mutation relative to a neutral mutation ($\bar{u}$), and the proportion of mutations falling into five $N_e s$ categories (0.0–0.1, 0.1–1.0, 1.0–10.0, 10.0–100.0, >100.0) are estimated. $\overline{N_e s}$ and $\bar{u}$ are important quantities for several questions,

including inferring the proportion of mutations fixed by positive selection and the rate of adaptive relative to neutral evolution (*i.e.*, α and ω_α, respectively; Eyre-Walker and Keightley 2009; Gossmann *et al.* 2010). $\overline{N_e s}$ was calculated by taking the arithmetic average of the selection coefficients over the range of s between 0 and 100 (*i.e.*, the $N_e s$ range was between 0 and 10^4, for $N_e = 100$). $\bar{u}$ was calculated by integrating over the DFE, as in Eyre-Walker and Keightley (2009),

$$\bar{u} = \int_0^\infty 2N_e u(N_e, s) f\langle s|\Theta\rangle ds, \tag{7}$$

where $u(N_e, s)$, is the fixation probability of a new deleterious mutation (Fisher 1930; Kimura 1957, 1962).

To assess the accuracy in recovering the properties (X) of the simulated distributions, we compared estimates (X_i) *vs.* true values (X_{true}). For $\bar{N}_e s$ and $\bar{u}$, we calculated the relative error as

$$\text{rel.error}(X) = \frac{X_i - X_{true}}{X_{true}}. \tag{8}$$

We compared the goodness of fit between models by comparing their likelihoods and by comparing Akaike information criterion (AIC) scores. The AIC score penalizes parameter-rich models as

$$\text{AIC} = 2k - 2\log(L), \tag{9}$$

where k is the number of parameters in the model, and L is the maximum likelihood for the estimated model. We considered an AIC difference >2 as significant when comparing models. For the spike/step models we increased the number of fitted spike/steps until an improvement of <2 AIC units was obtained.

Drosophila and house mouse data sets

We analyzed polymorphism data for protein-coding genes of *D. melanogaster* and *M. m. castaneus* using the six approaches described above. We also fitted a simple demographic model of a step change in population size. For *D. melanogaster*, we analyzed a data set of 17 genomes from individuals originating in East Africa (haploid Rwanda lines from the Drosophila Population Genomics Project (DPGP; release v. 2.0, http://www.dpgp.org/dpgp2/DPGP2.html; Pool *et al.* 2012). The data set was compiled as in Campos *et al.* (2012), but we did not use a minimum quality cut-off. It included polymorphism data for 8367 autosomal genes orthologous between *D. melanogaster* and *D. yakuba*. For *M. m. castaneus*, we used a data set of 20 genomes from individuals sampled in northwest India (Halligan *et al.* 2010; D.L. Halligan, A. Kousathanas, R.W. Ness, H. Li, B. Harr, L. Eory, T. M. Keane, D. J. Adams, P. D. Keightley, unpublished data). The data set included polymorphism data for 18,671 autosomal genes orthologous between *M. m. castaneus* and rat. CpG dinucleotides have substantially higher mutation rates in mammals (Arndt *et al.* 2003) and their frequencies differ between coding and noncoding DNA. Therefore for *M. m. castaneus*, we restricted the analysis to non-CpG-prone sites (sites not preceded by C or followed by G). To calculate α and ω_a we used the divergences at nonsynonymous and synonymous sites between *D. melanogaster* and *D. yakuba* and between *M. m. castaneus* and rat, as follows,

$$\alpha = \frac{d_N - d_S \bar{u}}{d_N}, \tag{10}$$

$$\omega_a = \frac{d_N - d_S \bar{u}}{d_S}, \tag{11}$$

where d_N and d_S are the nucleotide divergences between the focal species and the outgroup at nonsynonymous and synonymous sites, respectively.

Results

We simulated SFS data sets, choosing the parameters of the simulated distributions to create three different scenarios for their complexity (*i.e.*, unimodality, bimodality, and multimodality). We also aimed at generating distributions that were biologically plausible. We then examined the performance of several models incorporating parametric or nonparametric distributions. We considered four main criteria for evaluating the performance of the tested models: the log-likelihood score, the accuracy in estimating the mean effect of a new mutation ($\overline{N_e s}$), the accuracy in estimating the average fixation probability of a new mutation ($\bar{u}$), and the accuracy in estimating the proportion of mutations in five $N_e s$ ranges. Estimates for the parameters of each of the six tested models for each simulation set (SIM1, SIM2, SIM3) are shown in Supporting Information, Table S1.

A gamma distribution simulated (SIM1)

To approximate a realistic scenario for protein-coding loci, where current information suggests a leptokurtic DFE and most sites under strong negative selection, we simulated a gamma DFE with scale $a = 0.05$ and shape $b = 0.5$ (SIM1; Figure 1). As expected, the gamma model gave the best fit to the data, accurately estimating $\overline{N_e s}$ (SIM1; Table 2). The lognormal model performed poorly, overestimating $\overline{N_e s}$ and underestimating $\bar{u}$, while the beta model gave a good fit (ΔAIC from the best-fitting model was -0.5) and accurately estimated $\overline{N_e s}$ and $\bar{u}$ (SIM1; Figure 2, A and B, respectively). Based on their AIC scores, the best-fitting variable spike and variable steps models were the two-spike and two-step models, respectively (SIM1; Table 2), and these models fitted only slightly worse than the gamma model. However they did not recover $\overline{N_e s}$ and $\bar{u}$ as accurately as the gamma (SIM1; Figure 2, A and B, respectively). All models tested performed well in accurately recovering the proportions of mutations in the $N_e s$ ranges we examined

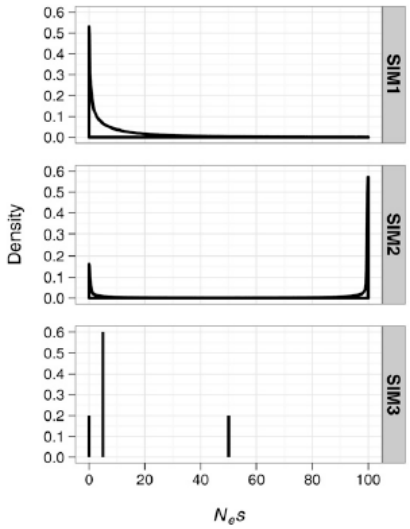


Figure 1 The simulated DFEs. For SIM1, we simulated a gamma DFE with scale $a = 0.05$ and shape $b = 0.5$. For SIM2, we simulated a beta DFE with shape parameters $k_1 = 0.2$ and $k_2 = 0.1$ scaled to the $N_e s$ interval [0, 100]. For SIM3, the DFE was composed of three selection coefficients, $N_e s_1 = 0$, $N_e s_2 = 5$, $N_e s_3 = 50$, with probability densities $p_1 = 0.2$, $p_2 = 0.6$, $p_3 = 0.2$.

(Figure 3). However, the lognormal and all the nonparametric models did not succeed in accurately assigning the proportions of mutations in the $N_e s$ ranges 0.0–0.1 and 0.1–1.0, presumably because there is little information to discriminate between these categories. In contrast, the gamma and beta models performed almost perfectly in assigning the proportions of mutations to these categories.

A bimodal beta distribution simulated (SIM2)

We then investigated a beta distribution with shape parameters $k_1 = 0.2$ and $k_2 = 0.1$ and scaled to the $N_e s$ interval [0, 100] (SIM2; Figure 1). For this distribution, ~10% of selected sites are under weak negative selection ($N_e s < 1$), another 10% are under moderately strong negative selection ($N_e s = 1$–10), and the remaining 80% are under very strong negative selection ($N_e s > 10$). Such a bimodal distribution is intended to model protein-coding loci where amino-acid changing mutations are either neutral or strongly deleterious, with relatively few mutations of intermediate effect. As expected, the beta model had the best AIC score (SIM2; Table 2), recovering $\overline{N_e s}$ and $\bar{u}$ accurately (SIM2; Figure 2, A and B, respectively). The unimodal lognormal and gamma models fitted the data very poorly (ΔAIC from beta = −597.2 for the lognormal and −89.9 for the gamma, SIM2; Table 2). $\overline{N_e s}$ was grossly overestimated by the lognormal and gamma models (SIM2; Figure 2A). However, $\bar{u}$ was estimated relatively accurately by these models (SIM2; Figure 2B). The estimate for $\overline{N_e s}$ can be heavily influenced by a long tail in the fitted distribution whereas $\bar{u}$ is mostly

Table 2 Goodness-of-fit statistics for the models tested for each simulation set

Simulation	Model	$\Delta\log L$	ΔAIC
SIM1 (gamma)	Lognormal	−13.9	−27.8
	Gamma	−0.02	0.0
	Beta	−0.3	−0.5
	Best spike (2)	−1.5	−4.9
	Best step (2)	0.0	−2.0
	Six-fixed spikes	−0.6	−7.1
SIM2 (bimodal beta)	Lognormal	−300.0	−597.2
	Gamma	−46.4	−89.9
	Beta	−1.4	0.0
	Best spike (3)	0.0	−3.1
	Best step (2)	−1.3	−1.8
	Six-fixed spikes	−3.5	−10.2
SIM3 (three-spike multimodal)	Lognormal	−29.5	−53.0
	Gamma	−6.9	−7.8
	Beta	−8.2	−10.4
	Best spike (3)	0.0	0.0
	Best step (3)	−0.7	−1.3
	Six-fixed spikes	−0.6	−1.3

The statistics reported are the mean log-likelihood and the mean AIC score difference from the highest scoring model over 100 simulation replicates. A sequencing effort of 64 alleles and 10^6 neutral and selected sites were assumed. Only results for the best-fitting spike and step model, based on the AIC criterion, are shown.

affected by effects in the $N_e s$ range 0–1. Therefore, the low accuracy of $\overline{N_e s}$ estimates from the lognormal and gamma models presumably reflects a bad fit to the "strong effects" part of the distribution (*i.e.*, $N_e s > 10$), but there is a reasonably good fit to the "nearly neutral effects" part of the distribution (*i.e.*, $0 < N_e s < 1$). The best-fitting three-spike and two-step models and the fixed six-spike model fitted almost as well as the beta distribution (SIM2; Table 2). These nonparametric models accurately estimated $\overline{N_e s}$ and $\bar{u}$ (SIM2; Figure 2, A and B, respectively). We observed that the lognormal, gamma, and nonparametric models assigned substantial proportions of mutations into the $N_e s$ >100 range (Figure 3), although the simulated distribution had a near-zero density in this range. Presumably, there is little information with which to precisely estimate the upper limit of the simulated distribution.

We also examined the performance of the models when varying the locations of the modes of a bimodal DFE. We investigated distributions with two classes of effects (two spike): The first class of mutations was assumed to be neutral with $N_e s_1 = 0$, and we varied the selection strength and probability density associated with the second class ($N_e s_2$ and p_2, respectively). We then fitted the gamma and the three-step models to these distributions and compared their performance. In Figure 4A we show the $\Delta\log L$ between the three-step and gamma models for different combinations of values for $N_e s_2$ and p_2. We found that for two-spike distributions, where $N_e s_2 \geq 10$ and $p_2 \geq 0.4$, the three-step model significantly outperformed the gamma model (Figure 4A). Additionally, we examined the performance of the

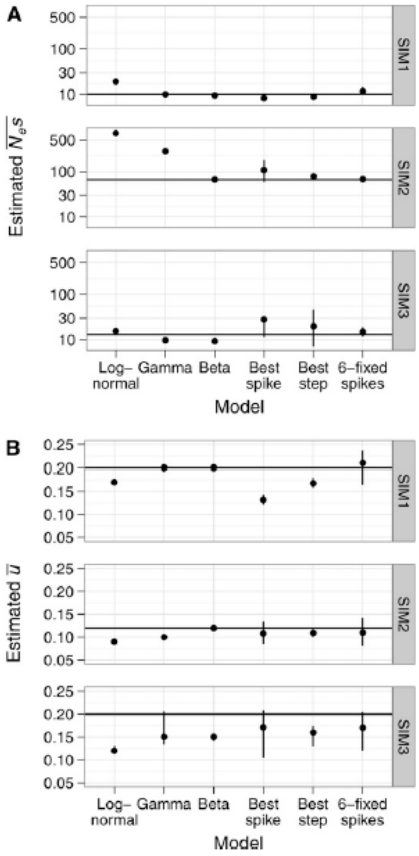


Figure 2 Summary statistics for the models tested for each simulation set. (A) Mean estimates of the mean effect of a new mutation ($\overline{N_e s}$) and (B) the probability of fixation of a new mutation ($\bar{u}$). Error bars are the 5th and 95th percentiles of estimates over 100 simulation replicates. The horizontal lines represent the simulated values. Only results for the best-fitting spike and step model, according to the AIC criterion, are shown. The *y*-axis is log scaled for panel A.

models in estimating $\overline{N_e s}$ and $\bar{u}$. We found that the gamma model overestimated $\overline{N_e s}$ when $N_e s_2 \geq 10$ and underestimated $\bar{u}$ for almost all parameter combinations of $N_e s_2$ and p_2 (Figure 4, B and C, respectively), while the three-step model overestimated $\overline{N_e s}$ and underestimated $\bar{u}$ when $N_e s_2 < 10$ (Figure 4, B and C, respectively).

A three-spike multimodal distribution simulated (SIM3)

To examine a case in which the true DFE is more complex, we simulated a DFE comprising three selection coefficients, $N_e s_1 = 0$, $N_e s_2 = 5$, $N_e s_3 = 50$, with probability densities $p_1 = 0.2$, $p_2 = 0.6$, $p_3 = 0.2$, respectively (SIM3; Figure 1). The choice of parameters was mainly based on generating three sufficiently distinct modes. As expected, a three-spike model gave the best fit according to the AIC criterion (SIM3; Table 2). The other nonparametric models fitted almost equally well (ΔAIC was -1.3 for both the three-step model and the fixed six-spike model, SIM3; Table 2). However, the lognormal, gamma and beta models gave a poorer fit than the nonparametric models (ΔAIC was -53, -7.8, and -10.4 for the lognormal, gamma, and beta models, respectively, SIM3; Table 2). However, we did not observe large differences in the accuracy of estimating $\overline{N_e s}$ and $\bar{u}$ between the models tested (SIM3; Figure 2, A and B, respectively). The lognormal, best spike, best step, and fixed six-spike models slightly overestimated $\overline{N_e s}$, whereas the gamma and beta models slightly underestimated $\overline{N_e s}$ (SIM3; Figure 2A). All models tested slightly underestimated $\bar{u}$ (SIM3; Figure 2B).

The effect of increasing the allele sequencing effort

The primary goal of this section was to examine whether the general trends in the performance of the six models tested hold for different allele sequencing efforts. We compared the performance of the models for 8, 16, 32, 64, 128, and 256 alleles sequenced. For the gamma distribution (SIM1), increasing the sequencing effort led to more accurate estimates of $\overline{N_e s}$ for all models (SIM1; Figure S1A). Accuracy of estimating $\bar{u}$ improved only marginally (SIM1; Figure S1B). For the beta distribution (SIM2), increasing the allele sequencing effort increased the accuracy of estimating $\overline{N_e s}$ (SIM2; Figure S1A), but the accuracy of estimating $\bar{u}$ did not increase for the spike, step, and fixed six-spike models and surprisingly decreased for the lognormal and gamma models (SIM2; Figure S1B). This decrease can be explained if we consider that the overall fit of the gamma and lognormal models improves as the number of alleles sequenced is increased, but the fit of the models to the $N_e s$ range 0–1 worsens (the good fit of the models to the $N_e s$ range 0–1 is crucial for an accurate estimate of $\bar{u}$). For the three-spike multimodal distribution (SIM3), we observed that the parametric lognormal, gamma, and beta models showed no improvement in accuracy for estimating $\overline{N_e s}$ and $\bar{u}$ when increasing the number of alleles sequenced (SIM3; Figure S1A and Figure S1B, respectively). The spike, step, and fixed six-spikes models at low sequencing efforts (8–32 alleles) had an inferior performance compared to the parametric models (SIM3; Figure S1A and Figure S1B). However, as the number of alleles sequenced was increased to 64 or greater, the performance of these models became superior to the parametric models (SIM3; Figure S1A and Figure S1B).

The effect of incorporating a population size change

We then examined whether population size changes can affect the performance of the nonparametric relative to the parametric models by simulating two population histories: an expansion and a bottleneck. The expansion was a threefold step change in population size. The bottleneck was a long-lasting 30% reduction in population size, followed by a short-lived fourfold step expansion. For the selected sites, we assumed a gamma DFE with scale $\alpha = 0.05$ and shape

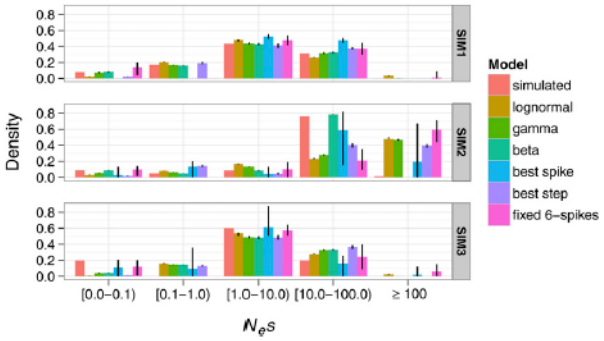


Figure 3 The mean estimated proportions of mutations in five $N_e s$ ranges for SIM1, SIM2, and SIM3. We assumed a sequencing effort of 64 alleles and 10^6 neutral and selected sites. Error bars are the 5th and 95th percentiles of estimates over 100 simulation replicates.

$b = 0.5$ (as for SIM1). Since our method can incorporate a model of a step change in population size, we fitted this model to the neutral data for both simulated histories. For the expansion scenario, the demographic parameters of the step change were accurately estimated and the performance of the different selection models was similar to SIM1 (Table S2). For the bottleneck scenario, the two-epoch demographic model appeared to mostly capture the second change in population size (Table S2). However, the nonparametric two-spike and two-step selection models fitted the data better than the parametric models (Table S2). Therefore, a long-lasting bottleneck followed by rapid expansion can produce a signal in the data that is not fully accounted for by the fitted two-step demographic scenario and can cause the spike and step models to overfit the data and produce spurious evidence for multimodality. Other population histories such as a bottleneck followed by long-lasting recovery or expansion gave similar results to the two-step expansion scenario (result not shown).

The effect of linkage and selection

In our simulations we have assumed that sites are unlinked, but genomes of real organisms can exhibit various amounts of linkage. We performed simulations assuming a range of recombination rates between sites to examine how linkage can affect the performance of the three-step model in detecting a bimodal DFE. This performance is assessed by a significantly better fit of the three-step model than the gamma model.

First, we investigated whether background selection alone could produce a spurious signature of a bimodal DFE by simulating a gamma DFE with $a = 0.05$ and $b = 0.5$. We observed a better fit of the three-step model than the gamma model for high levels of linkage (Figure S1C, top). However, when we fitted a demographic model of a step change to the neutral sites, a procedure that has been suggested to control for the effects of linkage (Messer and Petrov 2012), the three-step and gamma models fitted the data equally well at all levels of linkage (Figure S1C, bottom).

Second, we examined whether positive selection could produce a signature of a bimodal DFE. We simulated a gamma DFE with $a = 0.05$ and $b = 0.5$ for negatively selected mutations and a single spike for positively selected mutations with selection strength $N_e s_a = 10$ and probability density $p_a = 0.03$, which is similar to what has been observed for protein-coding genes in *D. melanogaster* (Schneider *et al.* 2011). We observed very similar results to those we obtained by assuming only negative selection (Figure S1D). Therefore fitting a demographic model to the neutral sites is essential for controlling the effects of linkage in producing spurious evidence of a bimodal DFE.

Third, we investigated whether linkage could affect our power to detect a multimodal DFE with the nonparametric steps model. We simulated a bimodal two-spike DFE with $N_e s_1 = 0$, $N_e s_2 = 10$ with probability densities $p_1 = 0.2$, $p_2 = 0.8$, respectively. We found that strong linkage can reduce the $\Delta \log L$ between three-step and gamma models (Figure S1E, top). The results were similar when we also fitted a demographic model of a step change to the neutral sites (Figure S1E, bottom). Therefore, a true bimodal DFE would be harder to detect in genomic regions that exhibit strong linkage.

Analysis of protein polymorphism data sets from D. melanogaster and M. m. castaneus

To account for demographic effects on our inferences of selection we fitted a step change in population size to synonymous sites. The step-change model inferred a population expansion for both *D. melanogaster* and *M. m. castaneus* (Table S3) and fitted very well to the data (Figure S2). We then fitted the lognormal, gamma, beta, variable spike, variable step, and fixed six-spike models to nonsynonymous sites. For each data set, we computed $\Delta \log L$, ΔAIC scores, the proportions of mutations falling into four $N_e s$ ranges (0–1, 1–10, 10–100, >100), $\overline{N_e s}$, and $\bar{u}$ (Table 3).

For *D. melanogaster*, we found that the best-fitting model according to the AIC citerion was the lognormal model, the gamma model having a slightly worse fit (ΔAIC from the lognormal was −5.1 units; Table 3). However, the estimated proportion of mutations in the examined $N_e s$ ranges, $\overline{N_e s}$ and $\bar{u}$, were very similar between these two models (Table 3). All models estimate that ~2–7% of new mutations are

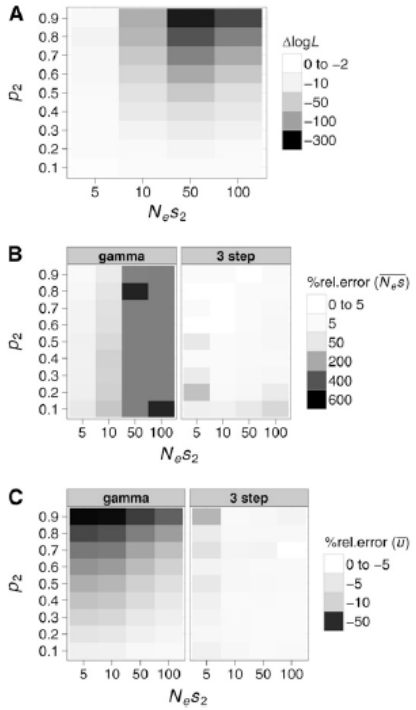


Figure 4 The performance of the gamma and three-step models when fitted to bimodal DFEs. We simulated two-spike DFEs with one spike fixed at $N_e s_1 = 0$ and we varied the selection strength ($N_e s_2$) and probability density (p_2) of the second spike. (A) $\Delta \log L$ between the three-step and gamma models fitted to the simulated DFEs as a function of $N_e s_2$ and p_2. We also compared the % rel. error in estimating (B) $\overline{N_e s}$ and (C) $\bar{u}$. Positive and negative values of % rel. error signify overestimation and underestimation of these parameters, respectively.

nearly neutral ($N_e s$ 0–1), a further ~4–20% are moderately to strongly deleterious ($N_e s$ 1–100), and ~80–90% are very strongly deleterious ($N_e s$ >100). The beta and six-fixed spike models gave a substantially poorer fit than the lognormal model (ΔAIC to lognormal was −187 units; Table 3). The main discernible difference was a ~10 times lower estimated $\overline{N_e s}$ for the beta and fixed six-spikes models than the lognormal model. The beta and fixed six-spike models do not allow selection strength $N_e s > N_e$ and their poor fit may be a consequence of a substantial proportion of mutational effects lying in that range.

For *M. m. castaneus*, the best-fitting model according to the AIC criterion was the three-spike model (Table 3). The estimated parameter values were $N_e s_1 = 2.3 \times 10^{-12}$, $N_e s_2 = 16.4$, $N_e s_3 = 1056$, with probability densities $p_1 = 0.19$, $p_2 = 0.12$, $p_3 = 0.69$, respectively (Table S3). The fixed six-spike, two-step, and beta models fitted only slightly worse than the three-spike model, while the lognormal and gamma models had substantially worse fits (Table 3). The parameter estimates of the three-spike model together with the good fit of the beta model support a bimodal DFE in *M. m. castaneus*. The DFE is inferred to have a peak at near neutrality ($N_e s$ 0–1) of density ~20%, and another peak at very strongly deleterious to lethal effects ($N_e s > 100$) with density ~70% (Table 3). Intermediate effects ($N_e s$ 1–100) are inferred to have a density of ~10% (Table 3).

The average fixation probability of a new deleterious mutation ($\bar{u}$) is an important quantity, since it can be used to estimate the fraction of adaptive substitutions between two species (Eyre-Walker and Keightley 2009). We calculated α and ω_a (Equations 10 and 11) by using the estimated $\bar{u}$ for each model (Table 3). For *D. melanogaster*, we obtained values of α in the range 0.47–0.7 and ω_a 0.063–0.1 from the different models (Table 3). For *M. m. castaneus*, the lognormal and the gamma models gave slightly lower estimates for $\bar{u}$ and therefore higher estimates for α and ω_a (0.30 and 0.070, respectively; Table 3) than the best-fitting three-spike model (0.20 and 0.047, respectively; Table 3).

Discussion

In this study, we have examined the performance of several models incorporating parametric and nonparametric distributions for inferring the properties of the DFE. Since the true DFE is of unknown complexity and can have multiple modes, our purpose was to examine the performance of the different models when the true DFE was unimodal, bimodal, or multimodal. We investigated parametric distributions, including the unimodal lognormal and gamma distributions, which are widely used to model the DFE, and the beta distribution, which can also take a bimodal shape. We also examined the performance of custom nonparametric models, including discretized distributions, where the selection coefficients are modeled as point masses, or uniform distributions, that are either variable or fixed. Spike or step models with two or more classes of effects performed almost as well as the gamma model for cases in which the true DFE was a gamma distribution. When the true DFE was a bimodal beta distribution, we found that the lognormal and gamma models fitted poorly and produced inaccurate estimates of $\overline{N_e s}$, $\bar{u}$, and the density in several $N_e s$ ranges, most notably mutations with $N_e s > 100$. When we simulated a more complex DFE, the biases affecting estimates of $\overline{N_e s}$ and $\bar{u}$ from the lognormal and gamma models were not as pronounced. Accuracy in estimating $\overline{N_e s}$ and $\bar{u}$ seems to depend mostly on the density of the extreme tails of the DFE, irrespectively of its complexity. In our simulations, we frequently observed that a particular model could have a good overall fit, but perform relatively poorly for parts of the DFE that are crucial for estimating $\overline{N_e s}$ or $\bar{u}$. For example, we consistently observed that $\bar{u}$ was not estimated with high accuracy if the models fitted were different from that simulated. Presumably, the SFS contains limited information about mutations with very small selective effects in the $N_e s$ range 0–1

Table 3 Results from the analysis of protein-coding loci in *D. melanogaster* and *M. m. castaneus*

				$N_e s$							
Species	Model	Δ log L	ΔAIC	[0-1)	[1-10)	[10-100)	≥100	$\overline{N_e s}$	$\bar{u}$	α	ω_a
D. melanogaster	Lognormal	−0.8	0.0	0.044	0.064	0.11	0.78	1359.2	0.050	0.62	0.082
	Gamma	−3.3	−5.1	0.049	0.055	0.12	0.78	1624.1	0.054	0.59	0.079
	Beta	−94.2	−187.0	0.064	0.025	0.043	0.87	94.6	0.066	0.50	0.067
	Best spike (3)	0.0	−4.5	0.063	0.00	0.10	0.84	275.2	0.063	0.52	0.069
	Best step (2)	−3.2	−7.0	0.023	0.097	0.058	0.82	289.4	0.039	0.70	0.10
	six-fixed spikes	−72.3	−144.6	0.070	0.00	0.048	0.88	96.8	0.070	0.47	0.063
M. m. castaneus	Lognormal	−23.9	−41.8	0.17	0.052	0.061	0.72	1298.9	0.16	0.30	0.070
	Gamma	−21.2	−36.4	0.17	0.050	0.065	0.71	1840.1	0.16	0.29	0.069
	Beta	−4.4	−2.9	0.18	0.016	0.022	0.78	141.2	0.18	0.22	0.052
	Best spike (3)	0.0	0.0	0.19	0.00	0.12	0.69	755.4	0.19	0.20	0.047
	Best step (2)	−2.8	−1.6	0.18	0.0098	0.10	0.71	237.4	0.19	0.20	0.047
	Six-fixed spikes	−2.9	−5.8	0.19	0.0053	0.02	0.78	142.6	0.19	0.20	0.046

Log-likelihood and AIC score differences from the highest scoring model, estimated proportion of mutations falling into four $N_e s$ ranges, estimated mean effects of a new mutation ($\overline{N_e s}$), estimated mean probability of fixation of a new mutation ($\bar{u}$), and estimates of α and ω_a are shown. Only results for the best-fitting spike and step models, based on the AIC criterion, are shown.

implying that estimation of $\bar{u}$ strongly depends on the properties of the distribution assumed. Since $\bar{u}$ can be used for calculating the proportion of adaptive substitutions (α) and the rate of adaptive evolution (ω_a), underestimation of $\bar{u}$ would lead to overestimation of α and ω_a (and *vice versa*). When we examined a series of bimodal DFEs in which we varied the locations and densities of the two modes of the DFE, we observed substantial underestimation of $\bar{u}$ by the gamma model for cases where one mode of the DFE was at $N_e s = 0$ with density <30% and the other mode was at a weakly to moderately deleterious effect with density >70%. Therefore, if the true DFE is bimodal, underestimation of $\bar{u}$ by the gamma model would be expected for genomic regions in which most of the sites are under selection, such as protein-coding genes or conserved noncoding elements, but not for genomic regions in which most of the sites are evolving neutrally such as UTRs and introns.

We also applied the parametric and nonparametric models to infer the DFE for amino acid-changing mutations in *D. melanogaster* and the house mouse *M. m. castaneus*, based on data from several thousand autosomal protein-coding genes. In *D. melanogaster*, we found that the lognormal model gave the best fit to the data, a result that is consistent with a previous study (Loewe and Charlesworth 2006). The estimate for $\overline{N_e s}$ was 1360 by the best-fitting lognormal model. This estimate is similar to estimates obtained from a smaller data set of Shapiro *et al.* (2007) analyzed by Keightley and Eyre-Walker (2007). If we assume that the DFE for amino acid-changing mutations in *Drosophila* is lognormal and that N_e is of the order 0.7×10^6 (Halligan *et al.* 2010), then the mean selection coefficient of new deleterious amino-acid changing mutations for *D. melanogaster* is of the order 2×10^{-3}. We also estimate that α and ω_a are 0.62 and 0.082, respectively. Reassuringly, the choice of the distribution to model the DFE does not strongly affect $\bar{u}$ and consequently α and ω_a. Regardless of the model assumed, $\alpha > 0.47$ *and* $\omega_a > 0.063$, supporting the presence of highly effective positive selection in *D. melanogaster*, as several other researchers have inferred (Sella *et al.* 2009).

In *M. m. castaneus*, we found that a three-spike model gave the best fit to the SFS. The beta distribution also fitted almost as well as the three-step model, while the lognormal and gamma models gave substantially poorer fits. These observations suggest that the DFE for new deleterious amino-acid changing mutations in *M. m. castaneus* is bimodal, with 20% of the distribution's density attributable to weakly deleterious mutations ($N_e s$ 0–1) and 70% to very strongly deleterious mutations ($N_e s > 100$). We also obtained estimates for α and ω_a, of 0.20 and 0.046, respectively. We observed differences among the estimates of α and ω_a between different models, the lognormal and gamma models producing higher estimates than the best-fitting three-spike and beta models. Underestimation of $\bar{u}$ by the gamma and lognormal models was observed in simulations in which the true DFE was a bimodal beta of similar properties to the inferred DFE for *M. m. castaneus*. It seems likely that fitting a lognormal or a gamma distribution to the DFE leads to overestimation of α and ω_a. Halligan *et al.* (2010), who fitted a gamma distribution to a small gene sample from *M. m. castaneus*, obtained estimates for α larger (α = 0.37 for non-CpG-prone sites and using rat as outgroup) than those obtained in the present study.

There are some potential caveats to our study. First, our models do not incorporate genetic linkage in the inference method. We investigated whether linkage and background or/and positive selection can affect inferences from the models tested and found that under moderate linkage, spurious evidence for multimodality can be produced (assessed by a better fit of spike/step models to data than unimodal distributions). We can account for the effects of linkage, however, by fitting a simple demographic model to the neutral class of sites (as is also suggested by Messer and

Petrov 2012). Second, our two-epoch demographic model is not sufficient for more complex demographic histories, such as bottlenecks. Assuming a more realistic population history of a long-lasting bottleneck followed by a rapid expansion, we found that the spike/step models can overfit the data, producing spurious evidence for multimodality of the DFE. Therefore, when inferring the DFE using spike/step models it is necessary to fit a three-epoch model to data from populations that have experienced bottlenecks. A three-epoch model can be incorporated into the inference procedure of our method, but due to computational limitations it was not feasible to investigate its performance in simulations. However, a three-epoch model fitted only slightly better to the folded synonymous SFS for *D. melanogaster* and *M. m. castaneus* than a two-epoch model ($\Delta \log L$ between the two-epoch and three-epoch model was 3 and 7, respectively; result not shown). Therefore, we do not expect a substantial effect of the demographic history on our inferences of selection in these populations. Third, the fact that we infer a bimodal DFE for *M. m. castaneus* does not necessarily rule out a more complex DFE. It appears that there is limited information in the SFS, and our simulations indicate that at best three modes can be inferred, even for very large data sets. It is likely that the precise shape of the DFE cannot accurately be determined based on SFS data alone, as has been shown for the demographic history of a population (Myers *et al.* 2008).

In conclusion, we have shown that nonparametric discretized models, such as the spike and step models, can perform as well or better than parametric distributions, such as the gamma. They produce accurate estimates of the important parameters, notably $\overline{N_e s}$ and $\bar{u}$, and increasing the numbers of alleles sequenced will increase their performance. These models can also help in determining whether the DFE has multiple modes. We note that we have examined only one particular case of each type of distribution (unimodal, bimodal, multimodal) and we do not consider the particular simulated examples as representatives of all possible unimodal, bimodal, and multimodal distributions. However, our results are relevant in showing the limitations of fitting relatively inflexible distributions, such as the gamma distribution to the DFE, and illustrate the advantages of using a more general model such as the spike or step model to infer the DFE. Fitting the spike or the step model with different numbers of classes of mutational effects can be informative about the complexity of the DFE and identifying which $N_e s$ ranges we have little information on.

Acknowledgments

We thank Dan Halligan, Adam Eyre-Walker, Brian Charlesworth, Laurence Loewe, and two anonymous reviewers for helpful comments on earlier versions of the manuscript and for helpful discussions. We thank Jose Campos for compiling the DPGP2 protein-coding data. We acknowledge funding from grants from the Biotechnology and Biological Sciences Research Council (BBSRC) and the Wellcome Trust. A.K. is funded by a BBSRC postgraduate studentship.

Literature Cited

Arndt, P. F., D. A. Petrov, and T. Hwa, 2003 Distinct changes of genomic biases in nucleotide substitution at the time of mammalian radiation. Mol. Biol. Evol. 20: 1887–1896.

Boyko, A. R., S. H. Williamson, A. R. Indap, J. D. Degenhardt, R. D. Hernandez *et al.*, 2008 Assessing the evolutionary impact of amino acid mutations in the human genome. PLoS Genet. 4: e1000083.

Campos, J. L., K. Zeng, D. J. Parker, B. Charlesworth, and P. R. Haddrill, 2012 Codon usage bias and effective population sizes on the X chromosome *vs.* the autosomes in *Drosophila melanogaster*. Mol. Biol. Evol., http://mbe.oxfordjournals.org/content/early/2013/01/20/molbev.mss222.

Charlesworth, B., 1996 The good fairy godmother of evolutionary genetics. Curr. Biol. 6: 220.

Ewens, W. J., 1979 *Mathematical Population Genetics*. Springer-Verlag, Berlin.

Eyre-Walker, A., and P. D. Keightley, 2007 The distribution of fitness effects of new mutations. Nat. Rev. Genet. 8: 610–618.

Eyre-Walker, A., and P. D. Keightley, 2009 Estimating the rate of adaptive molecular evolution in the presence of slightly deleterious mutations and population size change. Mol. Biol. Evol. 26: 2097–2108.

Eyre-Walker, A., P. D. Keightley, N. G. C. Smith, and D. Gaffney, 2002 Quantifying the slightly deleterious mutation model of molecular evolution. Mol. Biol. Evol. 19: 2142–2149.

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

Fisher, R. A., 1930 *The Genetical Theory of Natural Selection*. Clarendon Press, Oxford.

Foxe, J. P., V.-N. Dar, H. Zheng, M. Nordborg, B. S. Gaut *et al.*, 2008 Selection on amino acid substitutions in *Arabidopsis*. Mol. Biol. Evol. 25: 1375–1383.

Gossmann, T. I., B.-H. Song, A. J. Windsor, T. Mitchell-Olds, C. J. Dixon *et al.*, 2010 Genome wide analyses reveal little evidence for adaptive evolution in many plant species. Mol. Biol. Evol. 27: 1822–1832.

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.

Keightley, P. D., 2012 Rates and fitness consequences of new mutations in humans. Genetics 190: 295–304.

Keightley, P. D., and A. Eyre-Walker, 2007 Joint inference of the distribution of fitness effects of deleterious mutations and population demography based on nucleotide polymorphism frequencies. Genetics 177: 2251–2261.

Keightley, P. D., and A. Eyre-Walker, 2010 What can we learn about the distribution of fitness effects of new mutations from DNA sequence data? Philos. Trans. R. Soc. B. Biol. Sci. 365: 1187–1193.

Kimura, M., 1957 Some problems of stochastic processes in genetics. Ann. Math. Stat., 882–901.

Kimura, M., 1962 On the probability of fixation of mutant genes in a population. Genetics 47: 713–719.

Kousathanas, A., F. Oliver, D. L. Halligan, and P. D. Keightley, 2011 Positive and negative selection on noncoding DNA close toprotein-coding genes in wild house mice. Mol. Biol. Evol. 28: 1183–1191.

Loewe, L., and B. Charlesworth, 2006 Inferring the distribution of mutational effects on fitness in Drosophila. Biol. Lett. 2: 426–430.

Loewe, L., B. Charlesworth, C. Bartolomé, and V. Nöel, 2006 Estimating selection on nonsynonymous mutations. Genetics 172: 1079–1092.

Messer, P. W., 2013 SLiM: simulating evolution with selection and linkage. *arXiv:1301.3109*. http://arxiv.org/abs/1301.3109.

Messer, P. W., and D. A. Petrov, 2012 The McDonald–Kreitman test and its extensions under frequent adaptation: problems and solutions. *arXiv:1211.0060*. http://arxiv.org/abs/1211.0060.

Myers, S., C. Fefferman, and N. Patterson, 2008 Can one learn history from the allelic spectrum? Theor. Popul. Biol. 73: 342–348.

Nelder, J. A., and R. Mead, 1965 A Simplex method for function minimization. Comput. J. 7: 308–313.

Nielsen, R., and Z. Yang, 2003 Estimating the distribution of selection coefficients from phylogenetic data with applications to mitochondrial and viral DNA. Mol. Biol. Evol. 20: 1231–1239.

Piganeau, G., and A. Eyre-Walker, 2003 Estimating the distribution of fitness effects from DNA sequence data: implications for the molecular clock. Proc. Natl. Acad. Sci. USA 100: 10335–10340.

Pool, J. E., R. B. Corbett-Detig, R. P. Sugino, K. A. Stevens, C. M. Cardeno *et al.*, 2012 Population genomics of sub-Saharan *Drosophila melanogaster*: African diversity and non-African admixture. PLoS Genet. 8: e1003080.

Sawyer, S., and D. L. Hartl, 1992 Population genetics of polymorphism and divergence. Genetics 132: 1161–1176.

Sawyer, S., R. Kulathinal, C. Bustamante, and D. Hartl, 2003 Bayesian analysis suggests that most amino acid replacements in *Drosophila* are driven by positive selection. J. Mol. Evol. 57: S154–S164.

Schneider, A., B. Charlesworth, A. Eyre-Walker, and P. D. Keightley, 2011 A method for inferring the rate of occurrence and fitness effects of advantageous mutations. Genetics 189: 1427–1437.

Sella, G., D. A. Petrov, M. Przeworski, and P. Andolfatto, 2009 Pervasive natural selection in the Drosophila genome? PLoS Genet. 5: e1000495.

Shapiro, J. A., W. Huang, C. Zhang, M. J. Hubisz, J. Lu *et al.*, 2007 Adaptive genic evolution in the Drosophila genomes. Proc. Natl. Acad. Sci. U.S.A 104: 2271–2276.

Slotte, T., J. P. Foxe, K. M. Hazzouri, and S. I. Wright, 2010 Genome-wide evidence for efficient positive and purifying selection in *Capsella grandiflora*, a plant species with a large effective population size. Mol. Biol. Evol. 27: 1813–1821.

Williamson, S. H., R. Hernandez, A. Fledel-Alon, L. Zhu, R. Nielsen *et al.*, 2005 Simultaneous inference of selection and population growth from patterns of variation in the human genome. Proc. Natl. Acad. Sci. USA 102: 7882–7887.

Wilson,, D. J., R. D. Hernandez, P. Andolfatto, and M. Przeworski, 2011 A population genetics–phylogenetics approach to inferring natural selection in coding sequences. PLoS Genet. 7: e1002395.

Communicating editor: S. I. Wright