University of Sussex

Investigating genome wide patterns of natural selection in eukaryotes

Toni Ingolf Goßmann

Submitted for the degree of Doctor of Philosophy

University of Sussex

August 2012

UNIVERSITY OF SUSSEX

Toni Ingolf Gossmann, Doctor of Philosophy

Investigating genome wide patterns of natural selection in eukaryotes

Summary

Mutations are the ultimate source of new genetic information and they can be neutral, harmful or beneficial. The ultimate fate of all mutations is either to be lost or to eventually become fixed in a population. In this thesis I investigate genome wide traces of natural selection in eukaryotes. I focus on the most common type of mutations, point mutations, in protein coding genes.

I investigated whether there is adaptive evolution in 11 plant species comparisons by applying an extension of the McDonald Kreitman (MK) test and found little evidence of adaptive evolution. However, most of the investigated plant species have low effective population sizes (N_e) and the rate of adaptive evolution is thought to be correlated to N_e. I therefore extended my study using additional data from mammals, drosophilids and yeast to investigate the relationship between the rate of adaptive evolution and N_e. I found a highly significant correlation between the rate of adaptive evolution relative to the rate of neutral evolution (ω_a) and N_e.

It has been proposed that evidence of adaptive evolution can be an artifact of fluctuating selection. I simulated a model of fluctuating selection, in which the average strength of selection acting upon mutations is zero. Under this model adaptive evolution is inferred using MK-type tests. However, the mutations which become fixed are on average positively selected. The signal of adaptive evolution is therefore genuine.

N_e can not only vary between species but also across genomes. However, how much variation there is, and whether this affects the efficiency of natural selection, is unknown. I analysed 10 species and show that variation in N_e is widespread. However, this variation is limited, amounting to a few fold variation in N_e between most genomic regions. This is never-the-less sufficient to cause variation in the efficiency of selection.

Acknowledgements

"People who solved things usually had lots of persistence and some good luck"

Charles Bukowski (1920-1994)

This thesis is dedicated to those who have always encouraged me to continue my way. I want to thank my family, friends and colleagues. In particular I want to thank Adam Eyre-Walker who has been a great supervisor and mentor over the last years as well as David Waxman who contributed his supervision even from China. I also want to thank the John Maynard Smith foundation and the University of Sussex for the financial support I have received.

It has been a pleasure to share my office with Alan, Falk, Nina, Maria, Tangjie, Romain, Pierre, Sebastian, Stephanie, Ying, Guilhemne, Guilhaume and Yann and all the others which I did not intend to forget but perhaps are not listed here for letting me have a good time in room 5B23. I also want to thank the Coca-Cola company for providing all those cans of coke which probably made me become a diabetes case.

Preface

The research presented here was carried out at the University of Sussex. Parts of this thesis have been submitted and accepted for scientific publication. Details are as follows:

Chapter 2

Gossmann, T. I., Song, B.-H., Windsor, A. J., Mitchell-Olds, T., Dixon, C. J., Kapralov, M. V., Filatov, D. A., and Eyre-Walker, A. (2010). Genome wide analyses reveal little evidence for adaptive evolution in many plant species. *Mol Biol Evol*, 27(8):1822–1832

Chapter 3

Gossmann, T. I., Keightley, P. D., and 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*

Chapter 5

Gossmann, T. I., Woolfit, M., and Eyre-Walker, A. (2011). Quantifying the variation in the effective population size within a genome. *Genetics*, 189(4):1389–1402

Contents

List of Tables

List of Figures

Chapter 1

General Introduction

In this general introduction I will briefly summarize the biological concepts of genome structure, genetic variation and effective population size. Secondly, I will review the variety of test statistics that have been developed to infer the role of selection in molecular evolution, mainly focusing on implementations of the McDonald Kreitman (MK) test statistic (McDonald and Kreitman, 1991). Thirdly, I will review the current state of genome wide estimates of the rate of adaptive evolution in eukaryotic species based on the MK test framework and will discuss them in relation to the effective population size.

1.1 Biological concepts

1.1.1 Eukaryotic genome structure

The eukaryotic genome is divided into chromosomes, which already visually, under the light microscope, show structural characteristics, most apparently the telomeres, centromeres and heterochromatin. However along the linear chromosomal DNA molecule further functional categories can be distinguished, such as genic and intergenic regions and within protein coding DNA, introns and untranslated regions (UTRs). In terms of molecular evolution large focus has been centred on the coding parts of the DNA, owed to the possibility of functional characterization of the corresponding proteins and conservation among species resulting in easier annotation and comparability between species. Mutations occurring within protein coding DNA can either alter the amino acid, a so called non-synonymous change, or leave the amino acid unchanged, a synonymous change. The fact that synonymous mutations do not change the amino acid resulted in the concept of neutrality or nearly neutrality of synonymous mutations (King and Jukes, 1969). The selective neutrality of synonymous mutations has been used to infer selective effects of

amino acid altering mutations (Nei and Gojobori, 1986; Hughes and Nei, 1988; McDonald and Kreitman, 1991; Goldman and Yang, 1994; Yang and Bielawski, 2000). In recent analyses intergenic regions, which also harbour functional and conserved elements, have also been in the focus of evolutionary analysis. However, in this thesis we will focus our studies on those parts of the genome that encode for proteins and assume that a selection of loci represents an unbiased sample of the whole protein coding genome of a particular species.

1.1.2 Mutations as the source on which selection can act on

Mutations are one of the most important genetic processes and they are ultimately the only source of new genetic information (Hodgkinson and Eyre-Walker, 2011). Mutations are introduced at the level of DNA and lead to genetic variation between cells, individuals and species (Charlesworth, 2010; Hodgkinson and Eyre-Walker, 2011). The rate at which mutations appear throughout the genome varies considerably at various scales (Lynch, 2010; Hodgkinson and Eyre-Walker, 2011). Point mutations are the most common type of mutation (Charlesworth, 2010). In a very general classification they can be distinguished as either advantageous, harmful or without any effect relative to the nonmutated type. The fate of a particular mutation depends on the evolutionary forces acting on it. A mutation which has arisen in the DNA and is contributing to the genetic material of the offspring starts segregating in the population. Ultimately it will either become fixed or lost from the population depending on a manifold of evolutionary processes including selection, random genetic drift, migration and biased gene conversion. A fundamental question in population genetics is the relative impact of such forces acting on mutations and thereby the determinants of the rate of molecular evolution. In this thesis I investigate the role of positive and negative natural selection acting on point mutations in protein coding parts of eukaryotic genomes.

1.1.3 Effective population size (N_e)

The size of a population (N) is a fundamental quantity in ecology and molecular evolution. However in terms of population genetics N may be misleading, because for a particular individual the genetic contribution to the next generation may be limited or biased. Factors that cause deviations from an ideal population are for example age structure, unequal sex ratios, inbreeding or spatial structure (for an extensive review see Charlesworth (2009)). Therefore, the concept of the effective population size (N_e) was introduced by Wright

(1931); this corresponds to the number of individuals in an ideal population that would yield the same variance in allele frequency or inbreeding coefficient. Usually $N_e << N$ and the difference would depend on the extent to which various factors affect the population size such as variation in the offspring number or episodes of low population size. In population genetics the quantity of N_e is important for two major reasons. First, the product of N_e and the mutation rate per generation μ is an estimate of the neutral diversity in a population (Kimura, 1991). Second, the product of N_e and the intensity of selection, s, of a mutation determines the effectiveness of selection. A mutation is effectively neutral when its intensity of selection is less than the inverse to the effective population size (Kimura, 1983; Woolfit, 2009). Therefore the effective population size is predicted to affect the rate of molecular evolution as well as diversity within a population.

1.2 Identification of past selection using divergence data

Considering that the probability of fixation is higher for advantageous mutations and smaller for harmful mutations one expects to see an excess and decrease, respectively, of such mutations relative to neutrality depending on the predominant type of selection acting on the newly arising mutations (Suzuki, 2010). In regions were purifying selection is the predominant evolutionary force acting on new mutations one would expect that amino acid changing mutations fix more rarely than synonymous changes and therefore contribute much less to the divergence between species. One can compare these two measurements by quantifying the non-synonymous per site divergence (d_n) to the synonymous per site divergence (d_s) and indeed for major parts of the protein coding genome $d_n << d_s$ (Li, 1997).

In order to detect recurrent positive selection one can restrict the analysis to shorter time scales, particular sites under investigation, or certain branches of the phylogeny. Determining the rate of divergence can be obtained by statistical frameworks such as Bayesian approaches, most parsimonious or maximum likelihood models (Nei and Kumar, 2000; Yang and Bielawski, 2000). Assuming that $d_s = 2\mu t$ and $d_n = 2\mu t f/(1-\alpha)$ where μ is the mutation rate per generation, t the time of divergence and f and α are the proportion of neutral and advantageous nonsynonymous mutations, respectively. Under such a model d_n can only exceed d_s if $f > (1-\alpha)$, e.g. if there is frequent adaptive evolution (Eyre-Walker, 2006). However the inferences of selection solely based on divergence data for relatively long time-scales results in rather inaccurate estimates of d_n and d_s (Ota and Nei, 1994;

Hahn, 2009). Furthermore the restriction to comparable sites and therefore more conserved parts of the genome maybe misleading, because the gene age is a major predictor of rates of evolution, e.g. in humans (Cai and Petrov, 2010). Challenges with divergence based inferences include biases in the mutation rate, inhomogeneous selective effects over time as well as linkage between sites and demographic changes. Usually underlying models assume constant N_e and s which is rather unrealistic over relatively long time scales (Charlesworth, 2009; Fay, 2011).

1.3 Identification of recent and ongoing selection using polymorphism data

Once a selectively neutral mutation starts segregating in a population its fate will be determined by random genetic drift. Under the assumption of a constant mutation rate there will be a steady state between mutations entering the population and those becoming fixed or getting lost resulting in a diversity within a species. Classical population theory suggests that this diversity for selectively neutral mutations is the product $4N_e\mu$ where N_e is the effective population size and μ the mutation rate per generation. Furthermore, the probability of observing a mutation at frequency p in the population is proportional to $1/p$ (Charlesworth, 2010), resulting in the neutral site frequency spectrum (SFS). Deviations from this neutral SFS are the result of other evolutionary forces acting on mutations, such as selection, biased gene conversation or demographic changes. There are test statistics that use the information of the deviation from the neutral model to infer the underlying evolutionary forces acting on mutations.

1.3.1 Tajima's D

For example Tajima (1989) developed a test statistic D which contrasts two measurements of within population diversity, the average number of nucleotide differences (Nei and Li, 1979) and the expected number of segregating sites (Watterson, 1975). A negative test value indicates an excess of rare variants which could be caused by population expansion or purifying selection. A positive value indicates a decrease of high and low frequency polymorphisms which can be caused by a population decrease or balancing selection. However, when assumptions of population equilibrium are strongly violated then exceptionally low D values could be an artifact resulting from sampling schemes and sampling sizes (Ptak and Przeworski, 2002).

1.3.2 Genetic hitchhiking and background selection

Another possibility to identify regions within the genome that deviate from the neutral model based on polymorphism data is to include information on the diversity of nearby sites. Under a model of positive selection one would expect genetic hitch-hiking because the increase in the frequency of the selected variant would be associated with an increase of nearby polymorphisms that are linked to it (Smith and Haigh, 1974). The extent of this linkage depends upon the rate of recombination, which means that if positive selection is acting at different parts of the genome it should result in lower diversity in regions with low rates of recombination (Begun and Aquadro, 1992). The fixation of the selected variant along with the linked neutral variants leads to a selective sweep (Berry et al., 1991). A similar phenomenon causes background selection, if purifying selection is the predominant selective force acting upon mutations, linked neutral variation will be purged from the population together with the selective variant (Charlesworth et al., 1993) resulting in a reduced diversity. However it is possible to discriminate between genetic hitchhiking and background selection by using the SFS. For example Fay and Wu (2000) developed a test statistic that uses the excess of high frequency variants as a signature of positive selection. However this excess is hard to detect if the selective sweep arose from standing genetic variation (Przeworski et al., 2005). Moreover such single locus test statistics are sensitive to perturbations from the null structure frequency distribution, especially when they lead to an increase in the allele frequency (Garrigan et al., 2010). A manifold of test statistics have been developed addressing different patterns which can be observed associated with selective sweeps, such as the variation of linkage disequilibrium (Kim and Nielsen, 2004; Stephan et al., 2006), the effect of demographic changes (Li and Stephan, 2006) and structured populations based on population differentiation (Beaumont and Balding, 2004; Riebler et al., 2008) or haplotypes frequencies in subpopulations (Sabeti et al., 2002; Tang et al., 2007; Sabeti et al., 2007). These methods have been successfully applied especially for human and Drosophila and can even locate the target of selection (Charlesworth, 2010; Stephan, 2010). However it has been noted that there is little overlap in the loci identified as being under positive selection using different methods (Akey, 2009; Mallick et al., 2009).

1.3.3 Distribution of fitness effects (DFE)

Mutations can be roughly classified as harmful, beneficial or neutral, but it is more realistic to assume a spectrum of selective effects (Kimura, 1983; Gillespie, 1991b). Such a distribution of fitness effects (DFE) provides insights to major question of evolutionary

biology such as the maintenance of genetic variation (Charlesworth et al., 1995), the evolution of sex and recombination (Peck et al., 1997) and the impact of effective population sizes (Charlesworth, 2009). In principle the selective effects of mutations can be directly investigated by mutation accumulation and mutagenesis experiments. However under such experimental conditions the selective effects have to be moderately strong which is unlikely for the majority of mutations (Davies et al., 1999). In contrast, DNA sequence data can be used to infer characteristics of the DFE, for example by contrasting patterns of diversity or divergence data of species with different effective population sizes (N_e) (Nielsen and Yang, 2003; Loewe and Charlesworth, 2006). Another approach is to fit a distribution of selective effects to the SFS (Eyre-Walker et al., 2006; Keightley and Eyre-Walker, 2007; Boyko et al., 2008; Schneider et al., 2011). Under such a model two sets of sites are taken into consideration, one which is assumed to be under selection and an additional category of sites which is assumed to behave effectively neutral (e.g. introns or synonymous sites) to estimate the mutation rate and control for the impact of demography. Although such methods assume free recombination, the violation of this assumption appears to be severe only if linkage is very strong (Boyko et al., 2008; Eyre-Walker and Keightley, 2009; Keightley and Eyre-Walker, 2010). One has to note that it is not clear whether the DFE is truly captured by a relatively simple distribution (e.g. gamma, log-normal, Eyre-Walker and Keightley, 2007). It is also difficult to infer the DFE for individuals genes unless they are very large and a substantial number of individuals has been sequenced (Keightley and Eyre-Walker, 2010). However this may become feasible in the near future due to advances in sequencing technology.

1.4 Tests to identify selection combining divergence and polymorphism data

Test statistics have been developed which contrast patterns of diversity and divergence to infer the impact of natural selection. Here I present arguably the two most important test statistics applied in population genetics with emphasise on the McDonald Kreitman (MK) test.

1.4.1 The HKA Test

Under the neutral theory it is predicted that the ratio of divergence and diversity should be constant, and any departure from this prediction may be interpreted as the consequence

of selection. The HKA test (Hudson et al., 1987) uses neutral diversity and divergence for a goodness-of-fit statistic for a number of loci. It assumes that all mutations considered are selectively neutral and no recombination occurs within loci and free recombination between loci. Therefore this test is rather conservative regarding the effect of recombination, however scenarios of demographic changes or changes in the mutations rate may also lead to a rejection of the neutral hypothesis (Hudson et al., 1987). In their analysis Hudson et al. (1987) interpreted the departure of the neutral model as the consequence of balanced polymorphisms when comparing diversity and divergence of the *Adh* locus and its 5' flanking region in multiple *Drosophila melanogaster* indiviuals and one *D. sechellia* individual (Strobeck, 1983; Hudson et al., 1987). Even though this initial analysis was restricted to two genomic regions the HKA test can, in principle, be applied to an arbitrary number of distinct loci. However there are two limitations associated with the test. First, if a certain genomic region under investigation cannot be classified *a priori* into discrete regions the HKA test is not applicable. This may be circumvented by using a sliding window approach to identify regions where expectation and observation deviate (McDonald, 1996, 1998). Second, the contribution of a particular locus to the departure of the neutral model cannot be determined. To circumvent this problem marginal distributions may be used or the results of pairwise HKA tests (Moore and Purugganan, 2003). Alternatively, especially when the number of investigated genomic regions is large, a maximum likelihood model based on the HKA framework may be used a test for selection at a specific locus (Wright and Charlesworth, 2004).

1.4.2 The McDonald Kreitman (MK) test

The HKA test framework is a relatively complicated test statistic and incorporates a number of assumptions. A simpler statistic based on the same principle as the HKA test has been developed by McDonald and Kreitman (1991). In principle it contrasts diversity with divergence at two categories of sites that are interspersed relative to one another. For example, in a coding region substitutions and polymorphisms can be classified as either nonsynonymous or synonymous depending on the impact they have on the amino acid they code for. One would expect that the ratio of nonsynonymous to synonymous substitutions (D_n/D_s) is equal to the ratio of nonsynonymous to synonymous polymorphisms (P_n/P_s) if protein evolution is a neutral process. This hypothesis can be tested with a G-test (Sokal and Rohlf, 1981) because (non)synonymous sites are interspersed with each other along the gene and share therefore the same evolutionary history and sampling scheme. Table 1.1

	Polymorphisms	Substitutions
Nonsynonymous	$P_n = 2$	$D_n = 7$
Synonymous	$P_s = 42$	$D_s = 17$

Table 1.1: Genetic diversity and divergence in the *Adh* locus of three *Drosophila* species (McDonald and Kreitman, 1991). For an MK-test the equality of the ratio of nonsynonymous to synonymous polymorphisms and the ratio of of nonsynonymous to synonymous substitutions is tested.

illustrates an application of the MK-test where diversity and divergence in three *Drosophila* species at the *Adh* locus are compared (McDonald and Kreitman, 1991). Of particular interest is a deviation from the neutral assumption when the number of nonsynonymous substitutions is increased, which is inferred as positive directional selection. In this case the proportion of adaptive substitutions can be quantified as (Fay et al., 2001; Smith and Eyre-Walker, 2002):

$$\alpha = 1 - \frac{D_s P_n}{D_n P_s} \tag{1.1}$$

There is also interest to estimate the amount of adaptive changes on a genome wide scale using multiple genes (Fay et al., 2001; Bustamante et al., 2002; Smith and Eyre-Walker, 2002). Firstly, for technical reasons data across loci need to be combined because single gene estimate have large variances (Fay et al., 2001; Bustamante et al., 2002; Bierne and Eyre-Walker, 2004). Secondly, there is evidence that α itself varies across the genome (Fay et al., 2002; Obbard et al., 2009; Wilson et al., 2011) and between species (Bustamante et al., 2002).

1.4.3 Fitness effects of new mutations within the MK test framework

Under the classical McDonald Kreitman test framework advantageous and deleterious mutations are assumed to have relatively strong selective effects and therefore are rarely observed in a sample of polymorphisms. This is because strongly deleterious mutations will be quickly removed by purifying selection and strongly advantageous mutations will either be lost or segregate quickly through the population and reach fixation. Only mutations that behave neutrally or effectively neutrally will contribute to both divergence and diversity and this effect can be quantified by the comparisons to neutral sites which are used as a reference. However it is unrealistic to assume that mutations are either strongly deleterious or strongly advantageous (Kimura, 1983; Gillespie, 1991b). Rather a continuous distribution of fitness effects between those two extremes is expected. For example there are multiple lines of evidence that slightly deleterious mutations substantially contribute to polymorphisms (Charlesworth and Eyre-Walker, 2007). Neglecting the contribution of this class of polymorphisms leads to an underestimation of the proportion of nonsynonymous substitutions due to positive selection. Fay et al. (2001) dealt with this problem by removing polymorphisms from the analysis that segregate at low frequencies, thereby focusing on common variants, because slightly deleterious mutations are expected to lead to an excess of rare variants due to weak purifying selection. However this approach is only reasonably satisfactory if the rate of adaptation is high and the distribution of fitness effects is very leptokurtic (Charlesworth and Eyre-Walker, 2008). Additionally, removing polymorphisms that segregate at low frequency will lead to smaller entries in the 2x2 contingency table within a classic McDonald Kreitman test. This results in a reduced statistical power especially when diversity is low and/or the gene is short. An alternative approach to correct for the effect of slightly deleterious mutations is to estimate the distribution of fitness effects and obtain the proportion of slightly deleterious mutations. This estimate can then be incorporated into a McDonald Kreitman test framework (Boyko et al., 2008; Eyre-Walker and Keightley, 2009). However it is difficult to estimate the DFE for single genes (Keightley and Eyre-Walker, 2010) and consequently reliable estimates have to be obtained from multiple loci. Therefore incorporating the effect of slightly deleterious mutations is only feasible for large datasets. Moreover, if the amount of adaptive divergence is substantial then there is the expectation that slightly advantageous mutations may contribute to polymorphism (Sawyer et al., 2003; Charlesworth and Eyre-Walker, 2008). The degree would depend on the rate and strength of advantageous mutation; unfortunately this is challenging to disentangle based on inferences from the

SFS (Boyko et al., 2008; Schneider et al., 2011).

1.4.4 Potential biases in genome wide scans for adaptive evolution

Even though the MK test has proven to be powerful tool for the inference of positive selection in molecular evolution it relies on a number of assumptions and suffers from certain limitations. These challenges include biological limitations such as the impact of population size changes (McDonald and Kreitman, 1991; Eyre-Walker et al., 2002; Eyre-Walker and Keightley, 2009) as well as technical biases such as lack of statistical power or sampling scheme biases. Here I give an overview of limitations of the MK-test framework and summarize methods which provide modified versions to address some of the raised issues.

The lack of statistical power of the MK-test associated with little divergence or diversity is known and α is even undefined if $D_n = 0$ or $P_s = 0$ (Equation 1.1). It is not uncommon to exclude loci with little information from genome wide analysis (Hughes et al., 2008), however such exclusions neglect a substantial part of the available information and can lead to dramatic biases (Stoletzki and Eyre-Walker, 2011). The extent to which this may bias the analysis in terms of α has been investigated (Stoletzki and Eyre-Walker, 2011). To obtain genome wide estimates of α it is possible to combine information of multiple genes in summary statistics (Fay et al., 2001; Smith and Eyre-Walker, 2002; Bustamante et al., 2002; Shapiro et al., 2007). However, summing up values of neutral divergence and diversity across genes in a classical MK test framework can give an overestimate of the rate of adaptive evolution if N_e is negatively correlated to the proportion of neutral amino acid substitutions (Smith and Eyre-Walker, 2002). There is evidence for such a scenario in *Drosophila* (Shapiro et al., 2007) where N_e is thought to vary across the genome (Begun and Aquadro, 1992) as a consequence of background selection and genetic hitchhiking.

Demographic changes would alter the outcome and may lead to artifactual evidence of adaptive evolution. For example in a scenario where a population increase is affecting only divergence (Eyre-Walker et al., 2002). Before a population expansion a higher proportion of mutations would be slightly deleterious and possibly contribute to divergence. However there would be no evidence of such mutations segregating currently in the population and consequently there would be an inflation of the rate of adaptation. Unfortunately it is difficult to correct for such an effect because the change in population size is not

reflected in the polymorphism dataset. It is expected that for species such as human and *Drosophila*, recent population growth has shaped the patterns of diversity (Williamson et al., 2005). A population growth would lead to excess of rare variants in the population similar to weak purifying selection (Tajima, 1989). In contrast to distant demographic changes, recent demographic changes can be inferred from the neutral diversity and can be incorporated into a MK-test framework (Williamson et al., 2005; Keightley and Eyre-Walker, 2007; Eyre-Walker and Keightley, 2009). However such models assume relatively simple demographic models, such as a step change, to correct for the effect of population size change.

The MK-test is relatively robust against deviations of nonequilibrium demography and recombinations rates (Andolfatto, 2008; Eyre-Walker and Keightley, 2009). This is because nonsynonymous and synonymous sites are interspersed within a gene and share therefore the same evolutionary history. However this maybe violated when the two contrasted categories are unlinked or distant (Andolfatto, 2005; Begun et al., 2007). Interestingly variation in the rate of recombination may even affect single genes, especially when they are long (Andolfatto, 2008). Therefore highly structured population samples or bottlenecked populations should be treated with caution when applied to an MK-test framework.

One of the severe assumptions of the MK-test framework is that a category of sites behaves effectively neutral, such as synonymous sites. However there is evidence that this may not be the case in some species (Ingvarsson, 2008a). It is relatively difficult to distinguish between the effects of demography and selection on synonymous sites. However computer simulations suggest if biased gene conversion is neglected and the assumed demographic model is of little complexity then selection on codon usage and demography can be disentangled (Zeng and Charlesworth, 2009). The MK-test also makes the assumption that sites are independent, thereby neglecting epistatic effects and/or temporal changes in the selective constraint which could alter the outcome of the MK-test (Fay, 2011).

While extensive focus has been placed on various aspects of the polymorphic information within a MK test framework, much less attention has been given on the divergence estimates. There are systematic biases expected with estimates of divergence and the choice of the outgroup (Cai and Petrov, 2010; Keightley and Eyre-Walker, 2012). This is most apparent when divergence between two species is small and therefore divergence

estimates are inflated by polymorphisms (Welch, 2006). Usually divergence estimates are obtained by comparing two single alleles, one of the species for which polymorphisms data is available and one from the outgroup species. Under such a scenario it is possible is correct for the effect of polymorphisms on divergence (Keightley and Eyre-Walker, 2012). In principle this estimate can be improved by taking multiple alleles of both species similar to the approach of McDonald and Kreitman (1991). However often polymorphism information is only available for one species. Furthermore some of the MK methods rely on the correct classification of derived vs. ancestral state of a polymorphism (Boyko et al., 2008; Schneider et al., 2011). Unfortunately slight misinferences in the orientation of SNPs would artificially increase the amount of high frequency derived alleles (Williamson et al., 2005; Eyre-Walker et al., 2006). This minor deviation could have a large impact on the statistical outcome and lead to an overestimation of α, but is only relevant for methods which use the full site frequency spectrum. It also has been shown that gene age is one of the strongest predictors of the rate of evolution (Cai and Petrov, 2010) and this implies that distantly related outgroups would potentially lack genes for which higher rates of adaptive evolution are expected. The choice of the outgroup and an unbiased sampling scheme is therefore a crucial aspect for correctly estimating genome wide rates of adaptive evolution.

Table 1.2 summarizes the modified variants of the MK test which have been developed to take violations of assumptions of the classic MK test into account. Table 1.3 shows software packages available to conduct MK type of analyses.

Abbreviation	Source	Feature(s)
MK	McDonald and Kreitman (1991)	Classical single locus MK-test
FWW	Fay et al. (2001)	Exclusion of low frequency polymorphisms
		Summary statistic across genes
BNS+	Bustamante et al. (2002)	Multiple gene analysis using hierachical Baysian approach
SEW	Smith and Eyre-Walker (2002)	Less biased when combining information of multiple genes
BEW	Bierne and Eyre-Walker (2004)	Maximum Likelihood approach using also genes with little polymorphism
W	Welch (2006)	Refined ML version of BEW
LCB+	Loewe and Charlesworth (2006)	Inferring DFE from species with different N_e
BWI+	Boyko et al. (2008)	Incorporation of simultaneously estimation of DFE and demography
EWK	Eyre-Walker and Keightley (2009)	Incorporation of simultanously estimation of DFE and demography
OWK+	Obbard et al. (2009)	Refined ML version of W
SCE+	Schneider et al. (2011)	Contribution of advantageous mutations to polymorphism
WHA+	Wilson et al. (2011)	Fine scale variation in selection pressures within genes

Table 1.2: A list of the McDonald Kreitman test derivates developed

Software	Implemented methods	Source	Type/Platform
DoFe v3.0	BEW, EWK	`http://www.lifesci.sussex.ac.uk/home/Adam_Eyre-Walker/Website/Software.html`	Win, Mac, Linux
DFE-alpha/DFE-adaptive	EWK, SCE+	`http://homepages.ed.ac.uk/eang33/`	Web interface
MK test v2.0	W, OWK+	`http://www.gen.cam.ac.uk/research/welch/GroupPage/Software.html`	Linux/Mac
MKPRF	BNS+, BWI+	`http://cbsuapps.tc.cornell.edu/mkprf.aspx`	Web interface

Table 1.3: A list of the McDonald Kreitman test software application available

1.5 Genome-wide estimates of adaptive evolution in the context of the MK test

1.5.1 Drosophila

The genus *Drosophila* is perhaps the most studied group of organisms within the MK test framework. This is partly for historical reasons, since the fruitly is one of the model organisms in experimental genetics and has been used extensively in population genetics. Secondly, Drosophila species have relatively high levels of diversity and therefore a moderate number of sequences will provide sufficient data to analyse. Thirdly, due to the publication of the genomes of 12 different Drosophila species it has become possible to include divergence to quantify the rate of adaptation for multiple Drosophila species (Drosophila 12 Genomes Consortium et al., 2007). Fourthly, there is multiple evidence that adaptive evolution is widespread in numerous *Drosophila* species based on estimates within the MK framework even though these species have different demographic histories and population structures (Maside and Charlesworth, 2007; Shapiro et al., 2007; Betancourt et al., 2009; Eyre-Walker and Keightley, 2009; Petit and Barbadilla, 2009; Haddrill et al., 2010).

Initial attempts to quantify genome wide rates of adaptation within a MK test framework for the genus Drosophila have shown that the proportion of fixed differences due to positive selection in *D. melanogaster* and *D. simulans* might be around 50% for non-synonymous sites (Bustamante et al., 2002; Smith and Eyre-Walker, 2002; Bierne and Eyre-Walker, 2004) or even substantially higher (Sawyer et al., 2003). It is generally believed that the extent of adaptation is determined by the effective population size of a species. This because in a larger population more (advantageous) mutations enter the population and a higher proportion of mutations is effectively selected because the selected effect scales with N_e. Therefore later studies emphasized this aspect by focusing on species pairs which vary in their effective population sizes (Table 1.4), ranging from *D. miranda* with a relatively low effective population size to *D. melanogaster* and *D. pseudoobscura* with moderately high effective population sizes to *D. simulans* with the largest effective population size (Bachtrog, 2008; Haddrill et al., 2010; Andolfatto et al., 2011a). However the results are inconclusive and do not seem to support the conjecture of adaptation determined by the effective population size. An initial attempt has found similar rates of adaptation in *D. melanogaster* and *D. miranda*, even though those two species differ

by 5-fold in their effective population sizes (Bachtrog, 2008). Methods which explicitly model the effect of slightly deleterious mutations by estimating the DFE of new mutations (Loewe and Charlesworth, 2006; Eyre-Walker and Keightley, 2009) may refine this result. For example for the two species *D. pseudoobscura* and *D. miranda* it has emerged that the rates of adaptation differ in a way that is expected by their difference in effective population sizes (Haddrill et al., 2010), but the differences were not significant. Including more complex assumptions of population size changes, similar rates of adaptive protein evolution of ≈80% have been found for much larger datasets in *D. melanogaster*, *D. simulans*, *D. miranda* and *D. pseudoobscura* (Andolfatto et al., 2011a; Jensen and Bachtrog, 2011).

Studies in these four Drosophila species contrasting the impact of differences in N_e have focused on African populations and X-linked loci. This is because chromosomal variation show a strong geographic pattern (Baudry et al., 2004, 2006) associated with differences in the diversity of X-chromosomes and autosomes. Furthermore *D. melanogaster* populations show an autosomal inversion that may drive patterns of recombination (Kauer et al., 2002). Studies on much larger datasets in *D. melanogaster* (Shapiro et al., 2007; Eyre-Walker and Keightley, 2009) and *D. simulans* (Begun et al., 2007) have shown for autosomal loci that adaptation is widespread, however lower than the estimates for the X-chromosomes. There is additional evidence that the rate of adaptation varies considerably across the genome. It has emerged that regions with low recombination lack the signal of adaptation (Haddrill et al., 2007; Shapiro et al., 2007) and that sex-biased genes, in particular male genes, show the strongest signal of adaptation (Pröschel et al., 2006; Begun et al., 2007). Also the rate of adaptation varies considerably within genes of the immune system and is on average elevated in contrast to the rest of the genome in *D. melanogaster* and *D. simulans* (Obbard et al., 2009). Moreover, estimates based on a MK test framework have revealed that non-coding DNA is subject to substantial positive selection for *D. melanogaster* and *D. simulans*, particulary for UTRs where α has been estimated to be ≈60% (Andolfatto, 2005; Haddrill et al., 2008). For introns this value seems to be much lower (Andolfatto, 2005; Haddrill et al., 2008).

Studies have also addressed other Drosophila species with different characteristics to overcome biological limitations associated with genome wide estimates of adaptive evolution. Such a species is *D. americana* which has a well established ecology, a presum-

ably stable population size and diversity estimates comparable to those of the other Drosophila species. It has also shown substantial evidence of adaptive evolution (Maside and Charlesworth, 2007; Betancourt et al., 2009). However for the dot chromosome, the smallest Drosophila chromosome, adaptive evolution is limited with estimates of $\alpha \approx 0$ caused by its exceptionally low N_e resulting in reduced efficiency selection (Betancourt et al., 2009). While there is evidence for adaptive evolution for sex biased genes in *D. melanogaster* no evidence has been found in *D. pseudoobsurra* based on divergence comparisons (Metta et al., 2006). To understand this discrepancy a study has investigated the patterns of evolution of sex biased genes in a closer relative of *D. melanogaster*, *D. ananassae*. This analysis did not reveal evidence for differences between the evolutionary rates of sex-biased genes versus unbiased genes but it showed substantial amount of adaptive divergence with estimates of $\alpha \approx 60\%$ (Grath et al., 2009). A comparison of multiple Drosophila species has found that patterns of selection efficiency are the result of varying effective population sizes (Petit and Barbadilla, 2009). However currently there is an ongoing debate about the role of N_e as a major determinant of the rate of adaptive evolution in Drosophila species (Venton, 2012). Taken together, these results indicate that adaptive divergence in Drosophila species is widespread and substantial. These high rates of adaptation urge for the need to take the impact of advantageous mutations on the patters of polymorphisms into account (Sella et al., 2009; Schneider et al., 2011).

Contrasted species pair	Outgroup	α	MK method	N_e fold difference*	Source
D. miranda	*D. pseudoobscura*	0.44	BEW	2.8-fold	Bachtrog (2008)
vs. *D. melanogaster*		0.51	BEW		
D. miranda	*D. affinis*	0.38	LCB+	4.2-fold	Haddrill et al. (2010)
vs. *D. pseudoobscura*		0.70	LCB+		
D. melanogaster	lineage specific	0.84	EWK	1.3-fold	Andolfatto et al. (2011a)
vs. *D. simulans*		0.86	EWK		
D. miranda	*D. pseudoobscura*	0.78	EWK	2.7-fold	Jensen and Bachtrog (2011)
vs. *D. pseudoobscura*	*D. miranda*	0.83	EWK		

Table 1.4: Estimates of α for four *Drosophila* species in four species pairs comparisons. All studies have been conducted for X-linked loci. * as measured by synonymous diversity ratio of the corresponding study. It is therefore a measurement of the current effective population size.

1.5.2 Mammals

Not unsurprisingly there is a great interest in investigating the role of natural selection in our own species for which the MK test supplies a possible framework. An initial study estimated the amount adaptive divergence for humans in protein coding genes to be $\alpha \approx 35\%$ (Fay et al., 2001). Later studies have failed to find strong evidence and have estimated that α for humans must be close to zero (Chimpanzee Sequencing and Analysis Consortium, 2005; Zhang and Li, 2005). The reasons for this discrepancy between these studies may be caused by differences in the sampling scheme, the impact of deleterious mutations or demographic factors (Eyre-Walker, 2006). Bustamante et al. (2005) have found that 9% of the genes show evidence of positive selection, however they had to drastically reduce their dataset because of too little information of a large fraction of the analysed genes, and later studies found that their results were highly influenced by the prior in their Bayesian analysis (Han et al., 2009). A low proportion of genes ($< 1\%$) that show a signature of adaptation has also been found by estimates based on divergence (Nielsen et al., 2005). More complex models that used a joint estimation of demography and DFE show that 10-20% of the amino acid differences between human and chimpanzee have been caused by positive selection (Boyko et al., 2008). However only little evidence for adaptive evolution has been found by a similar approach for a different dataset (Eyre-Walker and Keightley, 2009) but this estimate could be downwardly biased. Even though there is a discrepancy in precision of the proportion of adaptation in humans, the general consensus from these studies attributes a reduced role of positive selection in hominids. It has been proposed that this relatively limited role in comparison to Drosophila species is caused by differences in N_e. Our sister species chimpanzee (*Pan troglodytes*) has an effective population size approximately two to three times larger than for humans (Yu et al., 2004; Hvilsom et al., 2012) suggesting that the average N_e of human and chimpanzees is substantially lower than in rodents and Drosophila. Genome wide estimates reveal a very limited role of adaptation in chimpanzees with the exception of the X-chromosome (Hvilsom et al., 2012), for which estimate of $\alpha \approx 38\%$ in accordance with the faster X hypothesis (Mank et al., 2010).

Beside estimates of α in primates, adaptive divergence in mammals was poorly investigated until very recently. Because the effective population size is thought to be a major cause of the low α estimates in primates, the wild house mice (*Mus musculus castaneus*) appeared to be an appropriate species to test this hypothesis. It has an effective

population size at least one order of magnitude higher than for humans which is well in the range of what is observed for Drosophila species. And indeed there is evidence for substantial amount of adaptation this species underwent for protein coding DNA with $\alpha \approx 50\%$ (Halligan et al., 2010; Phifer-Rixey et al., 2012). It has also been shown that α estimates are much lower for regions up and downstream of genes (5%-10%) caused by a higher proportion of neutrally evolving sites (Kousathanas et al., 2011). In contrast, in ultraconserved elements α is estimated to be $\approx 78\%$ (Halligan et al., 2011). Despite this large variation in the α estimates for different regions of the *M. m. castaneus* genome, absolute rates of adaptation appear to be relatively similar (Halligan et al., 2011). Even though the comparison between humans and mice appears to be well in line with the effect of N_e on the rate of adaptation, those two species largely differ in their biological aspects. Therefore a comparison between closely related species with similar habitats and range distribution provide a comparable measurement to disentangle the effect of N_e on the rate of adaptation. Two house mouse subspecies (*M. m. musculus* and *M. m. domesticus*) with an N_e much lower than for *M. m. castaneus* show a reduced amount of adaptation ($\alpha \approx$ 12-13%, Phifer-Rixey et al., 2012). The collection of estimates of adaptive evolution in mammals have been extended by investigations in lagomorphs. The two European rabbit subspecies *Oryctolagus cuniculus algirus* and *O. c. cuniculus* show high levels of nucleotide diversity which implies they have large effective population sizes and for both species adaptive evolution seems to be widespread (Carneiro et al., 2012). For both subspecies the proportion of effectively neutral new mutations is remarkably low ($< 4\%$) and for *O. c. algirus* there seems to be an increase in the rate of adaptation for X chromosomal genes.

Taken together these results suggest that in mammals adaptation plays a limited role in some species, so far only for two mammalian species it appears that adaptive evolution is widespread. However it might well be that mammals tend to have in general relatively small N_e causing this reduced role of adaptation.

1.5.3 Plant species

Beside approaches in Drosophila and mammals, plant species have come into focus to determine the role of genome wide adaptation (Siol et al., 2010). Most studies in plant population genetics have been conducted for two plant families, *Poacea* and *Brassicaceae*. Among the *Poacea*, important crops such as rice, maize and sorghum contribute substan-

tially to the world food supply. Consequently there is an outstanding economical interest for these plant species and research in molecular genetics is progressing rapidly. On the other hand extensive fundamental research has been conducted on a *Brassicaceae*, *Arabidopsis thaliana* and its relatives. *A. thaliana* is the model plant species with the first ever sequenced plant genome (AGI, 2000) and a well established ecology and morphology. There are currently plans to sequence more than 2500 *A. thaliana* individuals (`http://www.1001genomes.org/`). Further research in the field of plant population genetics comes from plant species with a very distinct ecology, such as trees, endemic species or sunflowers. Beside the estimation of genome wide rates of adaptation, plant population genetics research has been conducted to establish range distributions, polymorphic pattern amongst varieties, demographic histories as well as population structure and admixture and the characterization of morphological traits (Schmid et al., 2006; Caicedo et al., 2007; Foxe et al., 2009; Brachi et al., 2011). In comparison to animals, plant genomes are enriched in paralogous genes because they have undergone whole genome duplication events. More than 50% of all plant genomes show evidence of at least one whole genome duplication (Soltis and Soltis, 2009).

An initial estimate of adaptive evolution in *A. thaliana* based on the MK test framework showed little evidence of it; it was suggested that this is a consequence of inbreeding (Bustamante et al., 2002). However the close relative of *A. thaliana*, *A. lyrata* also showed little evidence for genome wide adaptive evolution even though it is largely outcrossing (Foxe et al., 2008). Both species have a comparable N_e. Even though evidence for genome wide adaptation is limited, similar to Drosophila it appears that there is variation in the rate of adaptation across the *Arabidopsis* genome, with genes involved in disease resistance and abiotic stress showing high rates of adaptation (Slotte et al., 2011). The related species *Capsella grandiflora* has an N_e which is substantially higher than for *A. thaliana* and reveals evidence for adaptation (Slotte et al., 2010). Also the poplar tree *Populus tremula* and several sunflower species with an increased N_e show evidence of widespread adaptive evolution (Strasburg et al., 2009; Ingvarsson, 2010; Strasburg et al., 2011a). In contrast there is no evidence of adaptive evolution in crops (Hamblin et al., 2006; Ross-Ibarra et al., 2009) or wild tomatoes (Tellier et al., 2011).

1.5.4 Other species

Only for a limited number of other eukaryotic species genome wide rates of adaptive evolution have been estimated. This includes chicken (*Gallus gallus*) for which $\alpha \approx 20\%$, however using a very limited dataset (Axelsson and Ellegren, 2009). For yeast species there is no evidence of genome wide adaptation despite its presumingly large N_e (Elyashiv et al., 2010). However currently there is lack of genome estimates of adaptive evolution for non model organisms, however this gap will be closed relatively soon due to advances in sequencing technology. In prokayotes adaptive evolution seems to be widespread (Charlesworth and Eyre-Walker, 2006; Hughes et al., 2008) even though not for all species (Hughes et al., 2008).

1.6 Objectives of this thesis

In the previous sections I have summarized the current state of art in the field of the molecular evolution in the context of McDonald-Kreitman test types of analyses. It has revealed that currently the number of species for which genome wide estimates of adaptive evolution have been obtained is limited to few key species, such *Drosophila*, human, mice and some plants. However there is a need to extend the number of estimates and to improve the first estimates that were conducted on very limited datasets. Secondly genome wide estimates of α have relied on different approaches and are therefore biased by the underlying methodology in different ways. For a comparable measurement preferably the same method should be used. Chapter two aims at both of these problems by investigating the role of adaptive evolution in nine plant species using the method of Eyre-Walker and Keightley (2009). Population genetic theory predicts that the effective population size plays a major role in determining the rate of adaptive evolution. Surprisingly from current studies there is inconclusive evidence that this is the case in natural populations. In chapter three I investigate this aspect by conducting the largest cross species analysis to infer the effect of N_e on the rate of adaptive evolution overcoming several limitations from previous analyses. I will show that for between species comparisons the absolute rate of adaptation (ω_a) and not α should be used to determine the impact of N_e and the rate of adaptation, a matter only few studies have considered so far (Bierne and Eyre-Walker, 2004; Halligan et al., 2011). In chapter four a theoretical limitation of the MK test will be investigated. Under the MK test it is assumed that selection pressure is constant over time. It has been suggested that the high rates of adaptive evolution as found in *Drosophila*

species are artifactual and not caused by recurrent positive directional selection but rather the outcome of changing selective pressures over time (Huerta-Sanchez et al., 2008). It is also known that there is intragenomic variation of N_e due to the processes of genetic hitchhiking and background selection. In chapter five I will investigate if this variation is common in biological populations and will quantify the amount of variation there is.

Chapter 2

Genome wide analyses reveal little evidence for adaptive evolution in many plant species.

2.1 Abstract

The relative contribution of advantageous and neutral mutations to the evolutionary process is a central problem in evolutionary biology. Current estimates suggest that while *Drosophila*, mice and bacteria have undergone extensive adaptive evolution, hominids show little or no evidence of adaptive evolution in protein coding sequences. This may be a consequence of differences in effective population size. To study the matter further we have investigated whether plants show evidence of adaptive evolution using an extension of the McDonald-Kreitman test that explicitly models slightly deleterious mutations by estimating the distribution of fitness effects of new mutations. We apply this method to data from 9 pairs of species. Altogether more than 2400 loci with an average length of ≈280 nucleotides were analysed. We observe very similar results in all species; we find little evidence of adaptive amino acid substitution in any comparison except sunflowers. This may be because many plant species have modest effective population sizes.

2.2 Introduction

The contribution of adaptive evolution relative to genetic drift is a fundamental problem in molecular evolution (Kimura, 1983; Gillespie, 1991b). Several methods to estimate the fraction of adaptive substitutions, α, have been developed based on the

McDonald-Kreitman test (McDonald and Kreitman, 1991) that contrast polymorphism and divergence between selectively and neutrally evolving sites (Charlesworth, 1994; Fay et al., 2001; Smith and Eyre-Walker, 2002; Bierne and Eyre-Walker, 2004; Welch, 2006; Boyko et al., 2008; Eyre-Walker and Keightley, 2009). These methods have been applied to a variety of species. Estimates in Drosophila (Smith and Eyre-Walker, 2002; Bierne and Eyre-Walker, 2004; Welch, 2006; Bachtrog, 2008) and rodents (Halligan et al., 2010) suggest that ≈50% of all amino acid substitutions have been fixed as a consequence of adaptive evolution and for microorganisms estimates may be even higher (Charlesworth and Eyre-Walker, 2006, Liti et al., 2009). However, although analyses of DNA sequence diversity show signs of some adaptive evolution (Fay et al., 2001; Zhang and Li, 2005), overall the level of adaptive evolution in hominids appears to be very low (Chimpanzee Sequencing and Analysis Consortium, 2005, Zhang and Li, 2005; Boyko et al., 2008; Eyre-Walker and Keightley, 2009). The contrast between hominids and other animals has led to the suggestion that effective population size may be an important determinant of the rate of adaptive evolution since hominids typically have low effective population sizes in contrast to rodents, insects and bacteria (Eyre-Walker et al., 2002; Fraser et al., 2007; Halligan et al., 2010). However, some caution should be exercised since the level of adaptive evolution is typically measured as the proportion of substitutions that are adaptive and this depends both on the numbers of substitutions that are effectively neutral and the number that are advantageous. It has been shown that the number of effectively neutral mutations is negatively correlated to the effective population size in many species (Woolfit and Bromham, 2003, 2005; Popadin et al., 2007; Moran et al., 2008; Piganeau and Eyre-Walker, 2009). Hence the correlation between proportion of substitutions that are adaptive and N_e may be a consequence of the correlation between the proportion of effectively neutral mutations and N_e, and may not reflect any change in the absolute rate of adaptive evolution.

The rate of adaptive evolution has also been studied in plants. On a genome wide scale, previous studies in *Arabidopsis thaliana* have shown little evidence for adaptive evolution (Bustamante et al., 2002; Barrier et al., 2003; Schmid et al., 2005). This was attributed to the high frequency of inbreeding in *A.thaliana* and the reduction in effective population size that this caused. However, the out-crossing species *Arabidopsis lyrata* (Foxe et al., 2008; Barnaud et al., 2008), the partially out-crossing cultivated tropical grass *Sorghum bicolor* (Hamblin et al., 2006) and the mainly out-crossing *Zea*

species (Bijlsma et al., 1986; Ross-Ibarra et al., 2009) also show little evidence of positive selection. Instead all these species show evidence of slightly deleterious mutations segregating. In contrast, to the pattern in other plant species, Strasburg et al. (2009), Ingvarsson (2010) and Slotte et al. (2010) have recently estimated that ≈75%, ≈30% and ≈40% of fixed amino acid differences were driven by adaptive substitutions in sunflowers, aspen trees and some *Brassicaceae* species respectively.

There is evidence that slightly deleterious mutations (SDMs) contribute to variation in many populations (Cargill et al., 1999; Akashi, 1999; Fay et al., 2002; Hughes, 2005; Charlesworth and Eyre-Walker, 2006). They are subject to weak negative selection ($N_e s \approx 1$), segregate at lower frequencies than neutral mutations and contribute proportionally more to polymorphism than to divergence when compared to neutral mutations. The presence of SDMs in the analyses of *Arabidopsis sp.*, *S.bicolor* and *Zea sp.* may explain why there is so little apparent positive selection, because SDMs are expected to bias the estimate of adaptive evolution downwards if population sizes are either stationary or contracting - they can lead to an overestimate if population sizes have expanded (McDonald and Kreitman, 1991; Eyre-Walker, 2002). Here we apply a method to estimate the proportion of adaptive substitutions, that controls for the effects of SDMs by estimating the distribution of fitness effects (DFE) of new mutations (Eyre-Walker and Keightley, 2009). It estimates the DFE from the polymorphic data and predicts the expected number of substitutions originating from neutral and slightly deleterious mutations. If the observed number of substitutions is greater than the expectation inferred from the DFE, it can be attributed to advantageous substitutions, yielding an estimate of α. We present a new parametrisation of this method which allows us to estimate the rate of adaptive substitution relative to the rate of synonymous substitution, thereby allowing us to explore whether the rate of adaptive evolution depends upon N_e independent of the effects of N_e on the number of effectively neutral substitutions.

We also apply the method of Fay et al. (2001), who suggested controlling for SDMs by removing low frequency mutations from the analysis. We apply these methods to 11 datasets covering the divergence between 9 pairs of largely independent species. We find little evidence of adaptive amino acid substitution in any comparison except sunflowers. Moreover we estimate the effective population size and find that some of the plant species analysed have large effective population sizes suggesting that other factors may be more

important than population size in determining the rate of adaptive evolution.

2.3 Materials and Methods

2.3.1 Sequence data

Data were retrieved from Genbank (`http://www.ncbi.nlm.nih.gov/Genbank`) for *Oryza* spp. (GenBank IDs: EF000002-EF01059, Caicedo et al. (2007)), *Populus tremula* (EU752500-EU753117, Ingvarsson (2008b)), *Arabidopsis lyrata* (BV683158-BV686427; EF502173-EF502282; EF502359-EF502483; EF502558-EF502973, Ross-Ibarra et al. (2008); EU592234-EU592323, Foxe et al. (2008)), *Zea mays* (BV123534-BV144210; BV446558-BV447590, Wright et al. (2005)), *Sorghum bicolor* (DQ427111-DQ430705, Hamblin et al. (2006)), *Boechera stricta* (FJ573482-FJ577247, Song et al. (2009); GQ907358-GQ910665), *Schiedea globosa* (GU830974-GU831538) and *Helianthus petiolaris* and *Helianthus annuus* (Strasburg et al., 2009, Table A2.1). Polymorphic data for *Arabidopsis thaliana* were downloaded from `http://walnut.usc.edu/2010` and for *Populus balsamifera* from `http://www.popgen.uaf.edu/data` (Olson et al., 2010).

The annotated protein-coding genome of *A. thaliana* was obtained from TAIR (`ftp://ftp.arabidopsis.org/home/tair/Genes/TAIR8_genome_release`). The annotated *Populus trichocarpa* and *Sorghum bicolor* genomes were obtained from JGI (`http://genome.jgi-psf.org/Poptr1_1` (Tuskan et al., 2006) and `http://genome.jgi-psf.org/Sorbi1` (Paterson et al., 2009)). Predicted coding sequences of *Zea mays* were obtained from `http://magi.plantgenomics.iastate.edu/downloadall.html` (Fu et al., 2005) and `http://ftp.maizesequence.org/release-3b.50/sequences/`. The rice genome was downloaded from `ftp://ftp.plantbiology.msu.edu/pub/data/Eukaryotic_Projects/o_sativa/annotation_dbs/pseudomolecules/version_6.0/all.dir/`. We analyzed the polymorphism data for each of the 11 plant species. Additionally, for some datasets information about populations were available; 5 populations of *A.thaliana* (Nordborg et al., 2005), 6 populations of *A.lyrata* (Ross-Ibarra et al., 2008), 3 domesticated subspecies of *O.sativa*, one domesticated subspecies (*Z. mays mays*) and one wild subspecies (*Z.mays parviglumis*) for *Z.mays*, 2 populations of *B.stricta*, 3 populations from *S.globosa* and 3 populations from *P.balsamifera* (Keller et al., 2010).

2.3.2 Preparation of the data

Sequences were aligned using Clustalw using default parameter values (Thompson et al., 1994). Coding regions were assigned using protein coding genomic data or if given, taken from the GenBank input files. The outgroup was assigned using the best BLAST (Altschul et al., 1990) hit or, if given, taken from the polymorphism dataset. We only used polymorphism data for which we could assign an outgroup sequence with the exception of *A. thaliana* for which no outgroup data is currently available (the *A. lyrata* sequence is nearly complete but unpublished). For all analyses the number of synonymous sites served as the neutral standard. For computational reasons the method to estimate the distribution of fitness effects needs all sites to have been sampled in the same number of chromosomes for each species; because some loci had been sampled in more individuals than others and other loci had missing data, we reduced the dataset to a common number of chromosomes by randomly sampling the polymorphisms at each site without replacement. The numbers of synonymous and nonsynonymous sites and substitutions were computed using the F3x4 model implemented in PAML (Yang, 1997) in which codon frequencies are estimated from the nucleotide frequencies at the three codon positions. The proportion of sites estimated by PAML were also used to compute the numbers of synonymous and non-synonymous sites for the polymorphism data. For each species or population, data were summed across all genes, although analyses run on the unsummed data gave similar results (results not presented). Statistics concerning numbers of loci, numbers of sites as well as polymorphic sites are shown in Table 2.1.

The distribution of fitness effects of new mutations and the proportion of adaptive substitutions (α) were estimated using the method II of Eyre-Walker and Keightley (2009) which accounts for the segregation and fixation of slightly deleterious mutations. To estimate the rate of adaptive evolution in a manner that is independent of the number of effectively neutral substitutions we reparameterize the method as follows. In the original formulation the expected numbers of synonymous, D_s, and non-synonymous, D_n, substitutions were as follows:

$$\begin{aligned} D_s &= 2utL_s \\ D_n &= 2utL_n \frac{\int D(S)P(S)dS}{1-\alpha} \end{aligned}$$

where L_s and L_n are the numbers of synonymous and non-synonymous sites, u is the nucleotide mutation rate and t is the time of divergence. D(S) is the distribution of S, the

Plant family	Ingroup	Population	Outgroup	Alleles	Loci	Sites	Seg.	d_s	π_s
Brassicaceae	*Arabidopsis lyrata*		*A. thaliana*	24	72	22739	610	0.15	0.018
		Germany		14	51	14263	191		
		USA		14	47	16401	74		
		Russia		16	48	17332	93		
		Iceland		12	63	18651	142		
		Canada		16	51	19458	103		
		Sweden		8	62	20609	123		
	Arabidopsis thaliana		none	24	919	260737	3401		0.007
		USA		8	664	167990	888		
		N.Sweden		8	742	191492	902		
		S.Sweden		8	771	202016	1480		
		C.Europe		8	789	211398	1430		
		England		8	816	222405	1252		
	Boechera stricta		*A. thaliana*	24	134	40915	220	0.21	0.003
		North		16	130	38589	111		
		South		10	130	39800	123		
Poaceae	*Sorghum bicolor*[1]		*S. propinquum*	14	135	30241	129	0.03	0.006
	Oryza rufipogon		*S. bicolor*[1]	18	73	18176	121	0.78	0.005
	Oryza sativa[1]		*S. bicolor*[1]					0.78	0.006
		temperate[1]		18	73	18163	22		
		tropical[1]		18	71	17618	85		
		indica[1]		18	73	18197	68		
	Zea mays[1]		*S. bicolor*[1]					0.27	
		spp. mays[1]		8	478	74155	939		0.015
		spp. parviglumis		10	378	59944	1241		0.021
Salicaceae	*Populus tremula*		*P. trichocarpa*	24	69	32255	578	0.05	0.01
	Populus balsamifera		*P. trichocarpa*	24	508	174855	1446	0.01	0.004
		Central		14	501	172046	1108		
		East		6	519	178473	643		
		North		10	508	178743	919		
Asteraceae	*Helianthus annuss*		*H. petiolaris*	14	54	7183	259	0.05	0.029
	H.petiolaris		*H.annuus*	10	54	7231	175	0.05	0.027
Caryophyllaceae	*Schiedea globosa*		*S. adamantis*	24	23	8030	189	0.02	0.014
		Maui		14	23	8026	98		
		Molokai		12	23	8029	122		
		Oahu		12	23	8027	122		

[1] includes domesticated species

Table 2.1: Summary of data sets used for the analyses. Number of coding sites (Sites) and number of segregating sites (Seg.) are from the polymorphic data. Average divergence between the species pairs at silent sites (d_s) and nucleotide diversity for synonymous sites (π_s). Sample size (Alleles) is constant for each species for computational reasons.

strength of selection (multiplied by four times the effective population size), P(S) is the probability of fixation and α is the proportion of substitutions that are adaptive. We can reparameterize the expression for D_n as follows:

$$D_n = 2utL_n(\int D(S)P(S)dS + \omega_a)$$

where ω_a is the ratio of the rate of adaptive nonsynonymous substitution to synonymous substitution. This is similar to the parametrisations given by Bierne and Eyre-Walker (2004) and Obbard et al. (2009).

Levels of polymorphism were quantified using Watterson's estimator of $4N_e\mu$ (θ_W) (Watterson, 1975) and nucleotide diversity, π, (Tajima, 1983) for synonymous sites. Our estimates of π and θ_W differ slightly from previous published estimates due to the fact that we have excluded some data and alleles from our analysis. We estimated the effective population size using the level of synonymous site diversity and dividing this by an estimate of the mutation rate per generation. Estimates of the synonymous mutation rate per year and generation times were taken from the literature: for species *A.thaliana*, *A.lyrata* and *B.stricta* we assumed, as other have done, the mutation rate of $\mu = 1.5 \times 10^{-8}$ (9.9×10^{-9}-2.1×10^{-8}) per site per year as estimated in the *Brassicaceae* (Koch et al., 2000) and a generation time of 1 year for the annual *A.thaliana* (Koornneef et al., 2004); and 2 years for the perennials *B.stricta* (Dobes et al., 2004) and *A.lyrata* (Ross-Ibarra et al., 2008). For the genera *Oryza*, *Sorghum*, *Helianthus* and *Zea* we assumed the mutation rate to be $\mu = 1.0 \times 10^{-8}$ (6×10^{-9}-1.7×10^{-8}) and a generation time of 1 year (Swigonová et al., 2004; Strasburg and Rieseberg, 2008); for *Populus* we assumed a mutation rate of $\mu = 2.5 \times 10^{-9}$ (1.7×10^{-9}-3.5×10^{-9}) per site per year and a generation time of 15 years (Ingvarsson, 2008b; Tuskan et al., 2006; Koch et al., 2000); and for *Schiedea* we assumed a mutation rate of $\mu = 1.9 \times 10^{-8}$ (1.4×10^{-8}-4.6×10^{-8}) per site per year and a generation time of 5 years (Filatov and Burke, 2004; Wallace et al., 2009).

2.3.3 Simulations

To investigate the effects of population structure on our estimates of adaptive evolution we performed forward population genetic simulations using SFS_CODE (Hernandez, 2008). We investigated two scenarios. In the first we have migration between two populations that divided sometime in the past and have continued to exchange migrants, but we only sample from one of the populations; and in the second, we sample from both populations equally. To simulate these two scenarios we divided an ancestral population into three

equal sized populations of 500 individuals 20,000 generations in the past; with a mutation rate of $\mu = 5 \times 10^{-6}$ per site per generation this would give an expected divergence at neutral sites of 20%, which is similar to that seen in some of our datasets. We allowed one of the populations to evolve independently; this was the outgroup. The other two populations exchanged migrants at various rates of $4N_em$ ranging from 0.01 to 10. All populations were subjected to mutation at $4N_eu = 0.01$ at 100 uncoupled loci with a length of 1002 nucleotides each. Each simulation was repeated 100 times. Since SFS_CODE and convertSFS_CODE do not allow population mixing or the sampling of two populations we allowed the two populations to exchange migrants at a very high rate ($4N_em = 250$) for the last 5 generations of the simulation (Figure A2.1).

2.4 Results

2.4.1 Data

To estimate the rate of adaptive substitution in plant protein coding sequences we compiled polymorphism data from 11 species and aligned 10 of these to outgroups to analyse the divergence between 9 pairs of species. The datasets range dramatically in size from 23 to 919 loci per species and 6 to 24 sequences per gene (Table 2.1). Note that we treat *O.sativa* and *O.rufipogon* as the same species in our analysis, *O.sativa* being the domesticate of *O.rufipogon*. For most genera we have access to polymorphism data from wild populations (*Arabidopsis lyrata*, *Arabidopsis thaliana*, *Boechera stricta*, *Populus tremula*, *Populus balsamifera*, *Oryza rufipogon*, *Helianthus annuus*, *Helianthus petiolaris* and *Schiedea globosa*), for one species from both wild and domesticated populations (*Zea mays*) and for the species *Sorghum bicolor* and *Oryza sativa* we only have data from domesticated populations. We use the polymorphism data of 11 species to estimate the distribution of fitness effects of new mutations and use this to further estimate the proportion and relative rates of adaptive substitutions between species. Our comparisons for *Boechera*, *Populus balsamifera*, *Schiedea* and *Oryza* are the first large scale investigations of adaptive evolution in these groups. The available outgroup data only allow us to estimate the rate of adaptive substitution between 9 pairs of species. We include the polymorphism data from *A.thaliana* as a comparison for the results obtained from *A.lyrata*; note also that the divergences between the *Brassicaceae* species and the *Poaceae* species are not entirely independent because the same outgroup (*A.thaliana* and *S.bicolor*, respectively) is used for different species. Based

on average d_s values (Table 2.1) *A.thaliana* and *A.lyrata* share ≈30% of their divergence with *B.stricta* and *O.rufipogon* and *Z.mays* share ≈22% of their divergence with *S.bicolor*.

2.4.2 Distribution of effects of new mutations

Species wide (e.g. ignoring population information within species) estimates of the distribution of fitness effects show a remarkably consistent picture among the investigated species (Figure 2.1) with the exception of *B.stricta*, *P.balsamifera* and *S.globosa*. For all species the largest proportion of mutations have $N_e s > 100$ and hence are strongly deleterious, and for most species less than 25% of amino acid-changing mutations behave as effectively neutral ($0 < N_e s < 1$).

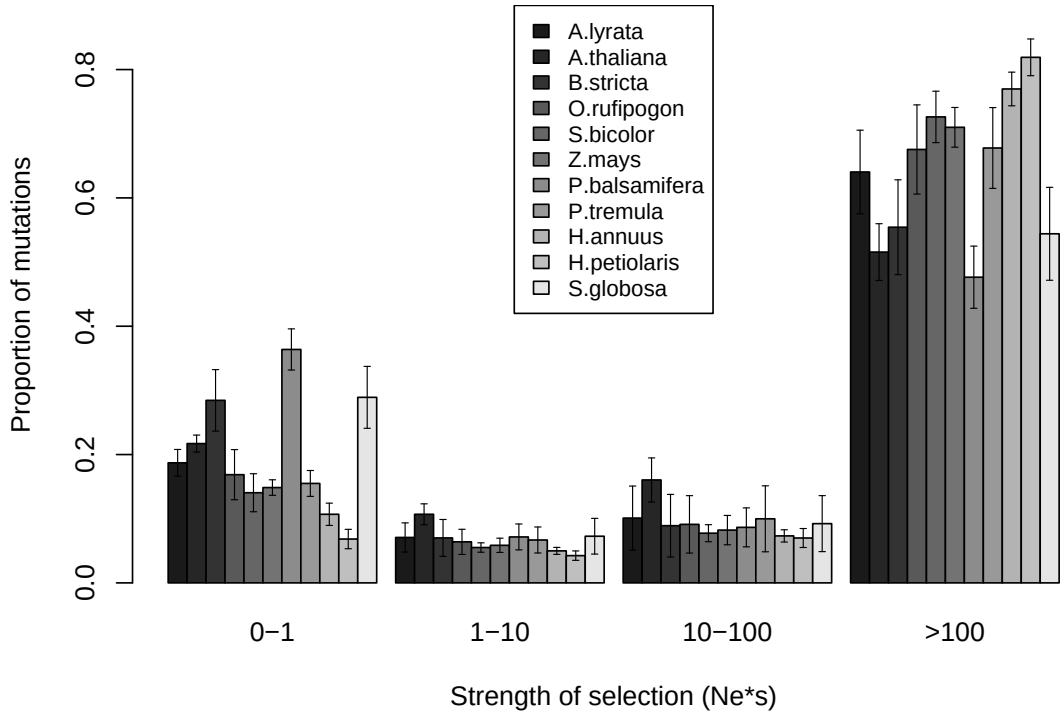


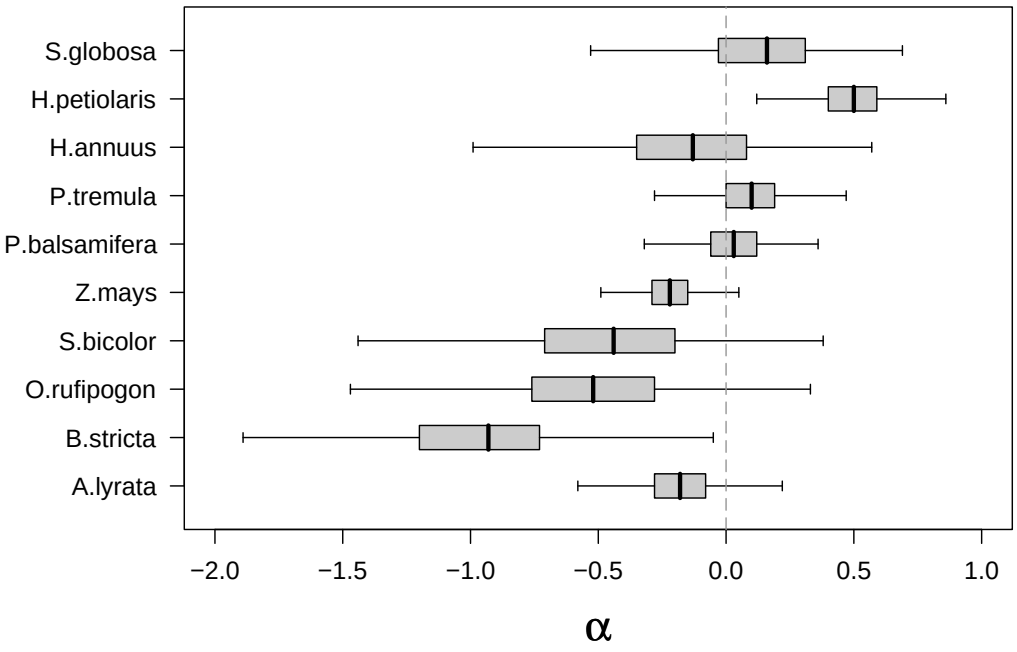
Figure 2.1: The distribution of fitness effects in 11 plant species. Estimates are given for the proportion of mutations in four different $N_e s$ ranges and standard errors. For *Z.mays* the results are shown for *Z.mays spp. parviglumis*

In contrast to this, *B.stricta*, *P.balsamifera* and *S.globosa* show an excess of neutral mutations (>25%) and a decrease of strongly deleterious mutations (<55%). The confidence intervals vary between species due to the varying sample sizes. Our estimates of

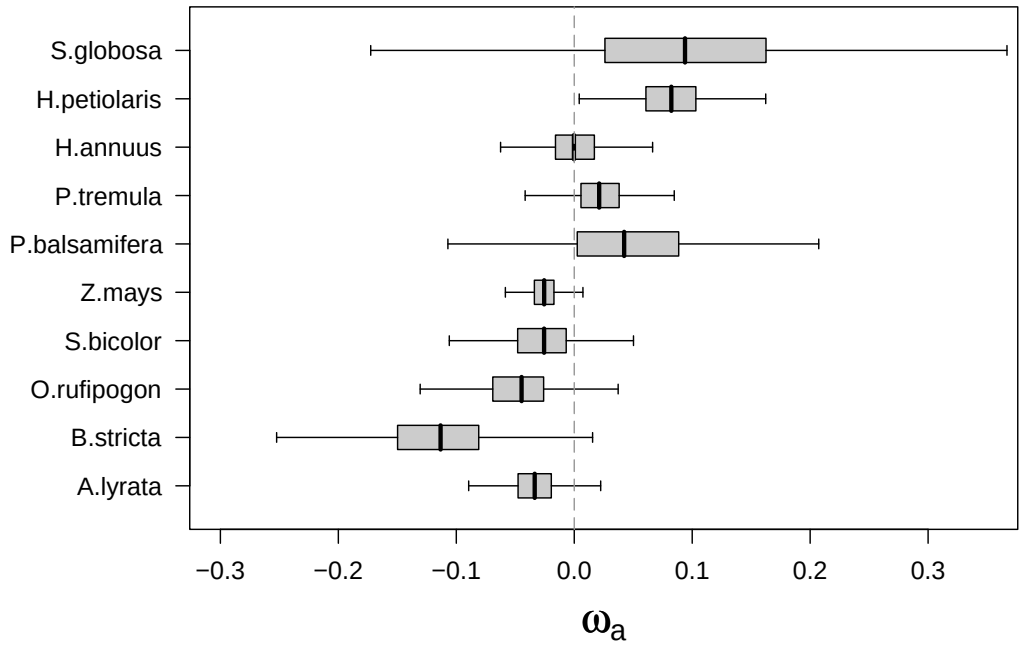
the DFE are not greatly affected by the removal of singletons suggesting that sequencing errors are not an issue (Figure A2.2). We additionally estimated the distribution of fitness effects from individual wild populations of *Arabidposis*, *Populus*, *Boechera* and *Schiedea*. For all populations of *A.thaliana* and *A.lyrata* (Figure A2.3) the distributions between and within these species are fairly similar to each other, although the confidence intervals for *A.lyrata* are larger because of the smaller data set. In all the investigated *Arabidopsis* populations more than 50% of the mutations are strongly deleterious ($N_e s>100$) and more than 18% behave as effectively neutral ($0<N_e s<1$). In contrast, populations of *B.stricta* and *S.globosa* show differences in their DFE, while populations of *P.balsamifera* are remarkably similar to each other (Figure A2.4). Additionally we find some differences between the DFE estimated from the wild and domesticated populations of rice (Figure A2.5). Both domesticated varieties of *O.japonica* show a higher proportion of effectively neutral mutations than *O.rufipogon*, which may reflect the lower effective population of the domesticated varieties; the DFE for the domesticated populations of *Oryza japonica* are more like that found in *B.stricta*, *P.balsamifera* and *S.globosa*. In contrast *O.indica* shows a very different DFE both to other *Oryza* species and to all other plant species, with many mutations inferred to be slightly or mildly deleterious ($1< N_e s <100$). However, the confidence intervals are large. The DFE inferred from domesticated *Zea mays* is similar to the wild population. These results suggest that the method to estimate the DFE copes well with the different demographic histories these populations have experienced, except possibly during some domestication events.

2.4.3 Adaptive substitutions

Surprisingly we find little evidence of adaptive evolution in any of the species we have considered (Figure 2.2), except in one sunflower comparison, despite the fact that we have controlled for the effects of slightly deleterious mutations, which tend to bias estimates of adaptive evolution downwards. The estimates of α are positive in the comparisons involving *Populus* species and *S.globosa* but these are less than 15% and non-significant; the estimate of α using the polymorphism data from *H.petiolaris* is significantly positive. All other comparisons are negative, with the estimate using *B.stricta* being significantly negative, but not after correction for multiple tests. We obtain similar results for estimates of α for 23 subpopulations of seven species (Table A2.1) with all positive estimates being non-significant.



(a)



(b)

Figure 2.2: Estimates of α and ω_a for 10 plant comparisons. Mean estimates as well as 50% and 95% confidence intervals are indicated by the box-plots. For *Z.mays* the results are shown for *Z.mays spp. parviglumis*. Note the *H.petiolaris* and *H.annuus* analyses are for the same divergence using polymorphism data from different species.

The mean estimate of α across all 10 species-wide comparisons is -0.18 (SE = 0.47). This is not significantly different from zero suggesting that plants generally go through very little adaptive evolution in their protein coding sequences, however it should be appreciated that we have overestimated the SE because of the nonindependence between some of the datasets. Similar results are obtained if we estimate ω_a, the rate of adaptive nonsynonymous substitution relative to the rate of synonymous substitution (Figure 2.2b).

Although the method, which we have used to control for slightly deleterious mutations in the estimation of α, appears to be robust (Eyre-Walker and Keightley, 2009), we also estimated the proportion of adaptive substitutions using a simple method based upon the McDonald-Kreitman test, which just uses the numbers of synonymous and non-synonymous polymorphisms and substitution summed across genes (Fay et al., 2001), rather than the full site frequency spectrum. Estimates of α are negative in all species except *H.petiolaris* and *S.globosa* when we consider all polymorphisms (Table A2.2). Furthermore the α estimates for the *Populus* species become positive if we remove polymorphisms below 15% to control for the effects of slightly deleterious mutations (Table A2.2). However all positive estimates based on the method of Fay et al. (2001) are non-significant.

2.5 Discussion

We have estimated the proportion of non-synonymous substitutions that are due to positive adaptive evolution in a range of plant species using an extension of the McDonald-Kreitman test. Surprisingly, we find little evidence of adaptive evolution in the plant species we have investigated; in four of the comparisons the estimates of α, the proportion of substitutions driven by positive selection, are greater than zero, but for three of them the estimates are less than 15% and not significantly different from zero. Similar results are obtained when we estimate ω_a, the rate of adaptive amino acid substitution relative to the rate of synonymous substitution. The low estimated rate of adaptive evolution is consistent with previous estimates in *Arabidopsis sp.*, *Sorghum* and *Zea* (Bustamante et al., 2002; Hamblin et al., 2006; Foxe et al., 2008; Ross-Ibarra et al., 2009); however in these previous analyses it was not possible to exclude the possibility that the low estimates of adaptive evolution were simply a consequence of the segregation of slightly deleterious mutations, which tend to bias estimates of adaptive evolution downwards (Fay et al., 2001; Charlesworth and Eyre-Walker, 2008). In our analyses we have used a method that explicitly models slightly deleterious mutations and

takes these into account. Our estimate of α in the *Helianthus* species is also broadly consistent with the estimate of 75% obtained by Strasburg et al. (2009). Using the same data, we find significant evidence of adaptive evolution when we consider the divergence between *H.annuus* and *H.petiolaris* using the polymorphism data from *H.petiolaris*, and a non-significant estimate of α when we use the polymorphism data from *H.annuus*. In both cases our estimate of α is lower than that obtained by Strasburg et al. (2009), but there are several methodological differences between the analyses so this is perhaps not surprising and the confidence intervals on all estimates are large. Our estimate of α is also lower than the ≈30% estimate obtained by Ingvarsson (2010) for the divergence between *P.tremula* and *P.trichocarpa* using a similar dataset but the method of Bierne and Eyre-Walker (2004); this seems to be largely due to the fact that we had to reduce those genes for which Ingvarsson had 38 haploid genomes to 24 so that all genes would have the same number of sampled chromosomes - this is a limitation of the Eyre-Walker and Keightley (2009) method used here. If we apply the method of Bierne and Eyre-Walker (2004) to our reduced dataset removing polymorphisms below a frequency of 15%, as Ingvarsson did, we estimate α to be 0.1303 (95% CIs: -0.2447, 0.3928), which is similar to the estimate obtained using the Eyre-Walker and Keightley (2009) method.

There are several reasons why the estimated rate of adaptive evolution might be low in plants. First, limited evidence in animals suggests that the proportion of adaptive substitutions may depend on the effective population size; primates appear to have relatively low rates of adaptive evolution, and also low effective population sizes, whereas mice, *Drosophila* and enteric bacteria have high rates of adaptive evolution (Smith and Eyre-Walker, 2002; Bierne and Eyre-Walker, 2004; Welch, 2006; Charlesworth and Eyre-Walker, 2006; Bachtrog, 2008) and also large effective population sizes (Andolfatto, 2001; Charlesworth and Eyre-Walker, 2006; Piganeau and Eyre-Walker, 2009). There is some evidence of this relationship in plants as well, with species such as *Capsella grandiflora* and *Helianthus petiolaris*, which have large effective population sizes (Strasburg and Rieseberg, 2008; Foxe et al., 2009), showing high rates of adaptive evolution (Strasburg et al., 2009; Slotte et al., 2010). This relationship between population size and the rate of adaptive evolution is expected if the rate of adaptive evolution is limited by the supply of advantageous mutations, because species with large populations produce advantageous mutations at a higher rate, and selection is also more effective on those that are weakly selected. To investigate whether plants have low effective population sizes like primates

we estimated the effective population size in our plant species using estimates of θ_W from the loci involved in our analyses and estimates of the mutation rate per generation and generation time from the literature (see material and methods). These estimates should be treated with caution since the mutation rate has not been directly estimated in plants, and actual demographic processes presumably deviate from the assumptions of these models.

The effective population sizes vary quite substantially between our species from *B.stricta* with an effective population size of just over 25,000 to *Helianthus annuus* with an N_e of more than 800,000, but most species have relatively modest population sizes of tens of thousands to just over one hundred thousand (Figure 2.3). Our estimates agree closely with those previous estimates for *O.rufipogon*, but our estimate is lower than the previous estimate from *P.tremula*; Ingvarsson (2008b) used the mode θ_W value and estimated N_e to be 118,000, which is roughly two times larger than our estimate. As expected our estimates for *Helianthus* species are mid-way between the estimates of the ancestral and derived population sizes for these species, which have undergone a recent population size expansion (Strasburg and Rieseberg, 2008).

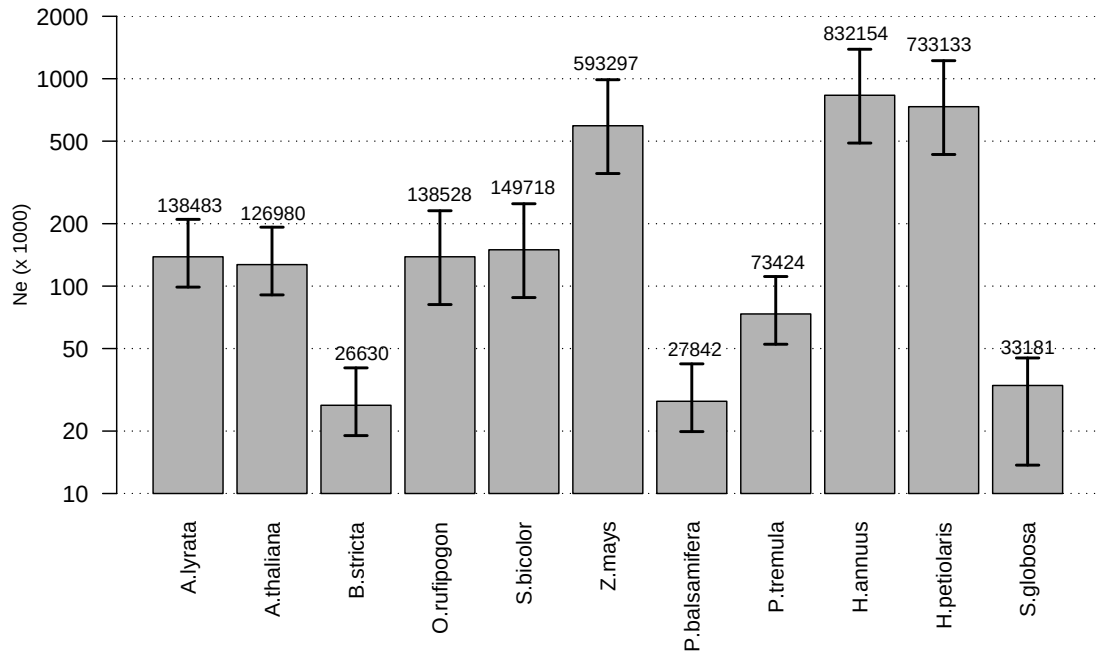


Figure 2.3: Estimates of the effective population sizes (N_e). Estimates of the effective population sizes (N_e) for 11 different plant species, based on θ_W. The confidence intervals are based on SE of the mutation rate μ.

The population specific estimates of N_e (Table A2.1) are similar to the overall estimate of N_e in *A.thaliana*, *B.stricta*, *P.balsamifera* and *S.globosa* and the estimates for the domesticated populations of rice and maize are substantially lower than for their wild relatives. In contrast to the other wild populations, in *A.lyrata* the overall N_e is considerably greater than the population specific estimates except the German population, reflecting the high level of population structure in this species (Ross-Ibarra et al., 2008). It therefore seems that most of these plants generally have higher effective population sizes than humans and chimpanzees, which have N_e values of 10,000 to 30,000 (Eyre-Walker et al., 2002). *Zea mays* has an effective population size of 500,000, which is of similar magnitude to that of rodents (*Mus domesticus* = 160,000, (Eyre-Walker et al., 2002), *Mus castaneus* = 500,000 (Halligan et al., 2010)), species for which α is estimated to be 60%, and yet it shows no evidence of adaptive evolution. Only the sunflower species show some evidence of adaptive evolution in our analysis and these do indeed have the largest effective population sizes; also *Capsella grandiflora* shows evidence of adaptive evolution (Slotte et al., 2010) and a large effective population size (Foxe et al., 2009). These results therefore suggest that the rate of adaptive evolution might be correlated to effective population size, but that many plants do not undergo high rates because they have relatively small population sizes.

Although there appears to be no relationship between the current estimate of the effective population size and the rate of adaptive evolution in plants, the low estimates of α and ω_a could be an artefact of a contracting population. If there are slightly deleterious mutations segregating and current population sizes are smaller than during the divergence between the species being considered, then α (and ω_a) will be underestimated (McDonald and Kreitman, 1991; Eyre-Walker, 2002), because some SDMs that currently segregate would not have segregated or become fixed in the past. The extent of the bias depends on the magnitude of the difference in effective population size between the polymorphism and divergence phases, the distribution of fitness effects of new mutations and the true level of adaptive evolution (Eyre-Walker, 2002; Eyre-Walker and Keightley, 2009). However, contracting effective population size seems an unlikely explanation for the patterns we observe in plants since we see little evidence of adaptive evolution across 9 comparisons. In contrast, an expansion in population size might explain why sunflowers show evidence of adaptive evolution since they have recently undergone population expansion (Strasburg and Rieseberg, 2008) and expansion can lead to an overestimate of adaptive evolution

(McDonald and Kreitman, 1991; Eyre-Walker, 2002; Eyre-Walker and Keightley, 2009).

A third possible explanation for the low levels of adaptive evolution in plants is population structure. Population structure can reduce the probability that an advantageous mutation will become fixed if migration rates are low because of the increased influence of drift within sub-divided populations and the added influence of population extinction (Charlesworth et al., 2003; Whitlock, 2003). Unfortunately, we do not have information about population structure for most of the species we have considered; *A.thaliana* and *A.lyrata* show contrasting levels of structure, so it is unclear what level of structure there was during the divergence of these species. Population structure could affect our estimates in two ways. First, it is possible that population structure could lead to biases in our method to estimate the distribution of fitness effects and hence the rate of adaptive evolution; this seems unlikely because our method gives similar estimates of the DFE for both the total population and each sub-population of *Arabidopsis*. Furthermore, it gives similar estimates to the method of Fay et al. (2001) when rare SNPs are excluded. However, to investigate the matter further we performed a series of simulations in which we had two subpopulations with varying levels of migration between them (Figures A2.1a/b). Our method successfully estimates the proportion of adaptive substitutions irrespective of the level of migration (Table A2.3) if the polymorphism data are sampled from one subpopulation only. The method of Eyre-Walker and Keightley (2009) overestimated α if the level of migration was low between the two subpopulations and the polymorphism data included samples from both subpopulations. This is due to an excess of fixed differences between the two subpopulations which will be treated as polymorphisms in a common dataset and hence strongly affect the SFS.

We may have also underestimated α because recent gene duplications are probably under-represented in our data because of problems of identifying orthologs. Plant genomes contain high frequencies of duplicated genes (AGI, 2000; IRGSP, 2005; Tuskan et al., 2006; Jaillon et al., 2007) due to polyploidization (Blanc and Wolfe, 2004; Soltis and Soltis, 2009) and tandem duplications (Rizzon et al., 2006). These have been shown to have higher levels of adaptive evolution than single copy genes, at least in mammals (Han et al., 2009). Thus we may be missing much of the adaptive evolution that goes in plants simply because it occurs in the divergence of paralogs which we have not sampled. There may also be a bias towards conserved genes, particularly

in the more divergent taxa, because of the need to be able to correctly identify the ortholog; if conserved genes undergo less adaptive evolution then α will be underestimated.

In the MK framework it is assumed that one category of mutations, in this analysis the synonymous mutations, are neutral. However, there is evidence in at least in *Populus* species that selection acts upon synonymous codon use (Ingvarsson, 2008a). Predicting the effects of selection, on the putatively neutral sites, on the estimation of the distribution of fitness effects and rates of adaptive evolution is not straightforward and deserves more investigation. If non-synonymous polymorphisms are on average more deleterious than synonymous mutations then the MK test and methods that estimate α from it, are expected to be conservative if populations sizes are stationary, this is because the nonsynonymous polymorphisms are less likely to be fixed than the synonymous polymorphisms. With non-stationary population size the predictions become complex. The fact that our method estimates a proportion of mutations to be slightly deleterious ($1< N_e s <10$) in all species suggests that non-synonymous polymorphisms are on average more deleterious than their synonymous counterparts. However, it is possible that selection at synonymous sites along with demographic changes is inducing an underestimate of α.

One potential problem with some of our comparisons is the very low level of divergence between the ingroup and outgroup species. It is intriguing that all three of the species pairs for which α is positive have a low level of synonymous divergence relative to nucleotide diversity: *H.petiolaris* (polymorphism)-*H.annuus*(outgroup) = 1.7; *S.globosa*-*S.adamantis* = 1.1; *P.tremula*-*P.trichocarpa*=4.7; *P.balsamifera*-*P.trichocarpa*=2.1. Only *H.annuus*-*H.petiolaris* and *S.bicolor*-*S.propinquum* have comparable ratios; all other comparisons have a ratio greater than 10. This may lead to either an over-estimation of α because the fixation time becomes important; advantageous and slightly deleterious mutations go to fixation faster than neutral mutations if they are due to fix (Kimura, 1983) so there may be a time after the divergence of species where sites subject to selection are more likely to show fixed differences than neutral sites because the mutations are more likely to have already spread to fixation (Bierne and Eyre-Walker, 2004). The application of the McDonald-Kreitman framework to closely related species requires more investigation. While some of the species are very closely related to each other rice is highly divergent from its outgroup sorghum; this could lead to an overestimate of α since we

might expect synonymous sites to become saturated. There is no evidence of this effect in our data because the estimate of α for the divergence between rice and sorghum is negative.

The lack of adaptive evolution in plants is particularly surprising in some genera, such as *Schiedea*. *Schiedea* is a Hawaiian endemic that adapted to a wide range of ecological conditions (from cool rainforest to arid desert-like conditions of coastal cliffs) and evolved extremely different morphology (ranging from vines to bushes) and reproductive biology (dioecy, gynodioecy, hermaphrodites) (Wagner et al., 2005). Given fairly low divergence between species in the genus ($K_s < 4\%$), all the morphological and ecological diversity in *Schiedea* has evolved surprisingly rapidly. Thus, one might expect that species in the genus evolved under fairly strong selection, which should be detectable in *Schiedea* genes. Indeed, phylogeny-based maximum likelihood analysis revealed that positive selection is more widespread in *Schiedea* genes, compared to mainland plant groups (Kapralov, Votintseva and Filatov, unpublished data). Thus, it is somewhat surprising that this rapidly evolving Hawaiian endemic does not show significantly positive α. However, in the current paper, the *Schiedea* dataset is the smallest of all, and adding more genes might provide evidence of positive selection in *Schiedea* genes. On the other hand, the current analysis is restricted to population-level evolutionary processes in only one of the species. Thus, the approach presented in the current paper cannot detect positive selection that might have acted during the early history of the genus while it adapted to diverse ecological conditions.

Although, the influence of effective population size on the rate of adaptive evolution in plants is unclear, there is a clear effect of N_e on the DFE; species with small N_e, such as *B.stricta*, *P.balsamifera* and *S.globosa* tend to have a relatively high proportion of mutations that are effectively neutral; there is a significant negative correlation between P_n/P_s and N_e, even when we take into account the non-independence between these two statistics (Figure 2.4; r = -0.75; p=7×10^{-3}, Spearman Rank Test) using the method of Piganeau and Eyre-Walker (2009) (note we do not add one to denominator, as suggested by Piganeau and Eyre-Walker since P_s is sufficiently large that this correction makes no difference). In contrast to the proportion of mutations that are effectively neutral, the proportion of mutations that are slightly deleterious, $1< N_es <10$, and to a lesser extent the proportion of mutations that are moderately deleterious, $10< N_es <100$, do not vary very much between species.

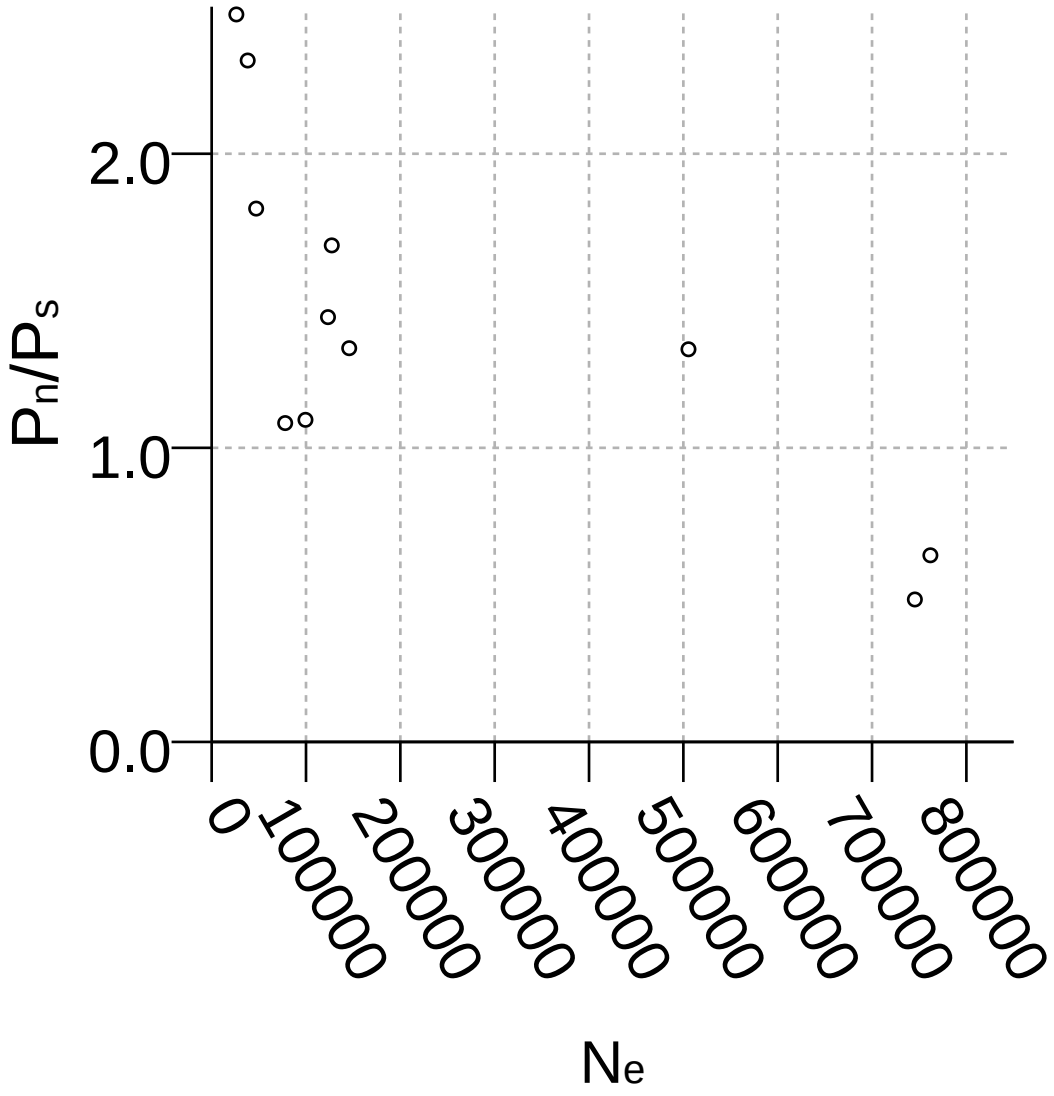


Figure 2.4: The correlation between P_n/P_s and N_e. P_n/P_s and N_e were estimated independently from two estimates of P_s that were obtained by splitting P_s into two independent halves according to the method suggested by Piganeau and Eyre-Walker (2009).

2.6 Conclusions

We have estimated the proportion of non-synonymous substitutions that are a consequence of positive adaptive evolution between 9 pairs of plant species; we find little evidence of adaptive evolution in any of them. This is in striking contrast to *Drosophila*, bacteria and mice, in which rates of adaptive evolution have been estimated to be substantial. The low estimate is unlikely to be a methodological artefact, and it does not appear due to low effective population size since some plants appear to have population sizes that rival that of mice. It therefore seems that the rate of adaptive evolution may be determined by other reasons as much as by effective population size and plants may have an outstanding role.

Chapter 3

The effect of variation in the effective population size on the rate of adaptive molecular evolution in eukaryotes.

3.1 Abstract

The role of adaptation is a fundamental question in molecular evolution. Theory predicts that species with large effective population sizes should undergo a higher rate of adaptive evolution than species with low effective population sizes, if adaptation is limited by the supply of mutations. Previous analyses have appeared to support this conjecture, since estimates of the proportion of non-synonymous substitutions fixed by adaptive evolution, α, tend to be higher in species with large N_e. However, α is a function of both the number of advantageous and effectively neutral substitutions, either of which might depend on N_e. Here we investigate the relationship between N_e and ω_a, the rate of adaptive evolution relative to the rate of neutral evolution, using nucleotide polymorphism and divergence data from 13 independent pairs of eukaryotic species. We find a highly significant positive correlation between ω_a and N_e. We also find some evidence that the rate of adaptive evolution varies between groups of organisms for a given N_e. The correlation between ω_a and N_e does not appear to be an artefact of demographic change or selection on synonymous codon use. Our results suggest that adaptation is to some extent limited by the supply of mutations and that at least some adaptation depends on newly occurring mutations, rather than standing genetic variation. Finally, we show that the proportion

of nearly neutral non-adaptive substitutions declines with increasing N_e. The low rate of adaptive evolution and the high proportion of effectively neutral substitution in species with small N_e are expected to combine to make it difficult to detect adaptive molecular evolution in species with small N_e.

3.2 Introduction

Population genetic theory predicts that the effective population size (N_e) of a species should be a major determinant of the rate of adaptive evolution if adaptive evolution is limited by the supply of new mutations. There are two reasons for this. First, the rate of adaptive evolution is expected to be proportional to $N_e s$ if $N_e s >> 1$, where s is the strength of selection. This is because the fixation probability of a new advantageous mutation is proportional to $N_e s/N$, where N is the census population size, if $N_e s >> 1$ and s is small (Kimura, 1983) and the rate at which new advantageous mutations occur is Nu; hence the rate of adaptive evolution is expected to be proportional to $Nu \times N_e s/N = uN_e s$. Second, in large populations, a higher proportion of mutations are expected to be effectively selected because a higher proportion are expected to have $N_e s >> 1$. Previous analyses have suggested that the proportion of adaptive substitutions (α) is correlated to the effective population size, since there is evidence of widespread adaptive amino acid substitutions in species such as Drosophila, house mice, bacteria and some plant species with large N_e (Smith and Eyre-Walker, 2002; Bustamante et al., 2002; Bierne and Eyre-Walker, 2004; Sawyer et al., 2003; Charlesworth and Eyre-Walker, 2006; Ingvarsson, 2010; Slotte et al., 2010; Haddrill et al., 2010; Strasburg et al., 2011b), whereas there is little evidence in hominids and other plant species that appear to have small N_e (Zhang and Li, 2005; Chimpanzee Sequencing and Analysis Consortium, 2005; Boyko et al., 2008; Eyre-Walker and Keightley, 2009; Gossmann et al., 2010). There are, however, some exceptions. Maize, for example, has a relatively large effective population size, approaching that of wild house mice, but shows little evidence of adaptive protein evolution (Gossmann et al., 2010), and the yeast *S. paradoxus*, which presumably has a very large N_e, also shows little evidence of adaptive protein evolution (Liti et al., 2009). Furthermore, *Drosophila simulans* does not appear to have undergone more adaptive evolution than *D. melanogaster* even though it is thought to have a larger N_e (Andolfatto et al., 2011b).

However, the correlation between α and N_e might be misleading, since α depends on

the rate of effectively neutral and advantageous substitution, variation in either of which could be caused by N_e (Gossmann et al., 2010); i.e. $\alpha = D_{adaptive}/(D_{adaptive}+D_{nonadaptive})$ where $D_{adaptive}$ and $D_{nonadaptive}$ are the rates of adaptive and non adaptive substitutions, respectively. There is evidence that the proportion of effectively neutral mutations is negatively correlated to N_e across many species (Popadin et al., 2007; Piganeau and Eyre-Walker, 2009), so a positive correlation between α and N_e might be entirely explained by variation in the number of effectively neutral substitutions. As a consequence, it has been suggested that ω_a, the rate of adaptive substitution relative to the rate of neutral evolution is a more appropriate measure of adaptive evolution for the purpose of comparison between genomic regions or species (Gossmann et al., 2010, see also Bierne and Eyre-Walker 2004; Obbard et al. 2009); i.e. $\omega_a = D_{adaptive}/D_{neutral}$ where $D_{neutral}$ is the substitution rate at sites that evolve neutrally. Contrary to expectation Gossmann et al. (2010) failed to find any evidence of a correlation between ω_a and N_e in plants; but many of the plant species they considered appeared to have low N_e, and there may have been insufficient information from species with larger N_e to reveal a significant positive correlation. In contrast, Strasburg et al. (2011b) have recently reported a significant positive correlation between ω_a and N_e within sunflowers, including some species that have very large N_e. There are two interpretations of a positive correlation between ω_a and N_e in sunflowers. Firstly, the correlation could be due to a higher rate of adaptive substitution or secondly it could be due to an artifact of population size change (Strasburg et al., 2011b). It has long been known that approaches to estimate adaptive evolution by methods related to the MK test are sensitive to changes in N_e, if there are slightly deleterious mutations (McDonald and Kreitman, 1991; Eyre-Walker, 2002; Eyre-Walker and Keightley, 2009). For example, if the population has recently expanded, then ω_a and α, will tend to be over-estimated because slightly deleterious mutations, which would have become fixed in the past when the population size was small, no longer segregate as polymorphisms. This bias might be a particular problem in the sunflower dataset, because each species was contrasted against a common outgroup species, so that each comparison shared much of its divergence with all other comparisons. Therefore any differences in N_e between the species must have occurred since they split, and may have caused a genuine or an artefactual increase in ω_a. It is difficult to differentiate between these effects.

In contrast to the pattern in sunflowers, Jensen and Bachtrog (2011) recently estimated the rate adaptive evolution in *D. pseudoobscura* and *D. miranda*; they estimated

that the two species probably had similar ancestral population sizes, but that *D. miranda* had gone through a recent severe bottleneck. Despite this the estimate of α along the two lineages was quite similar.

It is also evident that estimates of α or ω_a and N_e are not independent since N_e is usually estimated from the neutral diversity, which is also used to estimate α or ω_a. Sampling variation will therefore tend to induce a positive correlation between estimates of adaptive evolution and effective population size. This can be dealt with by randomly splitting the neutral sites into two halves, one of which is used to estimate N_e and the other to estimate the rate of adaptive evolution (Piganeau and Eyre-Walker, 2009; Stoletzki and Eyre-Walker, 2011). This correction is accurate whether or not the sites are linked (Piganeau and Eyre-Walker, 2009).

3.3 Materials and Methods

3.3.1 Preparation of data

Polymorphism data were retrieved from Genbank `http://www.ncbi.nlm.nih.gov/Genbank` or in case of *Arabidopsis thaliana* downloaded from `http://walnut.usc.edu/2010`. A summary of the analyzed datasets is shown in Table 3.1. Phylogenetic trees for the plant and Drosophila species used in our analysis are given in Supplementary Figures A3.1 and A3.2, respectively (Drosophila 12 Genomes Consortium et al., 2007; Tang et al., 2008; Stevens, 2010). Sequences were aligned using Clustalw using default parameter values (Thompson et al., 1994). Coding regions were assigned using protein coding genomic data coordinates, or, if given, derived from the information in the GenBank input files. An outgroup was assigned using the best BLAST (Altschul et al., 1990) hit against the outgroup genome, or, if included, taken from the GenBank Popset database (`http://www.ncbi.nlm.nih.gov/popset`). For all analyses synonymous sites served as the neutral standard. Because some loci had been sampled in more individuals than others and other loci had missing data, we obtained the site frequency spectra (SFS) for each number of chromosomes for each species (e.g. we obtained the SFS for those sites with 4, 5, ...etc chromosomes separately). As a consequence there was usually more than one SFS and its associated divergence data for each species. The estimation of the distribution of fitness effects (DFE) and ω_a was done jointly using all available SFS and divergence data for a given species. Summary statistics, such as π, were

Species	Outgroup	Loci	Dataset
Drosophila melanogaster	*D. simulans*	373	Shapiro et al. (2007)
D. miranda	*D. affinis*	76	Haddrill et al. (2010)
D. pseudoobscura	*D. persimilis*	72	Haddrill et al. (2010)
Homo sapiens	*Macaca mulatta*	445	EGP/PGA[1]
Mus musculus castaneus	*Rattus norvegicus*	77	Halligan et al. (2010)
Arabidopsis thaliana	*A. lyrata*	932	Nordborg et al. (2005)
Capsella grandiflora	*Neslia paniculata*	251	Slotte et al. (2010)
Helianthus annuus	*Lactuca sativa*	34	Strasburg et al. (2011b)
Populus tremula	*P. trichocarpa*	77	Ingvarsson (2008b)
Oryza rufipogon	*Oryza spp.*	106	Caicedo et al. (2007)
Schiedea globosa	*S. adamantis*	23	Gossmann et al. (2010)
Zea mays	*Sorghum bicolor*	437	Wright et al. (2005)
Saccharomyces paradoxus	*S. cerevisiae*	98	Tsai et al. (2008)

[1] EGP: http://egp.gs.washington.edu and PGA: http://pga.gs.washington.edu August 2010

Table 3.1: Summary of datasets used for the analyses.

calculated as weighted averages. The numbers of synonymous and nonsynonymous sites and substitutions were computed using the F3x4 model implemented in PAML (Yang, 1997) in which codon frequencies are estimated from the nucleotide frequencies at the three codon positions.

It is important in this type of analysis to count the numbers of synonymous and non-synonymous sites correctly and consistently across the divergence and polymorphism data. It is appropriate to use a "mutational opportunity" definition of a site (Bierne and Eyre-Walker, 2003) since we are interested in the relative numbers of mutations that can potentially occur at synonymous and non-synonymous sites. PAML provides estimates of the proportion of sites that are non-synonymous (and hence also synonymous) from the divergence data and these were used to calculate the number of non-synonymous and synonymous sites for the polymorphism data.

3.3.2 Estimation of N_e and ω_a

We assumed that synonymous sites were neutral, except when we estimated the strength of selection on synonymous mutations (see below). We estimated N_e from the level of nu-

cleotide diversity, π, at synonymous sites and estimates of the rate of nucleotide mutation per generation, μ, from the literature, since

$$N_e = \frac{\pi}{4\mu} \tag{3.1}$$

We estimated the mutation rate per generation in *Populus tremula* in the following manner. Tuskan et al. (2006) note that sequence divergence in putatively neutral sequences is approximately six times slower in *P. tremula* than in *A. thaliana*, and that the average generation time for *P. tremula* is $\approx$ 15 years. We therefore estimated the mutation rate per generation in *P. tremula* by multiplying the mutation rate estimated in *A. thaliana* from mutation accumulation lines by $15/6 = 1.75 \times 10^{-8}$.

The DFE and ω_a, the rate of adaptive substitutions relative to the rate of synonymous substitutions (Gossmann et al., 2010) were estimated using a modified version of the method of Eyre-Walker and Keightley (2009). First, the DFE and demographic parameters of the population are simultaneously estimated from the SFS of non-synonymous and synonymous sites using the method of Keightley and Eyre-Walker (2007). The DFE is then used to estimate the average fixation probability of mutations $\overline{f_n}$ at non-synonymous sites relative to that at neutral sites:

$$\overline{f_n} = \int_{-\infty}^{0} M(S)Q(S)\,dS \tag{3.2}$$

where $S = 4N_e s$, s is the strength of selection, $M(S)$ is the distribution of S as inferred by the method of Keightley and Eyre-Walker (2007) and

$$Q(S) = \frac{S}{1 - e^{-S}} \tag{3.3}$$

is the fixation probability of a new mutation relative to the fixation probability of a neutral mutation (Kimura, 1983). The rate of adaptive non-synonymous substitution relative to the rate of synonymous substitution, ω_a, can then be estimated as

$$\omega_a = \frac{d_n - d_s\overline{f_n}}{d_s} = \frac{d_n}{d_s} - \overline{f_n} \tag{3.4}$$

where d_n and d_s are the rates of non-synonymous and synonymous substitution respectively. The method of Eyre-Walker and Keightley (2009) does not take into account the fact that some substitutions between species are polymorphisms. This was taken into account in the following manner (Keightley and Eyre-Walker, 2012). The Keightley and Eyre-Walker (2007) method estimates the DFE and demographic parameters by

generating vectors representing the allele frequency distributions for synonymous and non-synonymous sites by a transition matrix approach and using these to calculate the likelihood of the observed SFS. Let the density of mutations at i of 2N copies be $v_n(i)$ and $v_s(i)$ for non-synonymous and synonymous sites respectively, and let us assume that we have sampled a single sequence from each species to estimate the divergence. The contribution of polymorphisms to apparent divergence is therefore

$$d'_n = 2\sum_{i=1}^{2N-i} \frac{i}{2N} v_n(i) \tag{3.5}$$

for non-synonymous sites, with an analogous expression for synonymous sites. The factor of two appears because polymorphism in both lineages contributes to apparent divergence and we assume that the diversity is the same in the two lineages. We can now estimated ω_a taking into account the contribution of polymorphism to divergence as

$$\omega_a = \frac{d_n - {d_n}'}{d_s - {d_s}'} - \overline{f_n} \tag{3.6}$$

We also estimated ω_a using a model in which there was negative selection upon synonymous mutations. We assume that all synonymous mutations are subject to the same strength of selection. Unfortunately, it is not possible to simultaneously estimate the demographic parameters and the strength of selection on synonymous mutations unless one includes information about which codons are preferred by selection (Zeng and Charlesworth, 2009), and this is not known for most of the species in our analysis. We therefore infer the strength of selection at synonymous sites from the SFS using the transition matrix approach described in Keightley and Eyre-Walker (2007) assuming a constant population size. The strength of selection at synonymous sites allows us to calculate the probability of fixation of synonymous mutations f_s and obtain a corrected estimate of ω_a as

$$\omega_a = \frac{d_n - {d_n}'}{d_s - {d_s}'} - \frac{\overline{f_n}}{f_s} \tag{3.7}$$

It is also necessary to adjust our estimate of N_e to take into account of the action of natural selection at synonymous sites. This was performed in two ways depending upon whether our estimate of the mutation rate was a direct estimate from a pedigree or mutation accumulation experiment, as in the *Drosophila* species, *Arabidopsis*, *Capsella*, *Populus* and *Saccharomyces*, or indirectly from phylogenetic analysis, as in *Mus*, *Helianthus*, *Oryza*, *Schieda* and *Zea*. Kimura (1969) showed that the nucleotide diversity at a site subject to recurrent mutation and semi-dominant selection, of strength s (positive s for advantageous mutations), relative to that at a neutral site is

$$H(S) = \frac{2(S - 1 + e^{-S})}{S(1 - e^{-S})} \tag{3.8}$$

For those species in which the mutation rate had been estimated directly we corrected the estimate of N_e obtained from equation 3.1, by dividing it by $H(S)$, where S is the strength of selection acting at synonymous sites; for those species in which the mutation rate came from a phylogenetic analysis we corrected for selection at synonymous sites by multiplying the estimates by $Q(S)/H(S)$. Synonymous codon bias was measured using the effective number of codons (ENC, Wright, 1990) and ENC taking into account base composition bias (ENC', Novembre, 2002). To investigate whether the proportion of effectively neutral non-synonymous mutations was correlated to N_e we calculated a variant on the ψ statistic suggested by Piganeau and Eyre-Walker (2009):

$$\psi = \frac{L_s P_n}{L_n(P_s + 1)} \tag{3.9}$$

where P_n and P_s are the numbers of non-synonymous and synonymous polymorphisms, and L_n and L_s are the numbers of non-synonymous and synonymous sites. ψ is expected to be less biased than P_n/P_s.

3.3.3 Creation of independent datasets

Estimates of ω_a and N_e are not independent because they both depend on neutral diversity, so sampling error will tend to induce a positive correlation between N_e and ω_a. We avoided this problem by splitting the synonymous site data into two independent sets (which is similar to splitting the dataset into odd and even codons as in Smith and Eyre-Walker, 2002; Piganeau and Eyre-Walker, 2009; Stoletzki and Eyre-Walker, 2011) by generating a random multivariate hypergeometric variable as follows:

$$\mathbf{P_{s1}} = multivariateHypergeometric(\mathbf{P}, 0.5 \times L_s) \tag{3.10}$$

$$\mathbf{P_{s2}} = \mathbf{P} - \mathbf{P_{s1}} \tag{3.11}$$

where L_s is the number of sites and $\mathbf{P}$ a vector consisting of the number of non-mutated sites and the site frequency spectrum so that $\sum \mathbf{P} = L_s$. We use $\mathbf{P_{s1}}$ and $\mathbf{P_{s2}}$ to compute two corresponding independent variables N_{e1} and ω_{a2}. Note, that N_{e2} and ω_{a1} could be obtained in a similar manner, however results were qualitatively comparable and we therefore only show results for N_{e1} vs ω_{a2}. The same strategy was used to investigate the relationship between ψ and N_e.

3.4 Results

To investigate the correlation between the rate of adaptive evolution and N_e, we compiled data from 13 phylogenetically independent pairs of species (Table 3.1, Figures A3.1 and

A3.2). We measured the rate of adaptive evolution using the statistic ω_a, which is the rate of adaptive substitution at non-synonymous sites relative to the rate of synonymous substitution, using a method that takes into account the contribution of slightly deleterious mutations to polymorphism and divergence (Eyre-Walker and Keightley, 2009; Keightley and Eyre-Walker, 2012). We estimated N_e by dividing the synonymous site nucleotide diversity by an estimate of the mutation rate per generation, taken from the literature. We also divided the synonymous sites into two groups when estimating ω_a and N_e in order to ensure that the estimates were statistically independent. Estimates of ω_a and N_e are given in Table 3.2.

								Selection on silent sites		
Species	π	$\mu \times 10^9$	N_e	ω_a	N_2/N_1	ENC	ENC'	$4N_es$	N_e^*	ω_a^*
Drosophila melanogaster	0.019	5.8 [1]	822351	0.03	2.31	53.56	54.42	-0.0002	822379	0.04
D. miranda	0.008	5.8 [1]	334502	-0.00	4.95	43.27	49.27	-0.0002	334513	0.01
D. pseudoobscura	0.019	5.8 [1]	798607	0.27	4.5	43.28	48.62	-0.0008	798714	-0.06
Homo sapiens	0.001	11 [2]	20974	-0.04	4.09	53.39	54.61	-1.2118	26127	0.02
Mus musculus castaneus	0.008	3.4 [3]	573567	0.18	2.79	52.95	54.51	-0.4946	483026	0.31
Arabidopsis thaliana	0.007	7 [4]	266769	-0.04	4.95	54.98	56.46	-0.0016	266840	0.03
Capsella grandiflora	0.018	7 [4]	641262	0.06	2.8	55.08	56.11	-0.0186	643257	0.04
Helianthus annuus	0.024	10 [5]	593436	0.11	4.5	57.23	58.92	-0.2328	548293	0.14
Populus tremula	0.011	17.4 [4,6]	156368	0.06	1.5	55.98	57.43	-0.0002	156373	0.08
Oryza rufipogon	0.005	10 [7]	131083	-0.07	10	59.10	58.83	-3.4624	28643	0.06
Schiedea globosa	0.013	95 [8,9]	34075	-0.12	4.5	56.58	57.62	-0.001	34054	-0.14
Zea mays	0.019	10 [7]	464010	-0.00	3.07	59.05	59.01	-2.4864	168117	0.03
Saccharomyces paradoxus	0.002	0.2 [10]	2562065	-0.02	4.5	53.31	56.85	-0.0002	2562150	-0.06

* corrected for the effect of selection on synonymous sites

Table 3.2: Summary of the nucleotide diversity for silent sites π, mutation rate per generation μ from the literature, estimates of effective population sizes N_e, ω_a, ENC and ENC' for the 13 analyzed species. ω_a was estimated under a simple demographic model assuming a step change of N_e ($k = N_2/N_1$), where the ratio of N_2/N_1 > 1 and < 1 indicate recent population size expansion and contraction, respectively. Estimates of the strength of selection on synonymous sites $4N_es$ and corresponding corrected estimates of N_e and ω_a. The strength of selection s on synonymous mutations was estimated assuming a constant population size. Literature Sources for mutation rates: [1] Haag-Liautard et al. (2007), [2] Roach et al. (2010), [3] Keightley and Eyre-Walker (2000), [4] Ossowski et al. (2010), [5] Strasburg and Rieseberg (2008), [6] Tuskan et al. (2006), [7] Swigonová et al. (2004), [8] Filatov and Burke (2004), [9] Wallace et al. (2009), [10] Fay and Benavides (2005)

There is a non-significant positive correlation between ω_a and N_e for the individual data points (Pearson's correlation $r = 0.16, P = 0.61$, Figure 3.1). However, there is also a positive correlation between the two variables for all groups for which we have two or more data points (Plants: $r = 0.74, P = 0.056$; *Drosophilidae*: $r = 0.55, P = 0.63$; Mammals: r=1.00, P not given since there are just two data points) suggesting that differences between taxonomic groups may obscure a significant correlation within the groups. To investigate this further, we performed an analysis of covariance (ANCOVA), grouping organisms as mammals, plants, Drosophila and fungi. In ANCOVA, a set of parallel lines are fitted to the data, one for each group. This enables a test of whether the common slope of these lines is significantly different from zero, and one can also investigate whether the groups differ in the dependent variable for a given value of the independent variable by testing whether the lines have different intercepts. Using ANCOVA, we find that ω_a and N_e are significantly positively correlated (P=0.017). Furthermore, there is significant variation between the intercepts (P=0.044). There is also a positive correlation between ω_a and $log(N_e)$ (P=0.018), although the difference between intercepts is no longer significant (P=0.12). The results therefore suggest that ω_a and N_e are positively correlated, and that the level of adaptive evolution may vary between groups for a given N_e.

The correlation between ω_a and N_e might be genuine, but it might also have arisen as an artefact, generated by changes in population size. For example, if species with large current N_e tend to have undergone population expansion and/or species with small N_e population size contraction, then a positive correlation between ω_a and N_e would be induced because population size expansion leads to an overestimate of ω_a and contraction to an underestimate if there are slightly deleterious mutations (Eyre-Walker, 2002). We investigated whether changes in population size explain the correlation between ω_a and N_e by taking advantage of the fact that the method we used to estimate ω_a simultaneously fits a demographic model to the data. In this model the population experiences a k-fold change in population size t generations in the past. The results of our analysis suggest that the correlation between the estimates of N_e and $\log(k)$ are weak and non-significant (Pearson: r = -0.41, P = 0.15; ANCOVA: slope P = 0.61) or between $\log(N_e)$ and $\log(k)$ (Pearson r = -0.15, P = 0.61; ANCOVA: slope P = 0.93); thus there is no evidence that species with large current N_e have undergone recent expansion and/or that species with small current N_e have undergone recent contraction. We also find little evidence that ω_a is correlated to $\log(k)$ (Pearson r = 0.17, P = 0.57; ANCOVA: P = 0.97), implying that

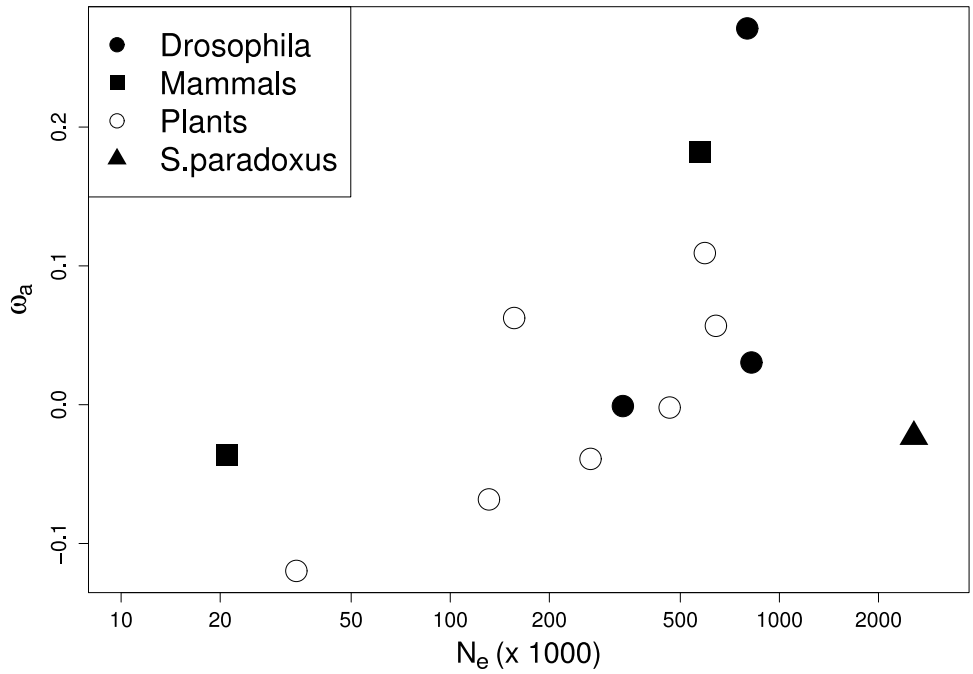


Figure 3.1: The rates of adaptive evolution (ω_a) versus the effective population size (N_e) for 13 species grouped into four phylogenetic sets. Details concerning the analysed species can be found in Table 3.1.

the correlation between ω_a and N_e is not an artefact of changes in population size. It should be noted, however, that this test is not definitive because MK-based approaches are sensitive to differences in the N_e experienced by the polymorphism and the divergence data (McDonald and Kreitman, 1991; Eyre-Walker, 2002; Eyre-Walker and Keightley, 2009). For example, a species might have experienced an expansion, that predates the origin of the polymorphism data, but is nevertheless recent in comparison to the overall divergence between the species being considered. In this case, there would be no evidence of expansion in the polymorphism data, but N_e for the polymorphism data would be greater than the average N_e during the divergence of the species. This would artefactually increase the estimate of ω_a.

A second explanation for the correlation between ω_a and N_e could be selection at synonymous sites. If the effectiveness of selection on synonymous sites increases with N_e, then this predicts a decrease in the level of synonymous divergence relative to polymorphism, leading to over-estimation of adaptive non-synonymous evolution. Although we might expect the effectiveness of selection on synonymous sites to increase with N_e, the evidence is mixed. Selection appears to be more effective on synonymous codon bias in *Drosophila simulans* than *D. melanogaster* (Akashi, 1996; McVean and Vieira,

2001), and N_e is thought to be larger in the former species (Aquadro et al., 1988; Akashi, 1996). However, in mammals selection appears to be more effective on synonymous sites in hominids than rodents (Eory et al., 2010), yet N_e is substantially larger in wild mice than hominids (Eyre-Walker, 2002; Halligan et al., 2010). Furthermore, selection on synonymous codon use appears to have little effect on estimates of α in *D. pseudoobscrura*, *D. miranda* and *D. affinis* (Haddrill et al., 2010).

To investigate whether the correlation between ω_a and N_e might be due to selection on synonymous sites, we performed two analyses. First, we investigated whether ω_a and our estimate of N_e were correlated to codon usage bias, as measured by the effective number of codons (ENC) and ENC taking into account base composition (ENC'). ω_a is negatively correlated to ENC and ENC', as expected if selection on synonymous codon use was causing an artifactual increase in ω_a, but in neither case was the correlation significant (ENC v ω_a: r = -0.481, P = 0.096; ANCOVA slope P = 0.40; ENC' v ω_a: r = -0.495, P = 0.085; ANCOVA slope P = 0.430). Furthermore, the correlation between N_e or $log(Ne)$, and ENC or ENC' are nonsignificant (ENC v N_e: r = -0.15, P = 0.61; ANCOVA slope P = 0.61; ENC' v N_e: r = -0.04, P = 0.89; ANCOVA slope P = 0.66; ENC v $log(N_e)$: r = -0.23, P = 0.44; ANCOVA slope P = 0.87; ENC' v $log(N_e)$: r = -0.15, P = 0.61; ANCOVA slope P = 0.86). Hence there is little evidence that the correlation between ω_a and N_e is a consequence codon usage bias.

In the second analysis we estimated ω_a while simultaneously estimating the strength of negative selection on synonymous sites. We also corrected our estimate of the effective population size for the effect of selection on synonymous sites. Estimates of N_e, ω_a and the strength of selection on synonymous mutations are given in table 3.2. The results of this analysis show some evidence of selection on synonymous sites in four species: *O. rufipogon*, *Z. mays*, human and mouse. There is independent evidence of selection in *H. sapiens* (Iida and Akashi, 2000; Hellmann et al., 2003; Chamary et al., 2006; Keightley et al., 2011) and mouse (Chamary and Hurst, 2004; Gaffney and Keightley, 2005; Keightley et al., 2011), but also in *P. tremula* (Ingvarsson, 2010), *D. melanogaster* (Zeng and Charlesworth, 2009), *D. pseudoobscura* (Akashi and Schaeffer, 1997; Haddrill et al., 2011) and *D. miranda* (Bartolomé et al., 2005; Haddrill et al., 2011) for which we do not find evidence of selection at synonymous sites. The failure to detect selection on synonymous sites may be due to the strength of the selection being weak, and furthermore

we have assumed a model with constant population size. This was necessary because it is not possible to simultaneously fit a model that allows demographic change and selection on synonymous codon use in the absence of detailed information about codons preferences (Zeng and Charlesworth, 2010). Correcting for selection on synonymous sites, we find that the correlation between ω_a and N_e is positive but not significant, whereas the correlation between ω_a and $\log(N_e)$ is positive and significant with ANCOVA (slope P = 0.028, intercept P = 0.032). Although not conclusive, these results suggest that the correlation between ω_a and N_e is not due to selection on synonymous codon use.

A third possible explanation for the correlation between ω_a and N_e is biased gene conversion (BGC). Like selection upon synonymous codon use, BGC can elevate the ratio of polymorphism to divergence relative to neutral expectations. However, it is less clear that this will affect synonymous sites preferentially.

We might expect that just as the number of adaptive substitutions increases with N_e, the number of effectively neutral substitutions will decline. We estimated the number of effectively neutral substitutions as $\omega_{\bar{a}} = \omega - \omega_a$, and found that $\omega_{\bar{a}}$ is significantly negatively correlated to N_e (r = -0.24, P = 0.43; ANCOVA slope P = 0.05; intercept P = 0.04) and $log(N_e)$ (r = -0.53, P = 0.06 ANCOVA slope P = 0.14; intercept P = 0.23). The slopes of the regression lines, from the ANCOVA, between $\omega_{\bar{a}}$ and N_e are similar in magnitude to those between ω_a and N_e (-2.4×10^{-8} versus 2.5×10^{-8}). We also investigated whether aspects of the DFE of deleterious mutations, as estimated from the polymorphism data, are correlated to N_e. We find a significant negative correlation between $\psi(= L_s P_n/(L_n(P_s + 1)))$ and N_e with ANCOVA controlling for the non-independence between these variables (Pearson r=-0.24, P=0.42; ANCOVA slope P = 0.014, intercepts P = 0.006), and between ψ and $\log(N_e)$ with Pearson (Pearson r=-0.64, P=0.018; ANCOVA slope P = 0.016; intercepts P = 0.041), but correlations between the shape parameter of the DFE and the mean value of $N_e s$ and N_e are nonsignificant. The lack of a significant correlation between mean $N_e s$ and N_e could be a consequence of the low precision of estimates mean $N_e s$ (Keightley and Eyre-Walker, 2007).

3.5 Discussion

We have presented evidence that the rate of adaptive protein evolution is positively correlated to N_e. We have shown that it is unlikely that this is due to recent demographic

changes or selection on synonymous sites. Such a result is not unexpected. If the rate of adaptive evolution is limited by the supply of new mutations, then species with larger N_e are expected to undergo more adaptive evolution than species with small N_e, because a greater number of advantageous mutations appear in the population and a higher proportion of these mutations are effectively selected.

The positive correlation between ω_a and N_e is consistent with a model in which the rate of adaptive evolution is limited by the supply of new mutations. The correlation seems less consistent with a model in which adaptation comes from standing genetic variation (Pritchard et al., 2010; Pritchard and Rienzo, 2010) for two reasons. First, although the level of advantageous, neutral and slightly deleterious genetic variation is expected to be correlated to N_e this correlation appears to be weak; levels of diversity, at least in mammalian mtDNA, are poorly correlated to effective population size (Piganeau and Eyre-Walker, 2009). This is probably due to a negative correlation between the rate of mutation per generation and the effective population size (Lynch, 2007; Piganeau and Eyre-Walker, 2009). Second, the level of diversity of strongly deleterious mutations is expected to be either independent of the effective population size, or negatively correlated to it, since species with long generation times, and small effective population size, appear to have higher rates of mutation per generation (Keightley and Eyre-Walker, 2000; Piganeau and Eyre-Walker, 2009).

We have shown that species with large N_e undergo more adaptive substitutions than species with small N_e. However this does not necessarily mean that these species adapt faster, though this is likely. This is because the total rate of adaptive evolution is a product of the number of adaptive substitutions and the effects of those substitutions. It is possible that species with large N_e undergo more adaptive substitutions, but that these are smaller in magnitude. We have also not considered adaptive evolution outside of protein coding genes.

The positive correlation between the rate of adaptive evolution and N_e implies that detecting the signature of adaptive evolution using MK approaches is likely to be difficult in species with small N_e, since they are expected to have undergone low levels of adaptive evolution. Furthermore, they are likely to have a higher proportion of effectively neutral mutations, which tends to obscure the signature of adaptive evolution. For example,

assume that we have two species with the same number of synonymous polymorphisms (20) and substitutions (100) in a sample of genes. Assume that the two species have undergone the same number of adaptive non-synonymous substitutions (15), but that species A has experienced no neutral mutations, whereas species B has undergone as many effectively neutral non-synonymous mutations as synonymous mutations. Under the assumption that adaptive mutations contribute little to polymorphism the MK tables for the two species would be as given in Table 3.3. It is evident that adaptive evolution would be detected in species A using a standard MK test (i.e. a χ^2 test of independence) but not in species B, because although both species have undergone the same amount of adaptive evolution, this is obscured by the large number of effectively neutral substitutions in species B. The fact that large numbers of effectively neutral substitutions obscure the signature of adaptive evolution means that it will be more difficult to detect adaptive evolution in poorly conserved regions of the genome, such as regulatory sequences.

We have found some evidence that the rate of adaptive evolution varies between groups of organisms for a given N_e. In particular, it is striking that the fungus *S. paradoxus* has the largest N_e amongst the species we have considered, but shows no evidence of adaptive evolution. If we remove *S. paradoxus* from the ANCOVA we find no evidence that the rate of adaptive evolution differs between groups (ANCOVA intercepts $P = 0.47$), although ω_a is correlated to N_e (ANCOVA slope $P = 0.017$). It is possible that *S. paradoxus* has a low rate of adaptive evolution, despite its large N_e, because it is largely asexual (Tsai et al., 2008). Consistent with this we note that there is a negative correlation between d_n/d_s and some measure of effective population in a number of non-recombining genetic systems. In mammalian mtDNA, d_n/d_s is correlated to body size (Popadin et al., 2007), which is believed to be correlated to N_e, and in both mammals and birds the largely non-recombining Y and W chromosomes, which are believed to have lower N_e than the autosomes, have higher d_n/d_s values (Wyckoff et al., 2002; Berlin and Ellegren, 2006). In contrast, we find no evidence of a significant correlation between d_n/d_s and N_e in our analysis (r = -0.37, P = 0.21; ANCOVA slope P = 0.34). This might be due to our small sample size, but it also may reflect a difference between recombining and non-recombining loci. In our analysis we find that the rate of adaptive substitution increases with N_e at a similar rate to the rate at which the effectively neutral substitutions decreases; this leaves the d_n/d_s uncorrelated to N_e. It might be that rates of adaptive evolution are lower in non-recombining systems and hence the decline in the number of

	Nonsynonymous sites		Synonymous sites	α	ω_a	MK-Test
	Adaptive	Effectively Neutral				P-Value
Species A (large N_e)						
Polymorphisms	n.a.	0	20			
Substitutions	15	0	100			
				100%	15%	0.024
Species B (low N_e)						
Polymorphisms	n.a.	20	20			
Substitutions	15	100	100			
				13%	15%	0.685

Table 3.3: Power to detect adaptive changes in species with different effective population sizes. Comparison between two hypothetical species (A and B) which have the same number of adaptive changes but different effective population sizes illustrated by a difference in the number of effectively neutral nonsynonymous sites.

effectively neutral substitutions dominates the relationship between d_n/d_s and N_e, and species such as *S. paradoxus* undergo little adaptive evolution.

Chapter 4

Fluctuating selection models and McDonald-Kreitman type analyses

4.1 Abstract

It is likely that the strength of selection acting upon a mutation varies through time due to changes in the environment. However, most population genetic theory assumes that the strength of selection remains constant. Here we investigate the consequences of fluctuating selection pressures on the detection and quantification of adaptive evolution using McDonald-Kreitman (MK) style approaches. In agreement with previous work, we show that fluctuating selection can generate evidence of adaptive evolution even when the expected strength of selection on a mutation is zero. However, we also find that the mutations, which contribute to both polymorphism and divergence tend, on average, to be positively selected during their lifetime, under fluctuating selection models. This is because mutations which fluctuate, by chance, to positive selected values, tend to reach higher frequencies in the population than those which fluctuate towards negative values. Hence the evidence of positive adaptive evolution detected under a fluctuating selection model by MK type approaches is genuine since fixed mutations tend to be advantageous on average during their lifetime. We show that this can even apply when the expected strength of selection on a mutation is negative.

4.2 Introduction

The environment for most organisms is constantly changing due to fluctuations in physical factors, such as temperature, and biotic factors, such as the prevalence of competitor

species and the density and genotype frequencies of other conspecific individuals. This is likely to lead to changes in the strength of selection acting upon a mutation through time (Bell, 2010); in the extreme this might mean that a mutation is advantageous at one time-point, but deleterious at another. Despite the likelihood that selection fluctuates through time there is relatively little evidence that this is the case. This is probably because measuring the strength of selection is difficult and detecting fluctuating selection requires analyses over several years. However, analyses of data from several species have suggested that some polymorphisms are subject to fluctuating selection (Fisher and Ford, 1947; Mueller et al., 1985; Lynch, 1987; O'Hara, 2005, reviewed by Bell, 2010). In these examples there are changes in the frequency of mutations that appear to be too great to be explained by either random genetic drift or migration. In most of these analyses the mean strength of selection acting upon a mutation appears to be close to zero. However, this might be a sampling artifact, a mutation subject to fluctuating selection in which the average selection coefficient is non-zero is more likely to be lost or fixed.

Fluctuating selection is likely to be more prevalent than the few well documented examples suggest and Bell (2010) has argued that fluctuating selection might help resolve why most traits show substantial heritability, even though selection on a short time-scale often appears to be quite strong.

Despite the likelihood that the strength of selection varies most population genetic theory has assumed that the strength of selection is constant through time. Exceptions are the work by Kimura (1954), Gillespie (1973, 1991a), Jensen (1973), Karlin and Levikson (1974), Takahata et al. (1975) and Huerta-Sanchez et al. (2008). Huerta-Sanchez et al. (2008) have investigated how fluctuating selection affects the allele frequency distribution, and hence the site frequency spectrum (SFS), and the probability of fixation. They investigated models in which the strength of selection acting upon a mutation changes every generation and models in which the strength of selection is correlated across generations, although in all models the expected strength of selection acting upon a mutation was zero. They show that models with and without auto-correlation are equivalent, in the sense that a model in which the strength of selection changes every generation behaves like one in which it changes every few generations, but with a larger variance in the selection coefficient. They demonstrate that although the expected strength of selection is zero, fluctuating selections leads to an increase in the probability of fixation, a decrease in

diversity and a change in the SFS. They argue that fluctuating selection might generate artifactual evidence of adaptive evolution under McDonald-Kreitman (MK) type analyses.

The McDonald-Kreitman (MK) test (McDonald and Kreitman, 1991), and its derivatives (Fay et al., 2001; Smith and Eyre-Walker, 2002; Bierne and Eyre-Walker, 2004; Eyre-Walker and Keightley, 2009) use the contrast between the levels of polymorphism and substitution at neutral and selected sites to infer the adaptive evolution in the divergence of species. Modified versions of the MK test allow one to quantify α, the proportion of nonsynonymous differences between species due to adaptive evolution (Fay et al., 2001; Smith and Eyre-Walker, 2002; Bierne and Eyre-Walker, 2004; Eyre-Walker and Keightley, 2009). The MK test has been widely applied to a number of species and estimates of α vary substantially from limited evidence ($\alpha \approx 0$ to 10%) in humans (Chimpanzee Sequencing and Analysis Consortium, 2005; Zhang and Li, 2005; Boyko et al., 2008) and many plant species (Gossmann et al., 2010) to more than 50% in Drosophila (Smith and Eyre-Walker, 2002; Charlesworth and Eyre-Walker, 2006), sunflowers (Strasburg et al., 2011a) and bacteria (Charlesworth and Eyre-Walker, 2006).

The MK test framework implicitly assumes a constant selection pressure on individual mutations and it is unclear how fluctuating selection pressures might alter the outcome of the MK test. Huerta-Sanchez et al. (2008) suggest that fluctuating selection will lead to artifactual evidence of adaptive evolution since fluctuating selection increases the ratio of divergence to diversity relative to neutral expectations. Here we investigate whether a MK type test infers adaptive evolution under a fluctuating selection model. We apply different versions of the MK test to simulated data under fluctuating selective conditions. We find, in agreement with earlier studies, that the fluctuating selection does lead to a signature of adaptive evolution. However, we also show that those mutations contributing to polymorphism and divergence are on average positively selected during their lives, even though the expected strength selection is zero or even negative. We therefore conclude that the signature of adaptive evolution is genuine.

4.3 Materials and Methods

4.3.1 Theoretical framework

We consider two alleles A and a in a haploid population of size N with the following fitness coefficients in generation t, where $t = 1, 2, \ldots$ (Karlin and Levikson, 1974).

Allele	A	a
Fitness in generation t	$1 + s_1^{(t)}$	$1 + s_2^{(t)}$

From the diffusion approximation the drift coefficient a and the variance b can be written in a simplified form as (Gillespie, 1973; Karlin and Levikson, 1974)

$$a(x) = x(1-x)(1 + \beta x(1-x)) \tag{4.1}$$

$$b(x) = x(1-x)(\delta + \beta(1/2 - x)) \tag{4.2}$$

where

$$\delta = N[E(s_1 - s_2) + \mathbf{Var}(s_2)/2 - \mathbf{Var}(s_1)/2] \tag{4.3}$$

$$\beta = N[\mathbf{Var}(s_1) + \mathbf{Var}(s_2)] \tag{4.4}$$

4.3.2 Random fluctuations

Huerta-Sanchez et al. (2008) consider a model of fluctuating selection where the strength of selection changes every generation. These are drawn from the same distribution with mean zero. We extent this model of fluctuating selection and use normal distributions with a variance of σ^2 and different means; i.e.

$$s_1 \sim \mathcal{N}(\mu, \sigma^2) \tag{4.5}$$

$$s_2 \sim \mathcal{N}(0, \sigma^2) \tag{4.6}$$

and obtain

$$\delta = N\mu \tag{4.7}$$

$$\beta = 2N\sigma^2 \tag{4.8}$$

In particular we obtain for the model of Huerta-Sanchez et al. (2008) ($\mu = 0$)

$$\delta = 0 \tag{4.9}$$

$$\beta = 2N\sigma^2 \tag{4.10}$$

Note that this β constant differs from β of Huerta-Sanchez et al. (2008) by factor 2. Also the δ value is different because of differences in the expressions of $a(x)$ and $b(x)$.

4.3.3 Probability of fixation

The probability of the fixation of a new mutation (i.e. a mutation which enters the population at frequency $1/N$) has the explicit expression (Karlin and Levikson, 1974):

$$p = \frac{\int_0^{1/N} f(\xi)\,\mathrm{d}\xi}{\int_0^1 f(\xi)\,\mathrm{d}\xi} \tag{4.11}$$

where

$$f(x) = e^{-\int_0^x \frac{2b(\eta)}{a(\eta)}\,\mathrm{d}\eta} \tag{4.12}$$

4.3.4 Simulations

To investigate the average strength of selection acting on each mutation which reaches a certain frequency in the population we conducted Monte Carlo simulations. We simulated a single locus subject to fluctuating selection in a haploid population of N individuals as follows:

1. Introduce a new mutation with frequency $1/N$ and generate two random variables using equations (4.5) and (4.6). Goto 3
2. Change the strength of selection according to equations (4.5) and (4.6)
3. Calculate the cumulative strength of selection experienced by the mutation, and from this calculate the mean strength of selection.
4. Calculate the expected frequency, f, of the mutation in the next generation.
5. Generate the actual frequency in the next generation, f', by generating a random binomial variable from the expected frequency, f, and the population size N
6. If $f' = 0$ or $f' = 1$ Goto 1
7. If $f' > 0$ and $f' < 1$ Goto 2

4.3.5 McDonald Kreitman test

We derive the absolute numbers of diversity and divergence from the simulated data as follows. For diversity we assume that $\theta = 1/50$. We run independent simulations for the neutral case ($\beta = 0$) and fluctuating case ($\beta > 0$). We assume a total number of 10000 neutral and selected sites, respectively. Based on θ one can calculate the proportion of mutated sites and derive the absolute SFS at a single time point. To retrieve divergence

estimates (D_n and D_s for selected and neutral divergence, respectively) we use the absolute number of fixations during the simulation process. Let us define a statistic, α', which measures the rate of substitution at sites subject to fluctuating selection relative to that at neutral sites:

$$\alpha' = 1 - \frac{D_s}{D_n} = \frac{D_n - D_s}{D_n} \tag{4.13}$$

A number of different methods based on the MK test have been proposed to estimate α, the proportion of substitutions driven by positive selection; this can also be viewed as the proportional increase in the substitution rate above neutral expectations. We use the method of Fay et al. (2001):

$$\alpha = 1 - \frac{D_s P_n}{D_n P_s} \tag{4.14}$$

where P_n and P_s are the number of polymorphism at selected and neutral sites, respectively and D_n and D_s are the number of substitutions at selected and neutral sites, respectively. We also apply the methods of Eyre-Walker and Keightley (2009) and Schneider et al. (2011) which incorporate the contribution of slightly deleterious mutations to polymorphism and divergence (Eyre-Walker and Keightley, 2007). The method of Eyre-Walker and Keightley (2009) assumes that advantageous mutations do not contribute substantially to polymorphism whereas the method of Schneider et al. incorporates the contribution of adaptive mutations to polymorphism (p_a) and jointly estimates both the strength of selection, s_a, and the proportion of mutations that are advantageous, p_a. It is difficult to simultaneously find the maximum likelihood values of s_a and p_a and so we ran the method of Schneider et al. for values of p_a of 0.001, 0.01, 0.1, 0.2, ..., 0.8 letting the method find the ML value of s_a; we then took the estimates of p_a and s_a which gave the highest likelihood; using p_a and s_a we could then estimate α as:

$$\alpha = p_a Q(s_a) + (1 - p_a) \tag{4.15}$$

with

$$Q(s_a) = S/(1 - e^{-S}) \tag{4.16}$$

where $S = 2Ns$ and $Q(s_a)$ is the probability of fixation of a mutation under selection strength s, relative to that of neutral mutation.

The methods of Eyre-Walker and Keightley (2009) and Schneider et al. (2011) are available online (http://homepages.ed.ac.uk/eang33/).

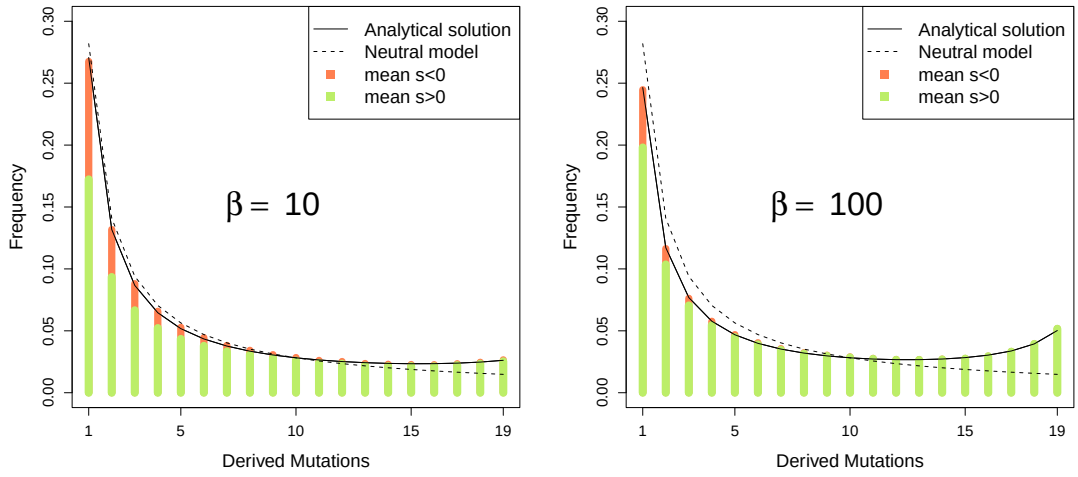
4.4 Results

Huerta-Sanchez et al. (2008) have previously shown that fluctuating selection pressures can alter both the SFS and the probability of fixation away from neutral expectations, even when the expected strength of selection acting upon a mutation is zero. They suggest that since fluctuating selection generates an increase in the ratio of divergence to diversity, relative to neutral expectations, that it would generate artefactual evidence of positive selection. Using simulated data and three MK-type methods we confirm that fluctuating selection does indeed lead to evidence of positive adaptive evolution (Table 4.1). We investigate three scenarios with varying intensities of fluctuations. In the first one we assume like Huerta-Sanchez et al. (2008) a fluctuation with an expected mean selection pressure of zero. In this case the methods of Fay et al. (2001) and Eyre-Walker and Keightley (2009) yield similar positive estimates of α, the proportion of substitutions inferred to be driven by positive adaptive evolution. The method of Schneider et al. (2011) yields estimates that are substantially higher. In the second scenario we investigate a pattern with fluctuating selection with a positive mean expected value and different fluctuation intensities. In this case the method of Fay et al. (2001) slightly underestimates and Schneider et al. (2011) slightly overestimates α while the method of Eyre-Walker and Keightley (2009) agrees very well. In the third scenario with a negative expected mean selection pressure we find that all methods tend to overestimate α. We also find that for low to moderate β values the ratio of divergence relative to the neutral expectation is not increased ($\alpha' < 0$), but all methods infer adaptive evolution. We also find evidence for increased divergence for high β values even though the expected mean is negative.

δ	β	α (MK[a])	α (MK[b])	α (MK[c])	α'
0	10	0.47	0.51	0.82	0.54
	20	0.65	0.68	0.89	0.70
	30	0.70	0.73	0.85	0.74
	40	0.75	0.78	0.93	0.79
	50	0.78	0.81	0.90	0.81
	100	0.87	0.90	0.94	0.90
10	10	0.85	0.94	0.97	0.94
	20	0.86	0.94	0.97	0.94
	30	0.87	0.94	0.98	0.94
	40	0.88	0.94	0.98	0.94
	50	0.88	0.94	0.98	0.94
	100	0.88	0.95	0.98	0.95
-10	10	-16	-0.95	0.21	n.d.
	20	-3.70	-0.77	0.01	-27.0
	30	0.09	0.46	0.55	-1.94
	40	0.10	0.4	0.73	-1.24
	50	0.47	0.65	0.88	-0.21
	100	0.78	0.82	0.89	0.72

[a]McDonald and Kreitman (1991)
[b]Eyre-Walker and Keightley (2009)
[c]Schneider et al. (2011)

Table 4.1: α estimates for different fluctuating conditions with a expected mean fitness of $\delta = Ns = 0, -10$ and 10. Estimates of adaptive divergence, α, for polymorphism and divergence simulated under varying random fluctuating selection. Three different MK type tests were used. The intensity of the fluctuation is denoted by β.



(a) (b)

Figure 4.1: SFS generated under fluctuating conditions ($\beta = 10$ and 100) with mean selective effect of zero. The proportion of mutations with positive and negative mean selection coefficients are shown in green and red, respectively.

The fact that a fluctuating selection model generates evidence of adaptive evolution even when the expected strength of selection is zero or negative suggests that fluctuating selection generates artifactual evidence of positive selection (Huerta-Sanchez et al., 2008). However, the mean strength of selection experienced by the mutation might not be zero, even though its expected value is zero; it might be that those mutations which spread to high frequency in the population are those, which just by chance have mean selective values that are positive, whilst those mutations which fluctuate to negative values are lost from the population. To investigate this we tracked the mean strength of selection of each mutation at each frequency up to when it was lost or fixed. From this analysis it is evident that the vast majority of mutations that contribute to the SFS are positively selected, except at very low frequencies and when fluctuations in the strength of selection are quite weak (low β values, Figure 4.1). This pattern is seen for the fluctuating selection model around zero and for those that involve a fluctuation with a net selective effect (Figure 4.2). The bias towards positive mean strengths of selection is even more extreme for those mutations that become fixed (Figure 4.3) even when the net selective effect is not zero (Figure 4.4).

If we track mutations that ultimately become fixed it is evident that those mutations that start off being slightly negative quickly become positive in their mean value (Figure

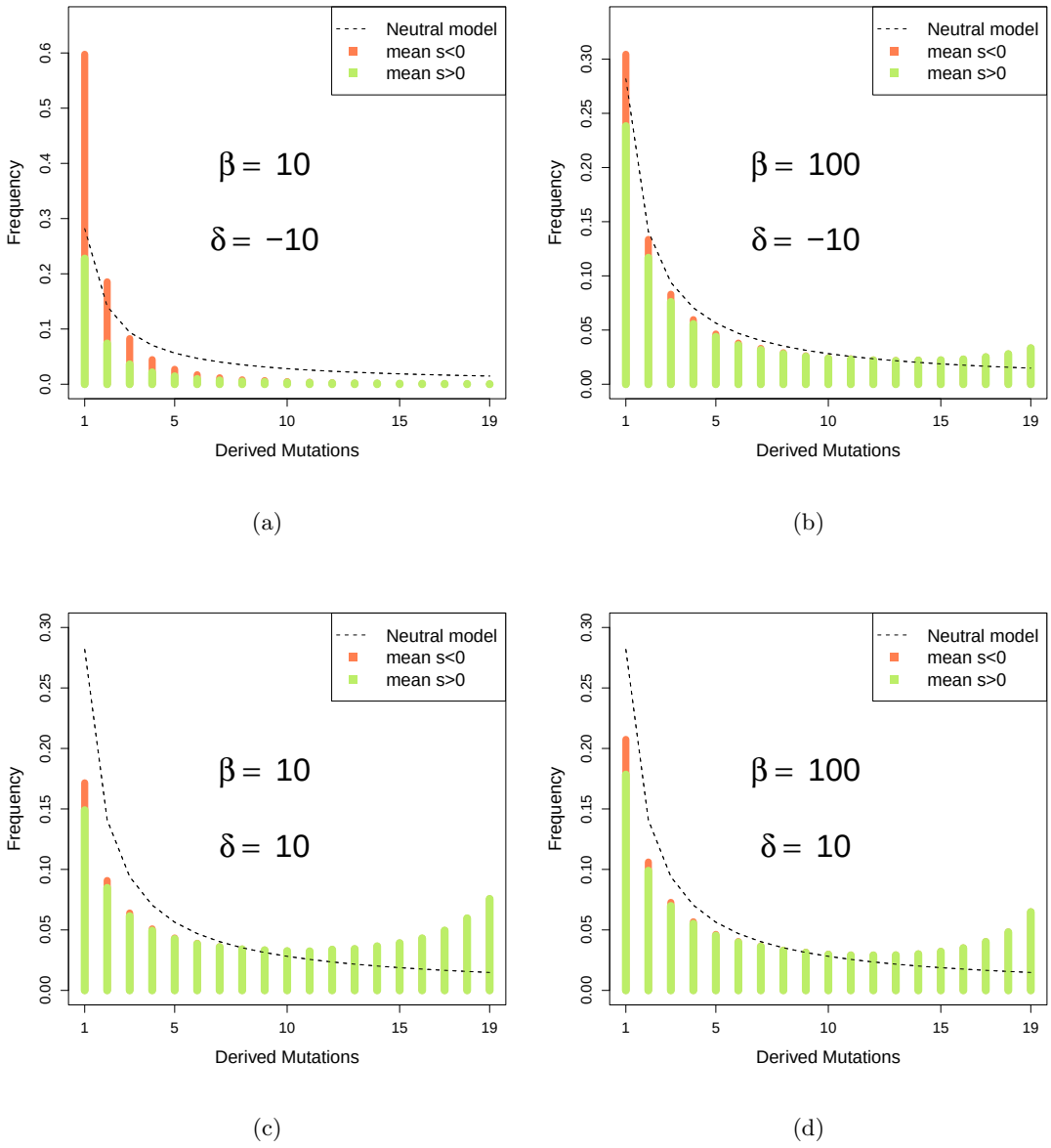
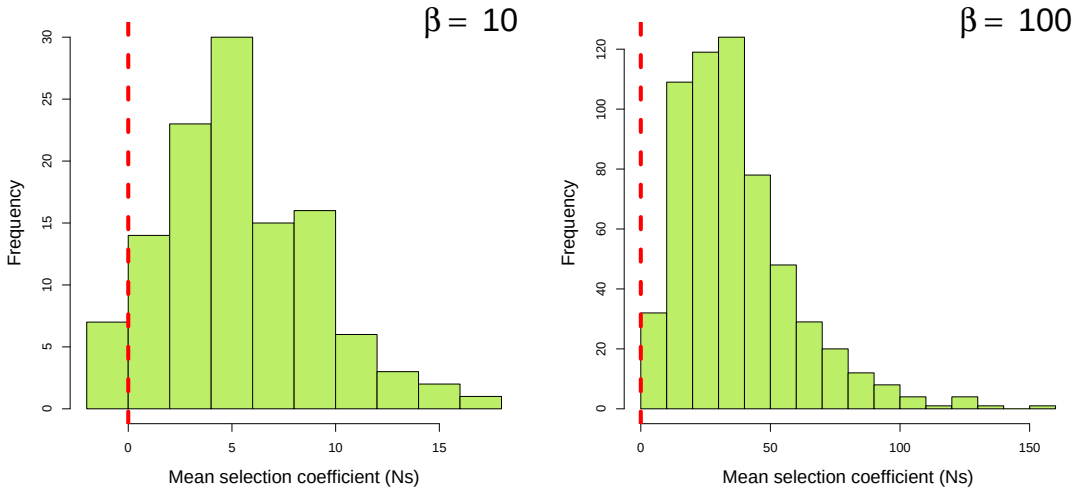
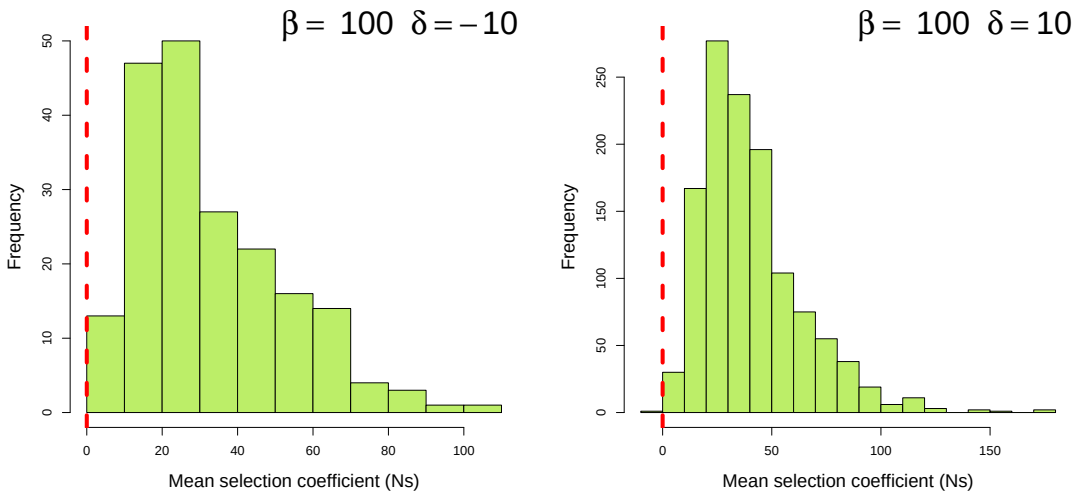


Figure 4.2: SFS generated under fluctuating conditions ($\beta = 10$ and 100) with mean selective effect of $\delta = Ns = -10$ and 10. The proportion of mutations with positive and negative mean selection coefficients are shown in green and red, respectively.



(a) (b)

Figure 4.3: Distributions of mean fitness effects of mutations at the time of fixation for fluctuating conditions ($\beta = 10$ and 100) with mean selective effect for all mutations of zero.



(a) (b)

Figure 4.4: Distributions of mean fitness effects of mutations at the time of fixation for fluctuating condition of $\beta = 10$ and mean selective effect for all mutations of $\delta = -10$ and 10.

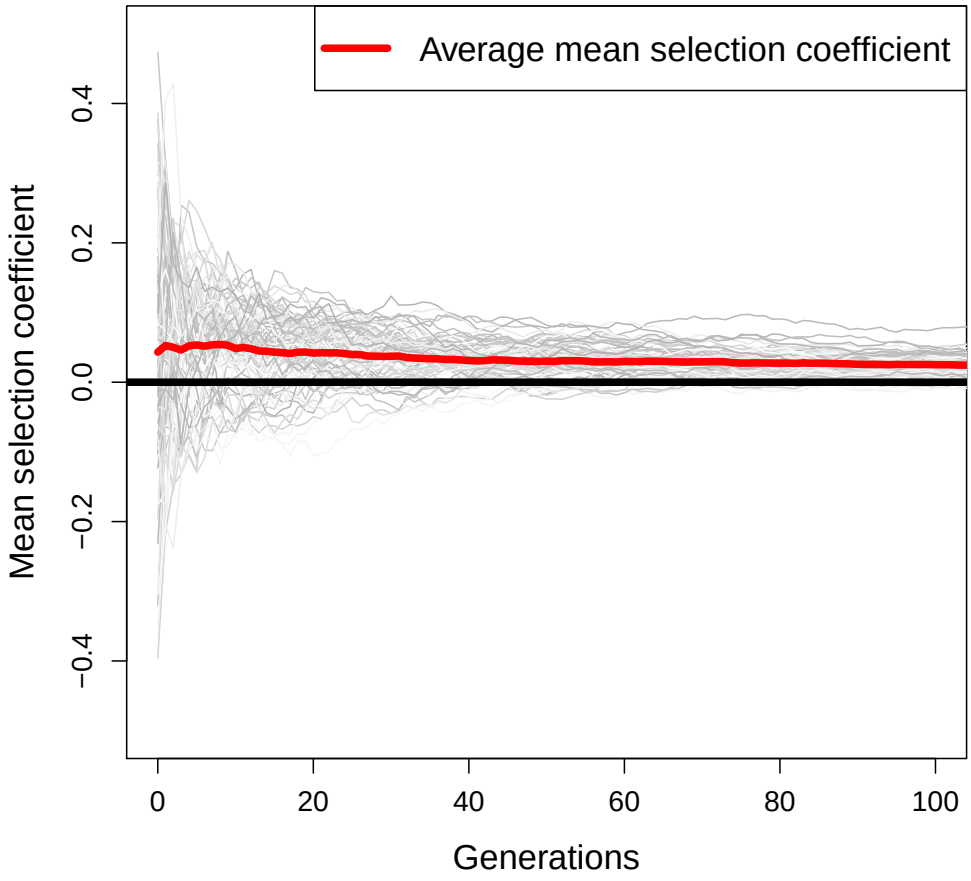


Figure 4.5: Average mean selection coefficient over time for fixed mutations under fluctuating conditions of $\beta = 100$ and no net selective effect ($\delta = 0$). Shown are the first 100 generations of 80 mutations that got fixed. The red line indicates the average mean selection coefficient, trajectories in grayscale indicate mean selection coefficients for individual mutations.

4.5). Interestingly those that start off being highly positive tend to decrease in mean selection coefficient as well; this is probably a consequence of averaging over many selective episodes, and hence approaching the expected value. We also find that the average mean selection coefficient for all mutations that get fixed declines with time. This is because the critical time for an advantageous mutation is when its rare because it is more likely to be lost. Those mutations that are strongly selected at an early stage have more chance of remaining in the population.

We used three different MK-style tests to estimate α'. The methods of Fay et al. (2001) and Eyre-Walker and Keightley (2009) tend to underestimate the true value of α

for fluctuation around zero ($\delta = 0$). The degree to which these methods underestimate α depends on the variance in the selection coefficient; the greater the variance the less bias there is. The method of Schneider et al. (2011) tends to overestimate the true α'. However it is perhaps not surprising that the method infers higher rates of adaptive evolution. The vast majority of fixations have an average positive selection coefficient and such mutations are likely to contribute to polymorphism, which the other two methods assume does not happen.

4.5 Discussion

We have confirmed, using simulation, that a model in which selection fluctuates around a value of zero behaves differently to a neutral model. In particular the ratio of divergence to diversity increases and this yields evidence of positive adaptive evolution under MK-type analyses. However, we show that the mutations contributing to diversity and divergence are on average positively selected, and hence any evidence of positive selection is not an artifact but genuine. We show that estimates of α, the increase in the rate of adaptive evolution, relative to the rate of neutral evolution, is slightly underestimated by the methods of Fay et al. (2001) and Eyre-Walker and Keightley (2009) and slightly overestimated by the method of Schneider et al. (2011) if fluctuations is around zero. Under a fluctuating model with a net positive selective effect we show that the method of Eyre-Walker and Keightley (2009) performs best. Under a fluctuating model with a net negative selective effect we show that all methods infer adaptive evolution even though the rate of selected divergence relative to neutrality is not increased. This effect is more severe for the methods of Eyre-Walker and Keightley (2009) and Schneider et al. (2011) indicating that the impact on the SFS relative to divergence is substantial. However even under such a model the signature of adaptation is genuine because the vast majority of fixations have a positive average selection coefficient.

The degree to which α is misestimated by each of these methods is possibly deceptive because we are assuming that every site is subject to fluctuating selection with an expected strength of selection of zero, but many sites will be subject to strong or weak consistently negative selection. To investigate the consequences of a model in which some sites are subject to consistent negative selection of various strengths and some sites to fluctuating selection, we ran a series of simulations in which we assumed that a proportion p_f of sites were subject to fluctuating selection and a proportion $1 - p_f$ were subject to

consistent selection, with the distribution of fitness effects at those sites being a gamma distribution with a shape parameter of 0.3 and a mean strength of selection of $Ns = 9,800$ (Keightley and Eyre-Walker, 2007), as has been estimated in Drosophila. Under this model we find that α is estimated to be positive if $p_f > 0.05$ although the ratio of selected divergence is not increased relative to neutrality (Table 4.2). Therefore the effect of fluctuating selection can be substantial even if only a few sites are subject to fluctuation.

Here we tried to estimate the true α using MK approaches as the quantity of increased divergence relative to neutrality. However α can also be seen as a measurement of the proportion of substitutions that are on average positively selected. We show that under fluctuating conditions almost all fixed mutations show an average positive selective effect which implies $\alpha \approx 1$. Such an interpretation of α explains the relatively high estimates of α by the method of Schneider et al. (2011).

p_f	β	α (MK[a])	α (MK[b])	Shape Γ	-NE(s)	α'
0	10	-0.98	-1.23	0.35	10838	< 0
0.01	10	-0.25	-0.02	0.35	21389	< 0
0.05	10	0.26	0.37	0.18	255862	< 0
0.1	10	0.36	0.44	0.13	∞	< 0
0.3	10	0.44	0.48	0.05	∞	< 0
0.5	10	0.46	0.51	0.05	64601	0.02
0.7	10	0.47	0.53	0.05	138.4	0.3
0.9	10	0.47	0.52	0.05	0.67	0.45

[a]McDonald and Kreitman (1991)
[b]Eyre-Walker and Keightley (2009)

Table 4.2: α estimates for a mixed model of selective effects with fluctuation and negative selection. The model assumes a proportion p_f of sites under fluctuating conditions with $\beta = 10$ and an expected mean fitness of zero and a proportion $(1 - p_f)$ of sites with consistent negative selection with gamma distributed selected effects with a shape of 0.3 and a mean of $Ns = 9,800$.

Chapter 5

Quantifying the variation in the effective population size within a genome.

5.1 Abstract

The effective population size (N_e) is one of the most fundamental parameters in population genetics. It is thought to vary across the genome as a consequence of differences in the rate of recombination and the density of selected sites due to the processes of genetic hitch-hiking and background selection. Although it is known that there is intragenomic variation in the effective population size in some species, it is not known whether this is widespread, nor how much variation in the effective population size there is. Here, we test whether the effective population size varies across the genome, between protein coding genes, in 10 eukaryotic species by considering whether there is significant variation in neutral diversity, taking into account differences in the mutation rate between loci by using the divergence between species. In most species we find significant evidence of variation. We investigate whether the variation in N_e is correlated to recombination rate and the density of selected sites in four species, for which this data is available. We find that N_e is positively correlated to recombination rate in one species, *Drosophila melanogaster* and negatively correlated to a measure of the density of selected sites in two others, humans and *Arabidopsis thaliana*. However, much of the variation remains unexplained. We then use a hierarchical Bayesian analysis to quantify the amount of variation in the effective population size and show that it is quite modest in all species - most genes have an N_e which is within a few fold of all other genes. Nonetheless we show

that this modest variation in N_e is sufficient to cause significant differences in the efficiency of natural selection across the genome, by demonstrating that the ratio of the number of non-synonymous to synonymous polymorphisms is significantly correlated to synonymous diversity and estimates of N_e, even taking into account the obvious non-independence between these measures.

5.2 Introduction

The effective population size (N_e) is one of the most fundamental quantities in population genetics, evolutionary biology and molecular ecology, since it determines the effectiveness of natural selection and the level of neutral genetic diversity that a population contains (Charlesworth, 2009). Populations and regions of the genome with small N_e tend to have low levels of genetic diversity, to be susceptible to the accumulation of deleterious mutations through genetic drift and to have potentially low rates of adaptive evolution (Charlesworth, 2009).

The effective population size is expected to vary across the genome as a consequence genetic hitch-hiking (Smith and Haigh, 1974) and background selection (Charlesworth et al., 1993). The action of both positive and negative natural selection, particularly in regions of the genome with low rates of recombination, is expected to reduce the effective population size leading to lower levels of genetic diversity and reduced effectiveness of selection. Hence variation in the rate of recombination and the density of selected sites is expected to generate variation in N_e.

The evidence that there is variation in N_e within a genome comes from three sources. First, it has been shown that levels of neutral genetic diversity are correlated to rates of recombination in *Drosophila* (Begun and Aquadro, 1992), humans (Hellmann et al., 2003) and some plant species (Tenaillon et al., 2004; Roselius et al., 2005). This could be due to variation in the mutation rate since neutral genetic diversity is proportional to the effective population size multiplied by the mutation rate. However, the level of neutral sequence divergence between species, which should be proportional to the mutation rate, is not correlated to the rate of recombination in *Drosophila* (Begun and Aquadro, 1992) and the plant species (Roselius et al., 2005) that have been investigated. Furthermore, although there is a correlation between neutral sequence divergence and recombination rate in humans, this correlation is not sufficient to explain the correlation

between diversity and the recombination rate (Hellmann et al., 2005). It is has also been shown that the Y and W chromosomes, which have no recombination over most of their length, have substantially lower diversity than other chromosomes, and that this cannot be attributed to differences in the mutation rate or the fact there are fewer Y and W chromosomes than autosomes (Montell et al., 2001; Filatov et al., 2001; Bachtrog and Charlesworth, 2002; Hellborg and Ellegren, 2004). It thus seems that the effective population size varies across genomes and is positively correlated to the rate of recombination.

Second, under the neutral theory of molecular evolution it is expected that levels of diversity and divergence should be proportional to each other, since both depend on the neutral mutation rate. Deviations from this hypothesis, caused for by variation in N_e, can be tested using the HKA test and derivatives of it (Hudson et al., 1987; Wright and Charlesworth, 2004; Ingvarsson, 2004; Innan, 2006). Evidence for departures from the neutral hypothesis, based on the HKA test, comes from multiple multilocus surveys in plants (Roselius et al., 2005; Schmid et al., 2005), the chicken Z chromosome (Sundström et al., 2004), humans (Zhang et al., 2002) and Drosophila (Moriyama and Powell, 1996; Machado et al., 2002).

Third, variation in the effective population size should manifest itself as variation in the effectiveness of selection and this has also been observed. In *Drosophila* it has been shown that codon usage bias is lower in the regions of the genome with very low rates of recombination (Hey and Kliman, 2002; Marais et al., 2003; Kliman and Hey, 2003). It has also been shown that the number of non-synonymous polymorphisms (P_n) relative to the number of synonymous polymorphisms (P_s) is higher in the low recombining parts of the *D. melanogaster* genome (Presgraves, 2005), that the rate of non-synonymous (d_N) relative to the rate of synonymous (d_S) substitution is positively correlated to the frequency of recombination (Betancourt and Presgraves, 2002) and that the overall efficiency of selection appears to be lower in the regions of the genome with low rates of recombination (Presgraves, 2005; Larracuente et al., 2008). Likewise it has been shown that d_N/d_S is higher on the Y or W chromosome than on the other chromosomes in humans (Wyckoff et al., 2002) and birds (Berlin and Ellegren, 2006) and the fourth chromosome of Drosophila species (Arguello et al., 2010). In contrast, Bullaughey et al. (2008) found no correlation between d_N/d_S and the rate of recombination in primates.

It is thought that the correlation between d_N/d_S or P_n/P_s and the rate of recombination is due to regions of the genome with little or no recombination having low effective population size and hence reduced effectiveness of natural selection (Betancourt et al., 2009). P_n/P_s is negatively correlated to the rate of recombination because regions with low effective population size allow more slightly deleterious mutations to segregate for longer. In contrast, d_N/d_S can either be positively or negatively correlated to the rate of recombination depending on the prevalence of advantageous mutations. If advantageous mutations are common then regions of the genome with high rates of recombination are expected to evolve faster because they have a higher effective mutation rate, and because selection is effective on a greater proportion of mutations. In contrast, if advantageous mutations are rare then regions of the genome with high rates of recombination may have low values of d_N/d_S because selection against slightly deleterious mutations is more effective.

Although it is well established that N_e varies across the genome in a few species, it is unclear whether this is true of all species and, more importantly, how much variation in N_e there is and whether this variation results in differences in the effectiveness of selection. Here we test whether there is variation in the effective population size by considering whether there is significant variation in neutral diversity, taking into account that this might be due to variation in the mutation rate by using the divergence between species to control for differences in the mutation rate. We also quantify the variation in N_e. We estimate N_e from the nucleotide diversity at putatively neutral sites, since this is expected to be equal to $4N_e\mu$ in a diploid organism, where N_e is the effective population size and μ is the mutation rate per generation. We use the divergence between two species at neutral sites as an estimate of the mutation rate per generation. Note that since we are comparing loci within a genome they all share the same generation time (unless they are on the sex chromosomes or in the mitochondrial DNA) and so this does not have to be explicitly taken into account. We can therefore estimate the effective population size for each locus. However, although each individual estimate is unbiased, the distribution of these values has a variance that is greater than the true variance because of sampling error; a locus might have a particularly low diversity just by chance, and not because its effective population size is particularly low. To get round this problem we use a hierarchical Bayesian framework to estimate the distribution of N_e across genes taking into account the sampling error associated with both the polymorphism and divergence

data.

We test for and investigate the variation in the effective population size in 10 eukaryotic species including humans, *Drosophila melanogaster*, *Arabidopsis thaliana* and *Saccharomyces paradoxus* (Table 5.1). We find that there is statistically significant variation in N_e across genes, but that it is rather modest in most of the organisms. We also investigate whether variation in N_e within a genome leads to variation in the proportion of effectively neutral mutations, by testing whether the ratio of the number of non-synonymous to synonymous polymorphisms is correlated to the effective population size, in a way which circumnavigates the obvious non-independence between the two variables. We find overall evidence for a correlation between these two parameters and hence conclude that even modest variation in the effective population size is sufficient to generate variation in the effectiveness of natural selection.

Species	Outgroup	Loci	Sites	Alleles	θ_s	d_s	Dataset
Drosophila melanogaster	*D.simulans*	302	40920	8	0.019	0.13	Shapiro et al. (2007)
Homo sapiens	*Macaca mulatta*	434	170441	32	0.001	0.08	EGP/PGA[1]
Mus musculus	*Rattus norvegicus*	66	5127	20	0.010	0.21	Halligan et al. (2010)
Arabidopsis thaliana	*A.lyrata*	918	64927	24	0.008	0.14	Nordborg et al. (2005)
Capsella grandiflora	*Neslia paniculata*	251	31273	8	0.019	0.16	Slotte et al. (2010)
Sorghum bicolor	*S.propinquum*	134	6799	14	0.004	0.02	Hamblin et al. (2006)
Boechera stricta	*A.thaliana*	129	10048	40	0.003	0.21	Song et al. (2009); Gossmann et al. (2010)
Arabidopsis lyrata	*A.thaliana*	66	5260	24	0.018	0.15	Ross-Ibarra et al. (2008); Foxe et al. (2008)
Capsella rubella	*A.thaliana*	49	5014	16	0.004	0.29	Foxe et al. (2009); Guo et al. (2009)
S.paradoxus	*S.cerevisiae*	94	28019	8	0.002	0.36	Tsai et al. (2008)

[1] EGP: http://egp.gs.washington.edu and PGA: http://pga.gs.washington.edu August 2010

Table 5.1: Summary of data sets used for the analyses. Number of synonymous sites (Sites) and nucleotide diversity (θ_s) are from the polymorphism data. Average divergence between the species pairs at silent sites (d_s).

5.3 Materials and Methods

5.3.1 Sequence data

We obtained data from different plant species, mouse, fruitfly and yeast using publicly available data from Genbank (http://www.ncbi.nlm.nih.gov/Genbank). Polymorphism data for *Homo sapiens* were downloaded from Enivironmental Genome Project (egp.gs.washington.edu) and Seattle SNPs (pga.gs.washington.edu) web-sites and for *Arabidopsis thaliana* from http://walnut.usc.edu/2010. The annotated protein-coding genome of *A. thaliana* was obtained from TAIR 8 (ftp://ftp.arabidopsis.org), the annotated *Arabidopis lyrata* genome was obtained from JGI (http://genome.jpi-psf.org). The annotated protein-coding genome of *Pan troglodytes*, *Macaca mulatta* and *Rattus norvegicus* were obtained from Ensembl (http://www.ensembl.org/info/data/ftp/index.html). The *S.cerevisae* genome chromosome III was obtained from http://www.yeastgenome.org. We restricted our analysis of *D. melanogaster* to data from the Zimbabwe population, from the *S.paradoxus* dataset to the European population and from the human dataset to African populations, since all of these represent the ancestral populations of the three species (Garrigan and Hammer, 2006; Stephan and Li, 2007; Liti et al., 2009). Qualitatively similar results were obtained in the three cases when using global data.

5.3.2 Preparation of the data

The analysis was performed using protein coding sequences. Coding regions were assigned using protein coding genomic data or if given, taken from the GenBank input files. Sequences were aligned using Clustalw using default parameter values (Thompson et al., 1994). The outgroup ortholog was assigned using the best BLAST (Altschul et al., 1990) hit or, if given, taken from the polymorphism dataset. We only used polymorphism data for which we could assign an outgroup sequence. For all analyses the number of synonymous substitutions and polymorphisms served as the neutral standard. For computational reasons all sites had to have been sampled in the same number of chromosomes within each species; because some loci had been sampled in more individuals than others and other loci had missing data, we reduced the dataset to a common number of chromosomes by randomly sampling the polymorphisms at each site without replacement. The numbers of synonymous and nonsynonymous sites and substitutions were estimated by randomly selecting one allele from the polymorphism data and comparing it against the outgroup using the F3x4 model implemented in PAML (Yang, 1997) in which codon frequencies are

estimated from the nucleotide frequencies at the three codon positions. The proportion of sites estimated by PAML was also used to compute the numbers of synonymous and non-synonymous sites for the polymorphism data. Although, how we choose to define a site can be important in some circumstances (Bierne and Eyre-Walker, 2003) this is not likely to be a problem in the current context because we use the same definition for both the divergence and polymorphism data; as such the number of sites effectively cancels out in most of our analyses (however see discussion of selection on synonymous codon bias below). Statistics concerning numbers of loci, numbers of sites as well as polymorphic sites are shown in Table 5.1.

5.3.3 Testing for variation in diversity and the effective population size

We investigated whether there was significant variation in the level of diversity across the genome using two tests. If we assume there is free recombination within and between loci (or no recombination within and between loci) then variation in diversity can be tested using a simple (2×k) χ^2 test of independence across the k loci within each species, where for each locus we have the number of sites with a polymorphism and the number of sites without a polymorphism. Note that this test is only valid when the same number of chromosomes have been sampled across all loci. However, some of the variation in diversity between loci might be due to variation in the genealogy if there is limited or no recombination between loci. We therefore applied a variant of the classic HKA test, but we removed the divergence information from the test. The test statistic X^2 is set up as follows:

$$X^2 = \sum_{i=1}^{L} (P_i - \hat{E}(P_i))^2 / \hat{V}ar(P_i) \tag{5.1}$$

where $\hat{E}(P_i)$ and $\hat{V}ar(P_i)$ are the expected value and variance of the number of segregating polymorphisms, P, in gene i:

$$\hat{E}(P_i) = M_i\theta \sum_{j=1}^{n-1} 1/j \tag{5.2}$$

$$\hat{V}ar(P_i) = \hat{E}(P_i) + (M_i\theta)^2 \sum_{j=1}^{n-1} 1/j^2 \tag{5.3}$$

with n being the number of alleles, L the number of loci, $\theta = 4N_e\mu$ and M_i the number of sites in gene i. Estimates of θ were obtained by minimizing the value of X^2. X^2 is expected to be χ^2 distributed with (L-1) degrees of freedom.

Any variation that we detect in diversity might be due to variation in the mutation rate or variation in the local effective population size. We therefore performed two

further analyses to investigate whether there was variation in diversity that could not be explained by variation in the mutation rate, as measured by synonymous divergence between species. The first test was a second approximate (2×k) χ^2 test of independence, performed as follows. For each locus we have the number of sites used to estimate the level of silent site divergence (L_d), the estimated number of substitutions (D), the number of sites used to estimate silent site diversity (L_p) and the number of sites with a polymorphism (P). Since L_d and L_p can be different we reduced the divergence or polymorphism dataset, whichever was larger, to the size of the other, resampling without replacement the numbers of substitutions or polymorphisms as appropriate; for example if L_d was half L_p, we sampled L_p sites from the divergence data to generate a sub-sample of the substitutions (D') over $L'_d = L_p$ sites. We can then perform a (2×k) χ^2 test where the cells for each gene are the number of sites of sites with a substitution (D') and the number of sites with a polymorphism (P'). Note that the dataset will be reduced using this method resulting in a loss of power. Furthermore this test is only approximate because we assume that the number of substitutions is binomially distributed, whereas in fact it has a more complex distribution because of the correction for multiple hits. Some of the expected values can be very small in both χ^2 tests: we therefore checked the p-values from the χ^2 tests by generating the null distribution for the test. This was performed by randomly assigning polymorphisms and substitutions across the contingency table preserving the marginal totals. We then recalculated the statistic and performed this 1000 times. The p-value was the proportion of such randomly generated values that exceeded the observed value. Generally we found that the p-value from randomisation, and the p-value assuming our test statistics were χ^2 distributed, were similar (Table A5.1). We therefore present the results from the standard χ^2 test.

This test assumes free recombination between sites within loci and loci (or no recombination between sites and loci). A more conservative test is the classic HKA test which tests for heterogeneity in the ratio of diversity divided by divergence between loci assuming no recombination within loci, but free recombination between loci. We performed the multiple locus HKA test using software provided by J. Hey (http://genfaculty.rutgers.edu/hey/software#HKA). To perform this test we had to exclude loci with zero divergence; for most species this constituted a small fraction of the total number of loci. However we had to exclude *S. bicolor* from the analysis because too many loci showed zero divergence.

5.3.4 Recombination and density of selected sites

We obtained estimates of recombination rate variation along chromosomes for *A. thaliana* (Singer et al., 2006), *D. melanogaster* (Hey and Kliman, 2002), *H. sapiens* (Kong et al., 2002) and *M. musculus* (Dumont et al., 2011). Gene density was estimated as the proportion of coding sites in window sizes of 50KB, 500KB and 5MB. Since results are qualitatively similar, we only discuss results for the window size of 500KB. Conservation scores (Siepel et al., 2005) were obtained from the UCSC genome browser (http://genome.ucsc.edu/) for *D. melanogaster* across 15 species, *H. sapiens* across 17 species and *M. musculus* across 30 species.

5.3.5 Bayesian analysis

To estimate the distribution of N_e we used a hierachical Bayesian analysis in which we estimate the parameters of the distribution of N_e (Figure A5.1). If we assume that the population size is stationary the expected number of polymorphisms segregating in a sample of n sequences, P_s, and the number of differences between the outgroup and a single sequence from the ingroup, D_s, are

$$P_s = 4\mu L_p N_e \sum_{i=1}^{n-1} 1/i \tag{5.4}$$

$$D_s = 2\mu t L_d \tag{5.5}$$

where L_p and L_d are the number of sites which can have a polymorphism or substitution respectively, μ is the nucleotide mutation rate per generation and t is the time of divergence. We are interested in the distribution of N_e. To estimate this distribution we assume that N_e and μ follow a log-normal or a gamma distribution. Assuming free recombination and using equations (5.4) and (5.5) above we can write the likelihood of observing $\hat{P}_s$ polymorphisms and $\hat{D}_s$ substitutions

$$L = \prod X(\hat{D}_s, D_s) X(\hat{P}_s, P_s) M(N_e \mid \sigma_{N_e}) M(\mu \mid \sigma_\mu) \tag{5.6}$$

where X(S,S(x)) is the Poisson distribution and $M(N_e \mid \sigma_{N_e})$ is the probability density of the distribution of N_e, and $M(\mu \mid \sigma_\mu)$ is the probability density of the distribution of the mutation rate; these distributions are parameterised such that the mean is fixed at unity leaving us to estimate the shape parameter. If there is no recombination within a locus

then we can rewrite equation (5.4) as

$$P_s = 4\tau\mu L_p N_e \sum_{i=1}^{n-1} 1/i \tag{5.7}$$

where τ is the length of the genealogy scaled such that $E[\tau] = 1$. We can rewrite equation (5.6), and the likelihood then becomes

$$L = \prod X(\hat{D}_s, D_s) X(\hat{P}_s, P_s) M(N_e \mid \sigma_{N_e}) M(\mu \mid \sigma_\mu) M(\tau \mid n) \tag{5.8}$$

To calculate the probability density distribution $M(\tau \mid n)$ of genealogy lengths we randomly simulated 10,000 genealogies, scaling them such that the average total length was unity. In theory it is possible to accommodate ancestral polymorphism into the method, however we found that the method rarely gave stable estimates of σ_{N_e}, particularly in the no recombination model. We therefore concentrated on datasets in which the influence of ancestral polymorphism was likely to be minimal - i.e. in which the average divergence was $> 5\times$ the average of θ_W (Table 5.1). If we assume that the ancestral N_e of a locus is correlated to the current N_e, we expect ancestral polymorphism to decrease the apparent variation in N_e.

To estimate the posterior distribution of the parameters σ_{N_e} and σ_μ we used a Monte-Carlo Markov chain running the Metropolis-Hastings algorithm (Hastings, 1970). Unfortunately because we have very few synonymous polymorphisms per gene this method tends to underestimate the true value of σ_{N_e}. For most datasets this underestimation is small, but it can be large. We therefore estimated the extent of bias by simulating data under a range of parameter values using the actual numbers of sites from the real data such that the expected numbers of polymorphisms and substitutions were equal to the mean values. For example, if we estimated σ_{N_e} to be 0.5 and σ_μ to be 0.1 we simulated data for σ_μ values of 0.1, 0.2 and 0.3 and for σ_{N_e} values between 0.4 and 1.0 in steps of 0.05. For each simulated dataset we estimated σ_{N_e} and using linear regression we inferred the relationship between σ_{N_e}(estimated) and σ_{N_e}(true). Using this relationship we inferred the true value of σ_{N_e} from the value estimated from the real data (Figures A5.2 and A5.3). To obtain a corrected SE we multiplied the observed standard error by the ratio of the corrected estimate of σ_{N_e} divided by the observed estimate of σ_{N_e}. This slightly underestimates the true SE since we have not taken into account the small amount of error associated with estimating the regression line. To test for heterogeneity in σ_{N_e} between species we assumed that the estimate of σ_{N_e} was normally distributed; under this assumption $(\sigma_{N_e} - \bar{\sigma}_{N_e})^2 / var(\sigma_{N_e})$ is χ^2 distributed with k-1 degrees of freedom for

k species. $\bar{\sigma}_{N_e}$ was calculated as a weighted average, where the weights were inversely proportional to the variance of the estimate (Eyre-Walker, 1996).

5.3.6 Variation of efficiency of selection

We tested whether the strength of selection on non-synonymous mutations was correlated to the effective population size, which can be seen as testing if the fraction of deleterious mutations varies with N_e. This can be done by considering the correlation of P_n/P_s and θ_s or P_n/P_s and N_e ($=\theta_s/(4\mu)$), where P_n and P_s are the numbers of non-synonymous and synonymous mutations respectively and N_e values are point estimates from the genetic diversity and mutation rates taken from the literature. However, P_s and θ_s are not independent. We overcome this problem by splitting P_s into two independent values by generating a random hypergeometric variable as follows (Piganeau and Eyre-Walker, 2009; Stoletzki and Eyre-Walker, 2011):

$$P_{s1} = Hypergeometric(P_s, L_s - P_s, 0.5P_s) \tag{5.9}$$

$$P_{s2} = P_s - P_{s1} \tag{5.10}$$

One of the P_s values is used to estimate P_n/P_s (see below) and the other one is used to estimate θ_S. There are two further problems to consider with this method, first, P_n/P_s can be an overestimate or underestimate of the true value of P_n/P_s and second the ratio P_n/P_s is undefined if $P_s = 0$. Both of these problems can be overcome by considering the correlation between ψ and θ_s (Piganeau and Eyre-Walker, 2009):

$$\psi = \frac{P_n}{P_s + 1} \tag{5.11}$$

Hence using our method to split P_s into independent values we have two independent pairs of θ_S and ψ; we only present results from one pair. Some of the datasets contain relatively little polymorphism which results in substantial variance of ψ. To overcome this problem we sum data across loci. For this we ranked loci according to their neutral diversity obtained from θ_{s2} and binned them into groups of size n (e.g. 2, 4, 8 and 16). For each group average θ_{s2} and corresponding N_{e2} values were calculated. Furthermore, for each group, the sums of P_n and P_{s1} were calculated in order to calculate ψ_1. Note, that ψ_2 can be obtained in a similar manner, however results were qualitatively comparable and we therefore only show results for ψ_1 vs θ_2 and N_{e2}. Also we only show results for group size 4 because results for group sizes > 2 were similar. The correlations were performed by calculating Spearman's rank correlation and probabilities were combined using the unweighted Z-method (Whitlock, 2005).

5.4 Results

To investigate variation in the effective population size within genomes, we assembled protein coding sequences from 10 species. The datasets are from six plant species, three animal species and one fungus. The datasets range in size from 66 to 918 loci per species and from 8 to 40 sequences per gene (Table 5.1). In all analyses we assume that synonymous mutations are neutral.

5.4.1 Variation of diversity and N_e within a genome using χ^2 and HKA tests

The level of genetic diversity appears to vary considerably within each genome (Figure 5.1); however, the number of polymorphisms per gene is generally quite low and hence this variation might be due to sampling error. To test whether the variation is significant we used two tests, which make different assumptions about the rate of recombination within loci - either free or no recombination. Both tests suggest that there is variation in the level of diversity in most species; all species are significant assuming free recombination and 6 out of 10 are significant assuming no recombination (Table 5.2). This variation in diversity between loci could be due to variation in the effective population size or to variation in the mutation rate. To investigate whether variation in the mutation rate might be responsible, we estimated the number of synonymous substitutions for each locus (D_S), between the species of interest and an outgroup species. In many species there is a significant positive correlation between D_S and P_S (Table 5.3) suggesting that part of the variation in diversity is due to variation in the mutation rate. However, if we test whether there is significant variation between loci taking into account the mutation rate, as estimated from the divergence between species, using either a χ^2 test of independence or the more conservative HKA test, then we find significant evidence in the majority of species, whether or not we assume free or no recombination within loci; 9 out of 10 loci for free recombination test and 6 out of 9 loci for the no recombination test (the HKA test could not be performed on *S.bicolor* due to the large number of genes in which the divergence was zero).

5.4.2 Correlates of N_e

The variation in N_e across the genome is likely to be due to genetic hitch-hiking and background selection. Both processes are expected to be stronger in regions of the genome

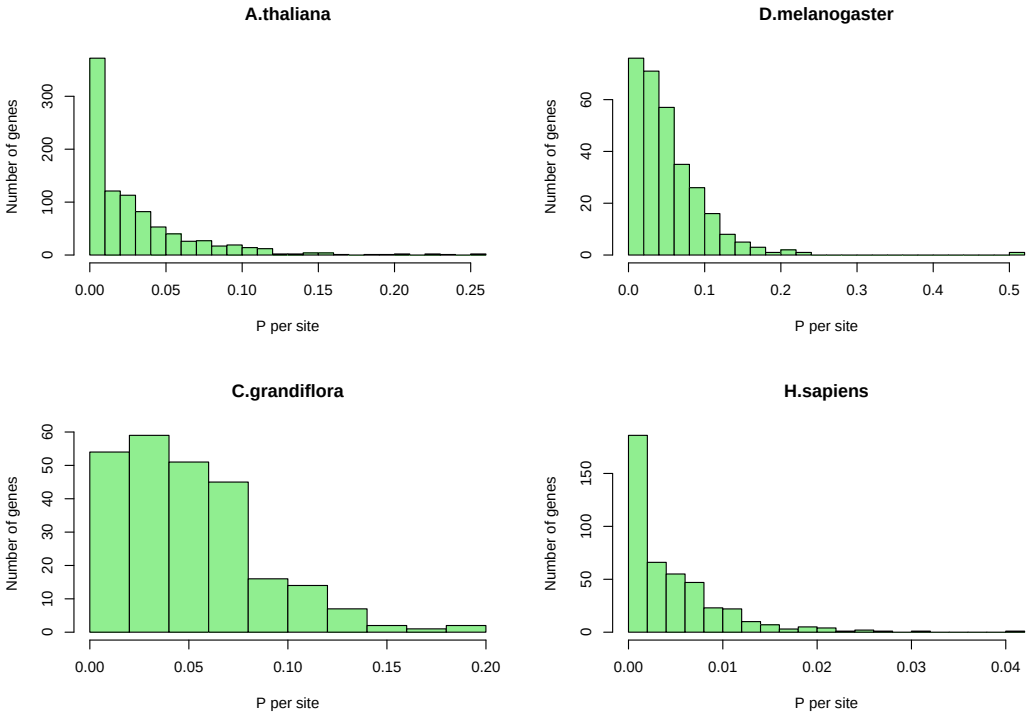


Figure 5.1: Distribution of the number of polymorphisms per site across genes for four species.

	Diversity		Diversity and Divergence	
Species	P-value (χ^2)	P-value (HKA)	P-value (χ^2)	P-value (HKA)
D.melanogaster	$< 1 \times 10^{-3}$	0.015	$< 1 \times 10^{-3}$	$< 1 \times 10^{-3}$
H.sapiens	$< 1 \times 10^{-3}$	$< 1 \times 10^{-3}$	$< 1 \times 10^{-3}$	$< 1 \times 10^{-3}$
M.musculus	$< 1 \times 10^{-3}$	0.432*	0.066*	0.429*
A.thaliana	$< 1 \times 10^{-3}$	$< 1 \times 10^{-3}$	$< 1 \times 10^{-3}$	$< 1 \times 10^{-3}$
C.grandiflora	$< 1 \times 10^{-3}$	0.462*	$< 1 \times 10^{-3}$	0.565*
S.bicolor	$< 1 \times 10^{-3}$	$< 1 \times 10^{-3}$	5.3×10^{-3}	n.a.
B.stricta	$< 1 \times 10^{-3}$	0.434*	6×10^{-3}	0.01
A.lyrata	$< 1 \times 10^{-3}$	5.4×10^{-3}	$< 1 \times 10^{-3}$	2.3×10^{-3}
C.rubella	$< 1 \times 10^{-3}$	$< 1 \times 10^{-3}$	$< 1 \times 10^{-3}$	$< 1 \times 10^{-3}$
S.paradoxus	1.9×10^{-3}	0.94*	$< 1 \times 10^{-3}$	0.35*

* not significant (P>0.05)

Table 5.2: Results of the χ^2 tests of independence and HKA tests for diversity and diversity/divergence data. For details see material and methods. P-values are given for each species.

	P_s vs D_s		P_s vs N_e	
Species	ρ	P-value	ρ	P-value
D.melanogaster	0.18	3.82×10^{-03}	0.46	1.87×10^{-17}
H.sapiens	0.29	1.62×10^{-06}	0.38	3.02×10^{-16}
M.musculus	0.38	5.98×10^{-03}	0.33	3.55×10^{-03}
A.thaliana	0.16	1.13×10^{-04}	0.44	5.94×10^{-45}
C.grandiflora	0.35	3.00×10^{-08}	0.52	2.53×10^{-19}
S.bicolor	0.54	3.17×10^{-03}	0.40	7.31×10^{-07}
B.stricta	0.10*	4.02×10^{-01}	0.14*	6.18×10^{-02}
A.lyrata	0.22*	1.03×10^{-01}	0.62	1.21×10^{-08}
C.rubella	0.12*	6.34×10^{-01}	0.65	2.35×10^{-07}
S.paradoxus	0.04*	7.91×10^{-01}	0.42	1.01×10^{-05}

* not significant (P>0.05)

Table 5.3: Results of the Spearman's rank correlates of P_s. The non-independence of P_s and N_e is taken into account by splitting the dataset into independent halves (see materials and methods). Correlation coefficients (ρ) and P-values are given for each species.

with low rates of recombination and a high density of sites subject to natural selection. To investigate which, if either of these factors is responsible for the variation in N_e, we investigated whether the variation in N_e was correlated to the rate of recombination and density of selected sites in four species for which this data was available: *D. melanogaster*, human, mouse and *A. thaliana*. We measured the density of selected sites as either the number of nucleotides in annotated exons (genic density), or the number of nucleotides in conserved regions (conserved site density), as annotated in the UCSC conservation track, in windows of size 50KB, 500KB and 5MB, where the window is centred on the gene from which the polymorphism data was taken (there is no conservation track for *A. thalaiana*, so in this species we just investigated the density of genic sites). Results for the different window sizes were generally consistent, so we present the results from the 500KB window size. We estimated N_e as the synonymous diversity divided by synonymous divergence.

In *D. melanogaster* we find, as others have done, that our estimate of N_e is positively correlated to recombination rate (Spearman's correlation coefficient r=0.45, $P < 0.01$). It is however also positively correlated to the density of conserved sites (r=0.24, $P < 0.01$), which is unexpected, though not genic sites (r=0.03, $P = 0.65$). The positive correlation with conserved site density might be due to the positive correlation that exists between the density of conserved sites and the rate of recombination (r=0.56, $P < 0.01$), and indeed if we perform a multiple regression we find that the correlation between N_e and the density of conserved sites disappears (P=0.74), while the positive correlation between N_e and recombination rate remains ($P < 0.01$).

In humans we find, as others have done, that both diversity (r=0.14, P=0.02) and divergence (r=0.18, $P < 0.01$) are positively correlated to the rate of recombination (Lercher and Hurst, 2002; Hellmann et al., 2005), and there is, as a consequence, no correlation between estimates of N_e and the rate of recombination (r=0.026, P=0.69). N_e is significantly negatively correlated to the density of genic sites (r=-0.19, $P < 0.01$), but not conserved sites (r=-0.085, $P = 0.17$). Using multiple regression does not alter this picture; N_e is only correlated to the density of genic sites.

In mouse we see no significant correlations between estimates of N_e and the rate of recombination (r=0.054, P=0.72), the density of genic (r=0.089, P=0.53) or conserved sites (r=0.093, P=0.51). This picture is unaffected by the use of multiple regression.

In *A. thaliana* we see a pattern like that in humans; both diversity (r=0.10, P=0.04) and to a lesser extent divergence (r=0.064, P=0.11) are positively correlated to recombination rate, and N_e is positively but not significantly correlated to recombination rate (r=0.080, P=0.11). N_e is significantly negatively correlated to genic density (r=-0.11, P=0.02). Unfortunately there is no data on conserved sites in this species.

5.4.3 Quantifying variation of N_e

Since we find evidence for variation in N_e in many of our species we attempted to quantify the amount of variation using a hierarchical Bayesian model. We assume underlying distributions for N_e and μ (e.g. log-normal distributions) and estimate the shape parameters σ_{N_e} and σ_μ and hence the variances of these distributions; the mean of each distribution is constrained to be equal to one (see materials and methods). We investigate two different models: in the first we assume free recombination and in the second we assume no recombination within loci, but free recombination between loci. These two models are likely to set the upper and lower bounds on the true level of variation in N_e. Under the free recombination model all the variation in diversity is attributed to variation in N_e, variation in the mutation rate and sampling error. In the model with no recombination, variation in diversity may additionally be due to variation in the coalescent process. Hence, the free recombination model gives an upper estimate on the variation in N_e and the no recombination model gives a lower bound.

We applied our method to the polymorphism data from each of the 10 eukaryotic species to estimate the variation of N_e within each genome along with the variation in the mutation rate, σ_μ (Table 5.4). As expected in all cases the estimate of σ_{N_e} is larger when free recombination is assumed, but the estimates from the two models are highly correlated (r=0.95). The estimate of σ_μ is unaffected by the model of recombination assumed. We find evidence that the value of σ_{N_e} varies between species for both the free and no recombination models ($P = 2.5 \times 10^{-9}$ and $P = 4.2 \times 10^{-8}$ respectively). We find that the level of variation of N_e is the lowest for *Mus musculus* and highest for *Capsella rubella* for both recombination models. The estimates of σ_{N_e} and σ_μ were of similar magnitude for each taxon suggesting that overall variation in the mutation rate and variation in the effective population size contribute a similar amount to the variation

Species	Free recombination		No recombination	
	σ_μ (Std)	σ_{N_e} (Std)	σ_μ (Std)	σ_{N_e} (Std)
D.melanogaster	0.370 (0.024)	0.743 (0.048)	0.372 (0.024)	0.516 (0.072)
H.sapiens	0.522 (0.021)	0.682 (0.07)	0.52 (0.02)	0.578 (0.11)
M.musculus	0.369 (0.045)	0.35 (0.119)	0.372 (0.045)	0.247 (0.15)
A.thaliana	0.419 (0.015)	0.83 (0.04)	0.423 (0.015)	0.809 (0.065)
C.grandiflora	0.355(0.021)	0.475 (0.043)	0.351 (0.021)	0.165 (0.067)
S.bicolor	0.689 (0.092)	0.903 (0.263)	0.710 (0.095)	0.675 (0.292)
B.stricta	0.441 (0.039)	0.503 (0.174)	0.443 (0.0379)	0.411 (0.178)
A.lyrata	0.276 (0.053)	0.729 (0.119)	0.278 (0.054)	0.651 (0.139)
C.rubella	0.263 (0.042)	1.191 (0.21)	0.258 (0.043)	1.126 (0.243)
S.paradoxus	0.23 (0.023)	0.566 (0.208)	0.23 (0.0218)	0.387 (0.131)

Table 5.4: Estimates of the variation of N_e in 10 eukaryotic species. Results are for an underlying Log-Normal distribution for N_e and μ assuming either free recombination or no recombination (see materials and methods). For each dataset the mean shape parameters σ_{N_e} and σ_μ and in parentheses their standard deviations (Std) obtained from the posterior distribution are given.

in diversity.

The level of variation in N_e we estimate using our method is quite modest. For example, *C.rubella* has the highest estimate of σ_{N_e}, but under this distribution the genes in the 90th percentile have an N_e that is only 7.2-fold greater than those in the 10th percentile, i.e. 80% of genes have an effective population size within 7.2-fold of each other. Four species have estimates of σ_{Ne} of less than 0.6 meaning that the difference between the 90th and 10th percentile is less than 4-fold.

The estimated distribution appears to fit the data reasonably well (Figure 5.2). We would not expect the fit to be perfect, particularly at the lower end of the distribution, since this is where sampling error is a major issue; e.g. many genes have no polymorphism because of sampling error, not because they have an effective population size of zero. It is possible that assuming a log-normal distribution places some unwanted constraints on the estimation procedure; in particular the probability density tends to zero for low N_e. We therefore also fitted a gamma distribution to the data (Table S3); with this distribution the probability density does not necessarily decline to zero near the origin. However, the estimated distributions are very similar to those obtained assuming a log-normal distribution (Figures A5.4 and A5.5). The species which show low variation in N_e are also those which tend to show little evidence of variation in N_e, as judged by the χ^2 or HKA tests. This implies that failure to detect variation in N_e is largely because there is limited variation in N_e rather issues with statistical power.

5.4.4 Variation in the efficiency of selection

Although we estimate the variation in the effective population size to be modest, it is of interest to investigate whether this translates into significant differences in the efficiency of natural selection across the genome. To investigate this we tested whether there was a correlation between $\psi = P_n/(P_s + 1)$ and either θ_s or N_e for each locus in a manner which controls for the obvious non-independence of the two variables (see materials and methods). We remove the non-independence by splitting P_s into two independent parts and we use ψ because it reduces the bias inherent in the estimation of P_n/P_s; furthermore it allows P_n/P_s to be calculated for all genes (Piganeau and Eyre-Walker, 2009). This test is not very powerful since ψ has a large variance; furthermore it is

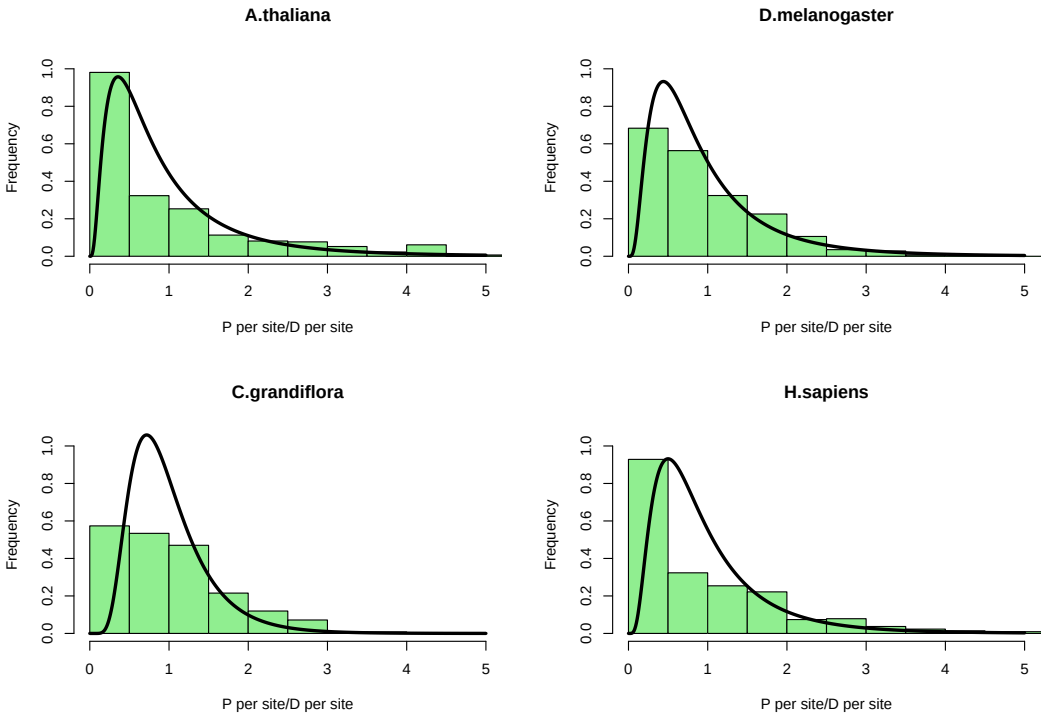


Figure 5.2: Distribution of the per site polymorphism/divergence ratio across genes for four species and corresponding distributions of N_e (solid line) estimated by the Hierarchical Bayesian analysis assuming a log-normal distribution.

statistically biased in a manner which tends to generate a positive correlation between ψ and θ_s or N_e. We therefore follow the approach suggested by (Piganeau and Eyre-Walker, 2009) and grouped genes according to their θ or N_e value. The results are qualitatively similar for groupings of 4, 8 and 16 genes, so we present the results for groups of 4. There is a significant negative correlation between both θ_s and ψ and N_e and ψ in *A. lyrata* and *C. grandiflora*, and a marginally significant correlation between ψ and θ_s in *D. melanogaster*, although only the correlations in *C. grandiflora* are significant after correction for multiple tests; otherwise the correlations are generally weak and non-significant. However, overall we find significant evidence for a negative correlation between ψ and θ_s or N_e if we combine probabilities: between ψ and θ_s P=0.043 and between ψ and N_e P=0.021.

The relationship between ψ and N_e can potentially yield information about the distribution of fitness effects (DFE; Loewe et al., 2006; Loewe and Charlesworth, 2006; Woolfit, 2006; Elyashiv et al., 2010). If we assume that the DFE for non-synonymous mutations is a gamma distribution, and that synonymous mutations are neutral, then P_n/P_s is expected to be proportional to $N_e^{-\beta}$, where β is the shape parameter of the gamma distribution (Welch et al., 2008). Hence we can estimate β by considering the slope of the regression line between $log(\psi)$ and $log(N_e)$. Since the log of zero is undefined we grouped genes in groups of size n such that no group had a zero estimate of ψ or N_e. We attempted to estimate β in the species which individually showed a significant correlation between ψ and N_e. However, we could not perform the analysis of *A. lyrata* because the diversity is so low that it was impossible to define groups that did not have zero values for both ψ and N_e. The estimates of β using this method are 0.41 (SE=0.15) in *C.grandiflora* and 0.23 (0.15) in *D. melanogaster*; these are similar to those obtained using an independent method that uses the site frequency spectrum (Keightley and Eyre-Walker, 2007): 0.27 (0.08) for *C. grandiflora* and 0.29 (0.07) for *D. melanogaster* (Table A5.2). This suggests that the gamma distribution is a reasonable approximation to the DFE, at least for mutations of weak effect.

5.5 Discussion

The effective population size (N_e) is one of the most important parameters in population genetics and evolutionary biology. It has been shown that N_e varies across the genome

of *Drosophila melanogaster* and some plant species, and it is thought that it might vary across the human genome (Hellmann et al., 2005). Here we have shown that it varies in most species that we have considered. However, the variation in N_e is not consistently correlated to either the rate of recombination or the density of selected sites. This might in part be because the variation in N_e is quite limited; most genes in a genome have an N_e which is within a few fold of most other genes. Nevertheless the variation is sufficient to cause differences in the effectiveness of natural selection on segregating non-synonymous polymorphisms.

There are a number of factors which might have led us to over- or underestimate the variation in N_e. First, we have assumed that there is either free recombination or no recombination within loci to estimate the variation in the effective population size. This is unsatisfactory since we know that recombination is one of the factors which generates variation in the effective population size, at least in species like *Drosophila*, in which there is a correlation between diversity and the rate of recombination. Unfortunately it is not easy to get around this problem. However, as we have noted earlier, the estimate assuming free recombination should give an upper estimate on the amount of variation, because under this method all variation in the diversity is assumed to arise from sampling error and variation in the mutation rate and N_e. In reality, some of the variation between genes will be a consequence of variation in the length of the genealogy in genes with little or no recombination.

Second, we have used the divergence between species as an estimate of the mutation rate, but if the mutation rate at a locus changes through time, for which there is evidence (Aguileta et al., 2006; Hodgkinson and Eyre-Walker, 2011), then we will tend to overestimate the variation in N_e; this is most easily seen by assuming there is variation in the mutation rate, but no variation in N_e; if the mutation rate has changed through time then the divergence will not be a perfect measure of the recent mutation rate and there will appear to be variation in N_e.

Third, we have assumed that synonymous mutations are neutral, but there is evidence of selection in humans (Iida and Akashi, 2000) and other species (Duret, 2002; Pond and Muse, 2005); although it is clear that selection has acted upon synonymous mutations in the past in *Drosophila melanogaster*, the evidence of selection currently acting is

contradictory (Akashi, 1996; McVean and Vieira, 2001; Zeng and Charlesworth, 2010) and biased gene conversion may be acting (Galtier et al., 2006; Zeng and Charlesworth, 2010). Most of the other species we have analysed have not been investigated in any detail. We need to consider two models. In the first model, let us assume that there is no variation in N_e but that there is variation in the strength of selection on synonymous codons. Such a model would generate apparent variation in N_e with the genes subject to the strongest selection apparently having the highest N_e, because negative selection affects divergence to a greater extent than polymorphism (Kimura, 1983). However, this would lead to the regions of the genome with the lowest diversity apparently having the highest effective population size. This is clearly not the case; if we split P_s into two independent samples, using a hypergeometric distribution, then we find a positive correlation between our estimate of N_e and P_s (Table 5.3). In the second model, let us imagine that there is variation in N_e and variation in the strength of selection on codon usage bias, but that they are uncorrelated to each other. In this case selection on codon usage bias will tend to generate an overestimate of the variation in N_e: as N_e increases selection becomes more effective, but this reduces the divergence more than the level of polymorphism, yielding a higher apparent effective population size. So genes in regions of high N_e will tend to have an exaggerated N_e. There is also another effect that needs to be considered. We have estimated the level of synonymous divergence using the method of Goldman and Yang (Goldman and Yang, 1994; Yang and Nielsen, 1998), which assumes that codon bias is due to mutation bias; however, this method will tend to overestimate the synonymous substitution rate if codon bias is due to selection, because it will incorrectly infer that genes with high bias have a small number of synonymous sites, and hence a relatively large number of substitutions (Bierne and Eyre-Walker, 2003; Yang, 2006). As a consequence the divergence in high biased genes will be overestimated, but at the same time the mutation rate will tend to be underestimated because of the action of selection. These two factors may cancel each other out.

Fourth, we have only applied our method to protein coding sequences, so we are estimating the variation in the effective population size that applies to the proteome; there might be further variation in N_e in regions that are relatively devoid of protein coding sequences, such as heterochromatin. Whether this is important depends on whether there are functional sequences within these regions. We have also only considered genes on the autosomes and occasionally the homogametic sex-chromosome (14 loci in *H.sapiens*). We

have not considered genes on the heterogametic sex chromosome, which often appear to have much lower effective population sizes. However, the heterogametic sex chromosome usually has very few genes (Graves, 2006).

Fifth, in estimating the variation in N_e we have assumed that there is either free recombination or no recombination and the population size has been stationary. Variation in population size can generate variation in diversity between loci, which may for example be mistaken for the signature of genetic hitch-hiking (Tajima, 1989; Pluzhnikov et al., 2002). In principle we could take this into account by estimating a demographic model from the polymorphism data while simultaneously estimating the variation in N_e. This is difficult and is beyond the scope of the current work.

Finally, we have not taken into account ancestral polymorphism within our method. Ignoring ancestral polymorphism will lead us to underestimate the variation in N_e because loci with large N_e will tend to have higher divergences than loci with small N_e and this will appear as though these loci have higher mutation rates; variation in N_e will therefore be underestimated because the mutation rate has been overestimated. In principle it is possible to include ancestral polymorphism within the method, but we observe a lack of convergence, probably because the number of polymorphisms for each gene was so low. However, we have chosen datasets in which divergence is generally considerably larger than diversity; for example, we chose macaque as the outgroup to humans because variation in N_e does appear to generate variation in the divergence between human and chimpanzee (McVicker et al., 2009).

Despite finding variation in N_e in many of the species we tested, we find no consistent evidence that N_e is correlated to either the rate of recombination or the density of selected sites, the two factors which we would have expected variation in N_e to depend upon. This is probably in part due to the fact that we are using synonymous diversity; as such our estimates of diversity are subject to considerable error. The lack of a strong correlation between recombination rate and N_e may also be due to the fact that the genetic maps in *A. thaliana* and mouse are relatively crude. Furthermore, for our mouse species we are using an F2 genetic linkage map constructed from intercrosses between *M. m. domesticus* and *M. m. castaneus* to infer recombination rates for *Mus musculus castaneus*. In humans it has previously been shown that diversity over

divergence is correlated positively to recombination rate (Hellmann et al., 2005) and that d_n/d_s is correlated to gene density (Bullaughey et al., 2008). In contrast to Hellmann et al. (2005) we do not find a significant correlation between N_e and recombination rate, but they used long non-coding sequences to investigate diversity over divergence; their estimates were therefore subject to much less error than ours. It is surprising that there is a correlation between genic density but not conserved site density in humans. This might be due to the fact that there is approximately twice as much variation in genic density as conserved site density (coefficient of variation 0.79 versus 0.30). It might also be due to differences in the DFE between the two types of sites; background selection is most effective when the strength of selection acting upon deleterious mutations is similar in magnitude to the rate of recombination (Nordborg et al., 1996).

In contrast genetic hitch-hiking depends upon the rate of advantageous mutation and sequences undergoing considerable adaptive evolution may not appear as conserved; the correlation between N_e and the density of genic sites may therefore suggest that hitch-hiking is more important in generating variation in N_e, than background selection. The lack of a correlation between N_e and the density of selected sites in Drosophila, once correlations to the rate of recombination have been taken into account may reflect the fact that the variation in N_e is generated by genetic hitch-hiking and a lot of adaptive evolution goes on outside coding sequences (Andolfatto, 2005).

Across species we find evidence that variation in N_e leads to variation in the effectiveness of natural selection on non-synonymous mutations across the genome (Table 5.5). However, this is individually significant for just two genomes: *C. grandiflora* and *A. lyrata*. A lack of a correlation in other genomes may be due to the fact that we have little power to detect the correlation since i) some of the datasets are quite small, ii) there is limited variation in N_e and iii) in most of these species the DFE is very leptokurtic. The kurtosis of the DFE is such that changes in effective population size do not greatly change the proportion of mutations that are effectively neutral. It can be shown that under a gamma DFE the proportion of effectively neutral mutations is proportional to $N_e^{-\beta}$ (Ohta, 1977; Kimura, 1979, 1983; Welch et al., 2008). Since β values are typically between 0.1 and 0.3 in most species (Table S2), changes in N_e tend to cause small changes in the proportion of effectively neutral mutations; for example a ten-fold increase in effective population size will reduce the proportion of effectively neutral mutations by only 37% if beta is

0.2. We find no evidence of a significant negative correlation between ψ and either θ_S or N_e in humans, in agreement with the work of Bullaughey et al. (2008). They found no evidence that the ratio of the non-synonymous (d_N) to the synonymous (d_S) substitution rate between human, chimpanzee and macaque was correlated to the rate of recombination.

Species	ψ vs. θ_s (groups of 4)			ψ vs. N_e (groups of 4)		
	n	ρ	P-value	n	ρ	P-value
D.melanogaster	77	-0.172	0.067	77	-0.1	0.194
H.sapiens	110	-0.068	0.239	110	0.016	0.564
M.musculus	18	-0.253	0.155	18	-0.261	0.147
A.thaliana	231	0.051	0.781	231	0.055	0.799
C.grandiflora	64	-0.357	0.002	64	-0.483	2.673×10^{-5}
S.bicolor	35	0.093	0.702	35	0.001	0.504
B.stricta	33	0.164	0.818	33	-0.168	0.175
A.lyrata	18	-0.477	0.022	18	-0.507	0.016
C.rubella	13	0.451	0.939	13	0.491	0.955
S.paradoxus	25	-0.219	0.146	25	-0.019	0.462
combined (Z-method)			0.043			0.021

Table 5.5: The correlation of $P_n/(P_s+1)=\psi$ and θ_s and N_e respectively in 10 eukaryotic species. The non-independence of ψ and θ_s is taken into account by splitting the dataset into independent halves (see materials and methods). Correlation coefficients (ρ) and P-values (one-tailed) are given for each species.

We find evidence that the amount of variation in N_e varies between species, however there are no obvious correlates of this variation. Both plants and animals have species with high and low levels of variation. Surprisingly we find no obvious effect of self-fertilization as suggested by previous studies (Cutter and Payseur, 2003; Roselius et al., 2005). *A.thaliana*, *C.rubella* and *B.stricta* are all self-fertile with selfing rates of around 0.95, 1 and 0.94 respectively (Charlesworth and Vekemans, 2005; Song et al., 2006; Foxe et al., 2009), whereas the closely related species *A. lyrata* and *C. grandiflora* are obligate outcrossing species. However, the variation in N_e seems to be relatively

low for *C.grandiflora* and *B.stricta* and similar for the two *Arabidopsis* species. It also should be noted that the confidence intervals on the estimate of N_e in *C.rubella* are very large and a substantial amount of variation is still shared between *C.grandiflora* and *C.rubella* so these estimates are not independent. Moreover, the lack of an effect for self-compatibility in our estimates of N_e for Arabidopsis may be not surprising as self-compatibility might have been evolved relatively recently in Arabidopsis (Bechsgaard et al., 2006; Tang et al., 2007). Furthermore, both *Arabidopsis* species have high sequence diversity in pericentromeric regions (Borevitz et al., 2007; Kawabe et al., 2008) which is not caused by varying mutation rates. Therefore this could be a major determinant of variation in N_e in those species and interfere with the effects of the breeding system.

Although the variation we observe in the effective population size appears to be modest, it does appear to influence both the level of neutral genetic diversity, and the effectiveness of selection. This potentially has important implications. If slightly deleterious mutations contribute substantially to phenotypic traits, then variation in the effective population size may affect where the genetic variation underlying fitness and other traits is distributed. For example, Rockman et al. (2010) have recently shown that expression QTLs (eQTLs) tend to be present in regions of the *C.elegans* genome with the highest rates of recombination and lowest density of genes, where N_e is expected to be largest. However, population genetic theory also suggests that such weakly selected mutations are unlikely to contribute much to the overall genetic variance in fitness unless the proportion of mutations under such weak selection is large (Eyre-Walker, 2010). Variation in the effective population might also affect the rate of adaptive evolution, as appears to be the case in *Drosophila* (Betancourt and Presgraves, 2002). Advantageous mutations can potentially come from three sources. They can be generated de novo, in which case we expect regions of the genome to adapt faster because the number of chromosomes an advantageous mutation can occur in is larger, and selection will be more effective on a greater proportion of the advantageous mutations. Advantageous mutations can also arise from standing genetic variation (Pritchard et al., 2010; Pritchard and Rienzo, 2010). If these mutations were previously strongly deleterious, the genetic variation is not expected to depend upon N_e, unless the mutations are highly recessive. If, however, the advantageous mutations were previously neutral or weakly selected, regions of the genome with high N_e are expected to have more genetic variation and hence adapt more rapidly.

Chapter 6

General conclusion and discussion

6.1 Selection as a shaping force of diversity and divergence in eukaryotic genomes

In this thesis we have investigated the role of selection as a shaping force of diversity and divergence in eukaryotic genomes. We have focused on two major quantities which play a key role in population genetics: (I) The rate of adaptive evolution using extensions of the McDonald Kreitman test and (II) Estimates of the effective population size and its variation within a genome. In chapter two, three and five we applied test statistics to biological sequence data, while in chapter four the role of fluctuating selection pressure was investigated using simulations. As shown in chapter two, the role of adaptive evolution in nine pairs of plants species is limited because we find little evidence for it in almost all comparisons. We extended this study in chapter three using a subset of the plant species pairs and adding available data from yeast, mammals and other plant species to investigate the relationship between adaptive evolution and effective population size and found a positive correlation between those two quantities. In chapter four we explored if high rates of adaptive evolution, such as found in some *Drosophila* species, might be falsely interpreted as positive selection and instead being rather the result of fluctuating selection pressure. Our findings suggest that fluctuation can cause high levels of adaptation, however this signal seems to be genuine. We also argue that a model with mean selective pressure with no net selective effect is likely to be unrealistic in real populations. In chapter five a Bayesian framework was used to quantify the variation of the effective population size within a genome taking the variation of mutation rate and recombination into account. Our findings suggest that this variation is relatively modest in all species but can cause significant differences in the efficiency of natural selection across the genome.

6.1.1 Genome wide estimates of adaptive evolution in plants

We have shown that there is only limited evidence for adaptive evolution in plants using data from nine pairs of plant species. To estimate the rate of adaptation we used a modified version of the McDonald Kreitman test which takes the distribution of fitness effects into account as well as a simple demographic model. This was done because slightly deleterious mutations can lead to an underestimate of the rate of adaptation, however our results suggest that this does not seem to cause the lack of evidence for adaptive evolution in plants. Other Studies for plant species have revealed that for species such as sunflowers, *Capsella grandiflora* and poplar adaptive evolution can be substantial. However, almost all of our species have relatively low effective population sizes, which could be a reasonable explanation for our findings. Some of our used datasets comprise a relatively small number of loci or few numbers of sequenced individuals which leads to a varying statistical power to infer adaptation for the different species. Surprisingly we failed to show evidence for adaptation in wild maize, a species with a large effective population size and for which the available dataset is also relatively large. Even though this result might be genuine there are two further explanations that may explain it. First we used *Sorghum bicolor* as an outgroup, which is a domesticated species and therefore the divergence leading to *Sorghum* should show little signature of genome wide adaptation as a consequence of the domestication process (Dillon et al., 2007). Secondly since the split from *Sorghum* maize has undergone a whole genome duplication which could lead to varying selection pressure between the resulting paralogs. Even though whole genome duplication events were frequent during angiosperm evolution, none of the other species shows such a recent whole genome duplication like maize. For *Populus tremula*, a species with a proposed large effective population size, we also fail to show adaptation, even though there is evidence from another study which is caused by methodological differences in the way to estimate the proportion of adaptive divergence. However unlike other species, poplars are currently the only trees for which estimates of adaptive divergence have been obtained. Those estimates rely all on divergence estimates of the same relatively close outgroup which may lead to an artifactual increase in the level of inferred adaptive evolution. Additionally the rough estimates of the generation time ($\approx$ 15 years) suggest overlapping generations and the excessive contribution of a few individuals to the offspring may lead to violation the assumed underlying Wright-Fisher coalescent process (Ingvarson, 2010). Poplar also shows strong signatures of selection on synonymous sites. This well studied example of the poplar trees illustrates that estimates of adaptive evolution for some plants may be

biased by technical or ecological factors and this has to be taken taken into account if a particular species is studied within a population genetics framework. Clearly additional data of more species will shed more light into the importance of adaptation for plant evolution. We also introduced an absolute measurement of adaptive evolution, ω_a, to circumvent the non-independence between the relative rates of adaptive evolution and the effective population size which we used for a subsequent study.

6.1.2 Effective population size as a determinant of the rate of adaptation

We extended our study of plant adaptation with data from further eukaryotic species to investigate the relationship between the rate of adaptive evolution and effective population size in 13 species pairs. This is currently the largest meta analyses across multiple taxa from different phylogenetic groups. To conduct the analysis we had to circumvent a number of technical challenges which previous analyses have suffered from. First, α, the proportion of adaptive divergence depends on the number of advantageous and neutral substitutions. This would lead to a positive correlation between α and N_e even when the absolute rates of adaptation are identical. Therefore we used the absolute rate of adaptive evolution ω_a (Gossmann et al., 2010). Secondly, α/ω_a and the effective population size N_e are not independent which we resolved by splitting the neutral variation into independent datasets. Thirdly, in previous analyses with multiple species, divergence was shared between the species pairs and resulted in non-independent estimates. We solved this problem by using only phylogenetic independent species pairs. Fourthly, we corrected for the effect of low divergence between species pairs which accounts for the contribution of polymorphisms into divergence. However, there are a number of concerns associated with this study (Venton, 2012). First, our estimates of the effective population size represents an estimate of the current effective population size. It can only be regarded as a point estimate of the long-term effective population size; estimating the long-term effective population size is difficult. Our estimates of N_e do not include demographic effects since the split between the species pairs and the estimates of lineage specific divergence would only partially reduce demographic effects on the divergence. However we argue even though this may be a problem for a particular species that it is unlikely that this would set up a systematic bias in our cross species analysis. If our estimate of the effective population is only poorly correlated with the real estimate we would not expect to observe any correlation between our estimate of N_e and ω_a. Second, we have not taken the effect of low recombining regions into account (Venton, 2012), which however is from the currently data hard to handle.

More complex analytic scenarios will be necessary to account for this effect. However, we do not expect this to bias our result systematically. Taken together, we conclude that it is rather surprising to find evidence for the correlation between ω_a and N_e even though currently available data suffer from shortcomings.

6.1.3 The impact of fluctuation on patterns of diversity and divergence

We investigated the impact of fluctuating selection on diversity and divergence and its outcome in a McDonald Kreitman test framework using simulations. We extended the model from Huerta-Sanchez et al. (2008) that uses random fluctuations and investigated average selective effects of individuals mutations. We find that signatures of adaptive evolution under fluctuating conditions are genuine because most mutations that reach fixation experience a positive mean selective effect even when the average mean effect of all mutations is zero or negative. Therefore it is not surprising that Huerta-Sanchez et al. (2008) have failed to distinguish a fluctuating model from a model of positive directional selection using a maximum likelihood approach based on information from the site frequency spectrum. However it remains unclear which values are realistic under biological conditions and we have shown that even when only few sites are under fluctuating selection the amount of adaptation inferred by an MK test framework can be substantial. Future studies will be focusing on accurately estimating reasonable mean fitness effects and their fluctuation intensities. Also Huerta-Sanchez et al. (2008) suggest that autocorrelated models behave equivalently to the random fluctuation model for which, however, they do not provide a formal analytical proof.

6.1.4 Estimating the variation in N_e within a genome

For a few species there is evidence that there is variation in N_e, but it was unknown how widespread this is. We investigated if variation in N_e is common or rather exceptional and limited to a few key species. We used synonymous diversity and divergence data from ten eukaryotic species and found that the variation is relatively modest but significant in all species. However much of the variation remains unexplained since we do not find consistent evidence that N_e is correlated to either the rate of recombination or density of selected in a subset of the ten species for which additional data are available. We also showed that even variation in N_e is relatively modest it is sufficient to cause significant differences in the efficiency of natural selection. Unfortunately our analysis suffers from several shortcomings which are difficult to resolve and will be of concern for future studies. We have considered

a model in which we investigate two extreme cases with either no or free recombination within loci. However there is evidence for species such as *Drosophila* that recombination rate variation generates variation in N_e. We also neglected the effects of changing mutation rates over time, demography, selection on synonymous sites and ancestral polymorphisms, which could either lead to an under- or overestimation the variation there is. Nevertheless this study provides the first large scale attempt to investigate if variation in N_e within a genome is widespread, which is the case, and also provides an initial attempt to quantify this variation.

6.2 Current limitations and perspectives

In this thesis we have focused on the comparison of coding regions of the genome, even though there is evidence that selection acts also in noncoding regions. However, availability of data from noncoding regions is currently limited. We have restricted our analysis to the most prominent type of mutations, single nucleotide polymorphisms and thereby neglected the contribution of other types of mutations to diversity and divergence. Currently within the MK framework estimates of α have been obtained for a few dozen species. In the near future data for much more species will be available due to advances in sequencing technology. Since newest sequencing approaches largely benefit from the existence of a reference genome, species for which whole genome information is available will be in main focus. But also for species for which estimates of adaptive evolution have been already obtained more sophisticated analyses from larger datasets will increase precision. This is will be of particular relevance for the distribution of fitness effects for which confidence intervals are relatively large at the moment.

Current studies that involve the estimation of adaptation using the MK test framework are usually restricted to a certain species and its close relatives because of the ecological and/or economical relevance. Under such a framework the ecological expertise improves the population sampling scheme and the consideration of population subdivision which will affect the MK analysis. However, estimates from such multiple closely related species are non-independent because they refer to the same set of genes or divergence information are shared between species. Also inferences of demographic events or population structure rely on similar kind of data that can be used in an MK framework and lead to potential biases in comparative studies. For example for a genetic study of a few closely related species one would for technical

reasons utilise a set of genes which is present in all species and which does therefore not represent a random sample of loci. Consequently large scale meta analysis as conducted in this thesis are necessary to consider such biases associated with the experimental set up.

Polymorphisms may lead to a misinference of the divergence between two species, mainly for two reasons. First, an actual polymorphism may appear as a substitution which is especially a problem when only few individuals are sequenced in at least of the two species under consideration. This is usually the case in a MK framework where only a single outgroup sequence is used. However, to some degree it is possible to correct for this effect. Furthermore closely related species share ancestral polymorphisms which may be falsely classified as divergence. The effect of ancestral polymorphisms can be reduced by using a more distantly related outgroup. Currently only a few studies have investigated the effect of different outgroups on the estimates of adaptive evolution for biological datasets (Halligan et al., 2010; Strasburg et al., 2011a). We have also shown that estimates for cross species comparison need to be statistically independent. Independence and precision of divergence estimates can be obtained by using lineage specific divergence, e.g. two outgroups - which few studies have already considered when additional data are available. Recent attempts use the unfolded (full) site frequency spectrum, in particularity to estimate the contribution of adaptive mutations to polymorphisms. These methods rely therefore on a correct reconstruction of the ancestral state of a particular site. Currently it is poorly understood what the effect of new sequencing approaches with regard to sequencing errors rates on the inferences of population genetic parameters is. In the near future such technical aspects will be investigated in greater detail.

Even though α, the proportion of adaptive substitutions or ω_a, the relative rate of adaptation are interesting quantities *per se* they do not provide any measure of the average selective advantage. To compare the role of adaptive evolution between species one might also ask the question if few mutations with relatively large effects or numerous mutations with little selective effects account for adaptive divergence. For this it would be necessary to estimate the distribution of fitness effects for new advantageous mutations. Even more challenging in this respect are models that assume that advantageous mutation arise from standing genetic variation or the inclusion of fluctuating conditions.

More progress is also needed to obtain precise estimates of N_e, e.g. identification of sites

which are truly neutral or developing methods to correct for the effects of demography, biased gene conversion and selection on neutral sites. Such methods have been implemented, however often it is difficult to disentangle between different effects. In this thesis we obtained measurements of the effective population size from the product of nucleotide diversity and the mutation rate per generation. However for some species estimates of the generation time and mutations rate are relatively poor and variation in these quantities are not considered under current models.

Bibliography

Aguileta, G., Bielawski, J. P., and Yang, Z. (2006). Evolutionary rate variation among vertebrate beta globin genes: implications for dating gene family duplication events. *Gene*, 380(1):21–29.

Akashi, H. (1996). Molecular evolution between Drosophila melanogaster and D. simulans: reduced codon bias, faster rates of amino acid substitution, and larger proteins in D. melanogaster. *Genetics*, 144(3):1297–1307.

Akashi, H. (1999). Inferring the fitness effects of DNA mutations from polymorphism and divergence data: statistical power to detect directional selection under stationarity and free recombination. *Genetics*, 151(1):221–238.

Akashi, H. and Schaeffer, S. W. (1997). Natural selection and the frequency distributions of "silent" DNA polymorphism in Drosophila. *Genetics*, 146(1):295–307.

Akey, J. M. (2009). Constructing genomic maps of positive selection in humans: where do we go from here? *Genome Res*, 19(5):711–722.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990). Basic local alignment search tool. *J Mol Biol*, 215(3):403–410.

Andolfatto, P. (2001). Contrasting patterns of X-linked and autosomal nucleotide variation in *Drosophila melanogaster* and *Drosophila simulans*. *Mol Biol Evol*, 18(3):279–290.

Andolfatto, P. (2005). Adaptive evolution of non-coding DNA in Drosophila. *Nature*, 437(7062):1149–1152.

Andolfatto, P. (2008). Controlling type-I error of the McDonald-Kreitman test in genomewide scans for selection on noncoding DNA. *Genetics*, 180(3):1767–1771.

Andolfatto, P., Wong, K. M., and Bachtrog, D. (2011a). Effective population size and the efficacy of selection on the X chromosomes of two closely related Drosophila species. *Genome Biol Evol*, 3:114–128.

Andolfatto, P., Wong, K. M., and Bachtrog, D. (2011b). Effective population size and the efficacy of selection on the X chromosomes of two closely related Drosophila species. *Genome Biol Evol*, 3:114–128.

Aquadro, C. F., Lado, K. M., and Noon, W. A. (1988). The rosy region of Drosophila melanogaster and Drosophila simulans. I. Contrasting levels of naturally occurring DNA restriction map variation and divergence. *Genetics*, 119(4):875–888.

Arguello, J. R., Zhang, Y., Kado, T., Fan, C., Zhao, R., Innan, H., Wang, W., and Long, M. (2010). Recombination yet inefficient selection along the Drosophila melanogaster subgroup's fourth chromosome. *Mol Biol Evol*, 27(4):848–861.

Axelsson, E. and Ellegren, H. (2009). Quantification of adaptive evolution of genes expressed in avian brain and the population size effect on the efficacy of selection. *Mol Biol Evol*, 26(5):1073–1079.

Bachtrog, D. (2008). Similar rates of protein adaptation in Drosophila miranda and D. melanogaster, two species with different current effective population sizes. *BMC Evol Biol*, 8:334.

Bachtrog, D. and Charlesworth, B. (2002). Reduced adaptation of a non-recombining neo-Y chromosome. *Nature*, 416(6878):323–326.

Barnaud, A., Trigueros, G., McKey, D., and Joly, H. I. (2008). High outcrossing rates in fields with mixed sorghum landraces: how are landraces maintained? *Heredity*, 101(5):445–452.

Barrier, M., Bustamante, C. D., Yu, J., and Purugganan, M. D. (2003). Selection on rapidly evolving proteins in the Arabidopsis genome. *Genetics*, 163(2):723–733.

Bartolomé, C., Maside, X., Yi, S., Grant, A. L., and Charlesworth, B. (2005). Patterns of selection on synonymous and nonsynonymous variants in Drosophila miranda. *Genetics*, 169(3):1495–1507.

Baudry, E., Derome, N., Huet, M., and Veuille, M. (2006). Contrasted polymorphism patterns in a large sample of populations from the evolutionary genetics model Drosophila simulans. *Genetics*, 173(2):759–767.

Baudry, E., Viginier, B., and Veuille, M. (2004). Non-African populations of Drosophila melanogaster have a unique origin. *Mol Biol Evol*, 21(8):1482–1491.

Beaumont, M. A. and Balding, D. J. (2004). Identifying adaptive genetic divergence among populations from genome scans. *Mol Ecol*, 13(4):969–980.

Bechsgaard, J. S., Castric, V., Charlesworth, D., Vekemans, X., and Schierup, M. H. (2006). The transition to self-compatibility in Arabidopsis thaliana and evolution within S-haplotypes over 10 Myr. *Mol Biol Evol*, 23(9):1741–1750.

Begun, D. J. and Aquadro, C. F. (1992). Levels of naturally occurring DNA polymorphism correlate with recombination rates in D. melanogaster. *Nature*, 356(6369):519–520.

Begun, D. J., Holloway, A. K., Stevens, K., Hillier, L. W., Poh, Y.-P., Hahn, M. W., Nista, P. M., Jones, C. D., Kern, A. D., Dewey, C. N., Pachter, L., Myers, E., and Langley, C. H. (2007). Population genomics: whole-genome analysis of polymorphism and divergence in Drosophila simulans. *PLoS Biol*, 5(11):e310.

Bell, G. (2010). Fluctuating selection: the perpetual renewal of adaptation in variable environments. *Philos Trans R Soc Lond B Biol Sci*, 365(1537):87–97.

Berlin, S. and Ellegren, H. (2006). Fast accumulation of nonsynonymous mutations on the female-specific W chromosome in birds. *J Mol Evol*, 62(1):66–72.

Berry, A. J., Ajioka, J. W., and Kreitman, M. (1991). Lack of polymorphism on the Drosophila fourth chromosome resulting from selection. *Genetics*, 129(4):1111–1117.

Betancourt, A. J. and Presgraves, D. C. (2002). Linkage limits the power of natural selection in Drosophila. *Proc Natl Acad Sci U S A*, 99(21):13616–13620.

Betancourt, A. J., Welch, J. J., and Charlesworth, B. (2009). Reduced effectiveness of selection caused by a lack of recombination. *Curr Biol*, 19(8):655–660.

Bierne, N. and Eyre-Walker, A. (2003). The problem of counting sites in the estimation of the synonymous and nonsynonymous substitution rates: implications for the correlation between the synonymous substitution rate and codon usage bias. *Genetics*, 165(3):1587–1597.

Bierne, N. and Eyre-Walker, A. (2004). The genomic rate of adaptive amino acid substitution in *Drosophila*. *Mol Biol Evol*, 21(7):1350–1360.

Bijlsma, R., Allard, R. W., and Kahler, A. L. (1986). Nonrandom Mating in an Open-Pollinated Maize Population. *Genetics*, 112(3):669–680.

Blanc, G. and Wolfe, K. H. (2004). Widespread paleopolyploidy in model plant species inferred from age distributions of duplicate genes. *Plant Cell*, 16(7):1667–1678.

Borevitz, J. O., Hazen, S. P., Michael, T. P., Morris, G. P., Baxter, I. R., Hu, T. T., Chen, H., Werner, J. D., Nordborg, M., Salt, D. E., Kay, S. A., Chory, J., Weigel, D., Jones, J. D. G., and Ecker, J. R. (2007). Genome-wide patterns of single-feature polymorphism in Arabidopsis thaliana. *Proc Natl Acad Sci U S A*, 104(29):12057–12062.

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

Brachi, B., Morris, G. P., and Borevitz, J. O. (2011). Genome-wide association studies in plants: the missing heritability is in the field. *Genome Biol*, 12(10):232.

Bullaughey, K., Przeworski, M., and Coop, G. (2008). No effect of recombination on the efficacy of natural selection in primates. *Genome Res*, 18(4):544–554.

Bustamante, C. D., Fledel-Alon, A., Williamson, S., Nielsen, R., Hubisz, M. T., Glanowski, S., Tanenbaum, D. M., White, T. J., Sninsky, J. J., Hernandez, R. D., Civello, D., Adams, M. D., Cargill, M., and Clark, A. G. (2005). Natural selection on protein-coding genes in the human genome. *Nature*, 437(7062):1153–1157.

Bustamante, C. D., Nielsen, R., Sawyer, S. A., Olsen, K. M., Purugganan, M. D., and Hartl, D. L. (2002). The cost of inbreeding in *Arabidopsis*. *Nature*, 416(6880):531–534.

Cai, J. J. and Petrov, D. A. (2010). Relaxed purifying selection and possibly high rate of adaptation in primate lineage-specific genes. *Genome Biol Evol*, 2:393–409.

Caicedo, A. L., Williamson, S. H., Hernandez, R. D., Boyko, A., Fledel-Alon, A., York, T. L., Polato, N. R., Olsen, K. M., Nielsen, R., McCouch, S. R., Bustamante, C. D., and Purugganan, M. D. (2007). Genome-wide patterns of nucleotide polymorphism in domesticated rice. *PLoS Genet*, 3(9):1745–1756.

Cargill, M., Altshuler, D., Ireland, J., Sklar, P., Ardlie, K., Patil, N., Shaw, N., Lane, C. R., Lim, E. P., Kalyanaraman, N., Nemesh, J., Ziaugra, L., Friedland, L., Rolfe, A., Warrington, J., Lipshutz, R., Daley, G. Q., and Lander, E. S. (1999). Characterization of single-nucleotide polymorphisms in coding regions of human genes. *Nat Genet*, 22(3):231–238.

Carneiro, M., Albert, F. W., Melo-Ferreira, J., Galtier, N., Gayral, P., Blanco-Aguiar, J. A., Villafuerte, R., Nachman, M. W., and Ferrand, N. (2012). Evidence for Widespread Positive and Purifying Selection Across the European Rabbit (Oryctolagus cuniculus) Genome. *Mol Biol Evol*, 29(7):1837–1849.

Chamary, J.-V. and Hurst, L. D. (2004). Similar rates but different modes of sequence evolution in introns and at exonic silent sites in rodents: evidence for selectively driven codon usage. *Mol Biol Evol*, 21(6):1014–1023.

Chamary, J. V., Parmley, J. L., and Hurst, L. D. (2006). Hearing silence: non-neutral evolution at synonymous sites in mammals. *Nat Rev Genet*, 7(2):98–108.

Charlesworth, B. (1994). The effect of background selection against deleterious mutations on weakly selected, linked variants. *Genet Res*, 63(3):213–227.

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

Charlesworth, B. (2010). Molecular population genomics: a short history. *Genet Res (Camb)*, 92(5-6):397–411.

Charlesworth, B., Charlesworth, D., and Barton, N. (2003). The effects of genetic and geographic structure on neutral variation. *Annu Rev Ecol Evol Syst*, 34:99–125.

Charlesworth, B., Morgan, M. T., and Charlesworth, D. (1993). The effect of deleterious mutations on neutral molecular variation. *Genetics*, 134(4):1289–1303.

Charlesworth, D., Charlesworth, B., and Morgan, M. T. (1995). The pattern of neutral molecular variation under the background selection model. *Genetics*, 141(4):1619–1632.

Charlesworth, D. and Vekemans, X. (2005). How and when did Arabidopsis thaliana become highly self-fertilising. *Bioessays*, 27(5):472–476.

Charlesworth, J. and Eyre-Walker, A. (2006). The rate of adaptive evolution in enteric bacteria. *Mol Biol Evol*, 23(7):1348–1356.

Charlesworth, J. and Eyre-Walker, A. (2007). The other side of the nearly neutral theory, evidence of slightly advantageous back-mutations. *Proc Natl Acad Sci U S A*, 104(43):16992–16997.

Charlesworth, J. and Eyre-Walker, A. (2008). The McDonald-Kreitman test and slightly deleterious mutations. *Mol Biol Evol*, 25(6):1007–1015.

Chimpanzee Sequencing and Analysis Consortium (2005). Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature*, 437(7055):69–87.

Cutter, A. D. and Payseur, B. A. (2003). Selection at linked sites in the partial selfer Caenorhabditis elegans. *Mol Biol Evol*, 20(5):665–673.

Davies, E. K., Peters, A. D., and Keightley, P. D. (1999). High frequency of cryptic deleterious mutations in Caenorhabditis elegans. *Science*, 285(5434):1748–1751.

Dillon, S. L., Shapter, F. M., Henry, R. J., Cordeiro, G., Izquierdo, L., and Lee, L. S. (2007). Domestication to crop improvement: genetic resources for Sorghum and Saccharum (Andropogoneae). *Ann Bot*, 100(5):975–989.

Dobes, C. H., Mitchell-Olds, T., and Koch, M. A. (2004). Extensive chloroplast haplotype variation indicates Pleistocene hybridization and radiation of North American *Arabis drummondii*, *A. x divaricarpa*, and *A. holboellii* (*Brassicaceae*). *Mol Ecol*, 13(2):349–370.

Drosophila 12 Genomes Consortium, Clark, A. G., Eisen, M. B., Smith, D. R., Bergman, C. M., Oliver, B., Markow, T. A., Kaufman, T. C., Kellis, M., Gelbart, W., Iyer, V. N., Pollard, D. A., Sackton, T. B., Larracuente, A. M., Singh, N. D., Abad, J. P., Abt, D. N., Adryan, B., Aguade, M., Akashi, H., Anderson, W. W., Aquadro, C. F., Ardell, D. H., Arguello, R., Artieri, C. G., Barbash, D. A., Barker, D., Barsanti, P., Batterham, P., Batzoglou, S., Begun, D., Bhutkar, A., Blanco, E., Bosak, S. A., Bradley, R. K., Brand, A. D., Brent, M. R., Brooks, A. N., Brown, R. H., Butlin, R. K., Caggese, C., Calvi, B. R., de Carvalho, A. B., Caspi, A., Castrezana, S., Celniker, S. E., Chang, J. L., Chapple, C., Chatterji, S., Chinwalla, A., Civetta, A., Clifton, S. W., Comeron, J. M., Costello, J. C., Coyne, J. A., Daub, J., David, R. G., Delcher, A. L., Delehaunty, K., Do, C. B., Ebling, H., Edwards, K., Eickbush, T., Evans, J. D., Filipski, A., Findeiss, S., Freyhult, E., Fulton, L., Fulton, R., Garcia, A. C. L., Gardiner, A., Garfield, D. A., Garvin, B. E., Gibson, G., Gilbert, D., Gnerre, S., Godfrey, J., Good, R., Gotea, V., Gravely, B., Greenberg, A. J., Griffiths-Jones, S., Gross, S., Guigo, R., Gustafson, E. A., Haerty, W., Hahn, M. W., Halligan, D. L., Halpern, A. L., Halter, G. M., Han, M. V., Heger, A., Hillier, L., Hinrichs, A. S., Holmes, I., Hoskins, R. A., Hubisz, M. J., Hultmark, D., Huntley, M. A., Jaffe, D. B., Jagadeeshan, S., Jeck, W. R., Johnson, J., Jones, C. D., Jordan, W. C., Karpen, G. H., Kataoka, E., Keightley, P. D., Kheradpour, P., Kirkness, E. F., Koerich, L. B., Kristiansen, K., Kudrna, D., Kulathinal, R. J., Kumar, S., Kwok, R., Lander, E., Langley, C. H., Lapoint, R., Lazzaro, B. P., Lee, S.-J., Levesque, L., Li, R., Lin, C.-F., Lin, M. F., Lindblad-Toh, K., Llopart, A., Long, M.,

Low, L., Lozovsky, E., Lu, J., Luo, M., Machado, C. A., Makalowski, W., Marzo, M., Matsuda, M., Matzkin, L., McAllister, B., McBride, C. S., McKernan, B., McKernan, K., Mendez-Lago, M., Minx, P., Mollenhauer, M. U., Montooth, K., Mount, S. M., Mu, X., Myers, E., Negre, B., Newfeld, S., Nielsen, R., Noor, M. A. F., O'Grady, P., Pachter, L., Papaceit, M., Parisi, M. J., Parisi, M., Parts, L., Pedersen, J. S., Pesole, G., Phillippy, A. M., Ponting, C. P., Pop, M., Porcelli, D., Powell, J. R., Prohaska, S., Pruitt, K., Puig, M., Quesneville, H., Ram, K. R., Rand, D., Rasmussen, M. D., Reed, L. K., Reenan, R., Reily, A., Remington, K. A., Rieger, T. T., Ritchie, M. G., Robin, C., Rogers, Y.-H., Rohde, C., Rozas, J., Rubenfield, M. J., Ruiz, A., Russo, S., Salzberg, S. L., Sanchez-Gracia, A., Saranga, D. J., Sato, H., Schaeffer, S. W., Schatz, M. C., Schlenke, T., Schwartz, R., Segarra, C., Singh, R. S., Sirot, L., Sirota, M., Sisneros, N. B., Smith, C. D., Smith, T. F., Spieth, J., Stage, D. E., Stark, A., Stephan, W., Strausberg, R. L., Strempel, S., Sturgill, D., Sutton, G., Sutton, G. G., Tao, W., Teichmann, S., Tobari, Y. N., Tomimura, Y., Tsolas, J. M., Valente, V. L. S., Venter, E., Venter, J. C., Vicario, S., Vieira, F. G., Vilella, A. J., Villasante, A., Walenz, B., Wang, J., Wasserman, M., Watts, T., Wilson, D., Wilson, R. K., Wing, R. A., Wolfner, M. F., Wong, A., Wong, G. K.-S., Wu, C.-I., Wu, G., Yamamoto, D., Yang, H.-P., Yang, S.-P., Yorke, J. A., Yoshida, K., Zdobnov, E., Zhang, P., Zhang, Y., Zimin, A. V., Baldwin, J., Abdouelleil, A., Abdulkadir, J., Abebe, A., Abera, B., Abreu, J., Acer, S. C., Aftuck, L., Alexander, A., An, P., Anderson, E., Anderson, S., Arachi, H., Azer, M., Bachantsang, P., Barry, A., Bayul, T., Berlin, A., Bessette, D., Bloom, T., Blye, J., Boguslavskiy, L., Bonnet, C., Boukhgalter, B., Bourzgui, I., Brown, A., Cahill, P., Channer, S., Cheshatsang, Y., Chuda, L., Citroen, M., Collymore, A., Cooke, P., Costello, M., D'Aco, K., Daza, R., Haan, G. D., DeGray, S., DeMaso, C., Dhargay, N., Dooley, K., Dooley, E., Doricent, M., Dorje, P., Dorjee, K., Dupes, A., Elong, R., Falk, J., Farina, A., Faro, S., Ferguson, D., Fisher, S., Foley, C. D., Franke, A., Friedrich, D., Gadbois, L., Gearin, G., Gearin, C. R., Giannoukos, G., Goode, T., Graham, J., Grandbois, E., Grewal, S., Gyaltsen, K., Hafez, N., Hagos, B., Hall, J., Henson, C., Hollinger, A., Honan, T., Huard, M. D., Hughes, L., Hurhula, B., Husby, M. E., Kamat, A., Kanga, B., Kashin, S., Khazanovich, D., Kisner, P., Lance, K., Lara, M., Lee, W., Lennon, N., Letendre, F., LeVine, R., Lipovsky, A., Liu, X., Liu, J., Liu, S., Lokyitsang, T., Lokyitsang, Y., Lubonja, R., Lui, A., MacDonald, P., Magnisalis, V., Maru, K., Matthews, C., McCusker, W., McDonough, S., Mehta, T., Meldrim, J., Meneus, L., Mihai, O., Mihalev, A., Mihova, T., Mittelman, R., Mlenga, V.,

Montmayeur, A., Mulrain, L., Navidi, A., Naylor, J., Negash, T., Nguyen, T., Nguyen, N., Nicol, R., Norbu, C., Norbu, N., Novod, N., O'Neill, B., Osman, S., Markiewicz, E., Oyono, O. L., Patti, C., Phunkhang, P., Pierre, F., Priest, M., Raghuraman, S., Rege, F., Reyes, R., Rise, (2007). Evolution of genes and genomes on the Drosophila phylogeny. *Nature*, 450(7167):203–218.

Dumont, B. L., White, M. A., Steffy, B., Wiltshire, T., and Payseur, B. A. (2011). Extensive recombination rate variation in the house mouse species complex inferred from genetic linkage maps. *Genome Res*, 21(1):114–125.

Duret, L. (2002). Evolution of synonymous codon usage in metazoans. *Curr Opin Genet Dev*, 12(6):640–649.

Elyashiv, E., Bullaughey, K., Sattath, S., Rinott, Y., Przeworski, M., and Sella, G. (2010). Shifts in the intensity of purifying selection: an analysis of genome-wide polymorphism data from two closely related yeast species. *Genome Res*, 20(11):1558–1573.

Eory, L., Halligan, D. L., and Keightley, P. D. (2010). Distributions of selectively constrained sites and deleterious mutation rates in the hominid and murid genomes. *Mol Biol Evol*, 27(1):177–192.

Eyre-Walker, A. (1996). Synonymous codon bias is related to gene length in Escherichia coli: selection for translational accuracy? *Mol Biol Evol*, 13(6):864–872.

Eyre-Walker, A. (2002). Changing effective population size and the McDonald-Kreitman test. *Genetics*, 162(4):2017–2024.

Eyre-Walker, A. (2006). The genomic rate of adaptive evolution. *Trends Ecol Evol*, 21(10):569–575.

Eyre-Walker, A. (2010). Evolution in health and medicine Sackler colloquium: Genetic architecture of a complex trait and its implications for fitness and genome-wide association studies. *Proc Natl Acad Sci U S A*, 107 Suppl 1:1752–1756.

Eyre-Walker, A. and Keightley, P. D. (2007). The distribution of fitness effects of new mutations. *Nat Rev Genet*, 8(8):610–618.

Eyre-Walker, A. and Keightley, P. D. (2009). Estimating the rate of adaptive molecular evolution in the presence of slightly deleterious mutations and population size change. *Mol Biol Evol*, 26(9):2097–2108.

Eyre-Walker, A., Keightley, P. D., Smith, N. G. C., and Gaffney, D. (2002). Quantifying the slightly deleterious mutation model of molecular evolution. *Mol Biol Evol*, 19(12):2142–2149.

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

Fay, J. C. (2011). Weighing the evidence for adaptation at the molecular level. *Trends Genet*, 27(9):343–349.

Fay, J. C. and Benavides, J. A. (2005). Evidence for domesticated and wild populations of Saccharomyces cerevisiae. *PLoS Genet*, 1(1):66–71.

Fay, J. C. and Wu, C. I. (2000). Hitchhiking under positive Darwinian selection. *Genetics*, 155(3):1405–1413.

Fay, J. C., Wyckoff, G. J., and Wu, C. I. (2001). Positive and negative selection on the human genome. *Genetics*, 158(3):1227–1234.

Fay, J. C., Wyckoff, G. J., and Wu, C.-I. (2002). Testing the neutral theory of molecular evolution with genomic data from *Drosophila*. *Nature*, 415(6875):1024–1026.

Filatov, D. A. and Burke, S. (2004). DNA diversity in Hawaiian endemic plant *Schiedea globosa*. *Heredity*, 92(5):452–458.

Filatov, D. A., Laporte, V., Vitte, C., and Charlesworth, D. (2001). DNA diversity in sex-linked and autosomal genes of the plant species Silene latifolia and Silene dioica. *Mol Biol Evol*, 18(8):1442–1454.

Fisher, R. and Ford, E. (1947). The spread of a gene in natural conditions in a colony of the moth Panaxia dominula L. *Heredity*, 1:143–174.

Foxe, J. P., Slotte, T., Stahl, E. A., Neuffer, B., Hurka, H., and Wright, S. I. (2009). Recent speciation associated with the evolution of selfing in Capsella. *Proc Natl Acad Sci U S A*, 106(13):5241–5245.

Foxe, J. P., un Nisa Dar, V., Zheng, H., Nordborg, M., Gaut, B. S., and Wright, S. I. (2008). Selection on amino acid substitutions in *Arabidopsis*. *Mol Biol Evol*, 25(7):1375–1383.

Fraser, C., Hanage, W. P., and Spratt, B. G. (2007). Recombination and the nature of bacterial speciation. *Science*, 315(5811):476–480.

Fu, Y., Emrich, S. J., Guo, L., Wen, T.-J., Ashlock, D. A., Aluru, S., and Schnable, P. S. (2005). Quality assessment of maize assembled genomic islands (MAGIs) and large-scale experimental verification of predicted genes. *Proc Natl Acad Sci U S A*, 102(34):12282–12287.

Gaffney, D. J. and Keightley, P. D. (2005). The scale of mutational variation in the murid genome. *Genome Res*, 15(8):1086–1094.

Galtier, N., Bazin, E., and Bierne, N. (2006). GC-biased segregation of noncoding polymorphisms in Drosophila. *Genetics*, 172(1):221–228.

Garrigan, D. and Hammer, M. F. (2006). Reconstructing human origins in the genomic era. *Nat Rev Genet*, 7(9):669–680.

Garrigan, D., Lewontin, R., and Wakeley, J. (2010). Measuring the sensitivity of single-locus "neutrality tests" using a direct perturbation approach. *Mol Biol Evol*, 27(1):73–89.

Gillespie, J. (1973). Natural selection with varying selection coefficients—a haploid model. *Genet. Res*, 21:115–120.

Gillespie, J. (1991a). *The causes of molecular evolution*, volume 2. Oxford University Press, USA.

Gillespie, J. H. (1991b). *The Causes of Molecular Evolution*. Oxford Series in Ecology and Evolution, Oxford.

Goldman, N. and Yang, Z. (1994). A codon-based model of nucleotide substitution for protein-coding DNA sequences. *Mol Biol Evol*, 11(5):725–736.

Gossmann, T. I., Keightley, P. D., and 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*.

Gossmann, T. I., Song, B.-H., Windsor, A. J., Mitchell-Olds, T., Dixon, C. J., Kapralov, M. V., Filatov, D. A., and Eyre-Walker, A. (2010). Genome wide analyses reveal little evidence for adaptive evolution in many plant species. *Mol Biol Evol*, 27(8):1822–1832.

Gossmann, T. I., Woolfit, M., and Eyre-Walker, A. (2011). Quantifying the variation in the effective population size within a genome. *Genetics*, 189(4):1389–1402.

Grath, S., Baines, J. F., and Parsch, J. (2009). Molecular evolution of sex-biased genes in the Drosophila ananassae subgroup. *BMC Evol Biol*, 9:291.

Graves, J. A. M. (2006). Sex chromosome specialization and degeneration in mammals. *Cell*, 124(5):901–914.

Guo, Y.-L., Bechsgaard, J. S., Slotte, T., Neuffer, B., Lascoux, M., Weigel, D., and Schierup, M. H. (2009). Recent speciation of *Capsella rubella* from *Capsella grandiflora*, associated with loss of self-incompatibility and an extreme bottleneck. *Proc Natl Acad Sci U S A*, 106(13):5246–5251.

Haag-Liautard, C., Dorris, M., Maside, X., Macaskill, S., Halligan, D. L., Houle, D., Charlesworth, B., and Keightley, P. D. (2007). Direct estimation of per nucleotide and genomic deleterious mutation rates in Drosophila. *Nature*, 445(7123):82–85.

Haddrill, P. R., Bachtrog, D., and Andolfatto, P. (2008). Positive and negative selection on noncoding DNA in Drosophila simulans. *Mol Biol Evol*, 25(9):1825–1834.

Haddrill, P. R., Halligan, D. L., Tomaras, D., and Charlesworth, B. (2007). Reduced efficacy of selection in regions of the Drosophila genome that lack crossing over. *Genome Biol*, 8(2):R18.

Haddrill, P. R., Loewe, L., and Charlesworth, B. (2010). Estimating the parameters of selection on nonsynonymous mutations in Drosophila pseudoobscura and D. miranda. *Genetics*, 185(4):1381–1396.

Haddrill, P. R., Zeng, K., and Charlesworth, B. (2011). Determinants of synonymous and nonsynonymous variability in three species of Drosophila. *Mol Biol Evol*, 28(5):1731–1743.

Hahn, M. W. (2009). Distinguishing among evolutionary models for the maintenance of gene duplicates. *J Hered*, 100(5):605–617.

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

Halligan, D. L., Oliver, F., Guthrie, J., Stemshorn, K. C., Harr, B., and Keightley, P. D. (2011). Positive and negative selection in murine ultraconserved noncoding elements. *Mol Biol Evol*, 28(9):2651–2660.

Hamblin, M. T., Casa, A. M., Sun, H., Murray, S. C., Paterson, A. H., Aquadro, C. F., and Kresovich, S. (2006). Challenges of detecting directional selection after a bottleneck: lessons from *Sorghum bicolor*. *Genetics*, 173(2):953–964.

Han, M. V., Demuth, J. P., McGrath, C. L., Casola, C., and Hahn, M. W. (2009). Adaptive evolution of young gene duplicates in mammals. *Genome Res*, 19(5):859–867.

Hastings, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57(1):97–109.

Hellborg, L. and Ellegren, H. (2004). Low levels of nucleotide diversity in mammalian Y chromosomes. *Mol Biol Evol*, 21(1):158–163.

Hellmann, I., Ebersberger, I., Ptak, S. E., Pääbo, S., and Przeworski, M. (2003). A neutral explanation for the correlation of diversity with recombination rates in humans. *Am J Hum Genet*, 72(6):1527–1535.

Hellmann, I., Prüfer, K., Ji, H., Zody, M. C., Pääbo, S., and Ptak, S. E. (2005). Why do human diversity levels vary at a megabase scale? *Genome Res*, 15(9):1222–1231.

Hernandez, R. D. (2008). A flexible forward simulator for populations subject to selection and demography. *Bioinformatics*, 24(23):2786–2787.

Hey, J. and Kliman, R. M. (2002). Interactions between natural selection, recombination and gene density in the genes of Drosophila. *Genetics*, 160(2):595–608.

Hodgkinson, A. and Eyre-Walker, A. (2011). Variation in the mutation rate across the mammalian genome. *Nat Rev Genet*, page in press.

Hudson, R. R., Kreitman, M., and Aguadé, M. (1987). A test of neutral molecular evolution based on nucleotide data. *Genetics*, 116(1):153–159.

Huerta-Sanchez, E., Durrett, R., and Bustamante, C. D. (2008). Population genetics of polymorphism and divergence under fluctuating selection. *Genetics*, 178(1):325–337.

Hughes, A. L. (2005). Evidence for abundant slightly deleterious polymorphisms in bacterial populations. *Genetics*, 169(2):533–538.

Hughes, A. L., Friedman, R., Rivailler, P., and French, J. O. (2008). Synonymous and nonsynonymous polymorphisms versus divergences in bacterial genomes. *Mol Biol Evol*, 25(10):2199–2209.

Hughes, A. L. and Nei, M. (1988). Pattern of nucleotide substitution at major histocompatibility complex class I loci reveals overdominant selection. *Nature*, 335(6186):167–170.

Hvilsom, C., Qian, Y., Bataillon, T., Li, Y., Mailund, T., Sallé, B., Carlsen, F., Li, R., Zheng, H., Jiang, T., Jiang, H., Jin, X., Munch, K., Hobolth, A., Siegismund, H. R., Wang, J., and Schierup, M. H. (2012). Extensive X-linked adaptive evolution in central chimpanzees. *Proc Natl Acad Sci U S A*, 109(6):2054–2059.

Iida, K. and Akashi, H. (2000). A test of translational selection at 'silent' sites in the human genome: base composition comparisons in alternatively spliced genes. *Gene*, 261(1):93–105.

Ingvarson, P. (2010). Nucleotide polymorphism, linkage disequilibrium and complex trait dissection in Populus. *Genetics and Genomics of Populus*, pages 91–111.

Ingvarsson, P. K. (2004). Population subdivision and the Hudson-Kreitman-Aguade test: testing for deviations from the neutral model in organelle genomes. *Genet Res*, 83(1):31–39.

Ingvarsson, P. K. (2008a). Molecular evolution of synonymous codon usage in *Populus*. *BMC Evol Biol*, 8:307.

Ingvarsson, P. K. (2008b). Multilocus patterns of nucleotide polymorphism and the demographic history of *Populus tremula*. *Genetics*, 180(1):329–340.

Ingvarsson, P. K. (2010). Natural Selection on Synonymous and Nonsynonymous Mutations Shapes Patterns of Polymorphism in Populus tremula. *Mol Biol Evol*, 27(3):650–660.

Initiative, T. A. G. (2000). Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature*, 408(6814):796–815.

Innan, H. (2006). Modified Hudson-Kreitman-Aguade test and two-dimensional evaluation of neutrality tests. *Genetics*, 173(3):1725–1733.

International Rice Genome Sequencing Project (2005). The map-based sequence of the rice genome. *Nature*, 436(7052):793–800.

Jaillon, O., Aury, J.-M., Noel, B., Policriti, A., Clepet, C., Casagrande, A., Choisne, N., Aubourg, S., Vitulo, N., Jubin, C., Vezzi, A., Legeai, F., Hugueney, P., Dasilva, C., Horner, D., Mica, E., Jublot, D., Poulain, J., Bruyère, C., Billault, A., Segurens, B.,

Gouyvenoux, M., Ugarte, E., Cattonaro, F., Anthouard, V., Vico, V., Fabbro, C. D., Alaux, M., Gaspero, G. D., Dumas, V., Felice, N., Paillard, S., Juman, I., Moroldo, M., Scalabrin, S., Canaguier, A., Clainche, I. L., Malacrida, G., Durand, E., Pesole, G., Laucou, V., Chatelet, P., Merdinoglu, D., Delledonne, M., Pezzotti, M., Lecharny, A., Scarpelli, C., Artiguenave, F., Pè, M. E., Valle, G., Morgante, M., Caboche, M., Adam-Blondon, A.-F., Weissenbach, J., Quétier, F., Wincker, P., and for Grapevine Genome Characterization, F.-I. P. C. (2007). The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. *Nature*, 449(7161):463–467.

Jensen, J. D. and Bachtrog, D. (2011). Characterizing the influence of effective population size on the rate of adaptation: Gillespie's Darwin domain. *Genome Biol Evol*, 3:687–701.

Jensen, L. (1973). Random selective advantages of genes and their probabilities of fixation. *Genet Res*, 21(3):215–219.

Karlin, S. and Levikson, B. (1974). Temporal fluctuations in selection intensities: Case of small population size. *Theoretical Population Biology*, 6(3):383 – 412.

Kauer, M., Zangerl, B., Dieringer, D., and Schlötterer, C. (2002). Chromosomal patterns of microsatellite variability contrast sharply in African and non-African populations of Drosophila melanogaster. *Genetics*, 160(1):247–256.

Kawabe, A., Forrest, A., Wright, S. I., and Charlesworth, D. (2008). High DNA sequence diversity in pericentromeric genes of the plant Arabidopsis lyrata. *Genetics*, 179(2):985–995.

Keightley, P. D., Eöry, L., Halligan, D. L., and Kirkpatrick, M. (2011). Inference of mutation parameters and selective constraint in mammalian coding sequences by approximate Bayesian computation. *Genetics*, 187(4):1153–1161.

Keightley, P. D. and Eyre-Walker, A. (2000). Deleterious mutations and the evolution of sex. *Science*, 290(5490):331–333.

Keightley, P. D. and Eyre-Walker, A. (2007). Joint inference of the distribution of fitness effects of deleterious mutations and population demography based on nucleotide polymorphism frequencies. *Genetics*, 177(4):2251–2261.

Keightley, P. D. and Eyre-Walker, A. (2010). What can we learn about the distribution of fitness effects of new mutations from DNA sequence data? *Philos Trans R Soc Lond B Biol Sci*, 365(1544):1187–1193.

Keightley, P. D. and Eyre-Walker, A. (2012). Estimating the Rate of Adaptive Molecular Evolution When the Evolutionary Divergence Between Species is Small. *J Mol Evol*.

Keller, S., Olson, M., Silim, S., Schroeder, W., and Tiffin, P. (2010). Genomic diversity, population structure, and migration following rapid range expansion in the balsam poplar, *Populus balsamifera*. *Molecular Ecology*, forthcoming.

Kim, Y. and Nielsen, R. (2004). Linkage disequilibrium as a signature of selective sweeps. *Genetics*, 167(3):1513–1524.

Kimura, M. (1954). Process Leading to Quasi-Fixation of Genes in Natural Populations Due to Random Fluctuation of Selection Intensities. *Genetics*, 39(3):280–295.

Kimura, M. (1969). The number of heterozygous nucleotide sites maintained in a finite population due to steady flux of mutations. *Genetics*, 61(4):893–903.

Kimura, M. (1979). Model of effectively neutral mutations in which selective constraint is incorporated. *Proc Natl Acad Sci U S A*, 76(7):3440–3444.

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

Kimura, M. (1991). The neutral theory of molecular evolution: a review of recent evidence. *Jpn J Genet*, 66(4):367–386.

King, J. L. and Jukes, T. H. (1969). Non-Darwinian evolution. *Science*, 164(3881):788–798.

Kliman, R. M. and Hey, J. (2003). Hill-Robertson interference in Drosophila melanogaster: reply to Marais, Mouchiroud and Duret. *Genet Res*, 81(2):89–90.

Koch, M. A., Haubold, B., and Mitchell-Olds, T. (2000). Comparative evolutionary analysis of chalcone synthase and alcohol dehydrogenase loci in *Arabidopsis*, *Arabis*, and related genera (*Brassicaceae*). *Mol Biol Evol*, 17(10):1483–1498.

Kong, A., Gudbjartsson, D. F., Sainz, J., Jonsdottir, G. M., Gudjonsson, S. A., Richardsson, B., Sigurdardottir, S., Barnard, J., Hallbeck, B., Masson, G., Shlien, A., Palsson, S. T., Frigge, M. L., Thorgeirsson, T. E., Gulcher, J. R., and Stefansson, K. (2002). A high-resolution recombination map of the human genome. *Nat Genet*, 31(3):241–247.

Koornneef, M., Alonso-Blanco, C., and Vreugdenhil, D. (2004). Naturally occurring genetic variation in *Arabidopsis thaliana*. *Annu Rev Plant Biol*, 55:141–172.

Kousathanas, A., Oliver, F., Halligan, D. L., and Keightley, P. D. (2011). Positive and negative selection on noncoding DNA close to protein-coding genes in wild house mice. *Mol Biol Evol*, 28(3):1183–1191.

Larracuente, A. M., Sackton, T. B., Greenberg, A. J., Wong, A., Singh, N. D., Sturgill, D., Zhang, Y., Oliver, B., and Clark, A. G. (2008). Evolution of protein-coding genes in Drosophila. *Trends Genet*, 24(3):114–123.

Lercher, M. J. and Hurst, L. D. (2002). Human SNP variability and mutation rate are higher in regions of high recombination. *Trends Genet*, 18(7):337–340.

Li, H. and Stephan, W. (2006). Inferring the demographic history and rate of adaptive substitution in Drosophila. *PLoS Genet*, 2(10):e166.

Li, W. H. (1997). *Molecular evolution.* Sinauer Associates Incorporated.

Liti, G., Carter, D. M., Moses, A. M., Warringer, J., Parts, L., James, S. A., Davey, R. P., Roberts, I. N., Burt, A., Koufopanou, V., Tsai, I. J., Bergman, C. M., Bensasson, D., O'Kelly, M. J. T., van Oudenaarden, A., Barton, D. B. H., Bailes, E., Nguyen, A. N., Jones, M., Quail, M. A., Goodhead, I., Sims, S., Smith, F., Blomberg, A., Durbin, R., and Louis, E. J. (2009). Population genomics of domestic and wild yeasts. *Nature*, 458(7236):337–341.

Loewe, L. and Charlesworth, B. (2006). Inferring the distribution of mutational effects on fitness in Drosophila. *Biol Lett*, 2(3):426–430.

Loewe, L., Charlesworth, B., Bartolomé, C., and Nöel, V. (2006). Estimating selection on nonsynonymous mutations. *Genetics*, 172(2):1079–1092.

Lynch, M. (1987). The consequences of fluctuating selection for isozyme polymorphisms in Daphnia. *Genetics*, 115(4):657–669.

Lynch, M. (2007). *The Origins of Genome Architecture*, volume 98. Sinauer Associates.

Lynch, M. (2010). Rate, molecular spectrum, and consequences of human mutation. *Proc Natl Acad Sci U S A*, 107(3):961–968.

Machado, C. A., Kliman, R. M., Markert, J. A., and Hey, J. (2002). Inferring the history of speciation from multilocus DNA sequence data: the case of Drosophila pseudoobscura and close relatives. *Mol Biol Evol*, 19(4):472–488.

Mallick, S., Gnerre, S., Muller, P., and Reich, D. (2009). The difficulty of avoiding false positives in genome scans for natural selection. *Genome Res*, 19(5):922–933.

Mank, J. E., Vicoso, B., Berlin, S., and Charlesworth, B. (2010). Effective population size and the Faster-X effect: empirical results and their interpretation. *Evolution*, 64(3):663–674.

Marais, G., Mouchiroud, D., and Duret, L. (2003). Neutral effect of recombination on base composition in Drosophila. *Genet Res*, 81(2):79–87.

Maside, X. and Charlesworth, B. (2007). Patterns of molecular variation and evolution in Drosophila americana and its relatives. *Genetics*, 176(4):2293–2305.

McDonald, J. H. (1996). Detecting non-neutral heterogeneity across a region of DNA sequence in the ratio of polymorphism to divergence. *Mol Biol Evol*, 13(1):253–260.

McDonald, J. H. (1998). Improved tests for heterogeneity across a region of DNA sequence in the ratio of polymorphism to divergence. *Mol Biol Evol*, 15(4):377–384.

McDonald, J. H. and Kreitman, M. (1991). Adaptive protein evolution at the Adh locus in *Drosophila*. *Nature*, 351(6328):652–654.

McVean, G. A. and Vieira, J. (2001). Inferring parameters of mutation, selection and demography from patterns of synonymous site evolution in Drosophila. *Genetics*, 157(1):245–257.

McVicker, G., Gordon, D., Davis, C., and Green, P. (2009). Widespread genomic signatures of natural selection in hominid evolution. *PLoS Genet*, 5(5):e1000471.

Metta, M., Gudavalli, R., Gibert, J.-M., and Schlötterer, C. (2006). No accelerated rate of protein evolution in male-biased Drosophila pseudoobscura genes. *Genetics*, 174(1):411–420.

Montell, H., Fridolfsson, A. K., and Ellegren, H. (2001). Contrasting levels of nucleotide diversity on the avian Z and W sex chromosomes. *Mol Biol Evol*, 18(11):2010–2016.

Moore, R. C. and Purugganan, M. D. (2003). The early stages of duplicate gene evolution. *Proc Natl Acad Sci U S A*, 100(26):15682–15687.

Moran, N. A., McCutcheon, J. P., and Nakabachi, A. (2008). Genomics and evolution of heritable bacterial symbionts. *Annu Rev Genet*, 42:165–190.

Moriyama, E. N. and Powell, J. R. (1996). Intraspecific nuclear DNA variation in Drosophila. *Mol Biol Evol*, 13(1):261–277.

Mueller, L. D., Barr, L. G., and Ayala, F. J. (1985). Natural selection vs. random drift: evidence from temporal variation in allele frequencies in nature. *Genetics*, 111(3):517–554.

Nei, M. and Gojobori, T. (1986). Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. *Mol Biol Evol*, 3(5):418–426.

Nei, M. and Kumar, S. (2000). *Molecular evolution and phylogenetics*. Oxford University Press.

Nei, M. and Li, W. H. (1979). Mathematical model for studying genetic variation in terms of restriction endonucleases. *Proc Natl Acad Sci U S A*, 76(10):5269–5273.

Nielsen, R., Bustamante, C., Clark, A. G., Glanowski, S., Sackton, T. B., Hubisz, M. J., Fledel-Alon, A., Tanenbaum, D. M., Civello, D., White, T. J., Sninsky, J. J., Adams, M. D., and Cargill, M. (2005). A scan for positively selected genes in the genomes of humans and chimpanzees. *PLoS Biol*, 3(6):e170.

Nielsen, R. and Yang, Z. (2003). Estimating the distribution of selection coefficients from phylogenetic data with applications to mitochondrial and viral DNA. *Mol Biol Evol*, 20(8):1231–1239.

Nordborg, M., Charlesworth, B., and Charlesworth, D. (1996). The effect of recombination on background selection. *Genet Res*, 67(2):159–174.

Nordborg, M., Hu, T. T., Ishino, Y., Jhaveri, J., Toomajian, C., Zheng, H., Bakker, E., Calabrese, P., Gladstone, J., Goyal, R., Jakobsson, M., Kim, S., Morozov, Y., Padhukasahasram, B., Plagnol, V., Rosenberg, N. A., Shah, C., Wall, J. D., Wang, J., Zhao, K., Kalbfleisch, T., Schulz, V., Kreitman, M., and Bergelson, J. (2005). The pattern of polymorphism in *Arabidopsis thaliana*. *PLoS Biol*, 3(7):e196.

Novembre, J. A. (2002). Accounting for background nucleotide composition when measuring codon usage bias. *Mol Biol Evol*, 19(8):1390–1394.

Obbard, D. J., Welch, J. J., Kim, K.-W., and Jiggins, F. M. (2009). Quantifying adaptive evolution in the Drosophila immune system. *PLoS Genet*, 5(10):e1000698.

O'Hara, R. B. (2005). Comparing the effects of genetic drift and fluctuating selection on genotype frequency changes in the scarlet tiger moth. *Proc Biol Sci*, 272(1559):211–217.

Ohta, T. (1977). Extension to the nearly neutral random drift hypothesis. *Kimura M (ed) Evolution and polymorphism, National Institute of Genetics, Mishima*, pages 148–167.

Olson, M. S., Robertson, A. L., Takebayashi, N., Silim, S., Schroeder, W. R., and Tiffin, P. (2010). Nucleotide diversity and linkage disequilibrium in balsam poplar (Populus balsamifera). *New Phytol*, 186(2):526–536.

Ossowski, S., Schneeberger, K., Lucas-Lledó, J. I., Warthmann, N., Clark, R. M., Shaw, R. G., Weigel, D., and Lynch, M. (2010). The rate and molecular spectrum of spontaneous mutations in Arabidopsis thaliana. *Science*, 327(5961):92–94.

Ota, T. and Nei, M. (1994). Variance and covariances of the numbers of synonymous and nonsynonymous substitutions per site. *Mol Biol Evol*, 11(4):613–619.

Paterson, A. H., Bowers, J. E., Bruggmann, R., Dubchak, I., Grimwood, J., Gundlach, H., Haberer, G., Hellsten, U., Mitros, T., Poliakov, A., Schmutz, J., Spannagl, M., Tang, H., Wang, X., Wicker, T., Bharti, A. K., Chapman, J., Feltus, F. A., Gowik, U., Grigoriev, I. V., Lyons, E., Maher, C. A., Martis, M., Narechania, A., Otillar, R. P., Penning, B. W., Salamov, A. A., Wang, Y., Zhang, L., Carpita, N. C., Freeling, M., Gingle, A. R., Hash, C. T., Keller, B., Klein, P., Kresovich, S., McCann, M. C., Ming, R., Peterson, D. G., ur Rahman, M., Ware, D., Westhoff, P., Mayer, K. F. X., Messing, J., and Rokhsar, D. S. (2009). The *Sorghum bicolor* genome and the diversification of grasses. *Nature*, 457(7229):551–556.

Peck, J. R., Barreau, G., and Heath, S. C. (1997). Imperfect genes, Fisherian mutation and the evolution of sex. *Genetics*, 145(4):1171–1199.

Petit, N. and Barbadilla, A. (2009). Selection efficiency and effective population size in Drosophila species. *J Evol Biol*, 22(3):515–526.

Phifer-Rixey, M., Bonhomme, F., Boursot, P., Churchill, G. A., Piálek, J., Tucker, P. K., and Nachman, M. W. (2012). Adaptive Evolution and Effective Population Size in Wild House Mice. *Mol Biol Evol*.

Piganeau, G. and Eyre-Walker, A. (2009). Evidence for variation in the effective population size of animal mitochondrial DNA. *PLoS One*, 4(2):e4396.

Pluzhnikov, A., Rienzo, A. D., and Hudson, R. R. (2002). Inferences about human demography based on multilocus analyses of noncoding sequences. *Genetics*, 161(3):1209–1218.

Pond, S. K. and Muse, S. V. (2005). Site-to-site variation of synonymous substitution rates. *Mol Biol Evol*, 22(12):2375–2385.

Popadin, K., Polishchuk, L. V., Mamirova, L., Knorre, D., and Gunbin, K. (2007). Accumulation of slightly deleterious mutations in mitochondrial protein-coding genes of large versus small mammals. *Proc Natl Acad Sci U S A*, 104(33):13390–13395.

Presgraves, D. C. (2005). Recombination enhances protein adaptation in Drosophila melanogaster. *Curr Biol*, 15(18):1651–1656.

Pritchard, J. K., Pickrell, J. K., and Coop, G. (2010). The genetics of human adaptation: hard sweeps, soft sweeps, and polygenic adaptation. *Curr Biol*, 20(4):R208–R215.

Pritchard, J. K. and Rienzo, A. D. (2010). Adaptation - not by sweeps alone. *Nat Rev Genet*, 11(10):665–667.

Przeworski, M., Coop, G., and Wall, J. D. (2005). The signature of positive selection on standing genetic variation. *Evolution*, 59(11):2312–2323.

Pröschel, M., Zhang, Z., and Parsch, J. (2006). Widespread adaptive evolution of Drosophila genes with sex-biased expression. *Genetics*, 174(2):893–900.

Ptak, S. E. and Przeworski, M. (2002). Evidence for population growth in humans is confounded by fine-scale population structure. *Trends Genet*, 18(11):559–563.

Riebler, A., Held, L., and Stephan, W. (2008). Bayesian variable selection for detecting adaptive genomic differences among populations. *Genetics*, 178(3):1817–1829.

Rizzon, C., Ponger, L., and Gaut, B. S. (2006). Striking similarities in the genomic distribution of tandemly arrayed genes in *Arabidopsis* and rice. *PLoS Comput Biol*, 2(9):e115.

Roach, J. C., Glusman, G., Smit, A. F. A., Huff, C. D., Hubley, R., Shannon, P. T., Rowen, L., Pant, K. P., Goodman, N., Bamshad, M., Shendure, J., Drmanac, R., Jorde, L. B., Hood, L., and Galas, D. J. (2010). Analysis of genetic inheritance in a family quartet by whole-genome sequencing. *Science*, 328(5978):636–639.

Rockman, M. V., Skrovanek, S. S., and Kruglyak, L. (2010). Selection at linked sites shapes heritable phenotypic variation in C. elegans. *Science*, 330(6002):372–376.

Roselius, K., Stephan, W., and Städler, T. (2005). The relationship of nucleotide polymorphism, recombination rate and selection in wild tomato species. *Genetics*, 171(2):753–763.

Ross-Ibarra, J., Tenaillon, M., and Gaut, B. S. (2009). Historical divergence and gene flow in the genus *Zea*. *Genetics*, 181(4):1399–1413.

Ross-Ibarra, J., Wright, S. I., Foxe, J. P., Kawabe, A., DeRose-Wilson, L., Gos, G., Charlesworth, D., and Gaut, B. S. (2008). Patterns of polymorphism and demographic history in natural populations of *Arabidopsis lyrata*. *PLoS ONE*, 3(6):e2411.

Sabeti, P. C., Reich, D. E., Higgins, J. M., Levine, H. Z. P., Richter, D. J., Schaffner, S. F., Gabriel, S. B., Platko, J. V., Patterson, N. J., McDonald, G. J., Ackerman, H. C., Campbell, S. J., Altshuler, D., Cooper, R., Kwiatkowski, D., Ward, R., and Lander, E. S. (2002). Detecting recent positive selection in the human genome from haplotype structure. *Nature*, 419(6909):832–837.

Sabeti, P. C., Varilly, P., Fry, B., Lohmueller, J., Hostetter, E., Cotsapas, C., Xie, X., Byrne, E. H., McCarroll, S. A., Gaudet, R., Schaffner, S. F., Lander, E. S., Consortium, I. H., Frazer, K. A., Ballinger, D. G., Cox, D. R., Hinds, D. A., Stuve, L. L., Gibbs, R. A., Belmont, J. W., Boudreau, A., Hardenbol, P., Leal, S. M., Pasternak, S., Wheeler, D. A., Willis, T. D., Yu, F., Yang, H., Zeng, C., Gao, Y., Hu, H., Hu, W., Li, C., Lin, W., Liu, S., Pan, H., Tang, X., Wang, J., Wang, W., Yu, J., Zhang, B., Zhang, Q., Zhao, H., Zhao, H., Zhou, J., Gabriel, S. B., Barry, R., Blumenstiel, B., Camargo, A., Defelice, M., Faggart, M., Goyette, M., Gupta, S., Moore, J., Nguyen, H., Onofrio, R. C., Parkin, M., Roy, J., Stahl, E., Winchester, E., Ziaugra, L., Altshuler, D., Shen, Y., Yao, Z., Huang, W., Chu, X., He, Y., Jin, L., Liu, Y., Shen, Y., Sun, W., Wang, H., Wang, Y., Wang, Y., Xiong, X., Xu, L., Waye, M. M. Y., Tsui, S. K. W., Xue, H., Wong, J. T.-F., Galver, L. M., Fan, J.-B., Gunderson, K., Murray, S. S., Oliphant, A. R., Chee, M. S., Montpetit, A., Chagnon, F., Ferretti, V., Leboeuf, M., Olivier, J.-F., Phillips, M. S., Roumy, S., Sallée, C., Verner, A., Hudson, T. J., Kwok, P.-Y., Cai, D., Koboldt, D. C., Miller, R. D., Pawlikowska, L., Taillon-Miller, P., Xiao, M., Tsui, L.-C., Mak, W., Song, Y. Q., Tam, P. K. H., Nakamura, Y., Kawaguchi, T., Kitamoto, T., Morizono, T., Nagashima, A., Ohnishi, Y., Sekine, A., Tanaka, T., Tsunoda, T., Deloukas, P., Bird, C. P., Delgado, M., Dermitzakis, E. T., Gwilliam, R., Hunt, S., Morrison, J., Powell, D., Stranger, B. E., Whittaker, P., Bentley, D. R., Daly, M. J., de Bakker, P. I. W., Barrett, J., Chretien, Y. R., Maller, J., McCarroll, S., Patterson, N., Pe'er, I., Price,

A., Purcell, S., Richter, D. J., Sabeti, P., Saxena, R., Schaffner, S. F., Sham, P. C., Varilly, P., Altshuler, D., Stein, L. D., Krishnan, L., Smith, A. V., Tello-Ruiz, M. K., Thorisson, G. A., Chakravarti, A., Chen, P. E., Cutler, D. J., Kashuk, C. S., Lin, S., Abecasis, G. R., Guan, W., Li, Y., Munro, H. M., Qin, Z. S., Thomas, D. J., McVean, G., Auton, A., Bottolo, L., Cardin, N., Eyheramendy, S., Freeman, C., Marchini, J., Myers, S., Spencer, C., Stephens, M., Donnelly, P., Cardon, L. R., Clarke, G., Evans, D. M., Morris, A. P., Weir, B. S., Tsunoda, T., Johnson, T. A., Mullikin, J. C., Sherry, S. T., Feolo, M., Skol, A., Zhang, H., Zeng, C., Zhao, H., Matsuda, I., Fukushima, Y., Macer, D. R., Suda, E., Rotimi, C. N., Adebamowo, C. A., Ajayi, I., Aniagwu, T., Marshall, P. A., Nkwodimmah, C., Royal, C. D. M., Leppert, M. F., Dixon, M., Peiffer, A., Qiu, R., Kent, A., Kato, K., Niikawa, N., Adewole, I. F., Knoppers, B. M., Foster, M. W., Clayton, E. W., Watkin, J., Gibbs, R. A., Belmont, J. W., Muzny, D., Nazareth, L., Sodergren, E., Weinstock, G. M., Wheeler, D. A., Yakub, I., Gabriel, S. B., Onofrio, R. C., Richter, D. J., Ziaugra, L., Birren, B. W., Daly, M. J., Altshuler, D., Wilson, R. K., Fulton, L. L., Rogers, J., Burton, J., Carter, N. P., Clee, C. M., Griffiths, M., Jones, M. C., McLay, K., Plumb, R. W., Ross, M. T., Sims, S. K., Willey, D. L., Chen, Z., Han, H., Kang, L., Godbout, M., Wallenburg, J. C., L'Archevêque, P., Bellemare, G., Saeki, K., Wang, H., An, D., Fu, H., Li, Q., Wang, Z., Wang, R., Holden, A. L., Brooks, L. D., McEwen, J. E., Guyer, M. S., Wang, V. O., Peterson, J. L., Shi, M., Spiegel, J., Sung, L. M., Zacharia, L. F., Collins, F. S., Kennedy, K., Jamieson, R., and Stewart, J. (2007). Genome-wide detection and characterization of positive selection in human populations. *Nature*, 449(7164):913–918.

Sawyer, S. A., Kulathinal, R. J., Bustamante, C. D., and Hartl, D. L. (2003). Bayesian analysis suggests that most amino acid replacements in Drosophila are driven by positive selection. *J Mol Evol*, 57 Suppl 1:S154–S164.

Schmid, K. J., Ramos-Onsins, S., Ringys-Beckstein, H., Weisshaar, B., and Mitchell-Olds, T. (2005). A multilocus sequence survey in *Arabidopsis thaliana* reveals a genome-wide departure from a neutral model of DNA sequence polymorphism. *Genetics*, 169(3):1601–1615.

Schmid, K. J., Törjék, O., Meyer, R., Schmuths, H., Hoffmann, M. H., and Altmann, T. (2006). Evidence for a large-scale population structure of Arabidopsis thaliana from genome-wide single nucleotide polymorphism markers. *Theor Appl Genet*, 112(6):1104–1114.

Schneider, A., Charlesworth, B., Eyre-Walker, A., and Keightley, P. D. (2011). A method for inferring the rate of occurrence and fitness effects of advantageous mutations. *Genetics*, 189(4):1427–1437.

Sella, G., Petrov, D. A., Przeworski, M., and Andolfatto, P. (2009). Pervasive natural selection in the Drosophila genome? *PLoS Genet*, 5(6):e1000495.

Shapiro, J. A., Huang, W., Zhang, C., Hubisz, M. J., Lu, J., Turissini, D. A., Fang, S., Wang, H.-Y., Hudson, R. R., Nielsen, R., Chen, Z., and Wu, C.-I. (2007). Adaptive genic evolution in the Drosophila genomes. *Proc Natl Acad Sci U S A*, 104(7):2271–2276.

Siepel, A., Bejerano, G., Pedersen, J. S., Hinrichs, A. S., Hou, M., Rosenbloom, K., Clawson, H., Spieth, J., Hillier, L. W., Richards, S., Weinstock, G. M., Wilson, R. K., Gibbs, R. A., Kent, W. J., Miller, W., and Haussler, D. (2005). Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res*, 15(8):1034–1050.

Singer, T., Fan, Y., Chang, H.-S., Zhu, T., Hazen, S. P., and Briggs, S. P. (2006). A high-resolution map of Arabidopsis recombinant inbred lines by whole-genome exon array hybridization. *PLoS Genet*, 2(9):e144.

Siol, M., Wright, S. I., and Barrett, S. C. H. (2010). The population genomics of plant adaptation. *New Phytol*, 188(2):313–332.

Slotte, T., Bataillon, T., Hansen, T. T., Onge, K. S., Wright, S. I., and Schierup, M. H. (2011). Genomic determinants of protein evolution and polymorphism in Arabidopsis. *Genome Biol Evol*, 3:1210–1219.

Slotte, T., Foxe, J. P., Hazzouri, K. M., and Wright, S. I. (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(8):1813–1821.

Smith, J. M. and Haigh, J. (1974). The hitch-hiking effect of a favourable gene. *Genet Res*, 23(1):23–35.

Smith, N. G. C. and Eyre-Walker, A. (2002). Adaptive protein evolution in *Drosophila*. *Nature*, 415(6875):1022–1024.

Sokal, R. and Rohlf, F. (1981). Biometry. The principles and practice of statistics in biological research.

Soltis, P. S. and Soltis, D. E. (2009). The role of hybridization in plant speciation. *Annu Rev Plant Biol*, 60:561–588.

Song, B.-H., Clauss, M. J., Pepper, A., and Mitchell-Olds, T. (2006). Geographic patterns of microsatellite variation in Boechera stricta, a close relative of Arabidopsis. *Mol Ecol*, 15(2):357–369.

Song, B.-H., Windsor, A. J., Schmid, K. J., Ramos-Onsins, S., Schranz, M. E., Heidel, A. J., and Mitchell-Olds, T. (2009). Multilocus patterns of nucleotide diversity, population structure and linkage disequilibrium in *Boechera stricta*, a wild relative of *Arabidopsis*. *Genetics*, 181(3):1021–1033.

Stephan, W. (2010). Genetic hitchhiking versus background selection: the controversy and its implications. *Philos Trans R Soc Lond B Biol Sci*, 365(1544):1245–1253.

Stephan, W. and Li, H. (2007). The recent demographic and adaptive history of Drosophila melanogaster. *Heredity*, 98(2):65–68.

Stephan, W., Song, Y. S., and Langley, C. H. (2006). The hitchhiking effect on linkage disequilibrium between linked neutral loci. *Genetics*, 172(4):2647–2663.

Stevens, P. F. (2010). Angiosperm Phylogeny Website, version 9.

Stoletzki, N. and Eyre-Walker, A. (2011). Estimation of the neutrality index. *Mol Biol Evol*, 28(1):63–70.

Strasburg, J. L., Kane, N. C., Raduski, A. R., Bonin, A., Michelmore, R., and Rieseberg, L. H. (2011a). Effective population size is positively correlated with levels of adaptive divergence among annual sunflowers. *Mol Biol Evol*, 28(5):1569–1580.

Strasburg, J. L., Kane, N. C., Raduski, A. R., Bonin, A., Michelmore, R., and Rieseberg, L. H. (2011b). Effective population size is positively correlated with levels of adaptive divergence among annual sunflowers. *Mol Biol Evol*.

Strasburg, J. L. and Rieseberg, L. H. (2008). Molecular demographic history of the annual sunflowers *Helianthus annuus* and *H. petiolaris*–large effective population sizes and rates of long-term gene flow. *Evolution*, 62(8):1936–1950.

Strasburg, J. L., Scotti-Saintagne, C., Scotti, I., Lai, Z., and Rieseberg, L. H. (2009). Genomic patterns of adaptive divergence between chromosomally differentiated sunflower species. *Mol Biol Evol*, 26(6):1341–1355.

Strobeck, C. (1983). Expected linkage disequilibrium for a neutral locus linked to a chromosomal arrangement. *Genetics*, 103(3):545–555.

Sundström, H., Webster, M. T., and Ellegren, H. (2004). Reduced variation on the chicken Z chromosome. *Genetics*, 167(1):377–385.

Suzuki, Y. (2010). Statistical methods for detecting natural selection from genomic data. *Genes Genet Syst*, 85(6):359–376.

Swigonová, Z., Lai, J., Ma, J., Ramakrishna, W., Llaca, V., Bennetzen, J. L., and Messing, J. (2004). Close split of sorghum and maize genome progenitors. *Genome Res*, 14(10A):1916–1923.

Tajima, F. (1983). Evolutionary relationship of DNA sequences in finite populations. *Genetics*, 105(2):437–460.

Tajima, F. (1989). Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics*, 123(3):585–595.

Takahata, N., Ishii, K., and Matsuda, H. (1975). Effect of temporal fluctuation of selection coefficient on gene frequency in a population. *Proc Natl Acad Sci U S A*, 72(11):4541–4545.

Tang, C., Toomajian, C., Sherman-Broyles, S., Plagnol, V., Guo, Y.-L., Hu, T. T., Clark, R. M., Nasrallah, J. B., Weigel, D., and Nordborg, M. (2007). The evolution of selfing in Arabidopsis thaliana. *Science*, 317(5841):1070–1072.

Tang, H., Bowers, J. E., Wang, X., Ming, R., Alam, M., and Paterson, A. H. (2008). Synteny and collinearity in plant genomes. *Science*, 320(5875):486–488.

Tellier, A., Fischer, I., Merino, C., Xia, H., Camus-Kulandaivelu, L., Städler, T., and Stephan, W. (2011). Fitness effects of derived deleterious mutations in four closely related wild tomato species with spatial structure. *Heredity (Edinb)*, 107(3):189–199.

Tenaillon, M. I., U'Ren, J., Tenaillon, O., and Gaut, B. S. (2004). Selection versus demography: a multilocus investigation of the domestication process in maize. *Mol Biol Evol*, 21(7):1214–1225.

Thompson, J. D., Higgins, D. G., and Gibson, T. J. (1994). CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res*, 22(22):4673–4680.

Tsai, I. J., Bensasson, D., Burt, A., and Koufopanou, V. (2008). Population genomics of the wild yeast Saccharomyces paradoxus: Quantifying the life cycle. *Proc Natl Acad Sci U S A*, 105(12):4957–4962.

Tuskan, G. A., Difazio, S., Jansson, S., Bohlmann, J., Grigoriev, I., Hellsten, U., Putnam, N., Ralph, S., Rombauts, S., Salamov, A., Schein, J., Sterck, L., Aerts, A., Bhalerao, R. R., Bhalerao, R. P., Blaudez, D., Boerjan, W., Brun, A., Brunner, A., Busov, V., Campbell, M., Carlson, J., Chalot, M., Chapman, J., Chen, G.-L., Cooper, D., Coutinho, P. M., Couturier, J., Covert, S., Cronk, Q., Cunningham, R., Davis, J., Degroeve, S., Déjardin, A., Depamphilis, C., Detter, J., Dirks, B., Dubchak, I., Duplessis, S., Ehlting, J., Ellis, B., Gendler, K., Goodstein, D., Gribskov, M., Grimwood, J., Groover, A., Gunter, L., Hamberger, B., Heinze, B., Helariutta, Y., Henrissat, B., Holligan, D., Holt, R., Huang, W., Islam-Faridi, N., Jones, S., Jones-Rhoades, M., Jorgensen, R., Joshi, C., Kangasjärvi, J., Karlsson, J., Kelleher, C., Kirkpatrick, R., Kirst, M., Kohler, A., Kalluri, U., Larimer, F., Leebens-Mack, J., Leplé, J.-C., Locascio, P., Lou, Y., Lucas, S., Martin, F., Montanini, B., Napoli, C., Nelson, D. R., Nelson, C., Nieminen, K., Nilsson, O., Pereda, V., Peter, G., Philippe, R., Pilate, G., Poliakov, A., Razumovskaya, J., Richardson, P., Rinaldi, C., Ritland, K., Rouzé, P., Ryaboy, D., Schmutz, J., Schrader, J., Segerman, B., Shin, H., Siddiqui, A., Sterky, F., Terry, A., Tsai, C.-J., Uberbacher, E., Unneberg, P., Vahala, J., Wall, K., Wessler, S., Yang, G., Yin, T., Douglas, C., Marra, M., Sandberg, G., de Peer, Y. V., and Rokhsar, D. (2006). The genome of black cottonwood, *Populus trichocarpa* (Torr. & Gray). *Science*, 313(5793):1596–1604.

Venton, D. (2012). Highlight–tracking adaptation's role: do larger populations evolve faster? *Genome Biol Evol*, 4(5):668–669.

Wagner, W., Weller, S., and Sakai, A. (2005). Monograph of *Schiedea* (*Caryophyllaceae* subfam. *Alsinoideae*). *Syst Bot Monogr*, 72:1–169.

Wallace, L. E., Weller, S. G., Wagner, W. L., Sakai, A. K., and Nepokroeff, M. (2009). Phylogeographic patterns and demographic history of *Schiedea globosa* (*Caryophyllaceae*) on the Hawaiian Islands. *Am. J. Bot.*, 96(5):958–967.

Watterson, G. A. (1975). On the number of segregating sites in genetical models without recombination. *Theor Popul Biol*, 7(2):256–276.

Welch, J. J. (2006). Estimating the genomewide rate of adaptive protein evolution in *Drosophila*. *Genetics*, 173(2):821–837.

Welch, J. J., Eyre-Walker, A., and Waxman, D. (2008). Divergence and polymorphism under the nearly neutral theory of molecular evolution. *J Mol Evol*, 67(4):418–426.

Whitlock, M. C. (2003). Fixation probability and time in subdivided populations. *Genetics*, 164(2):767–779.

Whitlock, M. C. (2005). Combining probability from independent tests: the weighted Z-method is superior to Fisher's approach. *J Evol Biol*, 18(5):1368–1373.

Williamson, S. H., Hernandez, R., Fledel-Alon, A., Zhu, L., Nielsen, R., and Bustamante, C. D. (2005). Simultaneous inference of selection and population growth from patterns of variation in the human genome. *Proc Natl Acad Sci U S A*, 102(22):7882–7887.

Wilson, D. J., Hernandez, R. D., Andolfatto, P., and Przeworski, M. (2011). A population genetics-phylogenetics approach to inferring natural selection in coding sequences. *PLoS Genet*, 7(12):e1002395.

Woolfit, M. (2009). Effective population size and the rate and pattern of nucleotide substitutions. *Biol Lett*, 5(3):417–420.

Woolfit, M. and Bromham, L. (2003). Increased rates of sequence evolution in endosymbiotic bacteria and fungi with small effective population sizes. *Mol Biol Evol*, 20(9):1545–1555.

Woolfit, M. and Bromham, L. (2005). Population size and molecular evolution on islands. *Proc Biol Sci*, 272(1578):2277–2282.

Woolfit, M. R. Q. (2006). *Effective population size and its effects on molecular evolution*. University of Sussex School of Life Sciences Department of Biology.

Wright, F. (1990). The 'effective number of codons' used in a gene. *Gene*, 87(1):23–29.

Wright, S. (1931). Evolution in Mendelian Populations. *Genetics*, 16(2):97–159.

Wright, S. I., Bi, I. V., Schroeder, S. G., Yamasaki, M., Doebley, J. F., McMullen, M. D., and Gaut, B. S. (2005). The effects of artificial selection on the maize genome. *Science*, 308(5726):1310–1314.

Wright, S. I. and Charlesworth, B. (2004). The HKA test revisited: a maximum-likelihood-ratio test of the standard neutral model. *Genetics*, 168(2):1071–1076.

Wyckoff, G. J., Li, J., and Wu, C.-I. (2002). Molecular evolution of functional genes on the mammalian Y chromosome. *Mol Biol Evol*, 19(9):1633–1636.

Yang and Bielawski (2000). Statistical methods for detecting molecular adaptation. *Trends Ecol Evol*, 15(12):496–503.

Yang, Z. (1997). PAML: a program package for phylogenetic analysis by maximum likelihood. *Comput Appl Biosci*, 13(5):555–556.

Yang, Z. (2006). *Computational Molecular Evolution*. Oxford University Press, USA.

Yang, Z. and Nielsen, R. (1998). Synonymous and nonsynonymous rate variation in nuclear genes of mammals. *J Mol Evol*, 46(4):409–418.

Yu, N., Jensen-Seaman, M. I., Chemnick, L., Ryder, O., and Li, W.-H. (2004). Nucleotide diversity in gorillas. *Genetics*, 166(3):1375–1383.

Zeng, K. and Charlesworth, B. (2009). Estimating selection intensity on synonymous codon usage in a nonequilibrium population. *Genetics*, 183(2):651–62, 1SI–23SI.

Zeng, K. and Charlesworth, B. (2010). Studying patterns of recent evolution at synonymous sites and intronic sites in Drosophila melanogaster. *J Mol Evol*, 70(1):116–128.

Zhang, J., Webb, D. M., and Podlaha, O. (2002). Accelerated protein evolution and origins of human-specific features: Foxp2 as an example. *Genetics*, 162(4):1825–1835.

Zhang, L. and Li, W.-H. (2005). Human SNPs reveal no evidence of frequent positive selection. *Mol Biol Evol*, 22(12):2504–2507.

Appendices

Species	Population	α	C.I.	N_e
Z.mays				
	spp. mays	-0.25	(-0.52,0.00)	94024
B.stricta				
	North	-0.92	(-1.87,-0.25)	16159
	South	-1.75	(-3.05,-0.78)	5726
O.sativa				
	temperate	-1.81	(-5.35,-0.06)	18123
	tropical	-1.13	(-2.37,-0.1)	101455
	indica	-0.16	(-1.99,0.59)	64193
A.lyrata				
	Germany	-0.43	(-1.06,0.02)	86937
	Iceland	-0.17	(-0.69,0.21)	43903
	Russia	-0.17	(-0.84,0.47)	29858
	Sweden	-0.30	(-0.90,0.16)	39759
	USA	-0.65	(-1.67,0.08)	27749
	Canada	-0.51	(-1.29,0.05)	25531
A.thaliana				
	USA			75708
	N.Sweden			68513
	S.Sweden			97815
	C.Europe			101394
	England			77511
P.balsamifera				
	Central	0.02	(-0.19,0.21)	34420
	East	-0.03	(-0.26,0.18)	23360
	North	-0.01	(-0.24,0.18)	28883
S.globosa				
	Maui	0.34	(-0.3,0.68)	25217
	Molokai	0.29	(-0.26,0.64)	32914
	Oahu	-0.05	(-1.01,0.49)	24217

Table A2.1: Estimates of α using the method of Eyre-Walker and Keightley (2009), their confidence intervals (C.I.) and effective population size (N_e) for subpopulations in our datasets.

Species	FWW	CI	FWW 15%	CI
A.lyrata	-0.5751	-0.9413 to -0.304	-0.4283	-0.807 to -0.0988
B.stricta	-1.2958	-2.0134 to -0.6588	-0.9725	-1.8707 to -0.2891
O.rufipogon	-1.2812	-2.3394 to -0.4452	-1.6851	-4.1069 to -0.4352
S.bicolor	-0.6648	-1.6187 to -0.0342	-0.6095	-1.7566 to 0.0588
Z.mays parviglumis	-0.4951	-0.773 to -0.2493	-0.2383	-0.577 to 0.0058
P.tremula	-0.1955	-0.5746 to 0.0974	0.091	-0.2901 to 0.3755
P.balsamifera	-0.1299	-0.3122 to 0.031	0.0336	-0.1297 to 0.1806
H.annuus	-0.3033	-1.3473 to 0.2374	-0.4058	-1.6573 to 0.2124
H.petiolaris	0.4337	-0.0658 to 0.6456	0.405	-0.1193 to 0.6162
S.globosa	0.0648	-1.2603 to 0.4854	0.2526	-1.5753 to 0.6103

Table A2.2: Estimates of α using the method of Fay et al. (2001) (FWW and FWW, removing polymorphisms segregating below 15% in the population) for 10 comparisons. CI, Confidence Intervals

Model a		
	Migration rate	α (C.I.)
	M=0.01	-0.02 (-0.26; 0.3)
	M=0.1	0 (-0.21; 0.25)
	M=1	0.07 (-0.2; 0.32)
	M=10	-0.04 (-0.34; 0.23)
Model b		
	Migration rate	α (C.I.)
	M=0.01	0.21 (0.07; 0.32)
	M=0.1	0.23 (0.05; 0.39)
	M=1	0.05 (-0.15; 0.24)
	M=10	0 (-0.2; 0.19)

Table A2.3: Estimates of α from simulated datasets with varying migration rates.

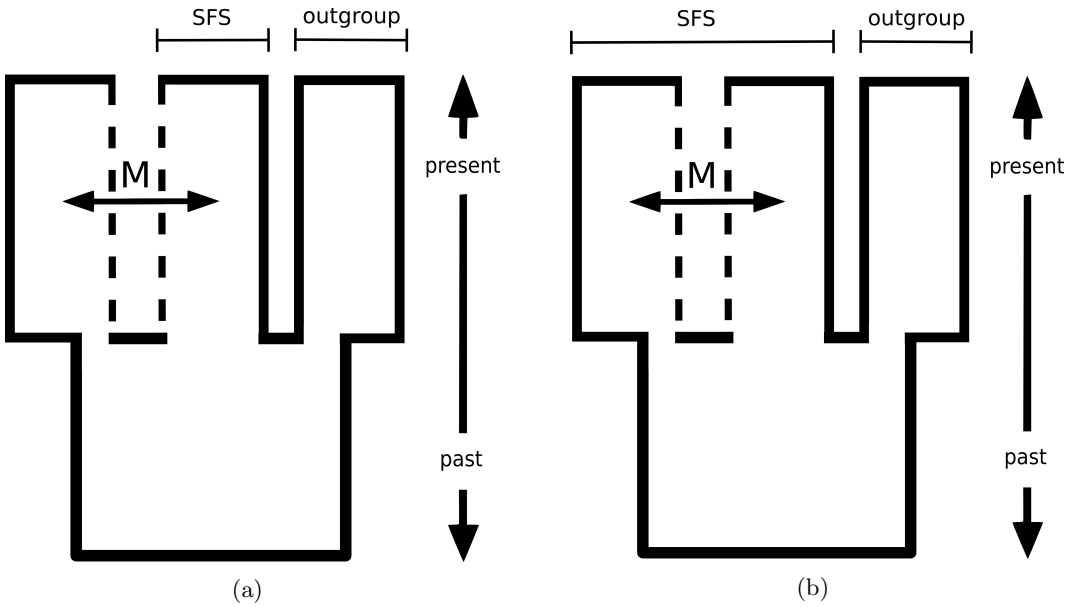


Figure A2.1: Schematic representations of divergence models simulated with SFS_code (Hernandez, 2008) to investigate the impact of population structure on α. Model (a) and model (b) differ only in the populations the site frequency spectrum (SFS) was obtained from.

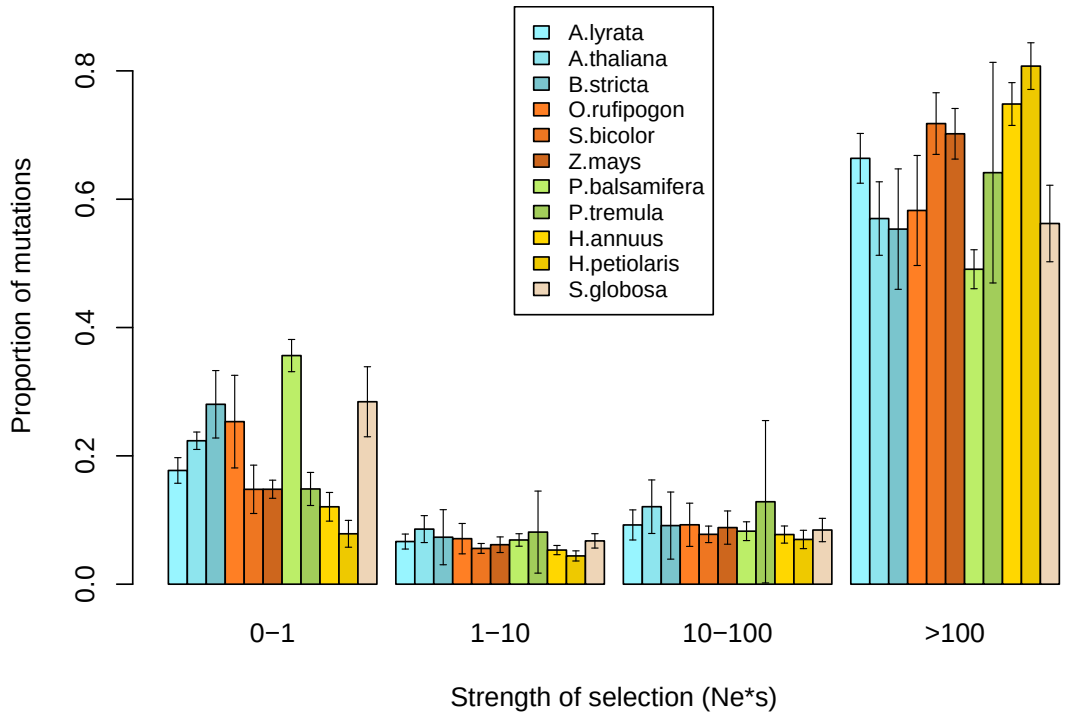


Figure A2.2: The distribution of fitness effects in 11 plant species excluding singletons. Estimates are given for the proportion of mutations in four different $N_e s$ ranges and standard errors. For *Z.mays* the results shown for *Z.mays spp. parviglumis*

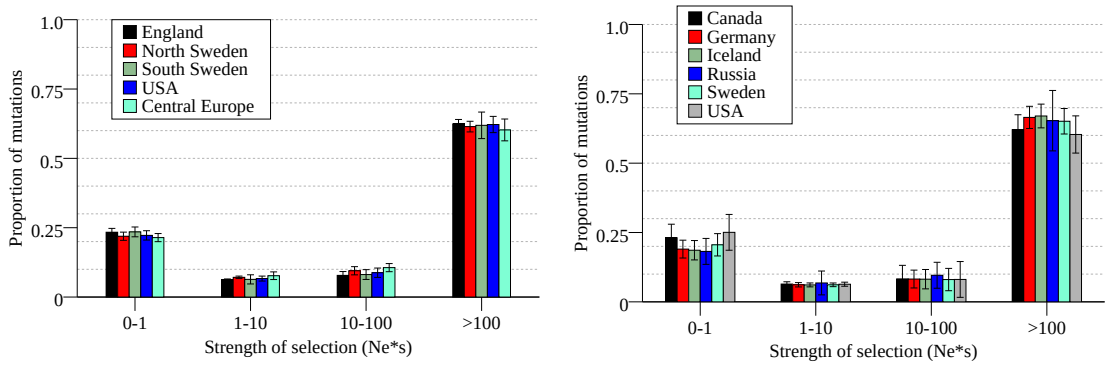


Figure A2.3: Distribution of fitness effects in different populations of *A.thaliana* (a) and *A.lyrata* (b).

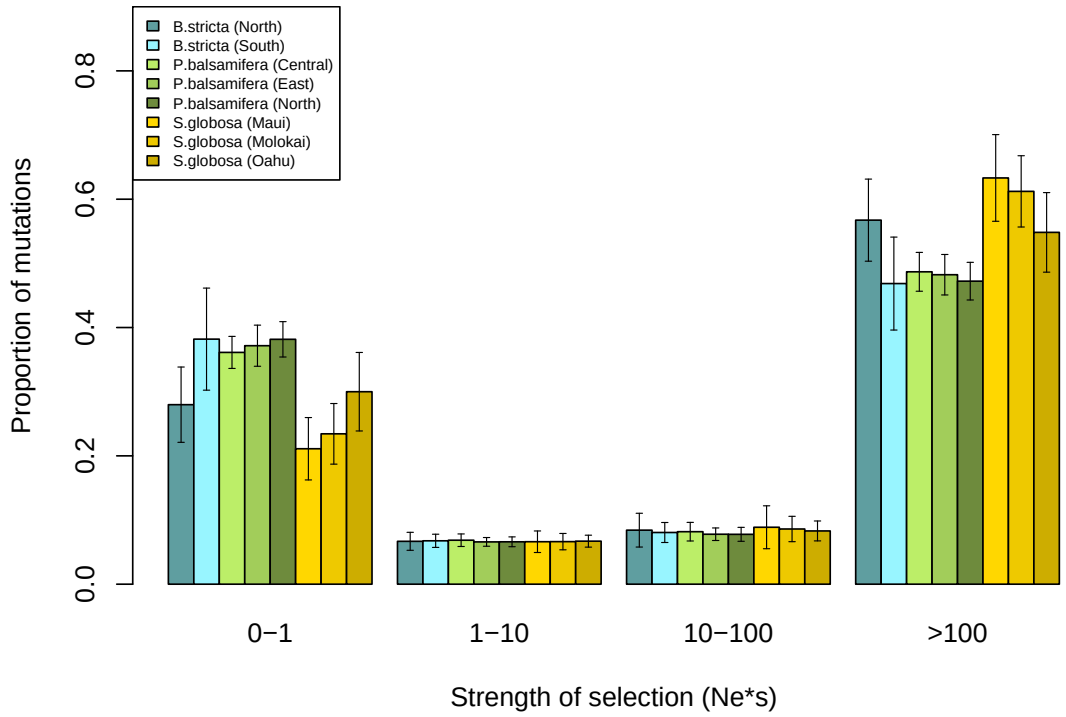


Figure A2.4: The distribution of fitness effects of eight plant subpopulations. Estimates are given for the proportion of mutations in four different N_es ranges and standard errors.

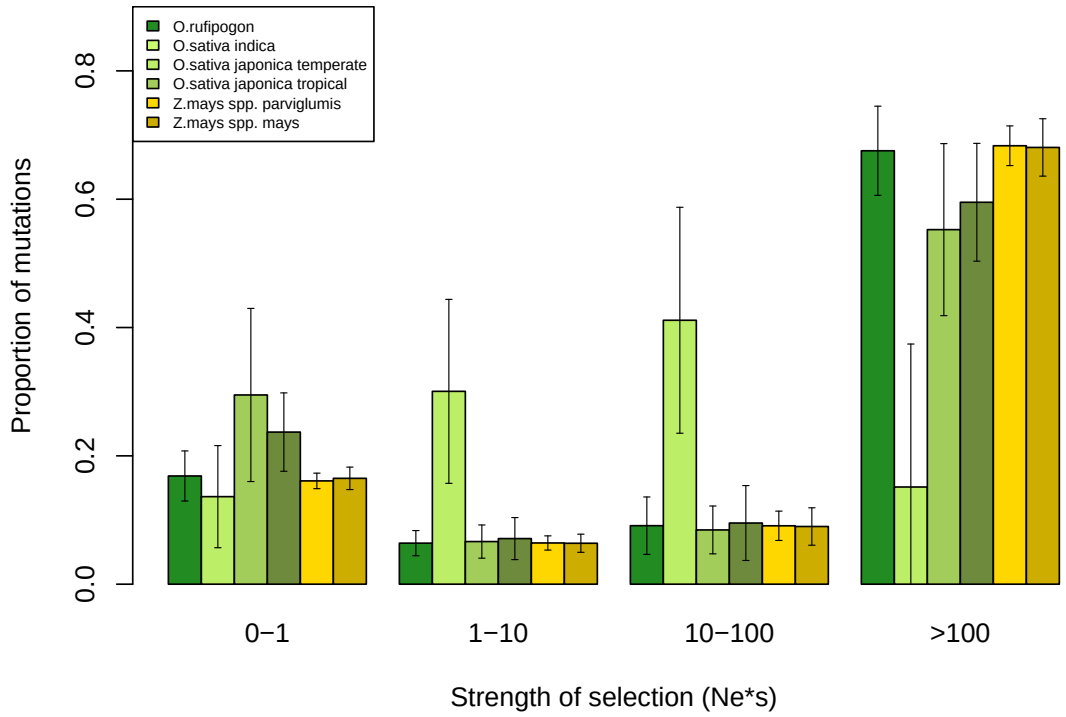


Figure A2.5: The distribution of fitness effects of four domesticated plant subpopulations as well as their two wild relatives. Estimates are given for the proportion of mutations in four different $N_e s$ ranges and standard errors.

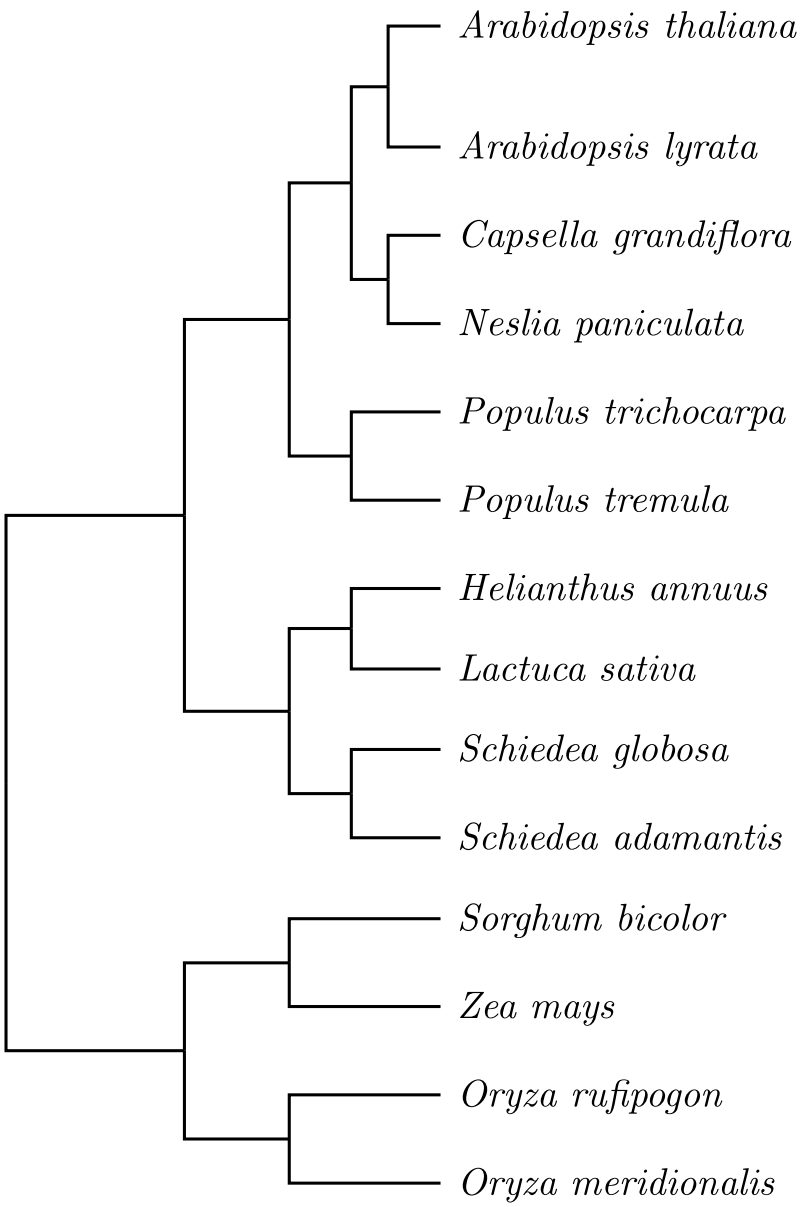


Figure A3.1: **Cladogram of the plant species used for analyses**

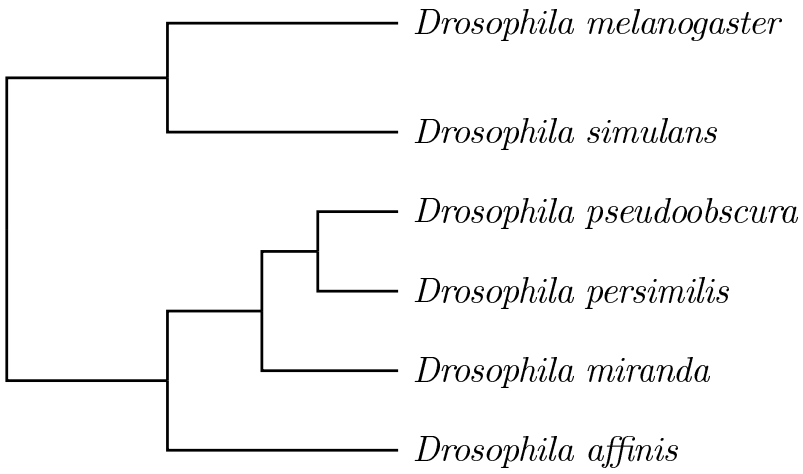


Figure A3.2: Cladogram of the Drosophila species used for analyses

Species	Diversity P-value (χ^2)	Diversity and Divergence P-value (χ^2)
D.melanogaster	$< 1 \times 10^{-3}$	$< 1 \times 10^{-3}$
H.sapiens	$< 1 \times 10^{-3}$	$< 1 \times 10^{-3}$
M.musculus	$< 1 \times 10^{-3}$	0.12^{*}
A.thaliana	$< 1 \times 10^{-3}$	$< 1 \times 10^{-3}$
C.grandiflora	$< 1 \times 10^{-3}$	$< 1 \times 10^{-3}$
S.bicolor	$< 1 \times 10^{-3}$	$< 1 \times 10^{-3}$
B.stricta	3×10^{-3}	0.031
A.lyrata	$< 1 \times 10^{-3}$	$< 1 \times 10^{-3}$
C.rubella	$< 1 \times 10^{-3}$	$< 1 \times 10^{-3}$
S.paradoxus	4×10^{-3}	$< 1 \times 10^{-3}$

* not significant (P>0.05)

Table A5.1: Results of the χ^2 tests of independence assuming the simulated null distribution for diversity and diversity/divergence data. For details see material and methods. P-values are given for each species.

Species	Estimated[1]		Literature[2]		
	β	CI	β	CI	Source
D.melanogaster	0.29	(0.15, 0.43)	0.30	(0.15, 0.49)	Keightley and Eyre-Walker (2007)
H.sapiens	0.12	(0.06, 0.18)	0.18	(0.16, 0.21)	Boyko et al. (2008)
			0.23	(0.19, 0.27)	Eyre-Walker et al. (2006)
M.musculus	0.14	(0.06, 0.22)	0.31		Halligan et al. (2010)
A.thaliana	0.16	(0.08, 0.24)	0.23	(0.11, 0.34)	Slotte et al. (2010)
C.grandiflora	0.27	(0.11, 0.43)	0.30	(0.18, 1.24)	Slotte et al. (2010)
S.bicolor	0.15	(0.11, 0.19)			
B.stricta	0.25	(0.01, 0.49)			
A.lyrata	0.13	(0.09, 0.17)			
C.rubella	0.13	(0.07, 0.19)			
S.paradoxus	0.14	(0.04, 0.24)			

[1] β estimates using the method from Keightley and Eyre-Walker (2007)
[2] β estimates published by others

Table A5.2: Shape parameter β estimates of the distribution of fitness effects (DFE) assuming a gamma distribution for 10 species.

Species	Free recombination				No recombination			
	β_μ	Std	β_{N_e}	Std	β_μ	Std	β_{N_e}	Std
H.sapiens	2.97	0.24	1.63	0.37	2.99	0.25	3.08	1.34
S.bicolor	1.70	0.50	9.55	8.53	1.85	0.54	13.64	8.26
B.stricta	5.57	0.99	18.64	13.08	5.61	1.00	18.81	10.03
D.melanogaster	7.63	0.93	1.89	0.25	7.46	0.91	5.00	1.75
M.musculus	8.30	2.07	23.17	13.50	8.50	2.19	27.22	13.93
C.grandiflora	8.06	1.01	3.67	0.73	8.10	1.00	25.06	10.92
C.rubella	17.63	6.05	0.72	0.28	18.81	6.61	0.96	2.83
A.lyrata	17.30	8.16	1.80	0.69	17.60	8.82	2.36	2.46
A.thaliana	5.81	0.44	1.37	0.14	5.79	0.45	1.94	0.31
S.paradoxus	21.66	4.11	4.84	3.91	21.84	4.09	15.93	8.08

Table A5.3: Estimates of the variation of N_e in 10 eukaryotic species. Results are for an underlying Gamma distribution for N_e and μ assuming either free recombination or no recombination (see materials and methods). For each dataset the mean shape parameter β_{N_e} and β_μ and their standard deviations (Std) are given.

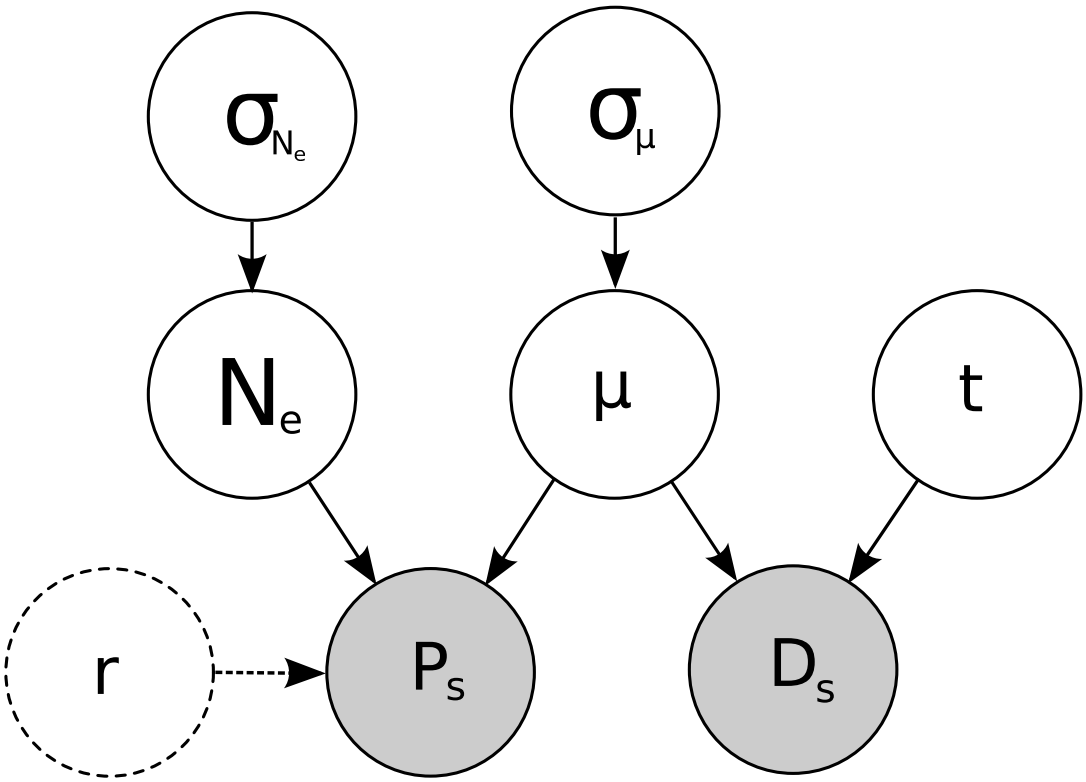


Figure A5.1: A directed acyclic graph to illustrate the depencies of the priors within the Bayesian framework. The dashed circle represents the optional parameter for the model with no recombination within loci.

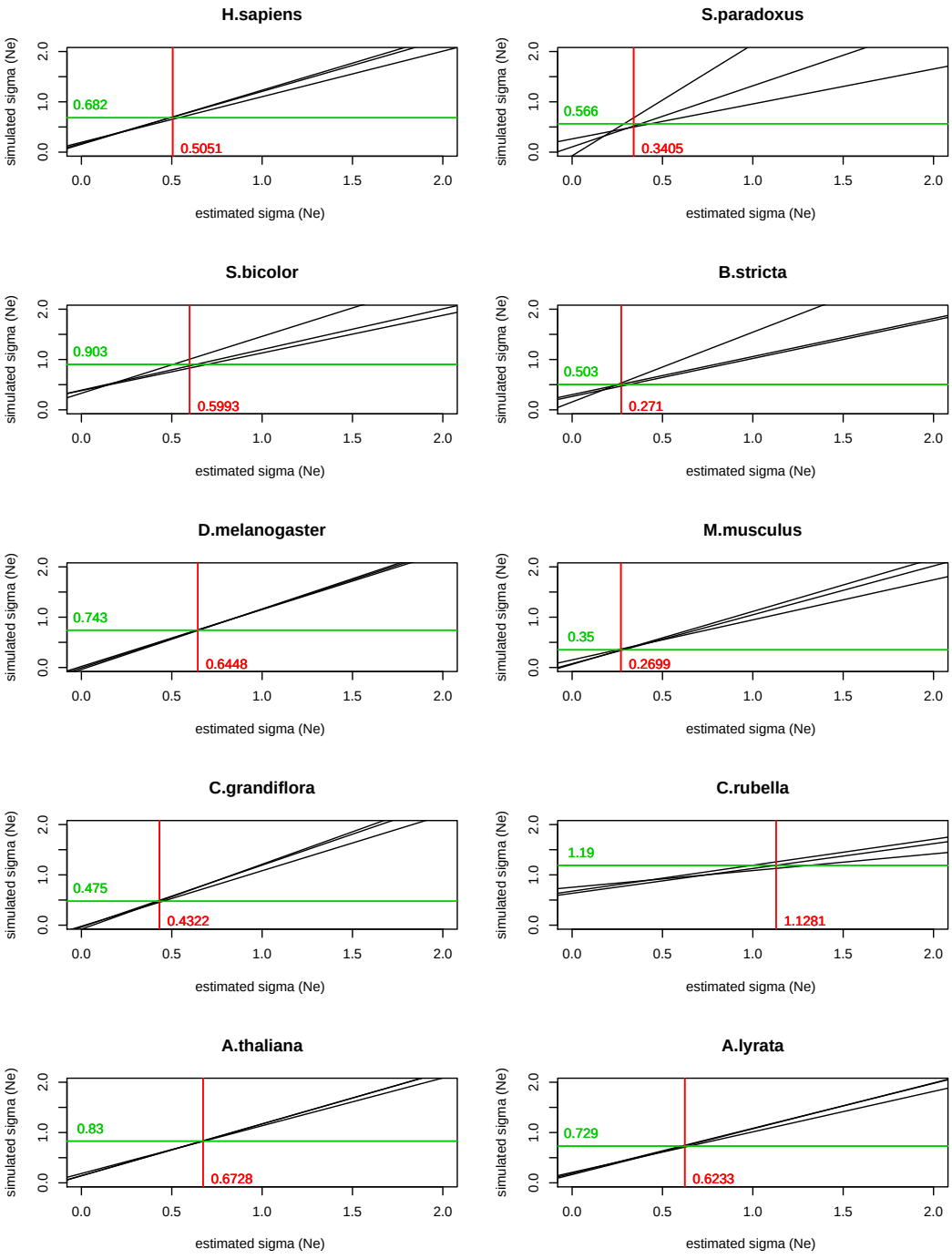


Figure A5.2: Linear Regression analyses for the relationship of σ_{N_e} (estimated) and σ_{N_e} (simulated) for free recombination. We estimated the extent of bias by simulating data under a range of parameter values for σ_{N_e} and σ_{μ} using the actual numbers of sites from the real data such that the expected numbers of polymorphisms and substitutions were equal to the mean values for each species. The obtained value of σ_{N_e} (Table 5.4, indicated by green solid line and value) is the mean estimate for the three regression lines for the initial σ_{N_e} of the Bayesian analysis (indicated by the red solid line and value).

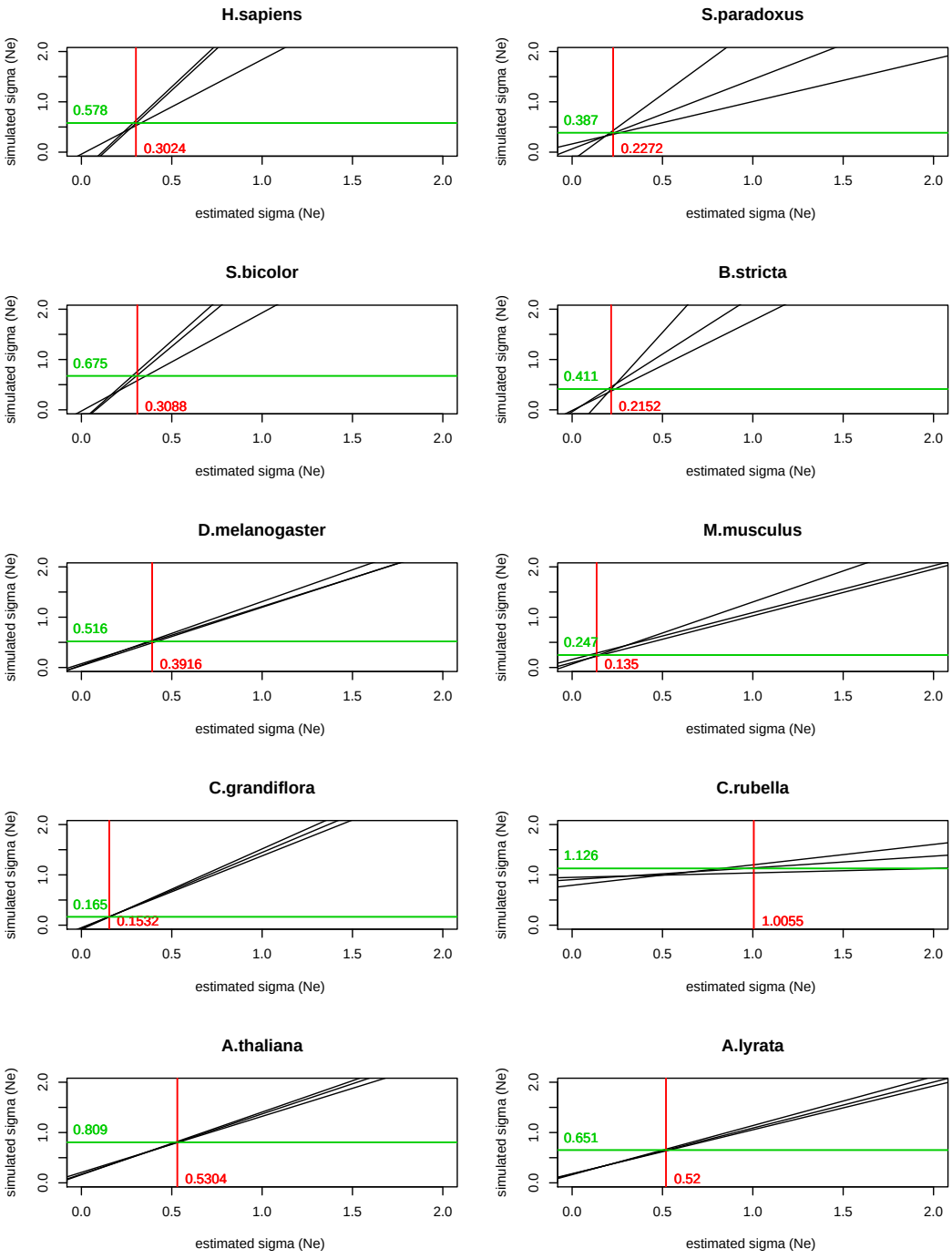


Figure A5.3: Linear Regression analyses for the relationship of σ_{N_e} (estimated) and σ_{N_e} (simulated) for no recombination. We estimated the extent of bias by simulating data under a range of parameter values for σ_{N_e} and σ_μ using the actual numbers of sites from the real data such that the expected numbers of polymorphisms and substitutions were equal to the mean values for each species. The obtained value of σ_{N_e} (Table 5.4, indicated by green solid line and value) is the mean estimate for the three regression lines for the initial σ_{N_e} of the Bayesian analysis (indicated by the red solid line and value).

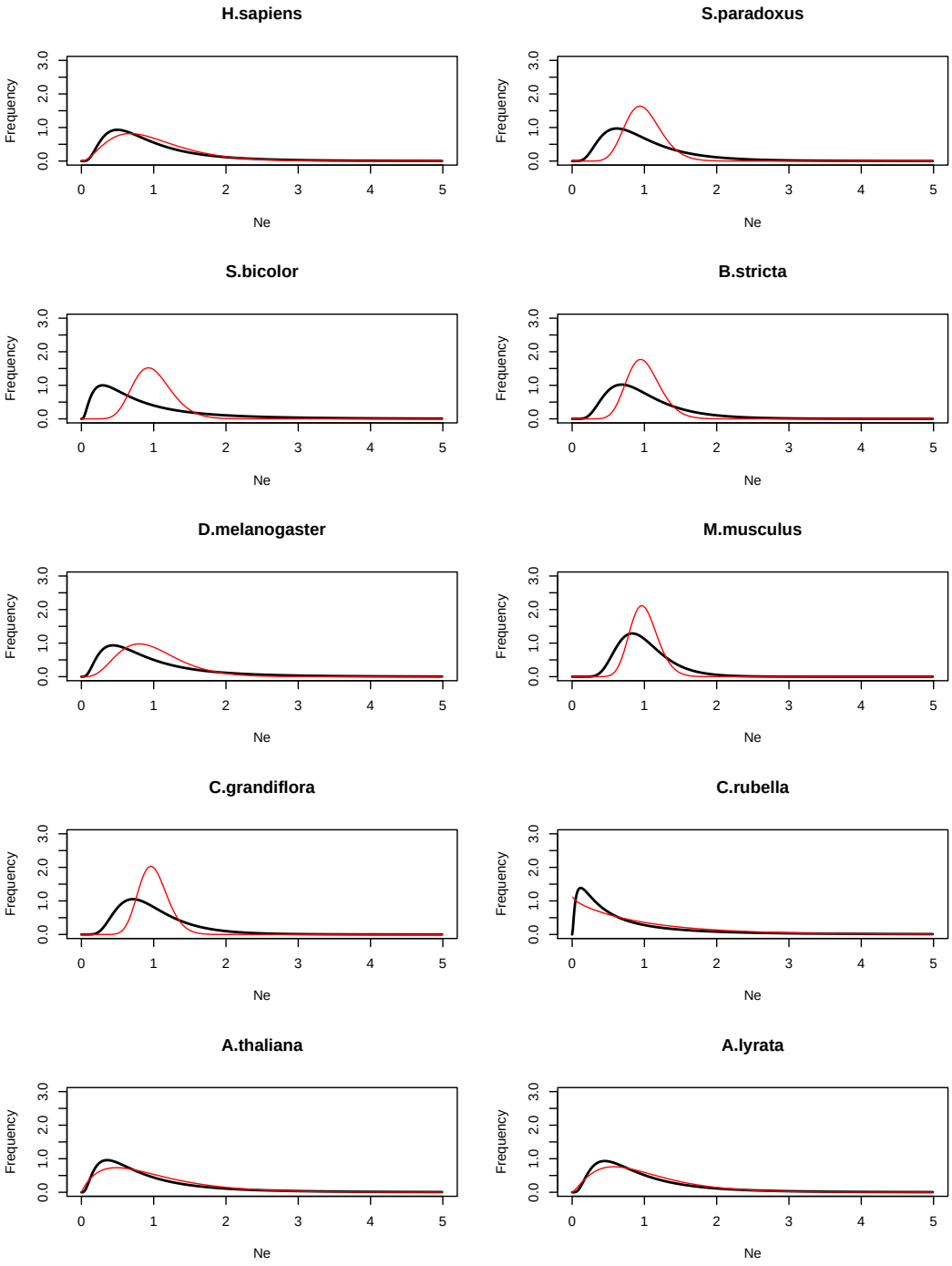


Figure A5.4: Distribution of N_e for 10 species with estimated mean shape parameters from the Bayesian analysis assuming free recombination for two different distributions: σ_{N_e} for log-normal (black line) and β_{N_e} for gamma (red line).

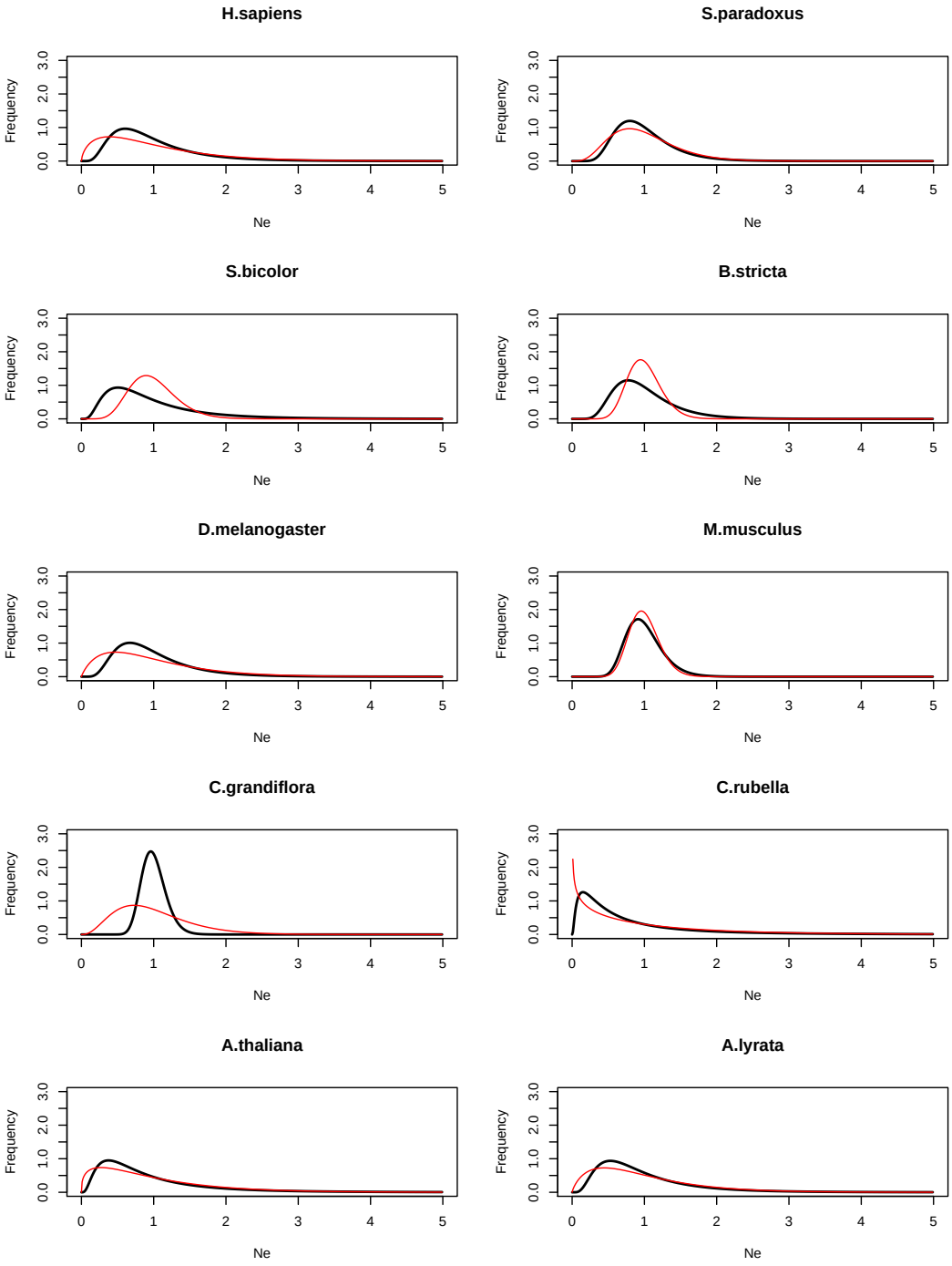


Figure A5.5: Distribution of N_e for 10 species with estimated mean shape parameters from the Bayesian analysis assuming no recombination for two different distributions: σ_{N_e} for log-normal (black line) and β_{N_e} for gamma (red line).