Patterns of natural selection across the human genome

Vivak Soni

Submitted for the degree of Doctor of Philosophy

University of Sussex

August 2021

UNIVERSITY OF SUSSEX

VIVAK SONI

DOCTOR OF PHILOSOPHY

PATTERNS OF NATURAL SELECTION ACROSS THE HUMAN GENOME

Natural selection is one of the key mechanisms by which evolution proceeds. It is the process by which different allelic types become more or less common in successive generations of a population due to differential genotypic responses to the environment. In this thesis I investigate genome wide patterns of natural selection in hominids.

I investigated the prevalence of balancing selection in the human genome using a novel method based on the McDonald-Kreitman (MK) test framework. Having shown that this test is robust to demographic change and that it can also give a direct estimate of the number of shared polymorphisms that are directly maintained by balancing selection, I applied this method to population genomic data from humans, finding that more than a thousand non-synonymous polymorphisms are subject to balancing selection.

It has been shown that the rate of adaptive evolution can be affected by numerous factors at the gene level and the site level. I correlated the rates of adaptive (ω_a) and non-adaptive (ω_{na}) evolution with four gene-level factors: recombination rate, gene age, gene length, and gene expression. For each factor I controlled for the other three factors in turn, finding a significant positive correlation between recombination rate and rates of adaptive and non-adaptive evolution.

I also investigated the correlation between the rates of adaptive and non-adaptive evolution and four site-level factors: relative solvent accessibility, amino acid volume difference, amino acid polarity difference and a measure of evolutionary dissimilarity, p_N/p_S, finding similar correlations to those found previously in *Drosophilids*, except in the case of p_N/p_S, where the slope of the relationship is significantly lower in hominids. This can be explained by contracting population size along the human and chimpanzee lineages. The statistic p_N/p_S is strongly correlated to the mean strength of selection acting against deleterious mutations, and this is expected to attenuate the relationship between the rate of adaptive evolution and p_N/p_S as we observe.

Effective population size (*Ne*) is an important quantity in determining the effectiveness of selection. It can vary not only between species but also across genomes. I investigate patterns of diversity across the human genome. Neutral diversity is expected to be a function of the mutation rate, effective population size and mean genealogy length. Surprisingly I find that the variation in diversity is less than the variation in the mutation rate, inferred from de novo mutation data. This suggests that the effective population size of a genomic region is negatively related to the mutation rate. I fit models and find that the effects of linked selection must be strong to explain the observed data.

Declaration

I declare that this thesis has been composed by myself and that the work has not been submitted for any other degree or professional qualification.

I confirm that the work submitted is my own, except where work which has formed part of jointly-authored publications has been included. Where relevant, co-authors have been listed in the preface section.

Vivak Soni, 2021

"To do science is to be a social actor engaged, whether one likes it or not, in political activity. The denial of the interpenetration of the scientific and the social is itself a political act, giving support to social structures that hide behind scientific objectivity to perpetuate dependency, exploitation, racism, elitism, colonialism."

Richard Levins and Richard Lewontin, *The Dialectical Biologist*.

This doctoral thesis is dedicated to my father, Harjindar Soni, who believed in my potential as a scientist before I ever did.

Acknowledgements

First of all, I would like to thank my PhD supervisor, Adam Eyre-Walker. Not only is he an outstanding population geneticist, but his constant patience and openness to ideas meant that my PhD experience was never an isolating one. Crucially, his emphasis on communicating complex ideas in accessible ways is perhaps the single most important lesson I have learned as a scientist.

I would also like to thank my secondary supervisor Maria Clara Castellanos. Despite working in a very different field she was always supportive and provided useful constructive criticism at our annual thesis committee meetings.

On the personal side of things, I would like to thank my partner Katherine, and our two dogs Mina and Leti. Wherever I am, I'm always happier when they are there with me. Together they constitute home.

I would like to thank my parents Harjindar and Damni for their constant and unrelenting support through my many years in academia. I would also like to thank my soon to be mother-in-law Louise, who housed us briefly and has always supported me, despite having no idea what I actually do.

Preface

The research presented here was carried out at the University of Sussex. Parts of this thesis have been submitted as pre-prints to Biorxiv. Details as follows:

Chapter 2

Soni, V., & Eyre-Walker, A. (2021). Factors that affect the rates of adaptive and non-adaptive evolution at the gene level in humans and chimpanzees. *BioRxiv*, 2021.05.05.442740. https://doi.org/10.1101/2021.05.05.442740

Chapter 3

Soni, V., Moutinho, A. F., & Eyre-Walker, A. (2021). Site level factors that affect the rate of adaptive evolution in humans and chimpanzees; the effect of contracting population size. *BioRxiv*, 2021.05.28.446098. https://doi.org/10.1101/2021.05.28.446098

Chapter 4

Soni, V., Vos, M., & Eyre-Walker, A. (2021). A new test suggests that balancing selection maintains hundreds of non-synonymous polymorphisms in the human genome. *BioRxiv*, 2021.02.08.430226. https://doi.org/10.1101/2021.02.08.430226

Contents

List of abbreviations

BGC	Biased gene conversion
DFE	Distribution of fitness effects
GO	Gene Ontology
MK test	McDonald-Kreitman test
N_e	Effective population size
RR	Recombination rate
RSA	Relative solvent accessibility
SFS	site frequency spectrum

1. General Introduction

1.1 Evolution at the molecular level

The field of population genetics arose out of the mathematical frameworks developed by Ronald Fisher, Sewell Wright and John Haldane in the early 20th century. The principal concern of the field is the study of genetic variation within and between populations; its origin; frequency; phenotypic significance; and distribution in space and time. In the following sections I will provide an overview the mechanisms that cause changes in allele frequencies over time: genetic drift, natural selection, gene flow and mutation.

1.1.1 Genetic drift

Genetic drift is the random sampling of individuals contributing to the next generation. In finite populations such random sampling can result in deviations from the previous generation's gamete frequencies. Under the Wright-Fisher model (Fisher, 1922; Wright, 1931) the probability that a neutral allele (i.e. in the absence of natural selection) goes to fixation is $\frac{i}{2N}$;

where i is the current frequency of the allele in the population, and *N* is the population size. Consequently, the probability that this allele is lost from the population is $1 - \frac{i}{2N}$. Across generations, the expected allele frequencies remain constant. The variance in allele frequency from generation to generation of allele p is $p(1-p)/N$, It is evident that the variance in allele frequency is larger in smaller populations (i.e. drift is much more effective in smaller populations than larger ones). Figure 1.1 shows results of Wright-Fisher simulations of genetic drift run in. As population size increases the distribution of allele frequencies becomes less noisy as drift becomes less effective at taking alleles to extreme frequencies, and ultimately to fixation or loss.

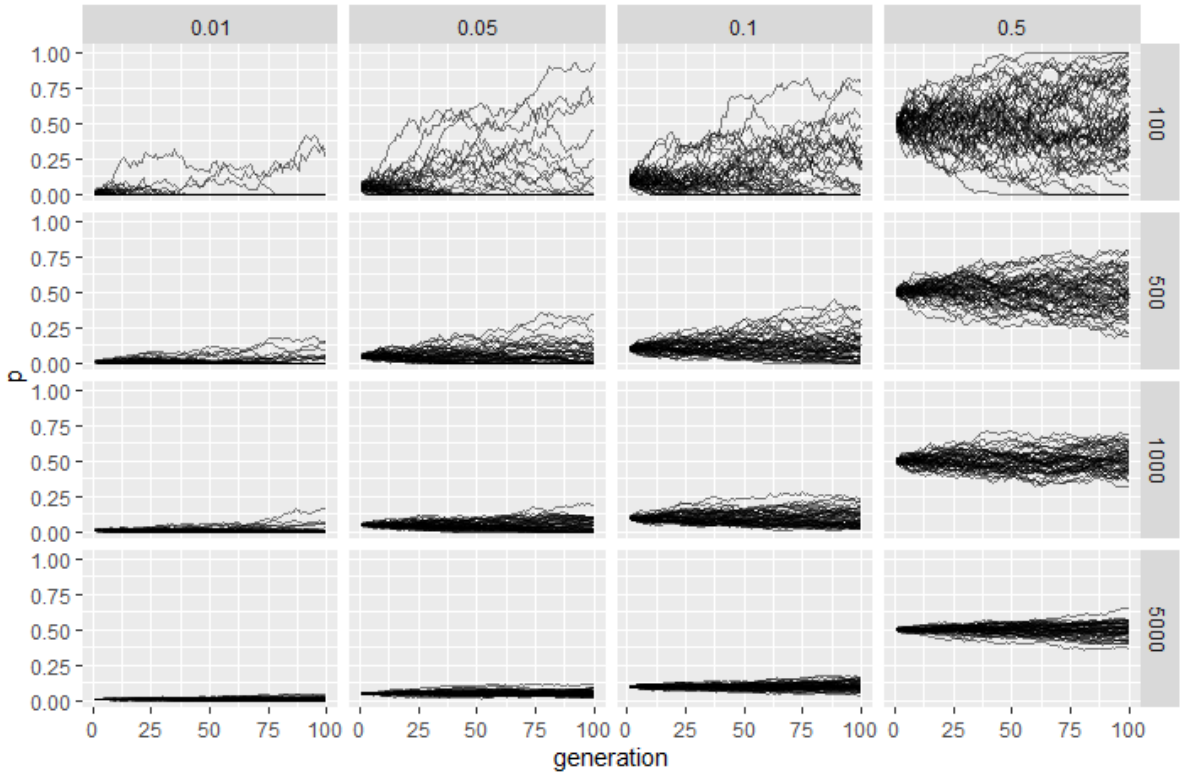


Figure 1.1: Simulations of genetic drift under the Wright-Fisher model, simulated for 100 generations (x-axis), with 50 replicates each. y-axis is the frequency of allele, p. Each column is a different starting allele frequency; each row is a different population size. Figure created in R (R Core Team, 2021), using GGPlot2 (Wickham, 2016).

There are numerous assumptions made by the Wright-Fisher model, including a constant population size, discrete and non-overlapping generations, panmixia and an equal sex ratio. Because every individual is equally likely to reproduce, the number of offspring is binomially distributed and this therefore determines the variance in allele frequency. In reality all natural populations will break many of these assumptions. For example, human populations undergo recombination and generations tend to overlap. Violating these assumptions has a significant effect on the evolutionary impact of genetic drift. So far we have discussed population size in terms of N, the census population size. The effective population size (N_e) is the size of an ideal population that has the same strength of genetic drift as the real, nonideal population. Drift is more effective if $N_e < N$ because the population is actually smaller than under the Wright-Fisher model. Where $N_e \neq N$ we can use the coalescent to determine N_e.

Where random loss of lineages forward in time is described by the process of genetic drift, the coalescent process describes the backward in time "coalescing" of genetic lineages. The history of a sample of size n comprises $n-1$ coalescent events. A coalescent event occurs when two genetic lineages fuse into a common ancestral lineage. The single lineage remaining at the final coalescent event is the most recent common ancestor of the sample (Kingman, 1982). The probability that two lineages coalesced t generations ago is,

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

Handily Ne can be used in place of N in equation 1.1 (Charlesworth, 2009), and derive the expected time to coalescence for two lineages as,

$$E(t) = \sum_{t=1}^{\infty} t \left(1 - \frac{1}{2N_e}\right)^{t-1} \frac{1}{2N_e} = 2N_e \tag{1.2}$$

The mean time to coalescence for two randomly selected neutral alleles is therefore $2N_e$ The total number of generations separating two alleles is $2t$, or $4N_e$ (because the expectation of t is $2N_e$). The probability that two alleles chosen at random differ is therefore given by,

$$\theta = 4N_e\mu \tag{1.3}$$

Where μ is the per site per generation mutation rate (Kimura, 1983). Finally, since there are $2N$ copies of any single mutational site in the gene pool, the total input of neutral mutations per generation is $2N\mu$. Because the probability of fixation of a neutral mutation is $\frac{1}{2N}$ the overall rate of neutral evolution is,

$$Rate\ of\ neutral\ evolution = 2N\mu \times \frac{1}{2N} = \mu \tag{1.4}$$

Although we have shown that the strength of drift is inversely proportional to population size, equation 1.4 shows that genetic drift is an important evolutionary force in all populations, regardless of their size.

1.1.2 Natural Selection

Although the theory of natural selection is attributed to Charles Darwin and Alfred Russel Wallace (Darwin and Wallace, 1858), several other scholars predating Darwin and Wallace published similar ideas, going as far back as Al-Jahiz in the 9th century (Zirkle, 1941). Natural selection is the biological driver of adaptation; the process by which species evolve towards the optima of their environment. The fitness of an organism indicates how close to the optimum it is. The visualisation of fitness as a multi-dimensional landscape was first conceived of by Sewall Wright (1932) and fitness (or adaptive) landscapes have become a cornerstone of the study of adaptation in population and quantitative genetics.

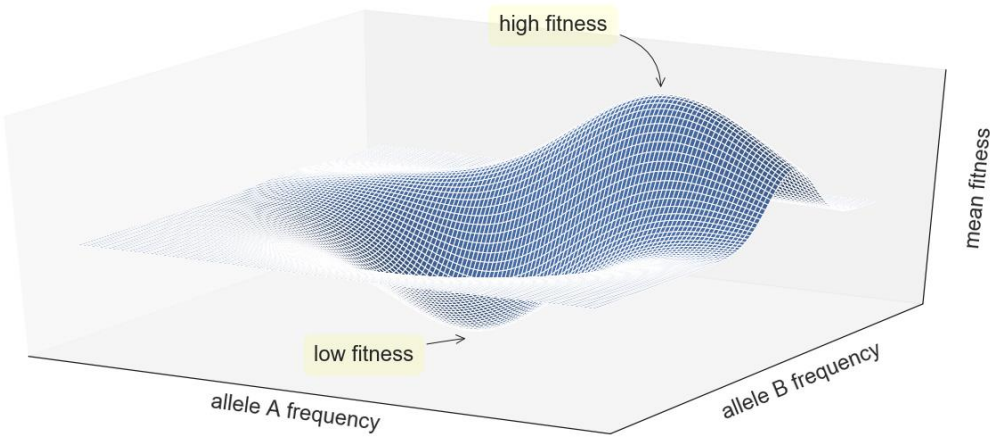


Figure 1.2: An example of an adaptive landscape. The mean fitness of the population is determined by the frequencies of allele A and allele B within the population. Figure created in R (R Core Team, 2021), using GGPlot2 (Wickham, 2016).

Figure 1.2 shows a simple 2-loci adaptive landscape model, where the mean fitness of a population is determined by the frequencies of alleles A and B within it. The area of high fitness is called a peak and represents the fitness optimum. Although figure 1.2 is particularly simplistic, in reality a fitness landscape will be multi-dimensional, with numerous alleles affecting the mean fitness within the population.

At the genotype level, the fitness of a genotype is determined by its selection coefficient, s, relative to the two other possible genotypes:

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

The dominance coefficient, h, determines the fitness of the heterozygous genotype. If $h > 1$ the heterozygous genotype (Aa) is fitter than either homozygous genotype (AA or aa). This is known as heterozygous advantage. Complete dominance occurs when $h = 1$; the dominant

allele completely masks the effect of the recessive allele. This is in contrast to incomplete dominance ($0 < h < 1$), where the dominant allele fails to completely mask the recessive allele, and an additive or blending of both allelic effects occurs. Finally $h = 0$ results in complete recessivity, where the A allele is completely recessive, and masked by the a allele.

It is worth briefly discussing the limits of adaptationism. Appearing in cultural theory as functionalism (Levins and Lewontin, 1977), and in evolutionary biology as an explainer and source of hypothesis generation, adaptationism conceives of the existence of certain problems to be solved by organisms (in the biological case). Any process by which species evolve towards the optima of their environment implies that there is a pre-existent form, problem or ideal to which the adaptation is taking place, and fails to account for the complex relationship between organism and environment (Bergelson et al. 2021); and between chance, contingency and necessity (Xie et al. 2021). With the application of methods such as approximate Bayesian computation to jointly estimate the effects of demographic history and the strength of selection (Johri et al. 2020), there is potential to develop null models that are able to capture some of this complexity. These methods are discussed in more detail in section 6.3.2.

Although there are many forms of natural selection, including directional, balancing, stabilising, diversifying, and purifying selection, in this thesis I consider both directional and balancing selection, and so will pay particular focus to both here.

1.1.3 Mutation and the distribution of fitness effects

The only source of new genetic information (Hodgkinson and Eyre-Walker, 2011), mutations 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). The fate of a particular mutation depends on the evolutionary forces acting on it. A mutation which is contributing to the genetic material of the offspring starts segregating in a population. Its fate (whether it fixes in the population; is lost; or continues to segregate as a polymorphism) is dependent on numerous evolutionary processes including those mentioned at the start of this section - genetic drift, selection, and gene flow.

Mutations can be broadly classified as harmful, beneficial or neutral, though it is more realistic to assume a distribution of fitness effects (Kimura, 1983; Gillespie, 1991), called a DFE. The DFE has been harnessed to gain insight into 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). DNA sequence data can be used to infer characteristics of the DFE by fitting a distribution of selective effects to the site frequency spectrum (which is the distribution of allele frequencies) (Eyre-Walker et al., 2006; Keightley and Eyre-Walker, 2007; Boyko et al., 2008; Schneider et al., 2011). Two sets of sites are considered, one which is assumed to be under selection and one which is assumed to be effectively neutral (commonly introns or synonymous sites are used). This latter category is used to estimate the mutation rate and control for the impact of demography.

There are two notable assumptions of such methods. It is unlikely that the DFE is truly captured by a relatively simple distribution (e.g. a gamma distribution is used to model the DFE in humans) (Eyre-Walker and Keightley, 2007). Secondly, free recombination is assumed. In the latter case, this assumption has been shown to hold in all scenarios except those in which linkage is very strong (Boyko et al., 2008; Eyre-Walker and Keightley, 2009; Keightley and Eyre-Walker, 2010). Finally, DFE inference tends to be limited to the genome-wide level. On rare

occasions it may be possible to infer the DFE for individual genes if they are extremely large, with a substantial number of sequenced individuals (Keightley and Eyre Walker, 2010).

1.1.4 The neutral theory of molecular evolution

Until the late 1960s almost all evolutionary changes were attributed to directional natural selection. Fisher had famously rejected any significant evolutionary role for genetic drift in the 1930s (Fisher, 1930). It was in this context that Motoo Kimura published his paper estimating the substitution rate of mutations occurring in protein-coding genes (Kimura, 1968) in humans as 1.8 nucleotide substitutions per year. Previously Haldane had estimated the substitution rate as 1.5 nucleotide substitutions per year (Haldane, 1957), which exceeded his own estimated upper limit (a problem known as "Haldane's dilemma"). Resolving this paradox had been the initial motivation for Kimura, and he posited that this excess of nucleotide substitutions was due to genetic drift. A year later King and Jukes had independently reached a similar conclusion (King and Jukes, 1969).

The neutral theory claimed that the observed variation within and between species was driven by the random fixation of selectively neutral mutations (Kimura, 1983). The majority of mutations are either neutral or strongly deleterious (and therefore efficiently removed from the population by selection). Positive selection therefore makes a negligible contribution to between-species divergence. The major extension of the neutral theory came in 1973 when Tomoko Ohta incorporated slightly deleterious mutations ($s \approx \frac{1}{2N_e}$) (Ohta, 1974). Ohta emphasised the role of population size in what is known as the nearly neutral theory, as drift is more effective in smaller populations.

Though there has been some debate in the period since the neutral theory was first published (including the recent reigniting of the selectionist vs neutralist debate (see Kern and Hahn, 2018 and Jensen et al. 2019), the neutral theory has become the central framework for generating null hypotheses. To demonstrate that a sequence is subject to selection, it must be shown that this sequence has not evolved neutrally, which forms the null hypothesis. In the next section I discuss the most commonly used tests that use the predictions of the neutral theory as their null hypothesis.

1.2 Detecting positive selection

Methods to detect selection can broadly be split into two categories – outlier and aggregate methods. Outlier methods compute statistics across a genomic region, and are commonly used to detect selective sweeps (see section 1.2.1). By contrast aggregate methods combine data from multiple sites, leveraging the additional power of a large number of loci. The trade-off of this increased power is that it is not possible to identify specific targets of selection with aggregate methods. Although I will discuss some outlier methods developed to detect balancing selection further on in this introduction, throughout this thesis I have used aggregate methods to detect selection.

1.2.1 Selective sweeps

The name selective sweep refers to the reduction in diversity that accompanies the increase in frequency of an advantageous mutation. Selective sweeps leave a complex spatial signature along the genome (see figure 1.3) that can be leveraged to develop novel neutrality tests, or improve the power of those that already exist. For example, using an explicit model of a

selective sweep, Kim and Stephan's method (Kim and Stephan, 2002) calculates the expected frequency spectrum for a site as a function of its distance from a beneficial mutation. It is then possible to estimate the location and strength of a selective sweep by fitting data to this model. Methods such as this allow researchers to identify genomic regions containing putative targets of selection by conducting genome wide scans.

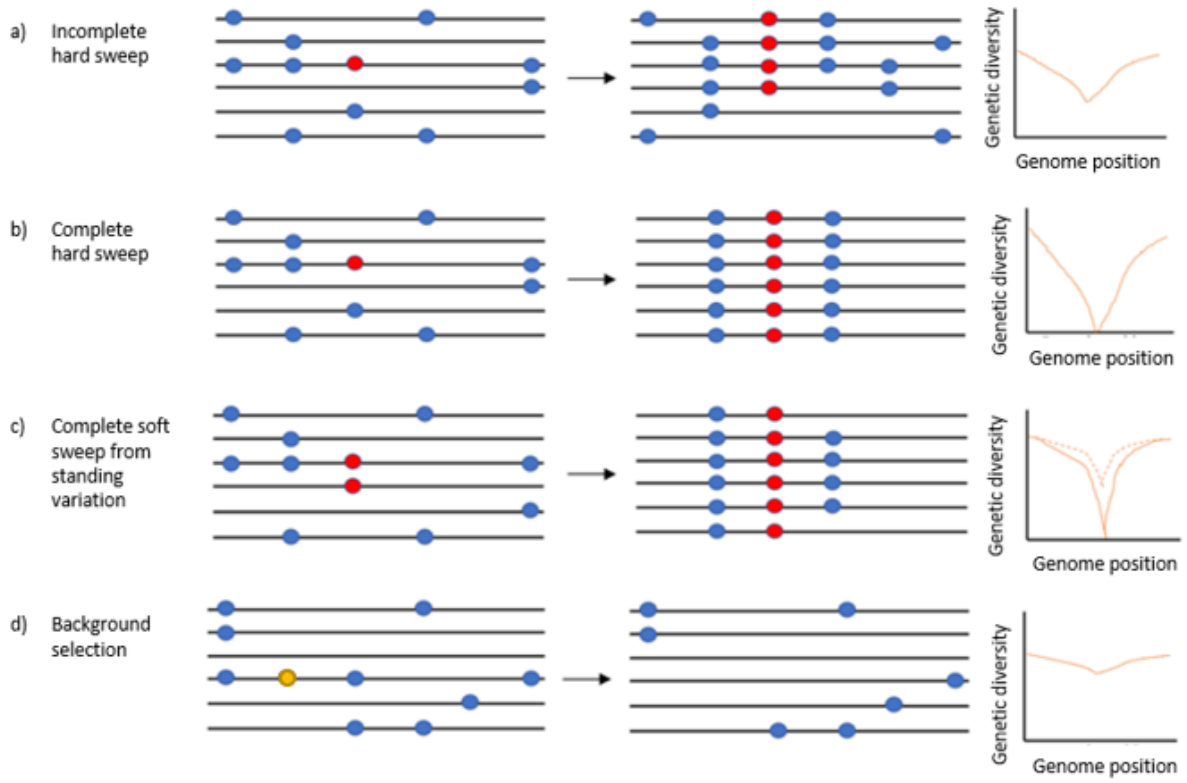


Figure 1.3: Selective sweeps and their effects on genetic diversity. a) and b) The classic model of a hard sweep. A new beneficial mutation (red) enters the population and rapidly increases in frequency. Eventually it fixes in the population (b). In doing so it drags some of the surrounding neutral genetic variation that is in linkage with it (i.e. its haplotype) to fixation, thereby significantly reducing neutral diversity around it. This phenomenon is known as genetic hitchhiking (Maynard-Smith and Haigh, 1974). c) If an allele that already exists in the population suddenly becomes beneficial it can spread through the population and fix, along with surrounding neutral variation. This is known as a soft sweep from

standing genetic variation. The reduction in genetic diversity around the selected site is dependent on the history of the causative variant (compare the dotted line and the filled line that show differing dips in diversity). d) Natural selection purges deleterious alleles (yellow), and in the process neutral variation linked to the deleterious alleles will also be purged, thereby reducing genetic diversity. If the removed variants are slightly deleterious, the site frequency spectrum is skewed towards rare alleles (Charlesworth et al. 1993). This is known as background selection. Blue circles represent neutral alleles, whilst red circles represent beneficial alleles.

Figure 1.3 shows how genetic diversity around the selected locus is affected by selection. As a positively selected mutation increases in frequency within a population, linked neutral variation increases in frequency too, reducing diversity at loci around the selected site. The severity of this decrease in genetic diversity is dependent on several factors including the ratio of the strength of positive selection to the recombination rate, and the nature of the sweep itself. If a positively selected mutation occurs in a population and sweeps to rapid fixation (figure 1.3a), the reduction in diversity is severe as linked neutral variation sweeps to fixation with it. However, if a sweep occurs from standing genetic variation (i.e. if a neutral or deleterious allele becomes positively selected for), the dip in diversity will be dependent on the frequency of the allele in the population. If it is at an intermediate frequency for example, the advantageous mutation has already recombined onto multiple backgrounds, and therefore linked variation is unlikely to experience the increase in frequency witnessed in a hard sweep. In genomic regions with restricted recombination and recurrent hard sweeps levels of diversity are expected to be lower because linkage maintains association between selected sites and surrounding neutral variation without being broken up by recombination. Several studies have shown that variation in Drosophila genomes is lower in regions of low recombination (Aguade et al. 1989; Stephan and Langley, 1989; Miyashita, 1990; Berry et al. 1991; Begun and Aquadro,

1991; Begun and Aquadro, 1992; Martin-Campos et al. 1992; Stephan and Mitchell, 1992; Langley et al. 1993), which could either be caused by genetic hitchhiking (Maynard-Smith and Haigh, 1974) - the process by which positive selection will reduce neutral variation linked to the selective locus – or by background selection (see section 1.2.2). However it is important to acknowledge that low recombination regions can produce an upward bias on detecting selection because of the increase in variance in most statistics in these regions. Selective sweeps also leave a much stronger signal in regions of low recombination, meaning that the statistical power of tests is a function of the recombination rate (Nielsen, 2005).

The dip in diversity caused by a selective sweep can also be caused by demographic change, and it can be challenging to distinguish selection from demographic history by solely looking at diversity. For example, a population bottleneck causes an increase in the variance of the levels of diversity within a population, making it more difficult to detect regional dips caused by selective sweeps.

1.2.2 Background selection

Background selection (figure 1.3d) has a qualitatively similar effect to hitchhiking, in that it reduces local diversity (Charlesworth et al. 1993) and skews the site frequency spectrum towards rare variants (Braverman et al. 1995) via negative selection against deleterious mutations. Initial work by Charlesworth et al. (1993) showed that diversity in a nonrecombining genomic regions will be reduced as a function of the proportion of copies of the region that contain deleterious mutations. Further work (Hudson and Kaplan, 1995; Nordberg et al. 1996) incorporated recombination rates into the equations derived by Charlesworth et al. (1993), quantifying how low recombining regions are affected by BGS. Recent studies show that BGS is a major factor affecting variation in nucleotide diversity across

large genomic windows (>100Kbp) in both Drosophila melanogaster (Comeron, 2014) and in humans (McVicker, 2009), to the extent that some (e.g. Comeron, 2014) have argued that BGS is the null model for detecting other modes of selection. The necessity of developing an appropriate null model that accounts for genetic drift (as modulated by a population's demographic history) and the distribution of fitness effects of direct and indirect purifying selection will be discussed at length in the discussion section of this thesis.

1.2.3 Evidence of decreases in diversity in humans

Selective sweeps are rare within humans. Hernandez et al. (2011) examined resequencing data from 179 human genomes for evidence of selective sweeps. Although they found that diversity decreases near exons and conserved non-coding regions, the dip in diversity around human-specific amino acid substitutions is no more pronounced than around synonymous substitutions. They also found that amino acid and putative regulatory sites are not significantly enriched in highly differentiated alleles between populations, relative to the genome background. These results were recapitulated in full by Fu and Akey (2013). Although these observations imply that selective sweeps have been rare in recent human history (over the past ~250,000 years), several other studies concluded that sweeps have been common. Williamson et al. (2007) identified 101 regions within the human genome with very strong evidence of selective sweeps, with as much as 10% of the genome affected by linkage to a selective sweep. In his review of 21 genome-wide scans for recent or ongoing positive selection in humans, Akey (2009) found that although ~14% of the genome (containing ~23% of genes) was identified as being under positive selection in at least one study, only 20% of those regions were identified in multiple studies, suggesting a high false positive rate. It is important to note that these signatures of selective sweeps are equally consistent with

background selection. There is currently no clear consensus as to which force is more important in the evolution of humans.

1.2.4 The McDonald-Kreitman test and its variants

Distinguishing between synonymous and non-synonymous substitutions in protein coding sequences (Li et al. 1985; Nei and Gojobori, 1986) forms the basis of many aggregate statistics, the simplest of which is ω, the ratio of non-synonymous to synonymous substitutions (d_N/d_S). It is important to note the assumption that synonymous mutations are neutral (see section 6.2.1 for further discussion of this assumption). In the absence of selection, $d_N = d_S$. Under purifying selection deleterious mutations are eliminated before they can go to fixation, and therefore $d_N < d_S$. It is only under positive directional selection, where non-synonymous mutations are favoured that $d_N > d_S$, indicating a higher rate of fixation in non-synonymous than synonymous substitutions. It is necessary for implementations of the d_N/d_S test to account for mutational bias (where certain mutations are more probable than others - e.g. in many species transition mutations occur more frequently than would be expected under neutrality (Stoltzfus and Norris, 2016) and multiple substitutions (which are more likely to occur with greater divergence times)). Because most sites are constrained during most of their evolution, it is unlikely that $d_N > d_S$ across entire protein coding regions within genes. It is therefore common to target only specific sites.

The d_N/d_S test only accounts for rates of substitution, but the neutral theory hypothesises that both the divergence between species (substitution) and the diversity within a species (polymorphism) are driven primarily by random genetic drift (Kimura, 1983). Thus, the proportions of non-synonymous to synonymous polymorphism (P_n/P_s) within a species are equal to the proportion of non-synonymous to synonymous substitutions (D_n/D_s) between

species under neutrality. This forms the null model of the McDonald-Kreitman (MK) test (1991), which compares variation at putatively neutral sites with variation at potentially selected sites. An elevation of D_n/D_s over P_n/P_s indicates an excess of fixed differences and is taken as evidence of positive directional selection (because advantageous mutations are expected to rapidly sweep through a population and become fixed differences between populations). A derivative of the MK test estimates the proportion of nucleotide substitutions at a class of sites that are driven by positive selection as (Charlesworth, 1994; Smith and Eyre-Walker, 2002),

$$\alpha = 1 - \frac{\mathrm{d}_S p_N}{d_N p_S}$$

(1.5)

Where p_N / p_S is the ratio of diversity at putatively functional (i.e. non-synonymous) sites and putatively neutral (i.e. synonymous) sites. α is therefore the proportion of substitutions fixed by natural selection. Estimates of α can be biased for several reasons. Slightly deleterious mutations can segregate in a population before being eliminated (Ohta, 1973), thereby leading to an underestimate of α by inflating p_N / p_S. A simple method for minimising the impact of slightly deleterious mutations is to exclude polymorphisms below a certain threshold frequency (Fay et al. 2001; Charlesworth and Eyre-Walker, 2008) or to calculate α for different frequency bins (Messer and Petrov, 2013). More sophisticated methods infer the DFE of deleterious (Fay et al. 2001; Smith and Eyre-Walker, 2002; Bierne and Eyre-Walker, 2004; Eyre-Walker et al. 2006; Keightley and Eyre-Walker, 2007; Eyre-Walker and Keightley, 2009; Stoletzki and Eyre-Walker, 2011) and beneficial (Galtier, 2016; Tataru et al. 2017) mutations from polymorphism data. By fitting a DFE to the SFS observed in a population sample, these methods explicitly model the contribution of deleterious mutations to polymorphism and divergence at specific frequency categories. The inferred DFE is used to predict the numbers of

substitutions originating from neutral and slightly deleterious mutations between the two species. If there is a greater number of observed substitutions than expected, the difference can be attributed to advantageous substitutions, yielding an estimate of α.

As table 1.1. shows, estimates of α vary greatly between different datasets, with the number of loci sampled and the choice of outgroup species contributing to this variation. Notably the highest estimate of α by Fay et al. (2001) used different genes for divergence than for polymorphism data. A clearer picture has started to emerge as larger datasets have become available. The largest study by Boyko et al. (2008) estimated that between 10 and 20% of protein coding loci are adaptively evolving.

Outgroup species	Locus type	Number of loci	α (% of adaptive substitutions)	Reference
Mouse	Protein	330	0	Zhang, 2005
Old-world monkey	Protein	149	0	Zhang and Li, 2005
	Protein	182/106[a]	35	Fay et al. 2001
Chimpanzee	Protein	13,500	0–9	Chimpanzee Sequencing and Analysis Consortium, 2005
	Protein	289	20	Zhang and Li, 2005
	5' flank	305	0.11[b]	Keightley et al. 2005
	3' flank	305	0.14[b]	Keightley et al. 2005
	Protein	4916	6	Bustamante et al. 2005
	Protein	47,576	10-20	Boyko et al. 2008
	Non-coding	255	0.11-0.14	Eyre-Walker and Keightley, 2009
	Protein	47,576	0.13	Messer and Petrov, 2013
	Protein	50,543	0.25	Zhen et al. 2021

Table 1.1: Estimates of α in humans. Where authors provided confidence intervals, α is given as a range. Otherwise only the point estimate is provided. [a] Numbers of genes differ for divergence (182) and polymorphism (106). [b] Authors split region into two, and calculated the average of the estimates given for the 1–500 base pair region and the 501–1000 region.

1.2.5 Rates of adaptive and non-adaptive evolution

Along with α, the other major statistic used to infer the rate of adaptive evolution is ω_a. ω_a is the rate of adaptive non-synonymous substitutions relative to the mutation rate and is given

by $\omega_a = \omega - \omega_{na}$ where ω_{na} denotes the portion of the ω ratio contributed by neutral and deleterious mutations. ω_{na} is referred to as the rate of non-adaptive evolution, and is a measure of negative constraint – the lower the rate of non-adaptive evolution, the more constraint there is on the locus in question. Because α is the proportion of adaptive amino-acid substitutions and is estimated as ω_a/ω, it is contingent on both ω_a and ω_{na}, making it unsuitable for disentangling the effects of positive and negative selection. Conversley, ω_a is normalised by the mutation rate (e.g. Castellano et al. 2016), and therefore cannot be used to evaluate the impact of mutation rate. The most appropriate statistic is therefore dependent on the question that is being addressed. In chapters 3 and 4, we seek to understand which factors affect the rates of adaptive and non-adaptive evolution, and therefore the most suitable statistics to estimate are ω_a and ω_{na}.

1.2.6 Balancing selection

How genetic variation is maintained, either in the form of DNA sequence diversity or quantitative genetic variation, remains one of the central problems of population genetics and the role that balancing selection plays in this process remains unknown. Balancing selection encapsulates several selective mechanisms that increase variability within a population. These include heterozygote advantage (also referred to as overdominance), frequency dependent selection, and spatio-temporal variability (Nielsen, 2005). These are expected to leave some similar signatures in genomic data that are detectable at different timescales, and upon which the various statistical tests for detection of the process are built. A balanced polymorphism originates in an ancestral population in one of two ways; either from a new mutation or from standing genetic variation due to a change in selection pressures. The rarer of the two alleles will increase to some intermediate frequency before the ancestral population divides to yield two or more descendent populations. If the balanced polymorphism originates from a de novo

mutation, then it will cause a partial selective sweep. In doing so it may potentially drag several neutral synonymous polymorphisms to high frequency in linkage disequilibrium (LD) (Fijarczyk and Babik, 2015). The pattern of increased homozygosity around the balanced locus (along with the elevated haplotype frequency) can be detected by linkage-based methods such as the Extended Haplotype Homozygosity (EHH) method (Sabeti et al, 2002). This method looks for alleles with unusually long-range LD when accounting for population frequency (because a partial selective sweep causes a rise in allele frequency that is rapid enough that recombination is not able to break down the haplotype on which selection occurs). Other methods have built on this logic, including the integrated haplotype score (iHS) (Voight et al. 2006). A major confounder however is that partial selective sweeps can indicate both positive directional selection and balancing selection, with the signatures being indistinguishable. This signal is only detectable for very recent balancing selection (up to 0.04Ne generations old (Fijarczyk and Babik, 2015), as recombination will eventually break up the long range associations generated via linkage.

A less transient signal is that of an excess of common polymorphism that builds up subsequent to the partial sweep that occurs as a result of balancing selection. The partial sweep skews the SFS from the expected L shape under neutrality towards an excess of alleles at intermediate frequencies. There are several methods that quantitatively measure this divergence from neutrality, including Fu and Li's F and D statistics (Fu and Li, 1993), and Tajima's D, which calculates the difference between the mean number of pairwise differences and the number of segregating sites (Tajima, 1989). Whilst this signal is maintained long enough to detect balancing selection that is older than $0.04N_e$ generations, these methods are extremely sensitive to demography, and so it is essential to either use a demographic model when forming a null hypothesis, or use an outlier approach (because the whole genome will be

affected). The signal of excess common polymorphism decays as the distance from the balanced polymorphism increases. DeGiorgio et al. (2014) introduced the composite likelihood ratio test, T_2, that models the effect of balancing selection on the genealogy at neutral loci that are linked to the target locus. The T_2 test utilises the allele frequency spectrum to calculate the conditional probability of observing a specific number of ancestral alleles within a specific region, accounting for the distance from the target site. Using simulations, the authors show that these tests have greater power than Tajima's D for detecting balancing selection under a range of demographic scenarios. Bitarello et al. (2018) use correlated allele frequencies (and therefore the aforementioned skew in the SFS) as the basis for their non-central deviation statistic, which measures the extent to which the local SFS deviates from expectations under balancing selection.

With older balancing selection the increased diversity around a selected locus is only distinguishable over a narrow genomic region, because recombination breaks down associations between the selected site and surrounding variation that it is linked to. A widely used test for this signature is the HKA test (Hudson et al. 1987). According to the neutral theory of molecular evolution, the within-species diversity is correlated with between-species divergence (Kimura, 1983). The HKA test compares the fit of polymorphism and divergence data against this null hypothesis. As with other methods that compare levels of polymorphism and substitution, the HKA test is susceptible to demography which must be accounted for when forming a null model. The T_1 test (sister to the T_2 test mentioned above) also uses this signature of increased diversity. This method estimates the composite likelihood that a site is under balancing selection given the distribution of polymorphisms around the target (DeGiorgio et al. 2014).

After a long enough divergence time between the two species, all shared neutral genetic variation will either have gone to fixation or been lost. Any remaining shared polymorphisms are being maintained by balancing selection. This is a signature of the oldest balancing selection, aka ancient balancing selection. Asthana et al. (2005) found a low incidence of ancestral polymorphism shared between humans and chimpanzees. The authors found eight SNPs that occur in the same genomic position in humans and chimpanzees with the same sequence changes (e.g. A to C in both genomes). Of these eight, only three shared polymorphisms occurred at non-synonymous sites. However, none of these sites show the signature of common polymorphism in the regions surrounding them. Four of these eight polymorphisms occurred at highly mutable CpG sites. Hodgkinson et al. (2009) have shown that the excess of coincident SNPs between humans and chimpanzees is due to recurrent mutation. To rule out ancestral polymorphism as a cause of this excess they looked for SNPs shared between humans and Macaques. Due to the much longer divergence times between humans and macaques (species that diverged 23-24 Mya compared to 6-10 Mya between humans and chimpanzees), the expectation is that very few polymorphisms will be shared between the two species. However, the authors identified a significant excess of shared SNPs, suggesting inheritance is not the cause of this excess of coincident SNPs. Subsequent work by Johnson and Hellmann (2011) found that the SFS for coincident SNPs is skewed towards rare variants; if most of these were ancestral polymorphisms they would have a relatively uniform SFS.

More recently, Leffler et al. (2013) identified multiple regions in which the same haplotypes were segregating in both humans and chimpanzees, reasoning that if a

polymorphism has been maintained since the ancestral split, a short ancestral segment should be preserved around the selected site. To filter out cases of recurrent mutation, they focused on cases with a minimum of two shared SNPs within 4 kilobases and in significant LD in both humans and chimpanzees. In six cases they found ancestral polymorphism shared between the two species.

The drawback of approaches that focus on the signature of trans-species polymorphism is that all shared neutral genetic variation that is not in linkage with the balanced polymorphism must have gone to fixation in at least one of the two populations. This makes the test weak because balancing selection must persist for a long enough time that all shared neutral polymorphisms are either lost or fixed in at least one of the two populations. In chapter 2 I demonstrate a simple solution to this problem using neutral genetic variation to inform us as to what to expect under neutrality.

1.3 Effective population size (N_e)

As mentioned in section 1.1.1, N_e has become one of the fundamental quantities in population genetics, determining the level of neutral genetic diversity and the efficacy of natural selection in a given population. The product of N_e and the mutation rate per generation is an estimate of the expected neutral diversity in a population (Kimura, 1991), whilst the product of N_e and the strength of selection, *s*, of a mutation determines the effectiveness of selection.

1.3.1 The effect of N_e on the rate of adaptive evolution

Population genetic theory predicts that α should be correlated to N_e. When $N_e = N$ (i.e. the effective population size is equal to the census population size), the probability of fixation of a new mutation is approximately,

$$2s/(1 - e^{-4Ns}) \tag{1.6}$$

where s is the strength of selection. From equation 1.6 we can see that as N_e increases so too does the probability to fixation, because the proportion of effectively neutral mutations increases. For an advantageous mutation in which $Ns >> 1$, the probability of fixation in equation 1.6 becomes approximately $2s$.

Mutation is the other relevant force here. Since there are $2N$ copies of any single mutational site in the gene pool, the total input of mutations per generation is $2N\mu$ (as discussed in section 1.1.1). If we then multiply the probability of fixation by the population mutation rate, which is the rate at which beneficial mutations occur, the rate of adaptation is then $4N\mu s$ per generation. Crucially, it is important to note that the effective population size is influencing the rate of adaptation in two ways – by affecting the proportion of mutations on which selection is effective, but also by affecting the population mutation rate.

Previous studies have suggested that the proportion of adaptive substitutions is correlated to the effective population size. Species with high effective population sizes, including *Drosophila*, house mice, bacteria, and some plant species show patterns of widespread adaptive amino acid substitution (Bustamante et al. 2002; Smith and Eyre-Walker, 2002; Sawyer et al. 2003; Bierne and Eyre-Walker, 2004; Charlesworth and Eyre-Walker, 2006; Haddrill et al. 2010;

Ingvarsson, 2010; Slotte et al. 2010; Strasburg et al. 2011; Moutinho et al. 2019), whilst several studies have found little evidence of adaptive substitution in species with small effective population sizes such as hominids (Chimpanzee Sequencing and Analysis Consortium, 2005; Zhang and Li, 2005; Boyko et al. 2008; Eyre-Walker and Keightley, 2009; Gossmann et al. 2010). There are notable exceptions, however. Despite being thought to have a larger effective population size, *Drosophila simulans* appears not to have undergone more adaptive evolution than *Drosophila melanogaster* (Andolfatto et al. 2011), and the yeast *Saccharomyces paradoxus* shows little evidence of adaptive evolution, despite having a presumably large N_e (Liti et al. 2009; Gossmann et al. 2012).

There is also evidence that the positive correlation between α and N_e can be explained by the variation in the number of effectively neutral substitutions, because the proportion of effectively neutral mutations is negatively correlated to N_e across numerous species (Popadin, 2007; Piganeau, 2009), and α is dependent on both the rates of effectively neutral and advantageous substitution because $\alpha = \frac{D_{adaptive}}{D_{adaptive}+D_{nonadaptive}}$, where $D_{adaptive}$ and $D_{nonadaptive}$ are the rates of adaptive and nonadaptive substitutions respectively. The methods used to estimate the rate of adaptive evolution by inferring the DFE of deleterious (Fay et al. 2001; Smith and Eyre-Walker, 2002; Bierne and Eyre-Walker, 2004; Eyre-Walker et al. 2006; Eyre-Walker and Keightley, 2009; Stoletzki and Eyre-Walker, 2011) and beneficial (Galtier, 2016; Tataru et al. 2017) mutations mentioned above allow estimation of both the rate of nonadaptive ($\omega_{na} = d_N^{na}/d_S$) and the rate of adaptive ($\omega_a = \omega - \omega_{na}$) nonsynonymous substitution, thereby disentangling the effects of positive and negative selection. Whilst previous studies (Gossman et al. 2010; Gossman et al. 2012; Galtier, 2016; Rouselle and Galtier, 2019) have found no correlation between ω_a and N_e in plants, Strasburg et al. (2011) found a significant positive correlation between ω_a and N_e in sunflowers. It is

notable that Gossman et al. mostly considered species with low N_e, whilst some species in the Strasbourg et al. dataset have a much larger N_e. This positive correlation could be explained by a higher rate of adaptive substitution, or by population size change. A smaller historic population size relative to the current population size will artifactually inflate α and ω_a.

A final consideration is that of population size changes. Because selection is more effective in larger populations, weakly deleterious mutations are purged by purifying selection. In smaller populations these weakly deleterious mutations are more likely to fix through genetic drift. If there has been population size expansion such that the effective population for the polymorphism data is much greater than for the divergence data, selection is more effective on weakly deleterious mutations in the current population (population phase) than the historic population (divergence phase) and hence the proportion of non-synonymous to synonymous polymorphisms is less than one would expect, giving the evidence that there is more constraint than there was during the divergence phase.

1.3.2 Variation in N_e across the genome

N_e can also vary across the genome due to the effects of selection at a focal site on the behaviour of variants at nearby sites, due to genetic hitchhiking (Smith and Haigh, 1974) and background selection (Charlesworth et al. 1993), both of which are expected to reduce N_e, resulting in lower levels of genetic diversity and the reduced effectiveness of selection. This effect is exacerbated in regions of low recombination because the associations due to linked selection are not broken up. There are three lines of evidence that show there is variation in N_e within a genome. First, genetic diversity has been shown to be correlated to recombination rate, but recombination rate is not correlated to neutral divergence. Levels of neutral diversity have been shown to correlate to the recombination rate in *Drosophila* (Begun and Aquadro,

1992), humans (Lercher and Hurst, 2002; Hellman et al. 2003), and some plant species (Tenaillon et al. 2004; Roselius et al. 2005), which might be explained by mutation rate (μ) variation (because the amount of neutral diversity is proportional to $N_e\mu$). However, these same studies (with the exception of Lercher and Hurst, 2002 and Hellman et al. 2003 who find that divergence and recombination are correlated) show that neutral sequence divergence between species (which is expected to be proportional to the mutation rate) is not correlated to the rate of recombination (in *Drosophila*: Begun and Aquadro, 1992; in plants: Roselius et al. 2005). In humans, Hellman et al. (2003) found a correlation between neutral divergence and recombination rate but show that this correlation is not sufficient to fully explain the correlation between diversity and the recombination rate. However, Smith et al (2018) found that almost all variation in diversity could be explained by variation in the mutation rate in humans.

The second line of evidence follows from the first, with deviations from the expectation that levels of neutral divergence and diversity are proportional to one another (since both depend on the neutral mutation rate) (Kimura, 1968; Kimura, 1983) being caused by variation in N_e. Using derivatives of the HKA test (Hudson et al. 1987; Ingvarsson, 2004; Wright and Charlesworth, 2004; Innan, 2006), deviations from the neutral expectation have been shown in plants (Roselius et al. 2005; Schmid et al. 2005), the Z chromosome in chickens (Sundstrom et al. 2004), humans (Zhang et al. 2002) and *Drosophila* (Moriyama and Powell, 1996; Machado et al. 2002).

Finally, variation in N_e should also manifest as variation in the efficacy of selection within a genome. In *Drosophila* it has been shown that the ratio of non-synonymous to synonymous polymorphisms, P_n/P_s, is higher in low recombination regions of the genome (Presgraves, 2005; Gossman et al. 2011; Castellano et al. 2018; Castellano et al. 2019), whilst the rate of non-synonymous to synonymous substitution, d_N/d_S, is positively correlated to the recombination

rate (Betancourt and Presgraves, 2002). These correlations are attributed to low recombination regions of the genome having a low effective population size, and therefore reduced efficacy of selection (Betancourt et al. 2009). P_N/P_S is negatively correlated to the recombination rate because regions with low N_e have less effective selection and therefore slightly deleterious mutations are able to segregate in the population. The direction of the correlation between d_N/d_S and the recombination rate is dependent on the prevalence of advantageous mutations. Where they are common, the correlation is expected to be positive due to the higher mutation rate and selection acting on a greater proportion of mutations. Where advantageous mutations are rare, the correlation between d_N/d_S and the recombination rate is likely to be negative because selection against slightly deleterious mutations is more effective in low recombination regions (Gossman et al. 2011).

Though the existence of variation in N_e across the genome is well established, there is still much work to be done in quantifying this variation. Gossmann et al (2011) found modest variation in N_e across genes in 10 eukaryotic species (including humans, *Drosophila melanogaster*, *Arabidopsis thaliana* and *Saccharomyces paradoxus*.

1.4 Human demographic history

A critical factor when using neutrality tests to detect selection is the confounding effect of demography. A well-known example is that of Tajima's *D* rejecting neutrality if a population bottleneck has occurred (Simonsen et al. 1995; Andolfatto and Przeworski, 2000; Przeworski et al. 2000; Nielsen, 2001; Stajich and Hahn, 2005; Wall et al. 2002), because a bottleneck can leave a similar footprint to a selective sweep, including a dip in diversity (Galtier et al. 2000; Barton, 1998). As discussed in section 1.2.2, the effective population size has an effect on estimates of the rate of adaptive evolution. It has also been shown that the MK test can generate artifactual evidence of adaptive evolution if some non-synonymous mutations are

slightly deleterious and the population in question has undergone recent expansion. This will result in the efficacy of selection being greater during the polymorphism phase (when the population is larger) than during the divergence phase (where the population size is smaller) (McDonald and Kreitman, 1991; Eyre-Walker, 2002). The reverse is also true, where population contraction can artifactually depress estimates of the rate of adaptive evolution.

Humans have undergone a complex demographic history. Although the effective population size in humans has increase since the Out-Of-Africa migration, this new effective population size is still considerably reduced from that in the human-chimpanzee ancestor (Hobolth et al. 2007; Burgess and Yang, 2008; Prado-Martinez et al. 2013; Schrago, 2014).

1.4.1 Inferring human population demography

There have been two types of approach used to infer population demographic history. Pairwise sequentially Markovian coalescent (PSMC)-type methods use whole genome sequence data from a small number of individuals (1-4) to infer the demographic history of the entire population (McVean and Cardin, 2005; Li and Durbin, 2011; Schiffels and Durbin, 2014). First, the local time to most recent common ancestor (TMRCA) is estimated for small genomic regions. The distribution of these coalescent times can then be used to infer an overarching demographic history. The main benefit of PSMC methods is that they only require a few individual genomes to infer the demographic history of the whole population. Examples of the application of PSMC methods applied to human populations include Li and Durbin (2011); Kidd et al. (2012); Schiffels and Durbin (2014); 1000 Genomes Project Consortium (2015); Henn et al. (2016); Malaspinas et al. (2016); Mallick et al. (2016) and Pagani et al. (2016). A notable concern over the accuracy of demographic models obtained using PSMC methods was raised by Mazet et al. (2015) who found that instead of estimating a measure of population size,

PSMC methods capture the inverse instantaneous coalescent rate (IICR), which only corresponds to the effective population size if the population is panmictic. It is therefore necessary to account for population structure and gene flow to avoid false positive signals of population expansion or contraction, which is a well-known issue of demographic inference (Ptak and Przeworski, 2002; Chikhi et al. 2010; Peter et al. 2010; Gattepaille et al. 2013; Heller et al. 2013; Mazet et al. 2015; Orozco-terWengel, 2016).

The second approach to inferring a population's demographic history is to from the SFS, which represents the distribution of alleles at varying frequencies in a sample of individuals from a population (Nielsen, 2000; Wakeley, 2009). The distribution of SNPs in a population is directly affected by that population's demographic history. For example, population expansion can lead to an excess of rare variants (Tajima, 1989; Slatkin and Hudson, 1991; Keinan and Clark, 2012). Unlike PSMC methods, SFS-based methods have been shown to accurately estimate population growth (Nelson et al. 2012; Tenessen et al. 2012; Gazave et al. 2014; Bhaskar et al. 2015; Gao and Keinan, 2016). However the main drawback is requiring a greater number of individuals to be sequenced than with PSMC methods.

Gutenkunst et al. (2009) and Gravel et al. (2011) inferred human demographic history using an SFS-based method using a diffusion approximation, finding that the Eurasian split is followed by a period of exponential growth in both the European and Asian populations, whilst the African population maintains a relatively stable demography. Beichman et al. (2017) tested the accuracy of the Gravel et al. (2011) model for three human populations from the 1000 genomes dataset (1000 Genomes Project Consortium, 2015): CUE, CHB and YRI by comparing the distribution of expected heterozygosity from data simulated under the Gravel et al. (2011)

model with empirical 1000 genomes data, as well as computing the observed and expected SFS. In both cases they found that the model fits the data well.

What emerges is a complex model of human demographic history, with the aforementioned population contraction from the human-chimpanzee ancestor, followed by later population expansion in the European and Asian populations. This means that demography is likely to affect estimates of positive selection both between humans and chimpanzees, and between human populations.

1.5 Thesis scope

In this thesis I focus on patterns of positive selection across the human genome, and the factors that affect these patterns. The analyses are conducted using human data from the 1000 genomes dataset (1000 Genomes Project Consortium, 2015).

In chapter 2 I develop and apply a novel method to determine the prevalence of balancing selection in the human genome. Where previous methods that interpret shared polymorphism between populations as a signal of balancing selection are limited by requiring a long enough divergence time to ensure that all shared neutral genetic variation has become fixed or lost in at least one of the two populations, this new method uses this neutral genetic variation as a null model, providing information on the expectation in the absence of balancing selection. I use forward simulations to develop an understanding of how demography affects the method. I apply this method to human continental populations in an attempt to understand the frequency of balancing selection in humans.

In chapter 3 I look at site level factors that affect the rates of adaptive and nonadaptive evolution in humans. I correlate the rates of adaptive and nonadaptive evolution with relative solvent accessibility, measures amino acid physiochemical dissimilarity (volume and polarity), and evolutionary dissimilarity (p_n/p_s). I also show how population contraction or expansion can attenuate the correlation between a factor and the rate of adaptive evolution, if that factor is also correlated to the mean strength of selection against deleterious mutations.

In chapter 4 I look at gene level factors that affect the rates of adaptive and nonadaptive evolution in humans. I correlate the rates of adaptive and nonadaptive evolution with four factors: recombination rate, gene age, gene length and gene expression. For each factor I individually control for each of the other three factors to understand which factors are driving evolution in humans. I also look at gene function by estimating rates of evolution in GO categories for both viral interacting proteins (VIPs) and non-viral interacting proteins (nonVIPs) to understand the extent to which viruses drive evolution in humans.

In chapter 5 I revisit the question of what factors determine the level of neutral diversity across the human genome using the number of SNPs from the 1000 genomes dataset (The 1000 Genomes Project Consortium, 2015) and de novo mutations (DNMs) from three datasets (Francioli et al. 2015; Wong et al. 2016; Jonson et al. 2017). We show that the inferred distribution of mutation rates is actually broader than the distribution of SNPs. This leads us to explore models in which the effects of linked selection are dependent upon the mutation rate.

2. A new test demonstrates that balancing selection maintains hundreds of non-synonymous polymorphisms in the human genome

2.1 Abstract

The role that balancing selection plays in the maintenance of genetic diversity remains unresolved. Here we introduce a new test, based on the McDonald-Kreitman test, in which the number of polymorphisms that are shared between populations is contrasted to those that are private at selected and neutral sites. We show that this simple test is robust to a variety of demographic changes, and that it can also give a direct estimate of the number of shared polymorphisms that are directly maintained by balancing selection. We apply our method to population genomic data from humans and conclude that more than a thousand non-synonymous polymorphisms are subject to balancing selection.

2.2 Introduction

How genetic variation is maintained, either in the form of DNA sequence diversity or quantitative genetic variation, remains one of the central problems of population genetics. Balancing selection encapsulates several selective mechanisms that increase variability within a population. These include heterozygote advantage (also referred to as overdominance), frequency dependent selection, and selection that varies through space and time (Nielsen, 2005). However, although there are some clear examples of each type of selection (Allison, 1956; Nosil et al. 2018), the overall role that balancing selection plays in maintaining genetic variation, either directly, or indirectly through linkage, remains unknown.

A number of methods have been developed to detect the signature of balancing selection (Hughes and Nei, 1988; Asthana et al. 2005; Bubb et al. 2006; Andres et al. 2009; Leffler et al. 2013; DeGiorgio et al. 2014; Gao et al. 2015; Hunter-Zinck and Clark, 2015; Fijarczyk and Babik, 2015; Sheehan and Song, 2016; Siewert and Voight, 2017; Bitarello et al. 2018). Application of these methods have identified a number of loci subject to balancing selection, largely in the human genome, in which most of this research has taken place. However, these methods are generally quite complex to apply, often leveraging multiple population genetic signatures of balancing selection and many require simulations to determine the null distribution. Furthermore, they do not readily yield an estimate of the number of polymorphisms that are directly subject to balancing selection, as opposed to being in linkage disequilibrium. Here we introduce a method that is simple to apply and which generates a direct estimate of the number of polymorphisms subject to balancing selection.

One signature of balancing selection that has been utilised in several studies is the sharing of polymorphisms between species (Asthana et al. 2004; Leffler et al. 2013; Gao et al. 2015). If

the species are sufficiently divergent that they are unlikely to share neutral polymorphisms, then shared genetic variation can be attributed to balancing selection. These studies have concluded that there are relatively few balanced polymorphisms that are shared between humans and chimpanzees (Asthana et al. 2004; Leffler et al. 2013). However, this test is likely to be weak because humans and chimpanzees diverged millions of years in the past and it is unlikely that any shared selection pressures will be maintained over that time period.

The major problem with approaches that consider the sharing of polymorphisms between species or populations is differentiating selectively maintained polymorphisms from neutral variation inherited from the common ancestor. This problem can be solved by comparing the number of shared polymorphisms at sites which are selected, to those that are neutral. We expect the number of shared polymorphisms at selected sites to be lower than at neutral sites because many mutations at selected sites are likely to be deleterious, and hence unlikely to be shared. However, we can estimate the proportion that are effectively neutral by considering the ratio of polymorphisms, which are private to one of the two populations or species, at selected versus neutral sites. Although the method can be applied to any group of neutral and selected sites that are interspersed with one another we will characterise it in terms of non-synonymous and synonymous sites. Let the numbers of polymorphisms that are shared between two populations or species be S_N and S_S at non-synonymous and synonymous sites respectively, and the numbers that are private to one of the populations be R_N and R_S respectively. Let us assume that synonymous mutations are neutral and non-synonymous mutations are either neutral or strongly deleterious. Then it is evident that $\frac{S_N}{S_S} = \frac{R_N}{R_S} = f$, where f is the proportion of the non-synonymous mutations that are neutral. However, if there is balancing selection acting on some non-synonymous SNPs and this selection persists

for some time such that the balanced polymorphisms are shared between populations then $\frac{S_N}{S_S} > \frac{R_N}{R_S}$. A simple test of balancing selection is therefore whether Z > 1 where

$$Z = \frac{S_N/S_S}{R_N/R_S} \tag{2.1}$$

This is a simple corollary of the McDonald-Kreitman test for adaptive divergence between species (McDonald and Kreitman, 1991). It can be shown, under some simplifying assumptions in which synonymous mutations are neutral and non-synonymous mutations are strongly deleterious, neutral or subject to balancing selection, that an estimate of the proportion of non-synonymous mutations subject directly to balancing selection is $\alpha_b = 1 - \frac{S_S R_N}{S_N R_S}$ (see results section). In this analysis, we perform population genetic simulations to investigate whether the method can detect the signature of balancing selection and assess whether the method is robust to demographic change. Second, we apply the method to human population genetic data. We show that the method is robust and we estimate that substantial numbers of non-synonymous polymorphisms are maintained by balancing selection in humans.

2.3 Methods and Materials

2.3.1 Human data

Human variation data was obtained from 1000 genomes Grch37.p13 vcf files (The 1000 Genomes Project Consortium, 2015)**.** Variants were annotated using Annovar's hg19 database (Wang and Li, 2010). The annotated data was then parsed to remove multi-nucleotide polymorphisms and indels. Because 1000 genomes data provides allele frequencies for the non-reference allele rather than the minor allele, the minor allele frequency for each superpopulation and also for the global minor allele frequency was calculated. We used 1000

genomes from the African, South Asian, East Asian and European populations. The American population was removed due to the fact that it is an admixed population. GO category information was obtained from Ensembl's BioMart data mining tool (Yates et al. 2019). We used pyrho demography-aware recombination rate maps (Spence and Song, 2019) for analyses that control for recombination rate.

2.3.2 Simulations

All simulations were run using the SLiM software platform (Haller and Messer, 2019). Parameter values were taken from human estimates. Almost all simulations were of a 288bp locus, this being the average size of a human exon (Yates et al. 2020). Unless otherwise stated, the scaled recombination rate and scaled mutation rate were set at $r = 1.1 \times 10^{-8}$ (Dumont and Payseur, 2008), $\mu = 2.5 \times 10^{-8}$ (Nachman and Crowell, 2000) in the ancestral population. The distribution of fitness effects was assumed to be a gamma distribution and the shape and mean strength of selection estimates for humans were taken from Eyre-Walker et al. (2006) (β = 0.23; mean $N_e s$ = 425). For *Drosophila* estimates were taken from Keightley and Eyre-Walker (2007) (β = 0.35; mean $N_e s$ = 1800); again these were values in the ancestral population. Unless dominance was fixed, it was calculated using the model of Huber et al. (2018), which was estimated from Arabidopsis. The Huber model varies the dominance coefficient depending on the selection coefficient of the mutation, where the dominance coefficient increases with the strength of selection. It's formula is $h = f(s) = \frac{1}{\frac{1}{\theta_{intercept}} - \theta_{rate} s}$ where $\theta_{intercept}$ defines the values of h at $s = 0$ and θ_{rate} determines how quickly h approaches zero with decreasing negative selection coefficient. We set $\theta_{intercept}$ to 0.5 so that all mutations with a selection coefficient of $s = 0$ have a dominance coefficient, $h = 0.5$, and θ_{rate} = 41225.56. This assumes an inverse relationship between h and s, which gives the highest log likelihood score of the relationships compared by Huber et al. (2018). For balancing selection simulations, we

assume a model of heterozygote advantage and the strength of selection was sampled from a distribution such that the equilibrium frequency was uniformly distributed between 0 and 1; however, it should be noted that some balanced polymorphisms with low equilibrium frequencies were lost in one of the descendent populations. These simulations were discarded. The balanced polymorphism is introduced at the centre of the 288bp region. Two million successful simulation runs were conducted for each model.

For the generic simulations (i.e. not those involving the human demographic model) the ancestral population size was set at 200. This was allowed to equilibrate for 15N generations before a balanced polymorphism was introduced 5N generations before the population was split into two. The descendant populations were then sampled every 0.05N generations up to 20N generations after the split. We ran five different generic simulations: (i) simulations in which the ancestral population was duplicated, (ii) vicariance simulations in which the ancestral population was divided between the daughter populations in splits of 0.5N-0.5N; 0.75N-0.25N; 0.9N-0.1N, (iii) variance simulations in which the descendant populations expanded, iv) dispersal simulations, in which some variable fraction (0.5N; 0.25N; 0.1N) of the ancestral population is duplicated to form the dispersal population, and the ancestral population continues as the other daughter population, and v) dispersal with population increase of the dispersal population. The dispersal population starts as 0.1N and expands exponentially 2 to 10x its original size after 21N generations. Scenarios ii-v were repeated with migration rates of 0.01N and 0.001N of the ancestral population size between the descendant populations.

We also ran some simulations under the human demographic model of Gravel et al. (2011); for details of the demographic structure of the simulation (Gravel et al. 2011). The distribution of

fitness effects for deleterious mutations was assumed to be a gamma distribution using the parameters estimated from the African superpopulation using the GammaZero model within the Grapes software (Galtier N. , 2016); the parameters are similar to those estimated by Eyre-Walker et al (2006), and used in the generic simulations (Gamma shape = 0.17 and Mean $N_e s$ = 1144). We chose to infer the DFE for the African superpopulation because this is currently the largest dataset available for a population that has been inferred to be relatively stable. Dominance was calculated using the Huber model discussed above. Sampling of all populations (African, East Asian and European) was conducted at the end of the simulation (i.e. the equivalent of the present day).

2.4 Results

2.4.1 Simulations

We propose a new test for balancing selection in which the ratio of selected to neutral polymorphisms is compared between those that are shared between populations or species and those that are private to populations or species. To explore the properties of our method to detect balancing selection we ran a series of simulations in which an ancestral population splits to yield two descendent populations. We initially simulated loci under a simple stationary population size model where the ancestral population is duplicated to form two equally sized populations (equal to each other and the ancestral population). This is an unrealistic scenario, but it has the advantage that it involves no demographic change in the transition from ancestral to descendent populations. We assume that synonymous mutations are neutral and we explore the consequences of different selective models for non-synonymous mutations. If all non-synonymous mutations are neutral, then as expected $Z = 1$ (figure 2.1a), and if we make some of the non-synonymous mutations deleterious, drawing their selection coefficients from a gamma distribution, as estimated from human polymorphism data (Eyre-Walker et al.

2006) we find that $Z < 1$ (figure 2.1a). Again, this is expected because slightly deleterious mutations (SDMs) are likely to contribute more to the level of private than shared polymorphism. If we simulate a locus in which most non-synonymous mutations are deleterious, drawn from a gamma distribution, but each locus contains a single balanced polymorphism that is shared between populations then $Z > 1$ (figure 2.1a). It is important to note that the density of balanced polymorphisms is high in these simulations because we have simulated a short exon, of just 288bp, the average length in humans, and each one contains a balanced polymorphism. If we were to reduce the density of balanced polymorphisms then Z can be less than one even if there is balancing selection operating.

Slightly deleterious mutations tend to depress the value of Z because they are more likely to segregate within a population, than to be shared between populations that diverged sometime in the past. There are two potential strategies for coping with this tendency. We can test for the presence of balancing selection as a function of the frequencies of the polymorphisms in the population, because SDMs will tend to be enriched amongst the rarer polymorphisms in the population. A similar approach has been used successfully to ameliorate the effects of SDMs in the classic MK approach for estimating the rate of adaptive evolution between species (Fay et al. 2001; Charlesworth and Eyre-Walker, 2008; Messer and Petrov, 2013). Or we can explicitly model the generation of shared and private polymorphisms under a realistic demographic and selection model to control for the effects of SDMs. We focus our attention here on the first of these strategies, although we touch on the latter strategy in the discussion. We apply the frequency filter to both the private and shared polymorphisms; this is necessary because if we applied the filter only to the private polymorphisms, we could be comparing high frequency private polymorphisms, with a low ratio of R_N to R_S, because SDMs have been excluded, to low frequency shared polymorphisms, which may contain many SDMs and hence

have a high value of S_N/S_S; this can yield artefactual evidence of balancing selection. This could be exacerbated if some of the SDMs are recessive. To investigate the effects of polymorphism frequency on our estimate of Z we divided polymorphisms into 5 bins of 0.1 (we did not orient SNPs); for shared polymorphisms we estimated their frequency as the unweighted mean frequency from the two populations.

If we simulate a population in which non-synonymous mutations are deleterious, whose effects are drawn from a gamma distribution, we find that $Z < 1$ but this is less marked for the high frequency categories, as we expect. For the lowest frequency category Z decreases as a function of the time to most recent common ancestor, whereas for the higher frequency categories it is either unaffected or increases slightly (figure 2.1b). If we include a balanced polymorphism, introduced prior to the population split and subject to strong selection, into the model, which still also includes deleterious mutations, we find that $Z > 1$ for all frequency bins except the lowest one (figure 2.1c). In each case Z increases as a function of the time since the population split; this is to be expected because Z is related to the proportion of shared polymorphisms that are subject directly to balancing selection (see below), and as time progresses, so neutral and SDMs go to fixation or loss in one or both of the descendant populations. Note, once again that the level of balancing selection in these simulations is high because every locus contains a balanced polymorphism.

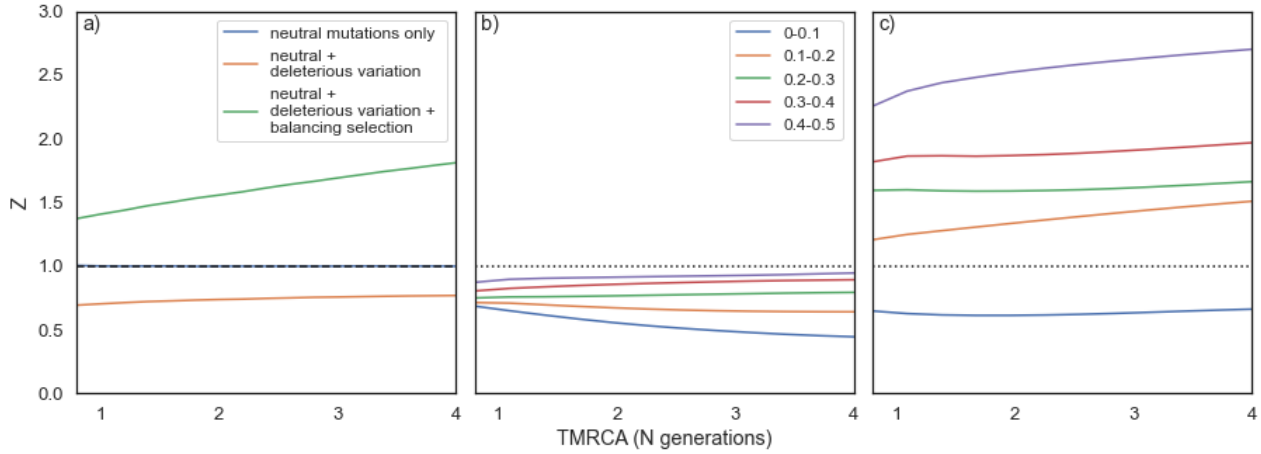


Figure 2.1: Stationary population size simulations, in which the ancestral population is duplicated to form two daughter populations of the same size to each other and the ancestor. Each simulation was repeated 2 million times. The time to the most recent common ancestor (tMRCA) is measured in N generations, where N is the population size. A Z value of greater than 1 indicates a greater proportion of shared non-synonymous polymorphisms than private non-synonymous polymorphism, which is a signal of balancing selection. For b-c) private polymorphisms have been binned by minor allele frequency, in bins of size 0.1. a) - simulations of neutral genetic variation only; a) – orange & b) neutral and deleterious variation; a) – green & c) neutral, deleterious and balanced polymorphisms.

The simulation above does not take into account the demographic effects that a division in a population involves. We therefore performed more realistic simulations which involve vicariance and dispersal scenarios with and without migration between the sampled populations (Appendix A, figures A1-A10). We also simulated with and without expansion after separation. We performed all simulations under two distributions of fitness effects (DFE), which were estimated from human and *Drosophila melanogaster* populations. In the vicariance scenario the ancestral population splits into two daughter populations of equal or unequal sizes. In the dispersal scenario a single daughter population is generated by duplicating part of the ancestral population, which remains the same size as it was before; we

vary the daughter population size. In both cases, we explore the consequences of expansion after separation of the populations.

None of the simulated demographic scenarios is capable of generating Z values greater than 1 under either DFE - the method does not seem to generate false positives (Appendix figures A1-A10). However, it is worth noting that a more severe difference in the size of the descendant populations results in depressed Z values in the smaller of the two populations, suggesting demography can affect the value of Z. In all cases the value of Z is smallest for the lowest frequency category, those polymorphisms with frequencies <0.1, and this frequency category often shows a dramatic difference to the other categories. We therefore suggest combining the polymorphisms above 0.1 when data is limited. As expected, we find that Z<1 in all simulations when we sum all polymorphisms with frequencies > 0.1 (Appendix figures A11 and A12).

2.4.2 Estimating the level of balancing selection

One of the great advantages of our method is that it gives an estimate of the number of polymorphisms that are directly affected by balancing selection, under a simple model of evolution. Let us assume that synonymous mutations are neutral and that non-synonymous mutations are strongly deleterious, neutral or subject to balancing selection; we further assume that all balanced polymorphisms arose before the two populations split. Then the expected numbers of non-synonymous, R_N, and synonymous, R_S, private polymorphisms are

$$R_s = \theta\rho W$$

$$R_N = \theta\rho W f \qquad (2.2)$$

where $\theta = 4N_e u$, N_e is the effective population size and u is the mutation rate per site per generation. ρ is the proportion of polymorphisms that are private to the population, W is Watterson's coefficient and f is the proportion of non-synonymous mutations that are neutral, (1-f) being deleterious or subject to balancing selection.

In deriving expressions for S_N and S_S we have to take into account that a balanced polymorphism can maintain neutral variation in linkage disequilibrium that may also be shared between populations. If we have b balanced non-synonymous polymorphisms and each of those maintains x neutral mutations in linkage disequilibrium, then the expected values of S_N and S_S are

$$S_S = \theta(1-\rho)W + bx$$

$$S_N = \theta(1-\rho)Wf + b + bxf \qquad (2.3)$$

It is then straightforward to show that the proportion of shared non-synonymous polymorphisms that are directly maintained by balancing selection is

$$\alpha_b = 1 - \frac{1}{Z} = 1 - \frac{S_S R_N}{S_N R_S} = \frac{b}{S_N} \qquad (2.4)$$

This is clearly an unrealistic model in several respects. First, it can be expected that there are slightly deleterious mutations in many populations and this will lead to an underestimation of α_b; second, it is likely that new balanced polymorphisms will be arising all the time and these will contribute to private polymorphism, increasing R_N/R_S and leading to a conservative estimate of α_b.

2.4.3 Data analysis - humans

We have shown that the method has the potential to detect balancing selection under realistic evolutionary models. We therefore applied our method to human data from the 1000 genomes project (The 1000 Genomes Project Consortium, 2015) focussing on four populations – Africans, Europeans, East Asians and South Asians. We find that Z >1 when private polymorphisms from the African population are used for all population comparisons if the frequency of private and shared polymorphisms >0.1; we also find that Z>1 in the South Asian and East Asian population comparison when using South Asian private polymorphisms and polymorphisms with frequencies > 0.1 (Figure 2.2). For several comparisons we have no polymorphisms at the relevant frequencies, and many of the confidence intervals on our estimates of Z are large. As a consequence, we combined the data for all polymorphisms with frequencies > 0.1 (Figure 2.2, right-most point in each panel). The patterns above are replicated; Z is significantly greater than one when we use private polymorphisms from Africa, and in the comparison between the East and South Asian populations, but Z<1 otherwise. In some cases, the Z value for the combined data can appear inconsistent with the Z values from the individual frequency categories – for example, in the European-South Asian comparison the combined Z value is greater than the Z value for 0.1-0.2 despite the fact that this is the only Z value above 0.1. This is because there are polymorphisms with frequencies >0.2, but there are not enough to yield a valid estimate of Z.

Although, the values of Z are not consistently > 1 across populations, the results suggest that there is balancing selection operating. Our simulations show that Z is consistently <1 when there is no balancing selection, and that the value of Z differs between the two population comparisons if the populations have undergone different demographies. We therefore infer that there is balancing selection maintaining polymorphisms between populations.

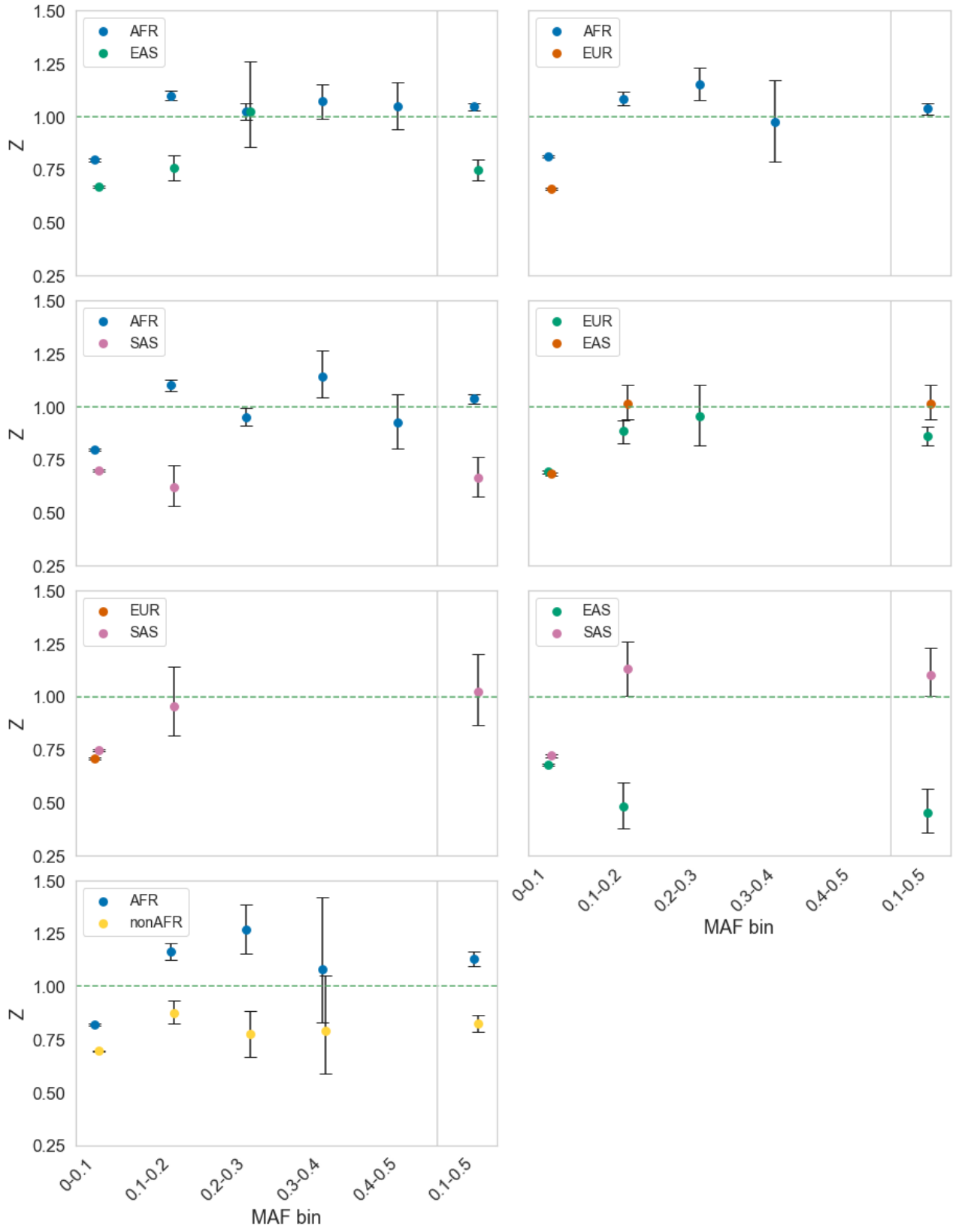

AFR
EAS
AFR
EUR
AFR
SAS
EUR
EAS
EUR
SAS
EAS
SAS
AFR
nonAFR
Z
1.50
1.25
1.00
0.75
0.50
0.25
0-0.1
0.1-0.2
0.2-0.3
0.3-0.4
0.4-0.5
0.1-0.5
MAF bin

Figure 2.2: The value of Z plotted against the frequency of shared and private polymorphisms, calculated for 1000 genome data from Africans (AFR), East Asians (EAS), Europeans (EUR) and South Asians (SAS). In each panel we show the value of Z for a comparison of two populations using the private polymorphisms from each, the population used being indicated in the plot legend. Data binned by minor allele frequency bins of size 0.1 on the x-axis. Final bin is 0.1-0.5 (i.e. all data minus the lowest frequency bin). Confidence intervals were generated by bootstrapping the data by gene 100 times. Only datapoints in which there were at least 20 polymorphisms for all polymorphism categories were plotted, because the confidence intervals were very large otherwise.

If we estimate α_b to those comparisons in which Z is significantly greater than 1, we estimate that approximately 12% of the non-synonymous shared polymorphisms between the African and other human populations are subject to balancing selection, as well as between the Asian populations (Table 2.1). These estimates are likely to be underestimates because there will still be SDMs segregating in our data, even though we have removed the lowest frequency variants (see simulation results). The proportions suggest that ~500 polymorphisms, which are shared between the African and other populations, are maintained by balancing selection. We estimate rather more are maintained between the two Asian populations, although the confidence intervals on this estimate are large. If we combine data across the non-African populations, we estimate that ~1400 polymorphisms are shared between African and non-African populations in total because of balancing selection (Table 2.1). The fact that the estimate from the African-non-African comparison is larger than the estimate from the African versus each individual population, suggests that many of the balanced polymorphisms shared between populations are unique to each population pair – i.e. there are balanced

polymorphisms shared between African and European populations, that are not shared between Africans and East Asians.

Target population	Comparative population	α_b	α_{b_low}	α_{b_high}	b	b_{low}	b_{high}
African	nonAfrican	0.12	0.09	0.14	1420	1077	1737
African	European	0.04	0.01	0.06	577	176	926
African	East Asian	0.05	0.03	0.06	657	401	882
African	South Asian	0.03	0.01	0.05	507	182	796
South Asian	East Asian	0.09	0.00	0.19	1273	18	2594

Table 2.1: Estimates of the proportion of shared non-synonymous polymorphisms under balancing selection, α_b, and the number of polymorphisms, b, being directly maintained by balancing selection for population comparisons in which Z>1. 95% confidence intervals were generated by bootstrapping the data by gene 100 times.

A concern in any analysis of human population genetic data is the influence of biased gene conversion (BGC). This process tends to increase the number and allele frequencies of AT>GC mutations, and reduce the number and allele frequencies of GC>AT mutations. If this process differentially affects synonymous and non-synonymous sites and shared and private polymorphisms, then it could potentially lead to Z>1. To investigate whether BGC has an effect we performed two analyses. In the first, we divided our genes according to whether they were in high and low recombining regions, dividing the data at the median recombination rate (RR). Our two groups differ substantially in their mean rate of recombination (mean RR in low group = 1.2×10^{-7} centimorgans per site and high group = 1.8×10^{-6} centimorgans per site). We find that Z is actually higher in the low RR regions, although not significantly so (Table 2.2), which suggests that BGC is not responsible for the comparisons in which Z > 1.

mean recombination rate	sN	sS	rN	rS	Z	Z_{low}	Z_{high}
1.20E-09	7023	7447	553	653	1.114	1.066	1.160
1.80E-08	7767	8514	604	702	1.060	1.021	1.103

Table 2.2: Estimates of *Z* for data split by median recombination rate. Confidence intervals were generated by bootstrapping genes 100 times.

In the second test of the influence of BGC on the value of Z we limited our analysis to mutations that are not affected by BGC – i.e. G<>C and A<>T mutations. This reduces our dataset by about 80%. As a consequence, we summed the data for all polymorphisms with frequencies >0.1. We find that our estimate is largely unchanged from that when all polymorphisms are included, however the confidence intervals are increased substantially so that Z is only significantly greater than one for the African-non-African, and the South versus East Asian comparisons (Table 2.3). Our two tests suggest that our results are not affected by BGC.

Target population	Comparative population	all polymorphism data			filtered for BGC		
		Z	Z_{low}	Z_{high}	Z	Z_{low}	Z_{high}
African	non-African	1.13	1.10	1.17	1.12	1.03	1.21
African	East Asian	1.05	1.03	1.07	1.00	0.95	1.05
African	European	1.04	1.01	1.07	1.01	0.93	1.10
African	South Asian	1.04	1.01	1.06	1.04	0.98	1.09
South Asian	East Asian	1.10	1.00	1.23	1.36	1.04	1.73

Table 2.3: Testing the effects of BGC for population comparisons which show Z>1 using all polymorphisms with frequencies > 0.1. To control for BGC the analysis was restricted to A<>T and G<>C SNPs. 95% confidence intervals were generated by bootstrapping the data by gene 100 times.

2.4.4 Groups of genes

We can potentially apply our test of balancing selection to individual genes or groups of genes, where we have enough data. Balancing selection has been implicated in the evolution of immune related genes (e.g. Bitarello et al. 2018, Weedel and Conway, 2010, Hughes and Nei, 1988, Hedrick, 2002), particularly major histocompatibility complex (MHC), or human leukocyte antigen genes (HLA) (Aguilar et al. 2004 and Paterson, 1998). To investigate whether we could detect this signature in our data, we split our dataset into HLA and non-HLA genes (The MHC sequencing consortium, 1999). Due to a lack of private polymorphisms, we combined all frequency categories >0.1. We found Z was significantly greater in HLA than non-HLA genes for all population comparisons except Europeans and South Asians using European private polymorphisms (Appendix figure A13). However, the value of Z is greater than one in population comparisons in which African private polymorphisms are used, consistent with balancing selection maintaining variation in the HLA region. We estimate that a very substantial proportion of non-synonymous genetic variation is being maintained by balancing selection, although the confidence intervals on our estimates are large; roughly 50% of the shared non-synonymous SNPs are being maintained by balancing selection between African and non-African populations in the HLA region and this equates to approximately 200 polymorphisms (Table 2.4)

However, the signature of balancing selection that we have detected across all genes is not simply due to the HLA genes. We find that Z>1 in non-HLA genes in most population comparisons in which Z>1 for all genes, except in the comparison of East and South Asian populations (Table 2.5); in most cases these estimates of Z are significantly greater than 1. We estimate that >1000 of non-synonymous polymorphisms are subject to balancing selection

amongst the polymorphisms shared by African and non-African populations, with several 100 shared between each of the populations and the African population.

target	comparative	α_b	α_{b_low}	α_{b_low}	*b*	b_{low}	b_{high}
AFR	nonAFR	0.702	-0.011	0.779	299	Na	332
AFR	EAS	0.287	-0.005	0.502	134	Na	234
AFR	EUR	0.731	0.317	0.820	338	147	379
AFR	SAS	0.529	0.307	0.694	247	143	324
EAS	SAS	-0.307	Na	0.292	Na	Na	133

Table 2.4: Estimates of the proportion of shared non-synonymous polymorphisms under balancing selection, α_b, and the number of polymorphisms being directly maintained by balancing selection, b, for population comparisons in the HLA region for population comparisons in which Z>1 when using all genes. Estimates for polymorphisms with frequency > 0.1. Confidence intervals were generated by bootstrapping the data by gene 100 times.

target	comparative	α_b	α_{b_low}	α_{b_low}	*b*	b_{low}	b_{high}
AFR	nonAFR	0.101	0.077	0.127	1193	907	1502
AFR	EAS	0.033	0.016	0.049	453	227	678
AFR	EUR	0.024	-0.001	0.049	355	Na	716
AFR	SAS	0.021	0.001	0.041	293	19	587
EAS	SAS	-1.254	-1.981	-0.833	Na	Na	Na

Table 2.5. Estimates of the proportion of shared non-synonymous polymorphisms under balancing selection, α_b, in non-HLA genes, and the number of polymorphisms being directly maintained by balancing selection, b, for population comparisons in which Z>1 when using all genes. Confidence intervals were generated by bootstrapping the data by gene 100 times.

If we run our analysis grouping genes by their GO category and restricting the analysis to those groups that have at least 100 polymorphisms with frequencies >0.1 we find 683 categories in which Z is significantly greater than 1 in at least population comparison and we list those significant in 5 or more population comparisons in Table 2.6. One of these GO categories, “nucleic acid binding” is shared across 7 of the 14 population comparisons, “endoplasmic reticulum membrane” across 6 population comparisons; amongst those categories shared among 5 are “viral process” and “immune system process”, but there are many others which are surprising including “protein import into the nucleus”. Eighty-five categories are shared between 4 or more population comparisons, and 155 amongst three or more population comparisons. These include 7 categories related to immunity (including immune system process which is significant in 5 population comparisons), and 40 categories that are linked to antigen presentation though not classified as immune-related categories. There are also 5 viral-related categories (including viral process which is significant in 5 population comparisons).

GO category	Number of population comparisons
nucleic acid binding	7
endoplasmic reticulum membrane	6
response to stimulus	5
viral process	5
zinc ion binding	5
DNA binding	5
nucleus	5
immune system process	5
chromatin binding	5
keratinization	5
intermediate filament	5
positive regulation of transcription by RNA polymerase II	5
chromosome	5

Table 2.6. GO categories in which Z is significantly greater than one in at least 5 population comparisons.

2.4.5 Individual genes

We applied our statistic to individual human genes, combining all frequency bins (0-0.5) due to a lack of polymorphism. We tested for significance using a one-tailed Fisher's exact test. Of the 14,261 genes we analysed 514 had Z>1 in at least one population comparison. Eighteen of these were nominally significant at $p<0.1$, but no gene was individually significant when we corrected for multiple testing using a Bonferroni correction. Eighteen genes have Z>1 in at least 9 population comparisons; note that since populations share polymorphisms, we cannot combine the evidence for balancing selection across these populations (Table 2.7). Four of these genes MUC4, RP1L1, PKD1L2 and ZAN have Z>1 in all population comparisons. We correlated Z with gene length to see if longer genes invariably have higher estimates of Z.

However we found no significant correlation when fitting a linear regression to the correlation between Z and gene length ($r=0.07$, $p>0.1$).

Gene symbol	Number of population comparisons in which Z>1	Gene length
MUC4	14	44,732
RP1L1	14	105,838
PKD1L2	14	119,495
ZAN	14	64,202
C1orf167	13	27,798
SPTBN5	12	45,907
MKI67	12	29,764
DNAH14	11	503,030
WDFY4	10	298,080
FAM230G	10	15,567
CMYA5	9	110,404
CRIPAK	9	4,442
SYNE2	9	464,534
FSIP2	9	94,486
GREB1	9	160,447
ALMS1	9	239,408
MUC19	9	177,437
CENPF	9	61,376

Table 2.7. Genes with Z>1 in 9 or more population comparisons. Z was estimated for all 14 population comparisons for each gene. Gene lengths are included for the 18 genes listed here.

If we use the 514 genes and do a GO enrichment analysis, we find multiple GO categories enriched for these genes including immune response categories with 3-fold enrichment. The most highly enriched categories are involved in energy production and conversion (including dynein binding) and intracellular transport (including microtubule motor activity).

2.5 Discussion

We propose a new method for detecting and quantifying the amount of balancing selection that is operating on polymorphisms, in which the numbers of non-synonymous and synonymous polymorphisms that are shared between populations and species are compared to those that are private. The method is analogous to the McDonald-Kreitman test used to test and quantify the amount of adaptive evolution between species (McDonald and Kreitman, 1991). We show that our test is robust to the presence of slightly deleterious mutations under simple demographic models of population division, expansion and migration. When we apply our method to data from human populations, we find evidence that hundreds of non-synonymous polymorphisms are being directly maintained by balancing selection in human populations.

Our method for detecting balancing selection is simple to apply and appears to be robust to changes in demography. The classic MK test of adaptive evolution between species can generate artefactual evidence of adaptive evolution if there are SDMs and there has been population size expansion (McDonald and Kreitman, 1991; Eyre-Walker, 2002); this is because SDMs that might have been fixed when the effective population size was small, no longer segregate once the population size is large. A similar bias does not appear to affect our test, although we have only investigated two DFEs and a limited number of demographic scenarios. Our test is likely to be more robust than the classic MK test because the shared polymorphisms are affected by the demographic changes that affect the private polymorphisms; i.e. if the population expands this will increase the effectiveness of natural selection on both the private and the shared polymorphisms. However, although our method seems to be relatively robust to changes in demography, in the sense that it does not generate artefactual evidence of balancing selection, it is evident that demography does affect the chance of balancing

selection being identified, because the values of Z depend on the demography and which population the private polymorphisms are taken from (Figure 2.2).

The method can in principle be applied to any pair of populations or species. However, the test is likely to be weak when the populations/species are closely related for two reasons. First, there will be relatively few private polymorphisms, and second, the proportion of shared polymorphisms that are subject to balancing selection is likely to be low, because so many neutral polymorphisms are shared between populations because of recent common ancestry. As the populations/species diverge so the number of private polymorphisms will increase, and the proportion of shared polymorphisms that are balanced will increase. Of course, as the time of divergence increases so the selective conditions that maintained the polymorphism are likely to change and the polymorphism might become neutral, or subject to directional selection. Our method is also likely, like all methods, to be better at detecting balanced polymorphisms that are common, because most populations are dominated by large numbers of rare neutral variants. Finally, our method requires that the neutral and selected sites are interdigitated; the method is therefore easy to apply to protein coding sequences, but may be more difficult to apply to other types of variation, such as that which affects gene expression.

A great advantage of our method is that it gives an estimate of the proportion and number of shared polymorphisms that are directly subject to balancing selection, under a set of simplifying assumptions. However, the method is likely to yield underestimates of the proportion of balanced polymorphisms, under a more realistic models of evolution. We have assumed, in deriving α_b, that all non-synonymous mutations are either strongly deleterious, neutral or subject to balancing selection. However, a substantial fraction of non-synonymous mutations appear to be slightly deleterious in humans (Cargill et al. 1999; Fay et al. 2001; Eyre-

Walker et al. 2002; Hughes et al. 2003; Asthana et al. 2007) and other species (Fay et al. 2001; Eyre-Walker et al. 2002; Hughes, 2005; Charlesworth and Eyre-Walker, 2006) – i.e. they are deleterious, but sufficiently weakly selected that they contribute to polymorphism. Under stationary population size assumptions – i.e. in which the ancestral population is duplicated to form the daughter populations - this will lead to an underestimate of Z because SDMs tend to contribute more to private than shared polymorphism, and hence inflate R_N/R_S relative to S_N/S_S (Figure 2.1). Under more realistic demographic models, in which at least one of the derived populations is reduced, this is expected to depress Z in the population that is being reduced because more SDMs will tend to segregate in smaller populations hence inflating R_N/R_S (compare Figure 2.1 to Appendix figures A2 and A3). Simulations suggest there is however a slight increase in Z using private polymorphisms from the larger of the populations but these Z values never exceed 1 (Appendix figures A2 and A3). The second reason that we are likely underestimating the number of balanced polymorphisms using our simple method is that we assume that there are no balanced polymorphisms that are private to each population; these would inflate R_N/R_S. Private balanced polymorphisms might arise from an ancestral polymorphism that is lost from one of the daughter populations, or one that arises de novo. A more realistic model of balancing selection is one in which balanced polymorphisms are continually generated with the selective forces persisting for some time before they dissipate (Sellis et al. 2011) and the balanced polymorphism is lost. The process of population division itself is likely to lead to the loss of many balanced polymorphisms as the environment shifts in the two daughter populations.

A potential solution to the tendency for our method to underestimate Z is to simulate data under a realistic demographic model assuming that there is no balancing selection, and interpret the observed values of Z that are greater than simulated values, as evidence of

balancing selection. A similar approach has been used to estimate the rate of adaptive evolution between species (e.g. Eyre-Walker and Keightley, 2009; Boyko et al. 2008; Schneider et al. 2011; Galtier, 2016; Tataru et al. 2017). However, there are challenges in this approach; in particular, we need an accurate demographic model. We have performed simulations under the commonly used human demographic model inferred by Gravel et al. (2011) estimating the DFE from the current African population; we chose the African population because it has been subject to relatively modest demographic change. Our observed Z values do not match the simulated values (Appendix figure A14); in particular we find that the observed values of Z are substantially greater than the simulated amongst the low frequency polymorphisms. However, the model of Gravel et al. does not fit the SFS of the individual populations of 1000 genome data; for example, in the African population there are far too many singleton SNPs even amongst the putative neutral synonymous mutations (Appendix figure A15). The lack of fit is perhaps not surprising; Gravel et al. inferred their model using 80 chromosomes per population, whereas the 1000 genome data contains >1000 chromosomes per population. Furthermore, the inference of a demographic model should take into account the influence of biased gene conversion and background selection, which appear to be pervasive factors in the human genome (Pouyet et al. 2018), so these simulations will be complex.

It might be argued that the evidence of balancing selection is weak because we typically find Z>1 using the private polymorphism from only one of the two populations. However, we have been unable to find a demographic model in which there is no balancing selection and Z>1 – note that we never observe Z>1 under the Gravel et al. model of human demography even when we change the parameters of the DFE and demographic model; furthermore, we find that simulations which involve different demographies in the two populations generate

different Z values for the two populations, so there is an expectation in many species that we will observe Z values that differ between populations.

Values of Z in excess of one could potentially be due to biased gene conversion; we expect BGC to increase the allele frequency of AT>GC mutations and to decrease the frequency of GC>AT mutations; this will tend to make AT>GC mutations more likely to be shared between populations than GC>AT mutations. Since the GC-content at the third codon position is typically higher than the GC content at the first two positions, this will mean that there are more non-synonymous AT>GC mutations than synonymous, and hence more shared non-synonymous than synonymous polymorphisms. However, our results do not appear to be affected by BGC; results are similar between genes in high and low recombination rate regions of the human genome (Table 2.2), and our point estimates of Z are largely unaffected by restricting the analysis to SNPs which are unaffected by BGC, although the confidence intervals increase substantially (Table 2.3).

We estimate that there ~500 balanced polymorphisms shared between the African and each of the other human populations, ~1250 shared between the Asian populations and ~1400 shared between the African and non-African populations; these are likely to be underestimates due to the presence of slightly deleterious mutations. The fact that we estimate that ~1400 shared non-synonymous polymorphisms are being maintained between African and non-African populations, but only ~500 between each of the individual populations and the African population suggests, that many of the polymorphisms shared between the African and each individual population are unique – i.e. many polymorphisms shared between Africans and Europeans, may not be shared between Africans and Asians. These numbers are substantial, but are consistent with those of Bitarello et al. (2018) who estimated that ~8% of human genes

show some evidence of long-term balancing selection, and many of these signatures are shared between populations. However, their method could not determine whether the balanced polymorphisms were coding or noncoding mutations.

As expected, we find evidence of balancing selection affecting the HLA or MHC genes (Table 2.4) (Hughes and Nei; 1988; Hedrick, 1998). However, we find evidence of balancing selection even when these genes are removed from the analysis (Table 2.5). The analysis of GO categories shows that numerous categories show evidence of balancing selection across multiple population comparisons. Some of these are expected, but many are not, such as "nucleic acid binding", which is significant in 7 of the 14 population comparisons.

No individual gene is significant when we control for multiple testing, however, several genes have Z>1 in multiple population comparisons including 10 which are shared across at least 10 of the 14 population comparisons. Three of these overlap with previous genome-wide scans of selection, namely the protein-coding gene DNAH14, implicated in brain compression and encoding axenomal dynein (Voight et al. 2006); MUC4, implicated in biliary tract cancer (Tennessen and Akey, 2011); and ZAN, which encodes a protein involved in sperm adhesion, previously implicated in balancing selection and positive selection in human populations (Gasper and Swanson, 2006). Two of these ten genes are associated with tumours. MKI67 expression is associated with a higher tumour grade and early disease recurrence (Yang et al. 2017), and WDFY4 plays a critical role in the regulation of certain viral and tumour antigens in dendritic cells (Theisen et al. 2019). PKD1L2 is associated with polycystic kidney disease and RP1L1 variants are associated with several retinal diseases including occult macular dystrophy (Davidson et al. 2013). SPTBN5 encodes for the cytoskeletal protein spectrin, that plays a role in maintaining cytoskeletal structure (Huh et al. 2001) and C1orf167 expresses open reading

frame protein that is highly expressed in the testis (Fagerberg et al. 2014). Finally, FAM230G is highly expressed in testes (Delihas, 2018).

Twenty-five of the 514 with Z>1 genes overlap with those genes identified by Bitarello et al. (2018), but this is similar to the level of overlap expected at random; i.e. they observed that 7.9% of protein coding genes overlapped regions identified by their method as being subject to balancing selection, and we identified 514 candidates; so we expect 0.079 x 514 = 41 by chance alone. The lack of a significant overlap is possibly not surprising; we have applied our method to non-synonymous variation, whereas the method of Bitarello et al. (2018) considers all variation. Furthermore, the method of Bitarello et al. (2018) is most powerful at detecting balancing selection over long time periods; in the case of humans, over periods of millions of years. In contrast, we have applied our method to populations that diverged 10,000's of years ago.

It is possible that the signature of balancing selection is caused by a form of associated overdominance, in which neutral alleles at a locus are linked to different deleterious recessive alleles at other loci. For example, let us imagine that we have two closely linked loci at which we have deleterious alleles; let the A2 allele be the recessive allele at the A locus and the B2 allele at the B locus. Now consider a third neutral locus with alleles C1 and C2. If C1 is in linkage disequilibrium (LD) with the A2 allele, and C2 is in LD with the B2 allele, then C1C2 heterozygous individuals will have higher fitness than C1C1 and C2C2 homozygotes. This form of selection can lead to the maintenance of genetic variation (Zhao and Charlesworth 2016) in low recombination rate regions. However, Z is not substantially greater in regions of low recombination so AOD seems an unlikely explanation (Table 2.2).

2.6 Conclusion

We present a new approach to test for the presence of balancing selection and to the number of polymorphisms that are directly affected by it. Our method appears to be robust to demographic change. Application of the method to human population genetic data suggests that 100s of non-synonymous polymorphisms shared between populations are being maintained by balancing selection.

3. Site level factors that affect the rate of adaptive evolution in humans and chimpanzees; the effect of contracting population size

3.1 Abstract

It has previously been shown in other species that the rate of adaptive evolution is higher at sites that are more exposed in a protein structure and lower between amino acid pairs that are more dissimilar. We have investigated whether these patterns are found in the divergence between humans and chimpanzees using an extension of the MacDonald-Kreitman test. We confirm previous findings and find that the rate of adaptive evolution, relative to the rate of mutation, is higher for more exposed amino acids, lower for amino acid pairs that are more dissimilar in terms of their polarity, volume and lower for amino acid pairs that are subject to stronger purifying selection, as measured by the ratio of the numbers of non-synonymous to synonymous polymorphisms (p_N/p_S). However, the slope of this latter relationship is significantly shallower than in Drosophila species. We suggest that this is due to the population

contraction that has occurred since humans and chimpanzees diverged. We demonstrate theoretically that population size reduction can generate an artefactual positive correlation between the rate of adaptive evolution and any factor that is correlated to the mean strength of selection acting against deleterious mutations, even if there has been no adaptive evolution (the converse is also expected). Our measure of selective constraint, p_N/p_S, is negatively correlated to the mean strength of selection, and hence we would expect the correlation between the rate of adaptive evolution to also be negatively correlated to p_N/p_S, if there is no adaptive evolution. The fact that our rate of adaptive evolution is positively correlated to p_N/p_S suggests that the correlation does genuinely exist, but that is has been attenuated by population size contraction.

3.2 Introduction

The rate of adaptive evolution in protein coding genes varies at several different levels. First, the rate of adaptive evolution appears to differ between species. Some species, including many plants (Bustamante et al. 2002; Barrier et al. 2003; Schmid et al. 2005; Gossman et al. 2010; also see Strasburg et al. 2009; Ingvarsson et al. 2010; Slotte et al. 2010) and the yeasts of the genus Saccharomyces (Gossman et al. 2012), appear to go through very little adaptive evolution, whilst many other species, including Drosophilids (Smith and Eyre-Walker 2002; Sawyer et al. 2003; Eyre-Walker and Keightley 2009; Haddrill et al. 2010), rodents (Halligan et al. 2010) and many multicellular animals (Galtier 2016; Rousselle et al. 2019), go through extensive adaptive evolution. The reasons for this variation remain unclear. It has been suggested that population size might be a factor; if adaptation is mutation limited, then one might expect species with large population sizes to adapt faster because they will generate the required mutation faster. There is some evidence that species with large population sizes undergo significantly faster adaptive evolution (Gossman et al. 2012; Bataillon et al. 2015;

Corbett-Detig et al. 2015; Rousselle et al. 2019), though in Galtier (2016) the correlation with ω_a is non-significant. Furthermore, it is unclear whether species are ever limited by the supply of mutations - there appears to be abundant genetic variation for most traits - and even if they are limited, species with large population sizes are predicted to be closer to their optimal fitness, and hence they may not have to adapt as much as species with small population sizes (Lourenco et al. 2013).

At the next level down, there appears variation in the rate of adaptation between genes. This is in part due to differences in function, with genes involved in immunity (Clark et al. 2003; Nielsen et al. 2005; Chimpanzee Sequencing and Analysis Consortium, 2005; Sackton et al. 2007; Obbard et al. 2009), interaction with viruses (Enard, et al. 2016) and male reproductive success (Proschel et al. 2006; Haerty et al. 2007) having high rates of adaptive evolution. Other factors also seem to be important, with the rate of adaptive evolution being higher in genes that recombine frequently (Presgraves, 2005; Betancourt et al. 2009; Arguello et al. 2010; Mackay et al 2012; Campos et al. 2014; Castellano et al. 2016; Moutinho et al. 2019), are located in regions of the genome with low functional DNA density (Castellano et al. 2016), have high mutation rates (Castellano et al. 2016) and reside on the X-chromosome (MacKay et al. 2012; Langley et al. 2012; Campos et al. 2014). Genes that have lower expression levels (Pal et al. 2001; Subramanian and Kumar, 2004; Wright et al. 2004; Rocha and Danchin, 2004; Lemos et al. 2005) or shorter coding sequence length (Zhang, 2000; Lipman et al. 2002; Liao et al. 2006), also seem to have higher rates of adaptation.

Finally, there appears to be variation at the site level. This variation has been widely documented in site-level tests that compare the rate of non-synonymous to synonymous

substitution (for example, Liberles et al. 2012). A number of factors seem to affect rates of adaptive evolution at the site level including protein secondary structure (Goldman et al. 1998; Guo et al. 2004; Choi et al. 2006) and the relative solvent accessibility (RSA) (Goldman et al. 1998; Choi et al. 2007; Lin et al. 2007; Franzosa and Xia 2009); RSA is a measure of how buried an amino acid is. In both *Drosophila* and *Arabidopsis* species, the rate of adaptive non-synonymous substitution is positively correlated to the relative solvent accessibility (RSA) (Moutinho et al. 2019). This suggests that amino acids on the surface of a protein have higher rates of adaptive substitution than those that are buried (Perutz et al. 1965; Overington et al. 1992; Goldman et al. 1998; Bustamante et al. 2000; Dean et al. 2002; Choi et al. 2006; Lin et al. 2007; Conant and Stadler 2009; Franzosa and Xia 2009; Ramsey et al. 2011). It has also been shown that amino acids that differ strongly in their physio-chemical properties, have lower rates of adaptive evolution than those that are more similar (Bergman and Eyre-Walker, 2019; though see Gojobori et al. 2007 and Chen et al. 2019). Finally, Bergman and Eyre-Walker (2019) also showed that amino acids pairs that are subject to high levels of negative selection have lower rates of adaptive substitution; they measured the level of negative selection using the ratio of the number of non-synonymous to synonymous polymorphisms, p_N/p_S.

In our analysis we consider whether the rate of adaptive evolution between humans and chimpanzees is correlated to several site level factors previously shown to be particularly important in other species - RSA and various measures of the difference between amino acids, and the overall level of negative selection acting on amino acid pairs. We find negative correlations between the rate of adaptive evolution and the difference in amino acid physio-chemical properties, and a positive correlation between the rate of adaptive evolution and RSA and our measure of negative selection.

3.3 Materials and methods

3.3.1 Data

We obtained gene sequences from Ensembl's biomart (Yates et al. 2020) for the human GRCh38.p13 genome build and for the Pan_tro_3.0 chimpanzee genome build. Orthologous genes were aligned using MUSCLE (Edgar, 2004). After filtering out genes with gaps that were not multiples of three we were left with 16,344 pairwise alignments. Numbers of synonymous and non-synonymous substitutions per site were obtained using PAML's codeml (Yang, 2007) program. We used polymorphism data from the African superpopulation of the 1000 genomes dataset (The 1000 Genomes Consortium, 2015) to construct our site frequency spectra, with rates of adaptive and non-adaptive evolution estimated using Grapes (Galtier, 2016), under the "GammaZero" model. We chose African data because the African population is thought to have undergone less complex demographic changes then other human populations (Gutenkunst et al. 2009; Gravel et al. 2011). We fitted a weighted regression to our estimates of the rate of evolution, weighting by the reciprocal of the variance for each estimate of ω_a and ω_{na}. The confidence interval and variance on our estimates of ω_a and ω_{na} were obtained by bootstrapping the dataset by gene 100 times.

3.3.2 RSA analysis

In order to obtain structural information for each protein sequence, we ran blastp (Schaffer, 2001) to assign each protein sequence to a PDB structure, and respective chain. Rather than setting a cut-off, we used sequences with the maximum identity by using the "pdbaa" library and an *E*-value threshold of 10^{-10}. We filtered sequences with the maximum identity and in instances of multiple matches, the match with the lowest *E*-value was kept. The corresponding PDB structures were further processed to only keep the corresponding chain per polymer. PDB manipulation and analysis were carried on using the R package "bio3d" (Grant et al. 2006).

Values for solvent accessibility (SA) per residue were obtained using the “dssp” program with default options. To map SA values to each residue of the protein sequence a pairwise alignment between each protein and the respective PDB sequence was performed with MAFFT, allowing gaps in both sequences in order to increase the block size of sites aligned. The final data set comprised a total of 7,984,041 sites with SA information. We computed the RSA by dividing SA by the amino-acid’s solvent accessible area (Tien et al. 2013), giving us a final dataset of 3,505,615 sites for which we have RSA information.

These sites were grouped into 20 RSA bins of roughly equal size in terms of the number of sites, with rates of adaptive and non-adaptive evolution estimated for each bin. These rates were correlated with the mean RSA of each bin.

3.3.3 Amino acid dissimilarity analysis

For the amino acid dissimilarity analysis we followed the methodology outlined in Bergman and Eyre-Walker (2019), with amino acid polarity and volume scores taken from data available in the AAindex1 database (Kawashima et al. 2008). We compared the SFS for a particular amino acid pair with synonymous data from 4-fold degenerate codons separated by the same mutational step. For example, alanine and glycine are separated by a single nucleotide change (C<>G at second position). Therefore, we used the SFS and divergence for all 4-fold degenerate codons separated by a single C<>G mutational step in estimating ω_a and ω_{na}. For amino acids separated by more than one mutational step (e.g. a C<>G or an A<>T mutational step), we used a weighted average SFS from the SFSs for the mutational types at 4-fold sites, weighting by the frequency of the respective codons as in Bergman and Eyre-Walker (2019).

For the analysis involving p_N/p_S we used a hypergeometric distribution to resample the SFS, and generate two SFSs, one used to estimate rates of adaptive and non-adaptive evolution, and one used to estimate p_N/p_S.

3.4 Results

3.4.1 Theory

It is well established that MK-type methods lead to biased estimates of the rate of adaptive evolution if the effective population size differs between the divergence and polymorphism phases (McDonald and Kreitman 1991; Eyre-Walker 2002). Could changes in effective population size also artefactually affect the relationship between the rate of adaptive evolution and another genomic variable, such as the difference in physico-chemical properties between two amino acids?

Let us assume that synonymous mutations are neutral and non-synonymous mutations are neutral or subject to negative selection. The ratio of the non-synonymous to synonymous substitution rates $\omega = \omega_a + \omega_{na}$ where ω_a and ω_{na} are the rate of adaptive and non-adaptive non-synonymous substitution relative to the rate of synonymous substitution, which is an estimate of the mutation rate under this model. Hence,

$$\omega_a = \omega - \omega_{na} \quad (1)$$

If we assume that all non-synonymous are deleterious with effects drawn from a gamma distribution then

$$\omega \approx \frac{k}{(N_d \bar{s})^{\beta}} \quad (2)$$

(Welch et al. 2008) where N_d is the effective population size during the divergence phase, k is a constant, β is the shape parameter of the gamma distribution and $\bar{s}$ is the mean absolute strength of selection acting against deleterious mutations.

We can also write a simple expression for ω_{na}. This is estimated in MK type approaches from polymorphism data, using the site frequency spectra (SFS) at synonymous and non-synonymous sites, to estimate the distribution of fitness effects (DFE) at non-synonymous sites. This DFE is then used to infer ω_{na}. Hence

$$\omega_{na} = \frac{k}{(N_p\bar{s})^{\beta}} \quad (3)$$

where N_P is the effective population size pertaining to the polymorphism data.

Substituting equation 2 and 3 into 1 we have

$$\omega_a = \frac{k}{(N_d\bar{s})^{\beta}} - \frac{k}{(N_p\bar{s})^{\beta}} = \frac{k\left((N_p\bar{s})^{\beta} - (N_d\bar{s})^{\beta}\right)}{(N_p\bar{s})^{\beta}(N_d\bar{s})^{\beta}} = \frac{k\left((N_p/N_d)^{\beta} - 1\right)}{(N_p\bar{s})^{\beta}} \quad (4)$$

From this equation it is evident that $\omega_a > 0$ if $N_p > N_d$, and $\omega_a < 0$ if $N_p < N_d$ as we expect. However, of more interest is the fact that the over- or under-estimation of ω_a depends on $\bar{s}$, the mean strength of selection acting against deleterious mutations. With population size expansion we predict that ω_a will be overestimated but that the magnitude of this overestimation will decrease as the mean strength of selection increases. Conversely, with population size contraction ω_a will be under-estimated and this underestimation will diminish as the mean strength of selection increases. Hence, under population size expansion we expect a negative correlation between ω_a and any variable that is correlated to the mean absolute strength of selection acting against deleterious mutations and a positive correlation with population contraction, if there is no adaptive evolution.

If we note that

$$\frac{p_N}{p_S} = \frac{m}{(N_p\bar{s})^{\beta}} \quad (5)$$

(Welch et a. 2006), where *m* is a constant which depends on how many chromosomes have been sampled, then equation 4 can be rewritten as

$$\omega_a = k\left(\frac{(N_p/N_d)^{\beta}-1}{m}\right)\frac{p_N}{p_S} \quad (6)$$

Hence, we expect ω_a to be positively and linearly correlated to p_N/p_S if there was been population size expansion and negatively correlated if there has been contraction, if there is no adaptive evolution occurring.

An alternative measure of the rate of adaptive evolution is the proportion of substitutions that are fixed by positive selection. Under our model this becomes

$$\alpha = \frac{\omega_a}{\omega} = 1 - \left(\frac{N_d}{N_p}\right)^{\beta} \quad (7)$$

As expected, if $N_p > N_d$ then $\alpha > 0$, and if $N_p < N_d$ then $\alpha > 0$, however the magnitude of this bias is independent of the strength of selection acting upon deleterious mutations.

What do we expect if there has been adaptive evolution? Let the rate of adaptive evolution, relative to the mutation rate, potentially be a function of the mean strength of selection acting against deleterious mutations, $A(\bar{s})$. Then equation 2 becomes

$$\omega \approx \frac{k}{(N_d\bar{s})^{\beta}} + A(\bar{s}) \quad (7)$$

which leads to a revision of equations 4 and 6

$$\omega_a = \frac{k\left((N_p/N_d)^{\beta}-1\right)}{(N_p\bar{s})^{\beta}} + A(\bar{s}) \qquad \omega_a = \left(\frac{(N_p/N_d)^{\beta}-1}{m}\right)\frac{p_N}{p_S} + A(\bar{s}) \quad (8)$$

Thus, if the rate of adaptive evolution is independent of the mean strength of selection acting against deleterious mutations, i.e. $A(\bar{s}) = a$, then it is evident that our predictions, derived under the assumption of no adaptive evolution, hold – e.g. population contraction will induce an artefactual positive correlation between ω_a and a variable that is correlated to the mean strength of selection against deleterious mutations. If the rate of adaptive evolution is correlated to the mean strength of selection, then this will tend to either increase or decrease the strength of the relationship.

3.4.2 Data analysis

Given the theoretical predictions derived above, is it of interest to examine patterns of adaptive evolution in the divergence of humans and chimpanzees, two species for which we know a substantial amount about their long-term demographic history; they appear to have undergone a population size contraction since they split. We have investigated whether several site-level factors affect the rate of adaptive and non-adaptive evolution in hominids – relative solvent accessibility (RSA), and measures of physio-chemical (volume and polarity) and an estimate of the average level of negative selection acting on mutations between two amino acids (p_N/p_S). We measure the rates of adaptive and non-adaptive evolution using the statistics ω_a and ω_{na}, which are respectively estimates of the rate of adaptive and non-adaptive evolution relative to the mutation rate. Both statistics were estimated using an extension of the McDonald-Kreitman method (McDonald and Kreitman, 1991), in which the pattern of substitution and polymorphism at neutral and selected sites is used to infer the rates of substitution, taking into account the influence of slightly deleterious mutations. We use the method implemented in GRAPES (Galtier, 2016), which is a maximum likelihood implementation of the second method proposed by Eyre-Walker and Keightley (2009).

3.4.3 Relative solvent accessibility

Previous studies have shown that amino acid residues at the surface of proteins evolve faster than those at the core (Goldman et al. 1998; Choi et al. 2006; Lin et al. 2007; Franzosa and Xia, 2009). These studies do not distinguish whether this higher substitution rate is due to reduced selective constraints on exposed residues or an increased rate of adaptive substitutions (or both). Moutinho et al (2019) disentangled these effects by estimating both the rates of adaptive and non-adaptive evolution across several RSA categories in *Drosophila* and *Arabidopsis*, finding positive correlations between RSA and the rates of both adaptive and non-adaptive substitution. Their findings suggest that both reduced negative selection and a higher rate of adaptive evolution operate on more exposed residues. We find a significant correlation between the rate of adaptive evolution and RSA ($r=0.486$, $p<0.001$) when we use a weighting by the reciprocal of the variance of the rate of adaptive or non-adaptive evolution. However, the correlation with the rate of non-adaptive evolution is non-significant ($r=0.001$, $p=0.324$) (figure 3.1). To check that our grouping scheme did not adversely affect our results, we repeated our analysis randomly allocating genes to RSA bins, estimating the rate of adaptive evolution and re-estimating the slope of the relationship between ω_a and ω_{na}; in none of 100 randomised datasets did we see a correlation as strong as that observed for ω_a in the real data.

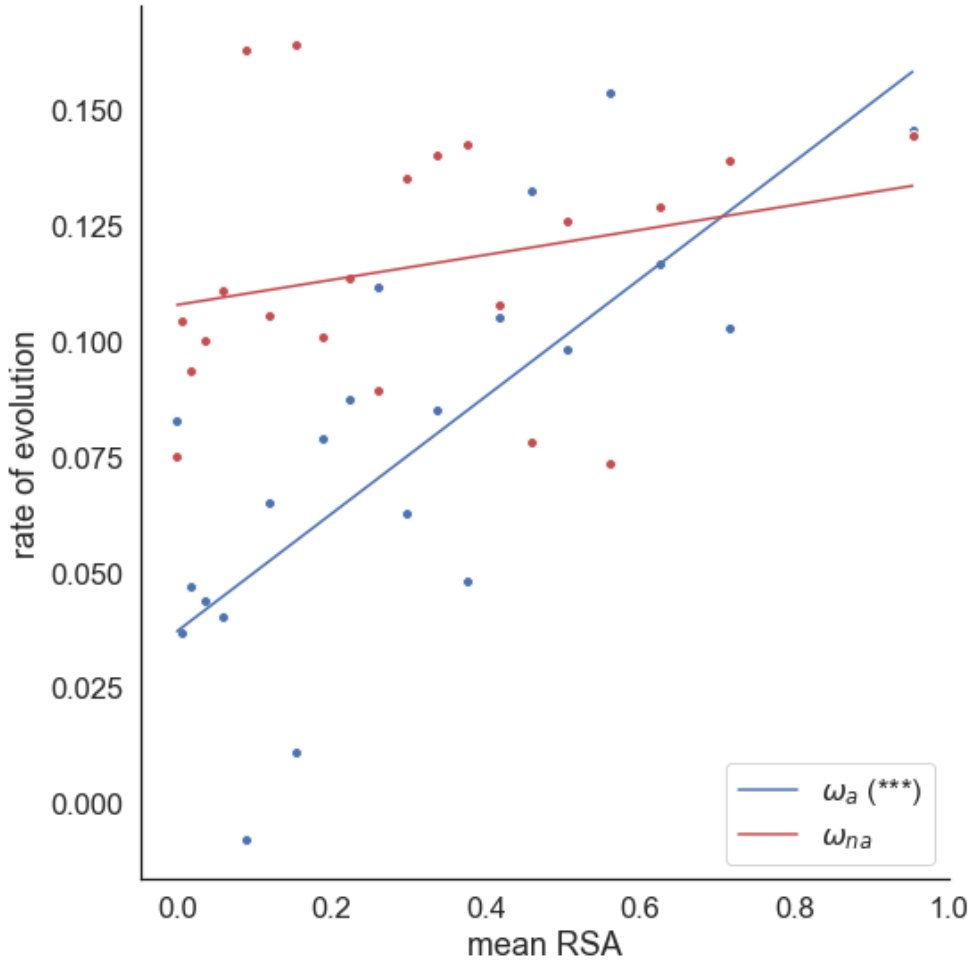


Figure 3.1: Estimates of ω_a and ω_{na} plotted against mean relative solvent accessibility. Data binned into 20 RSA bins of roughly equal number of sites. For each analysis, a weighted linear regression is fitted to the data. The respective significance of each correlation is shown in the plot legend (*P < 0.05; **P < 0.01; ***P < 0.001; "." 0.05 ≤ P < 0.10) for ω_a and ω_{na}). Regression is weighted by the reciprocal of the variance for each estimate of ω_a and ω_{na}, which were estimated by bootstrapping the data by gene 100 times for each data point.

3.4.4 Amino acid dissimilarity

To investigate whether the rates of adaptive and non-adaptive evolution are affected by amino acid dissimilarity, we estimated ω_a and ω_{na} between all 75 pairs of amino acids that are separated by a single mutational step in hominids. Bergman and Eyre-Walker (2019) found

negative correlations between measures of amino acid dissimilarity (differences in volume and polarity) and ω_a between *Drosophila* species. We find that the rate of adaptive substitution is significantly negatively correlated to Δvolume (r = -0.290, p = 0.018) and Δpolarity (r = -0.269, p = 0.027) (figures 2a and 2b) when we fit a weighted linear regression to the data, suggesting that the rate of adaptive evolution is higher between more physiochemically similar amino acids. Similar negative correlations are observed for the rate of non-adaptive evolution (Δvolume: r = -0.545, p < 0.001; Δpolarity: r = -0.170, p<0.001). The slopes are significantly steeper for ω_{na} (Table 3.1); however, this appears to be simply because rates of non-adaptive evolution are greater than rates of adaptive evolution; when we divide ω_a and ω_{na} by their means, the slopes are not significantly different (Table 3.1).

		ω_a		ω_{na}		
Statistic	**Rescaled**	**Slope**	**SE**	**Slope**	**SE**	**Sig.**
Δvolume	No	-0.00027	0.000098	-0.0010	0.00026	0.012
Δpolarity	No	-0.0064	0.0020	-0.022	0.0054	0.010
Δvolume	Yes	-0.0054	0.0020	-0.0051	0.0013	n.s.
Δpolarity	Yes	-0.13	0.042	-0.11	0.027	n.s.

Table 3.1. The slope of the relationship between ω_a and ω_{na} and the Δvolume and Δpolarity; rescaled values are where ω_a and ω_{na} have been divided by their means. Significance was measured using an analysis of variance.

The difference in polarity and volume are not significantly correlated to each other (r=0.122, p=0.258), so it seems likely that both Δvolume and Δpolarity have an influence over the rate of adaptive and non-adaptive evolution. A multiple regression confirms this for ω_{na} with both factors being highly significant and of similar influence, as judged by standardised regression

coefficients (Δvolume b_s = -0.29, p = 0.015; Δpolarity b_s = -0.31, p = 0.008). For ω_a, only Δpolarity is significant (Δvolume b_s = -0.19, p = 0.14; Δpolarity b_s = -0.27, p = 0.036); the loss of significance for Δvolume is probably due to a loss of power due to lack of data; in multiple regression we are effectively holding one variable constant and testing whether the other remains significant.

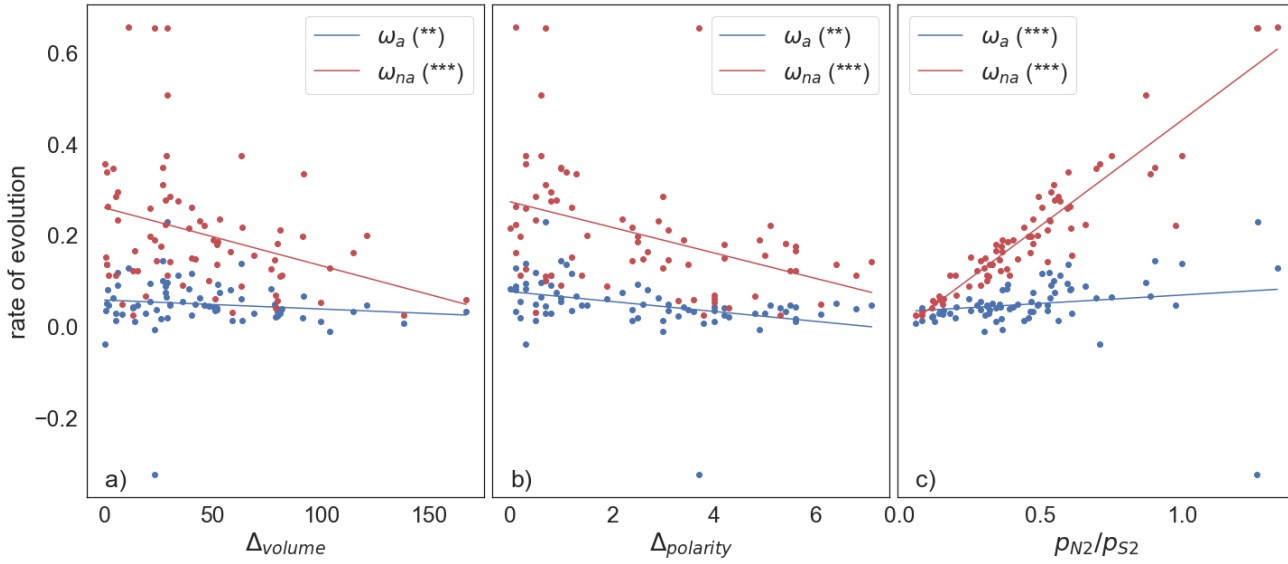


Figure 3.2: The adaptive and non-adaptive substitution rate plotted against the difference in a) volume, b) polarity and c) the ratio of nonsynonymous to synonymous polymorphisms, p_{N2}/p_{S2}. In c) the polymorphisms are split by sampling from a hypergeometric distribution, with one set used to calculate rates of adaptive and non-adaptive substitution and the other to estimate the polymorphism statistics. A weighted linear regression is fitted to the data, weighted by the variance of each estimate. The respective significance of each correlation is shown in the legend (*P < 0.05; **P < 0.01; ***P < 0.001; "." 0.05 ≤ P < 0.10).

Volume and polarity reflect only two of the multiple ways in which amino acids differ. As an alternative measure of amino acid dissimilarity Bergman and Eyre-Walker (2019) suggest using

the ratio of non-synonymous to synonymous polymorphism; p_N/p_S is expected to decrease as the strength of selection against deleterious mutations increases. We find that hominids are consistent with this expectation as p_N/p_S is negatively correlated with both amino acid volume difference ($r = -0.456$, $p < 0.001$) and polarity difference ($r = -0.269$, $p = 0.047$). Polymorphism data is used to estimate both the rates of adaptive and non-adaptive substitution, meaning that p_N/p_S is not statistically independent of either measure. To account for this source of sampling error we follow the method of Bergman and Eyre-Walker (2019), resampling the site frequency spectrum using a hypergeometric distribution to generate two independent spectra. One of these is used to estimate p_N/p_S (referred to as p_{N2}/p_{S2}) and the other is used to estimate ω_a and ω_{na}, therefore removing the nonindependence between p_N/p_S and ω_a and ω_{na}. We find that ω_a is positively correlated to p_{N2}/p_{S2} ($r = 0.419$, $p<0.001$) in hominids, consistent with previous findings in *Drosophila* (Bergman and Eyre-Walker, 2019). Consistent with our physicochemical dissimilarity correlations, ω_{na} is also shows a positive correlation with p_N/p_S, but a stronger one ($r = 0.882$, $p < 0.001$) (figure 3.2c).

It is possible that the correlations between ω_a and ω_{na} and various site level factors are interrelated; for example, the positive correlation between ω_a and RSA might be due to amino acids that are found exposed on the surface of proteins being one mutational step closer to similar amino acids. However, there is no correlation between the average RSA of an amino acid and the average difference in volume or polarity to its one mutation step neighbours (RSA-volume: $r=-0.171$, $p=0.471$; RSA-polarity: $r=0.059$, $p=0.803$ – supplementary figure B1).

3.4.5 Biased gene conversion

Biased gene conversion can potentially impact estimates of the rate of adaptive evolution, since it increases the fixation probability of Weak (W) to Strong (S) alleles relative to S>W neutral alleles, more than it increases levels of W>S polymorphisms relative to S>W polymorphisms; a problem exacerbated by differences in base composition between synonymous and non-synonymous sites (Galtier and Duret, 2007; Berglund et al. 2009; Ratnakumar et al. 2010; Rousselle et al. 2020). To investigate whether the correlation between the rates of adaptive and non-adaptive evolution and our measures of amino acid dissimilarity are due to BGC we restricted the analysis to polymorphisms and substitutions that involve nucleotide changes that are unaffected by BGC – i.e. A<>T and G<>C changes. This reduces our dataset substantially removing 80% of our substitutions and polymorphisms, and reducing the amino acid analysis to just 12 amino acid pairs. However we find that the correlations between ω_a, RSA, Δvolume and p_N/p_S all remain significant with only the correlation to Δ polarity becoming non-significant (RSA: r = 0.260, p < 0.05; Δvolume: r=-0.576, p<0.01; Δpolarity: r=-0.166, p<0.1; p_{N2}/p_{S2}: r =0.796, p<0.001); the correlations between the rate of non-adaptive evolution, ω_{na}, and Δvolume and p_{N2}/p_{S2} remain significant (RSA: r=0.011, p=0.370; Δvolume: r=0.513, p<0.01; Δpolarity: r=0.115, p=0.150; p_{N2}/p_{S2}: r=0.804, p<0.001).

3.4.6 Are the correlations artefactual?

In summary, we have shown that ω_a is significantly positively correlated to RSA and p_N/p_S, and negatively correlated to the difference in polarity and volume. Could these correlations be explained as an artefact of population size contraction. The method we have used to estimate ω_a generates an estimate of the mean absolute strength of selection acting against deleterious mutations. We find that $\log(|\overline{s}|)$ is positively correlated to Δvolume (r=0.205, p=0.08) and Δpolarity (r=0.310, p=0.008) and significantly negatively correlated to p_N/p_S (r=-0.880, p<0.001)

but there is no correlation with RSA (r=-0.088, p=0.704) (Figure 3.3). Thus, if there was no adaptive evolution, or the rate of adaptive evolution was independent of the variable being investigated (e.g. the difference in polarity), then we would expect ω_a to be positively correlated to the difference in volume and polarity, and negatively correlated to p_N/p_S. In fact, we observe the opposite pattern in each case suggesting that these correlations are not an artefact of population size contraction, but are genuine.

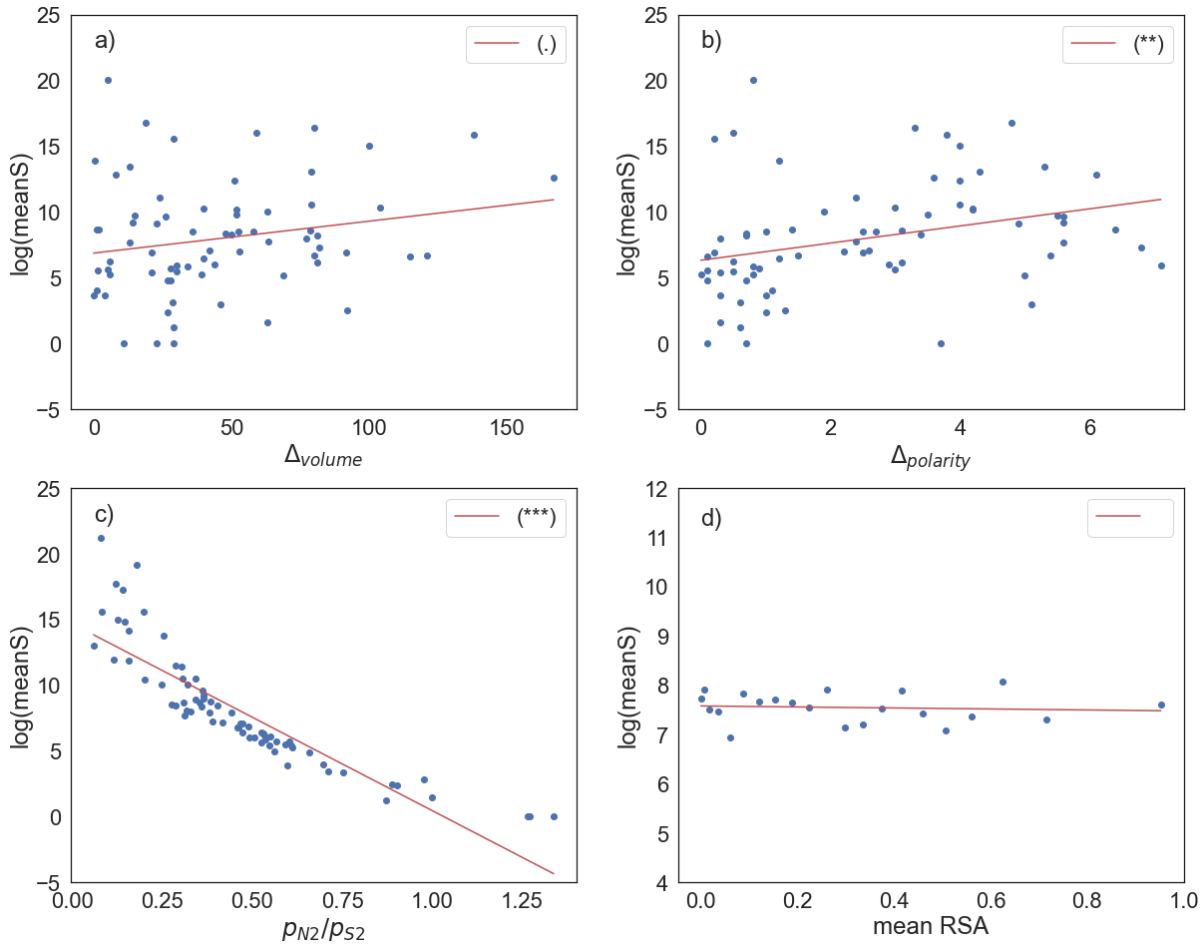


Figure 3.3: log(meanS) plotted against a) volume difference, b) polarity difference, c) p_{N2}/p_{S2}, d) mean RSA. The respective significance of each correlation is shown in the plot legend, (*P < 0.05; **P < 0.01; ***P < 0.001; "." 0.05 ≤ P < 0.10) based on an unweighted regression fit to the data.

3.4.7 Comparison to Drosophila

It is of interest to ask how the slopes of the relationships between ω_a and each factor compares to those previously estimated in *Drosophila* species (Bergman and Eyre-Walker 2019; Moutinho et al. 2019). We find that the slope is not significantly different for RSA, Δvolume and Δpolarity. However, the slope between ω_a and p_N/p_S is much steeper in Drosophilids than in hominids (Table 3.2). This might be because of population contraction. For each genomic variable, population size contraction is expected to reduce the slope of the relationship between ω_a and the factor in the human-chimp comparison, except for RSA which is not correlated to the mean strength of selection. However, the relationship between $\log(|\bar{s}|)$ and p_N/p_S is much stronger and steeper than for the other variables; if we standardise the variables by subtracting the mean and dividing by the standard deviation the slopes between $\log(|\bar{s}|)$ and each factor are: RSA = -0.101, Volume b = 0.862, Polarity, b= 1.30, p_N/p_S = -3.90. Hence, we might expect population contraction to have a disproportionate effect on the relationship between ω_a and p_N/p_S.

		Hominids (this analysis)		Drosophila (Bergman and Eyre-Walker 2019)		Sig.
Dataset	**Independent variable**	**Slope**	**SE of slope**	**Slope**	**SE of slope**	
Original	RSA	0.13	0.029	0.078	0.0065	n.s.
Original	ΔVol	-0.00026	0.00010	-0.00027	0.000061	n.s.
Original	ΔPol	-0.0064	0.0020	-0.0047	0.0011	n.s.
Original	p_N/p_S	0.061	0.019	0.29	0.029	<0.001
Rescaled	RSA	1.6	0.36	1.6	0.13	n.s.
Rescaled	ΔVol	-0.0054	0.0020	-0.011	0.0024	n.s.
Rescaled	ΔPol	-0.13	0.042	-0.18	0.041	n.s.
Rescaled	p_N/p_S	1.3	0.40	11	1.1	<0.001

Table 3.2. Slopes of the regressions between ω_a and measures of amino acid dissimilarity in hominid and *Drosophila* datasets. In the rescaled analyses, the ω_a values have been divided by their mean. The slopes for the *Drosophila* analysis were obtained from the results supplied by Bergman and Eyre-Walker (2019).

3.5 Discussion

One of the main weaknesses of methods that estimate the rate of adaptive evolution using a McDonald-Kreitman type approach, is their sensitivity to changes in the effective population size; with an expansion in population size, these methods overestimate the rate of adaptive evolution, and with a contraction they underestimate it (Eyre-Walker 2002). Here, we demonstrate an additional problem; MK-style methods are also susceptible to producing artefactual correlations between the rate of adaptive evolution, scaled relative to the mutation rate, and another variable, such as amino acid dissimilarity, if that variable is correlated to the mean absolute strength of selection acting against deleterious mutations. This then might call into question previous correlations of this type. For example, it has been observed that p_N/p_S, for pairs of amino acids separated by one mutational step, is negatively correlated to the mean strength of selection in Drosophilids (Bergman and Eyre-Walker 2019); hence the positive correlation between ω_a and p_N/p_S across pairs of amino acids in these species (Bergman and Eyre-Walker 2019) could simply be an artefact of population size expansion, although there is no evidence that population size expansion has affected the species involved. There might be no adaptive evolution, and if there is adaptive evolution, its rate may not be correlated to p_N/p_S. In future, attempts should be made to estimate the mean strength of selection acting against deleterious mutations and investigate whether this is correlated to the factor in question; for example, if we are investigating whether the rate of adaptive evolution is correlated to the rate of recombination, we should investigate whether the mean strength of selection is correlated to the rate of recombination. If it is, then we should be cautious about interpreting our results unless we know something about the demographic history of the species.

Humans and chimpanzees are potentially useful because both their ancestral and current effective population sizes have been estimated; analyses suggest that the human-chimp ancestral population size was considerably larger than the current effective population size of either species (Holboth et al. 2007; Burgess and Yang 2008; Prado-Martinez et al. 2013; Schrago, 2014). Given the correlations we observe between each factor we have considered and the mean strength of selection, we predict, under population size contraction, that the correlations should be opposite to those observed. Hence, it seems that the correlations between ω_a and RSA, Δvolume, Δpolarity and p_N/p_S are all genuine, in hominids at least, and this lends to support to the notion that similar correlations in *Drosophila* and *Arabidopsis* species are also real. However, some caution should be exercised because although we know something about the effective population of the ancestral and current populations of humans and chimpanzees, we know little about the population size between these two timepoints; it is possible the ancestral population contracted shortly after the species diverged and has subsequently re-expanded towards the present; under this scenario the effective population during the divergence phase could have been lower than that during the polymorphism phase.

Population contraction leads to an underestimate of the rate of adaptive evolution when using MK-style methods (McDonald and Kreitman 1990; Eyre-Walker 2002). As a consequence, Zhen et al. (2021) have argued that the rate of adaptive evolution between humans and chimpanzees has been underestimated, and that they have undergone higher rates of adaptive evolution than Drosophila species. In fact, the average of ω_a across amino acid pairs is significantly higher in hominids than Drosophila (hominids, mean ω_a = 0.0488 (SE = 0.0072; Drosophila mean ω_a = 0.0258 (SE = 0.0024); t-test t = 3.01, p<0.001), so hominids seem to be adapting faster relative to the mutation rate even without taking into account population

contraction. What is perhaps surprising is that ω_a is not negative even when we correlate it against factors that appear to influence it. The observed value of ω_a is expected to be equal to

$$\omega_a(obs) = \omega_a(true) + \omega_a(predicted) \quad (7)$$

Where $\omega_a(true)$ is the true value, and $\omega_a(predicted)$ is the value predicted in the absence of adaptive evolution from equation 4 or 6; i.e. it is the bias in the estimate due to the differences in the effective population size between the divergence and polymorphism phases. For example, ω_a is positively correlated to RSA, however, even those sites with very low RSA, have a positive estimate of ω_a.This seems surprising and suggests that adaptive evolution is more prevalent than we thought in hominids. However, predicting how much is difficult because we do not know how the effective population size has changed during the divergence of humans and chimpanzees.

We confirm the findings of Moutinho et al. (2019) with respect to RSA - more exposed amino acid residues have higher rates of adaptive evolution. Moutinho et al. (2019) also showed that the rate of non-adaptive evolution is positively correlated to RSA. These observations are consistent with two models of evolution; either the fitness landscape is relatively flat for more exposed residues, or the mutational steps are relatively small. It is difficult to differentiate between these models.

We also confirm the results of Bergman and Eyre-Walker (2018) – rates of adaptive and non-adaptive evolution are lower between more dissimilar amino acids. It seems likely that these correlations are due to the mutational steps being smaller and hence that adaptive evolution proceeds via small steps in this component of evolution. Chen et al. (2019) apparently came to a different conclusion, but their analysis largely focussed on a statistic that is related to the

proportion of substitutions that are adaptive, and hence conflates the pattern of adaptive and non-adaptive evolution. In fact, consistent with their findings and those of Bergman and Eyre-Walker (2018), we find the proportion of substitutions that are adaptive is uncorrelated to either the difference in volume or polarity (Δvolume: r=-0.012, p=0.707; Δpolarity: r=0.0003, p=0.314).

In summary, we demonstrate that population size change can lead to an artefactual correlation between a measure of adaptive evolution and any variable related to the mean strength of selection against deleterious mutations. Our analysis in hominids suggests that there are genuine negative correlations between ω_a and amino acid dissimilarity and positive correlations between ω_a and RSA and a measure of negative selection acting on mutations between pairs of amino acid mutations, because under population size contraction we would expect the opposite.

We set out to investigate whether several site-level factors affect the rate of adaptive and non-adaptive evolution in hominids – relative solvent accessibility (RSA), and measures of physio-chemical (volume and polarity) and the level of negative selection acting on mutations between two amino acids (p_N/p_S). We measure the rates of adaptive and non-adaptive evolution using the statistics ω_a and ω_{na}, which are respectively estimates of the rate of adaptive and non-adaptive evolution relative to the mutation rate. Both statistics were estimated using an extension of the McDonald-Kreitman method (McDonald and Kreitman, 1991), in which the pattern of substitution and polymorphism at neutral and selected sites is used to infer the rates of substitution, taking into account the influence of slightly deleterious mutations. We use the method implemented in GRAPES (Galtier, 2016), which is a maximum

likelihood implementation of the second method proposed by Eyre-Walker and Keightley (2009).

4. Factors that affect the rates of adaptive and non-adaptive evolution at the gene level in humans and chimpanzees

4.1 Abstract

The rate of amino acid substitution has been shown to be correlated to a number of factors including the rate of recombination, the age of the gene, the length of the protein, mean expression level and gene function. However, the extent to which these correlations are due to adaptive and non-adaptive evolution has not been studied in detail, at least not in hominids. We find that the rate of adaptive evolution is significantly positively correlated to the rate of recombination, protein length and gene expression level, and negatively correlated to gene age. The correlations remain significant when each factor is controlled for in turn, except when controlling for expression in an analysis of protein length; and they also remain significant, or marginally significant, when biased gene conversion is controlled for. However, the positive

correlations could be an artefact of population size contraction. We also find that the rate of non-adaptive evolution is negatively correlated to each factor, and all these correlations survive controlling for each other and biased gene conversion. Finally, we examine the effect of gene function on rates of adaptive and non-adaptive evolution; we confirm that virus interacting proteins (VIPs) have higher rates of adaptive and lower rates of non-adaptive evolution, but we also demonstrate that there is significant variation in the rate of adaptive and non-adaptive evolution between GO categories when removing VIPs. We estimate that the VIP/non-VIP axis explains about 5-8x more of the variance in evolutionary rate than GO categories.

4.2 Introduction

There is substantial variation in the rate of evolution between different genes within a genome; some genes, such as those coding for histones, evolve very slowly, whereas many genes involved in immunity evolve rapidly (Clark et al. 2003; Chimpanzee Sequencing and Analysis Consortium, 2005; Nielsen et al. 2005; Sackton et al. 2007; Obbard et al. 2009). The reasons for this variation have been extensively studied and a number of factors appear to influence or be correlated to the rate of protein evolution including function (e.g. Proschel et al. 2006; Haerty et al. 2007; Obbard et al. 2009), mutation rate (Taddei et al. 1997; Tenaillon et al. 1999; Giraud et al. 2001; Denamur and Matic, 2006; Lynch et al. 2016), recombination rate (RR), gene expression, gene length and position in the protein interaction network. Correlations with other factors, such as essentiality, appear to be less clear (Hurst and Smith, 1999). Any one of these patterns could be due to adaptive or non-adaptive evolution, but the relative roles of these two different evolutionary processes have rarely been studied.

At the functional level, genes involved in immunity, tumor suppression, apoptosis and spermatogenesis have been shown to have higher rates of adaptive evolution in hominids (Clark, et al., 2003; Nielsen, et al., 2005; Chimpanzee Sequencing and Analysis Consortium, 2005). Particularly striking is the amount of adaptive evolution that appears to occur in virus-interacting genes, which appear to account for 30% of all adaptive substitutions in hominids, whilst these genes only constitute 13% of the proteome by length (Enard et al. 2016). In *Drosophila* it has been shown that male-biased genes, such as testes specific genes, have higher rates of adaptive evolution (Proschel et al. 2006; Haerty et al. 2007), as do genes involved in immunity (Sackton et al. 2007; Obbard et al. 2009). The dominant role of VIPs in hominid adaptive evolution begs the question of whether there is variation between other categories of genes, and how much of the variation in the rate of adaptive evolution is partitioned between the VIP and non-VIP categories. The role of gene function in determining non-adaptive evolution has not been addressed in detail.

The rate of protein sequence evolution has been shown to be correlated to gene expression, with highly expressed genes having lower rates of protein evolution in both eukaryotes (Pal et al. 2001; Subramanian and Kumar, 2004; Wright et al. 2004; Lemos et al. 2005) and prokaryotes (Rocha and Danchin, 2004). Moutinho et al. (2019) has shown that this correlation is due to both adaptive and non-adaptive evolution in *Drosophila* suggesting that gene expression constrains the rate of adaptive substitution as well as the effect of purifying selection. In *Arabidopsis* the correlation with expression seems to be largely associated with non-adaptive evolution (Moutinho et al. 2019). The role of gene length has also been studied, with several studies showing that smaller genes evolve more rapidly (Zhang, 2000; Lipman et al. 2002; Liao et al. 2006). Again, this appears to be due to both adaptive and non-adaptive

evolution, in *Drosophila* species, but possibly only due to non-adaptive evolution in *Arabidopsis* (Moutinho et al. 2019).

Genes differ not only in function, expression, and length, but also in age (Lynch, 2002; Daubin and Ochman, 2004; Tautz and Domazet-Loso, 2011; Neme and Tautz, 2013). Multiple studies have shown that phylostraigraphically young genes (i.e. those genes whose recognised homologs are only present in closely related species (Domazet-Loso et al. 2007) evolve faster than old genes (Thornton and Long, 2002; Domazet-Loso and Tautz, 2003; Krylov et al. 2003; Daubin and Ochman, 2004; Alba and Castresena, 2005; Wang et al. 2005; Cai et al. 2006; Wolf et al. 2009; Cai and Petrov, 2010; Zhang et al. 2010; Vishnoi et al. 2010; Tautz and Domazet-Loso, 2011; Cui et al. 2015). Cai and Petrov (2010) found clear evidence for the role of non-adaptive evolution in this relationship but no evidence for adaptive evolution. However, there is an expectation that young genes will be further from their evolutionary optimum than old genes, and hence that they should undergo rapid adaptive evolution when they are born. There is some limited evidence for this; the *jingwei* gene, which appeared very recently in the *Drosophila* phylogeny is evolving very rapidly, with 80% of the amino acid substitutions estimated to have been due to adaptive evolution (Long and Langley, 1993).

Recombination is expected to affect the probability that both advantageous and deleterious mutations are fixed, due to its ability to reduce Hill-Robertson interference between selected mutations (Hill and Robertson 1966; Marais and Charlesworth, 2003). Rates of adaptation have been shown to be strongly positively correlated to recombination rate in *Drosophila* (Presgraves, 2005; Betancourt et al. 2009; Arguello et al. 2010; Mackay et al 2012; Campos et al. 2014; Castellano et al. 2016; Moutinho et al. 2019) and *Arabidopsis* (Moutinho et al. 2019),

and rates of non-adaptive evolution to be negatively correlated in both *Drosophila* and *Arabidopsis* species (Moutinho et al. 2019).

In summary, a number of factors have been shown to correlate to rates of protein evolution, and in some of these cases the relative roles of adaptive and non-adaptive evolution have been disentangled. However, relatively little work has been done on these questions in hominids. We addressed these questions by considering the role of gene age, RR, gene expression, protein length and gene function in determining rates of both adaptive and non-adaptive evolution. To disentangle the effects of adaptive and non-adaptive evolution we use an extension of the McDonald-Kreitman test which estimates these quantities taking into account the distribution fitness effects of new mutations.

4.3 Materials and methods

4.3.1 Data

We obtained orthologous human and chimpanzee gene sequences from Ensembl's biomart (Yates et al. 2019) for the human GRCh38.p14 and Pan_tro_3.0 genome builds. We aligned these orthologos using MUSCLE (Edgar, 2004). After filtering out genes with gaps that were not a multiple of 3 we were left with 16,344 pairwise alignments. Proportions of synonymous and non-synonymous substitutions were estimated using codeml from the PAML package (Yang, 2007) program. We used polymorphism data from the African superpopulation of the 1000 genomes dataset (The 1000 Genomes Consortium, 2015) to construct our site frequency spectra, with rates of adaptive (ω_a) and non-adaptive (ω_{na}) evolution estimated using Grapes (Galtier, 2016), under the "GammaZero" model. We used African SNPs because the African population has been subject to relatively simple demographic processes (Gravel et al. 2011)

Confidence intervals on our estimates of ω_a and ω_{na} were generated by bootstrapping the dataset by gene.

Gene ages were obtained from Litman *et al.* (Litman and Stein, 2019). In this dataset genes are ranked by phylostratigraphic category based on their earliest ortholog. Gene lengths were obtained by taking the total coding sequence length of each protein, whilst gene expression data was obtained from the Expression Atlas database (Papatheodorou et al. 2019). We estimated the mean expression value across tissues for each gene. Recombination rate maps were obtained from Spence and Song (2019), and the mean recombination rate was calculated for each gene. GO category information was obtained from Ensembl's Biomart (Ashburner et al. 2000; The Gene Ontology Consortium, 2021; Yates et al. 2019).

4.3.2 Correlating factors with rates of adaptive and non-adaptive evolution

To correlate the rates of adaptive and non-adaptive evolution with each of recombination rate, protein length and gene expression we binned our genes into 20 roughly equal sized bins. For gene age we binned data by phylostratigraphic category. To control for biased gene conversion in our recombination rate analysis we restricted the analysis to those polymorphisms and substitutions that are unaffected by biased gene conversion – i.e. A<>T and G<>C changes. This reduced our dataset to about 20% of its previous size.

We then reran the analysis for each factor, individually controlling for each of the other three factors in turn. We controlled for each factor by taking the values of the co-correlate close to the modal value. We took the modal value and 0.5 standard deviations either side which

reduces the standard deviation of the co-correlate within each analysis. Because this reduces the data set considerably, we also ran an analysis in which we predicted the correlation coefficient between Y and X under the assumption that they are only correlated to each other because they are both correlated to Z. If R(YZ) is the correlation between Y and Z, then R(YZ)^2 is the proportion of variance in Y explained by Z, and vice versa. Hence, the proportion of variance explained in Y by X, because of their mutual correlation to Z is R(YZ)^2 x R(XZ)^2. Hence the expected correlation coefficient between Y and X is

$$R(YX) = S * \mathrm{Sqrt}(r^2(YZ) * r^2(XZ)) \quad (1)$$

where S = +1 if r(YZ)*r(XZ) is positive and S = -1 if r(YZ)*r(XZ) > 0 is negative. To assess significance we grouped genes according to X variable, and then within each group we generated a bootstrap dataset. We estimated ω_a, ω_{na}, the mean value of X and Z for each group and the observed and predicted correlations between ω_a, ω_{na}, mean X and mean Z. We tabulated the number of bootstrap replicates in which predicted R(YX)/observed R(YX) > 1. We performed 100 bootstrap replicates for each analysis.

4.3.3 Gene function analysis

Genes were divided by GO category and rates of adaptive and non-adaptive evolution were estimated for each category (note genes can contribute to multiple categories). For the VIP analysis we split each GO category into two groups – VIP and non-VIP genes, as per (Enard et al. 2016). To test whether there was significant variation in ω_a and ω_{na} across GO categories we shuffled data between gene labels; i.e. for each gene we have its synonymous and non-synonymous site frequency spectra and numbers of synonymous and non-synonymous substitutions. This data was randomly assigned to gene labels, hence preserving the covariance structure of the data - i.e. the fact that a gene can contribute to multiple GO categories.

We are interested in the extent to which the rate of adaptive and non-adaptive evolution is determined by whether its a VIP gene versus other GO categorisations. We can do this by a partitioning the variance in a two-way analysis of variance where the dimensions are VIP/non-VIP, and GO category. However, to estimate the variances we need to balance the data so that the error variance is the same for all cells in the two-way ANOVA. We did this by downsampling the data using a hypergeometric distribution, such that each cell had 200,000 combined non-synonynous and synonymous sites. To estimate the error variance we split the SFS and substitution data into two halves using a hypergeometric distribution and estimated ω_a and ω_{na} for each set; hence we have for each combination of VIP/non-VIP and Go category two estimates of the rate of adaptive and non-adaptive evolution, where the error variances for these estimates should be approximately equal.

4.4 Results

We set out to investigate whether a number of gene-level factors affect the rate of adaptive and non-adaptive evolution in primates – the rate of recombination (RR), gene age, the level of gene expression, gene length and gene function. We measure the rates of adaptive and non-adaptive evolution using the statistics ω_a and ω_{na}, which are estimates of the rate of evolution relative to the mutation rate. We estimated both statistics using an extension of the MacDonald-Kreitman method, in which the pattern of substitution and polymorphism at neutral and selected sites is used to infer the rates of substitution, taking into account the influence of slightly deleterious mutations. We use the method implemented in Grapes (Galtier, 2016), which is a maximum likelihood implementation of the second method proposed by Eyre-Walker and Keightley (2009). Note that genes are grouped together according to the factors analysed, since most genes have relatively little polymorphism data, and this makes estimating the rate of adaptive evolution for individual genes is impractical.

We estimated ω_a and ω_{na} for 16,344 genes for the divergence between humans and chimpanzees using African SNPs from the 1000 genomes data. We find that the average rate of adaptive evolution is approximately five-fold lower than the rate of non-adaptive evolution (ω_a = 0.037 [0.035,0.039] versus ω_{na} = 0.192 [0.190,0.194]). The proportion of substitutions that are adaptive, alpha, is estimated to be 0.162, which is close to previous recent estimates (Eyre-Walker and Keightley, 2009; Boyko et al. 2008; Messer and Petrov 2013).

4.4.1 Adaptive evolution

The rate of adaptation is expected to be retarded in regions of low recombination because of Hill-Robertson interference, and we do indeed find that the rate of adaptive evolution is significantly positively correlated to the rate of recombination in hominids (Figure 4.1a; ($r=0.737$, $p<0.001$)). A similar positive correlation has previously been observed in *Drosophila* (Presgraves, 2005; Betancourt et al. 2009; Arguello et al. 2010; Mackay et al. 2012; Campos et al. 2014; Castellano et al. 2016). In the most detailed study of this relationship in *Drosophila*, Castellano et al. (2016) found that the rate of adaptive evolution increases with RR, but that it asymptotes, suggesting that above a certain level of recombination, Hill-Robertson interference has little effect. However, we do not observe an asymptote using our grouping scheme (figure 4.1a). However, there is a large difference in average recombination between the two groups with the highest recombination rate. We therefore repeated the analysis with 50 mean recombination bins; although we still observe a significant positive correlation between ω_a and RR ($r=0.582$, $p<0.001$), the analysis failed to reveal a clear signal of an asymptote (Appendix figure C1).

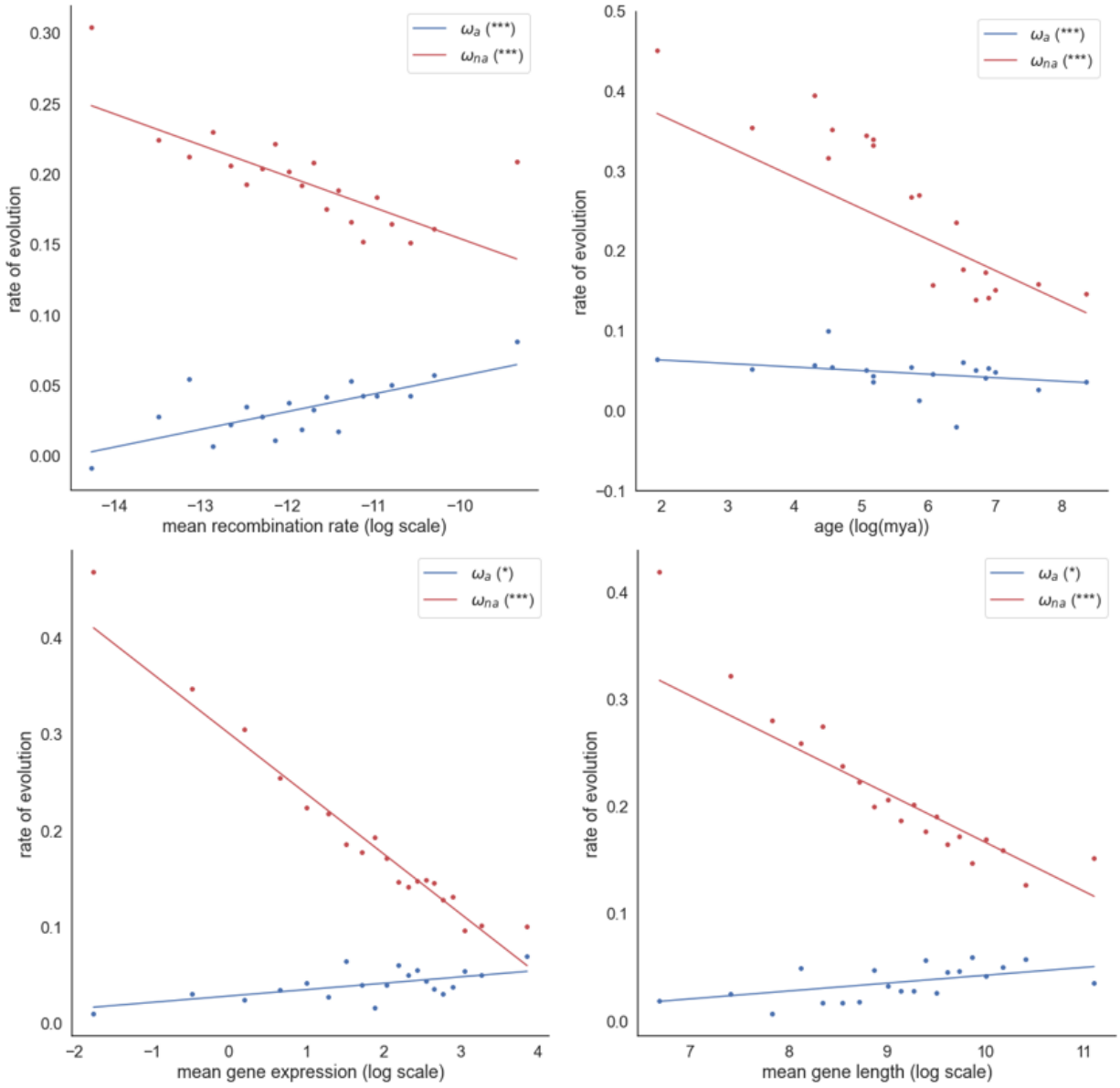


Figure 4.1: Estimates of ω_a and ω_{na} plotted against mean recombination rate (a), gene age (b), mean gene expression (c) and mean protein length (d). The respective significance of each correlation is shown in the plot legend, (*P < 0.05; **P < 0.01; ***P < 0.001; "." 0.05 ≤ P < 0.10) for ω_a and ω_{na}). An unweighted regression is fitted to the estimates of ω_a and ω_{na}. **a)** ω_a and ω_{na} plotted against the natural log of the mean recombination rate for genes binned into 20 recombination bins of equal size. **b)** ω_a and ω_{na} plotted against the natural log of the gene age for genes binned into 19 phylostratigraphic age bins. **c)** ω_a and ω_{na} plotted against the log of the mean gene expression for genes binned into

20 expression bins of equal size. **d)** ω_a and ω_{na} plotted against the log of the mean protein length for genes binned into 20 bins of equal size.

Young genes have been shown to evolve faster than old genes (Thornton and Long, 2002; Domazet-Loso and Tautz, 2003; Krylov et al. 2003; Daubin and Ochman, 2004; Alba and Castresena, 2005; Wang et al. 2005; Cai et al. 2006; Wolf et al. 2009; Cai and Petrov, 2010; Zhang et al. 2010; Vishnoi et al. 2010; Tautz and Domazet-Loso, 2011; Cui et al. 2015). There is an expectation that young genes will undergo faster rates of adaptive evolution because they are further from their adaptive optima (Wright, 1931; 1932), and we find a significant negative correlation between ω_a with gene age (r=-0.404, p=0.012) in hominids (figure 4.1b).

Highly expressed genes have been shown to exhibit lower rates of protein evolution in both eukaryotes (Pal, et al., 2001; Subramanian and Kumar, 2004; Wright, et al., 2004; Lemos, et al., 2005) and prokaryotes (Rocha and Danchin, 2004). Moutinho, et al. (2019) found significant negative correlations in *Drosophila* species between ω_a and both gene expression and protein length. Intriguingly, the correlations are reversed in hominids, with both correlations being significantly positive (gene expression: r=0.642, p=0.002; protein length: r=0.597, p=0.005) (figures 4.1c and 4.1d).

4.4.2 Independent effects

Our measure of adaptive evolution, ω_a, is significantly positively correlated to RR, expression and protein length, and negatively to gene age. However, the rate of recombination, gene age, gene expression and protein length are all significantly, or marginally significantly, correlated to each other (Table 4.1) so it is of interest to determine whether each factor has an

independent effect on the rate of adaptive evolution. The correlation between Y and X, might be due to the fact that each is correlated to a third factor Z, and with no variation in Z there is no correlation between Y and X. To investigate this, we conducted two analyses. In the first instance, we repeated our analyses controlling for each factor in turn by taking the values of the co-correlate close to the modal value. We took the modal value and 0.5 standard deviations either side; this significantly reduced the standard deviation of the co-correlate within each analysis, largely controlling for this factor (Table 4.1). However, controlling for each factor this way reduces the data set considerably, so we also ran an analysis in which we calculated the expected correlation between two variables assuming that the only reason they are correlated is because of their correlation to a third variable. It can be shown that if the correlation between Y and Z is rYZ and that between X and Z is rXZ, then expected correlation between Y and X due to the covariation with Z is rXY = Sign * Sqrt(rYZ^2 rXZ^2), where Sign is positive if both rYZ and rXZ are positive or negative, and negative otherwise.

	gene expression	**gene length**	**recombination rate**	**CV**	**CV of near modal values**
gene age	0.868 (***)	0.860 (***)	-0.621 (**)	1.385	0.381
gene expression		0.437 (***)	-0.035 (***)	1.451	0.411
gene length			0.101 (***)	1.727	0.496
recombination rate				1.143	0.325

Table 4.1: Correlations between the gene age, gene expression, gene length and recombination rate; logs were taken of all variables. The CV column is the coefficient of variation of the factor for all data. The final column is the CV of the restricted data (i.e. when we control for the factor in question by subsetting the dataset to include only genes with the modal value + 0.5 standard deviations).

Our two analyses suggest that there is a direct association between ω_a and RR; when we control for age and length, we find that although the correlation is no longer significant when we control for either variable, the correlation does remain positive, and the observed correlations are significantly greater than predicted correlation. The analysis also suggests that there is a direct association between ω_a and age, because the correlation remains significantly negative when we control for RR, and the predicted correlation is significantly smaller in magnitude than the observed correlation. However, the results with gene expression and length are less clear; when each variable is controlled for in the analysis of the other, the correlation becomes non-significant. The observed correlation between ω_a and expression is marginally significantly greater than the predicted correlation, using length as the covariate, whereas the opposite is not true; this would seem to suggest that there is a direct correlation between ω_a and expression, and that the correlation between ω_a and length is due to the fact that both are correlated to expression. However, the evidence is not strong in support of this hypothesis.

There is another factor that needs to be controlled for in any analysis of age - fast evolving genes are harder to identify in more distant species, and this can lead to an artefactual correlation between the age of a gene and the rate of evolution. The distribution of non-synonymous substitution rates is bimodal, with many genes having $d_N = 0$. We took genes around the second mode, those with rates between 0.002 and 0.007. This reduces our dataset from 15,439 to 4,961 genes, and as a consequence we had to combine multiple age categories together. We find no significant correlation between ω_a and age when we do this (r=0.413, p=0.270), suggesting that the correlation between ω_a and age might be an artifact of the problems in identifying fast evolving genes in older taxa.

Y variate	X variate	Observed r	Z variate	Observed r - controlling for Z	Predicted r	Predicted/observed > 1
ω_a	RR	0.74***	Age	0.25	0.15	0
ω_a	RR	0.74***	Length	0.43	0.086	0
ω_a	Age	-0.40*	RR	-0.58*	-0.093	0.02
ω_a	Expression	0.64**	Length	0.00	0.38	0.03
ω_a	Length	0.60**	RR	0.64**	0.091	0
ω_a	Length	0.60**	Expression	0.25	0.37	0.13
ω_{na}	RR	-0.73***	Length	-0.54*	-0.34	0
ω_{na}	Age	-0.91***	Expression	-0.76**	-0.76	0
ω_{na}	Age	-0.91***	Length	-0.87***	-0.75	0
ω_{na}	Expression	-0.98***	Age	-0.74***	-0.90	0
ω_{na}	Expression	-0.98***	Length	-0.61**	-0.95	0.01
ω_{na}	Length	-0.94***	RR	-0.91***	-0.42	0
ω_{na}	Length	-0.94***	Age	-0.49*	-0.88	0
ω_{na}	Length	-0.94***	Expression	-0.71***	-0.89	0

Table 4.2. The observed and predicted slope between Y and X assuming the relationship is solely due to the correlation between each variable and a third factor Z.

4.4.3 Controlling for BGC

Biased gene conversion can potentially impact estimates of the rate of adaptive evolution, either by increasing the fixation probability of S over W neutral alleles (Galtier & Duret, 2007; Berglund et al. 2009; Ratnakumar et al. 2010; Rousselle et al. 2020), or by promoting the fixation of slightly deleterious S alleles (Duret and Galtier, 2009; Glemin, 2010; Necsulea et al. 2011; Lachance and Tishkoff, 2014; Rousselle et al. 2019). To investigate whether BGC affects our results we can leverage some of the results above. The correlation between ω_a and either age and gene length remains if we control for RR (Table 4.2) (Appendix figures, C3a and C6a respectively), so it seems that BGC is unlikely to be responsible for these correlations. If we control for RR in the regression between ω_a and expression, we find that the correlation remains, suggesting that this correlation is also not due to BGC (r=0.449, p<0.001) (Appendix figure C5a).

To investigate whether the correlation between ω_a and RR is due to BGC we performed a different analysis restricting the analysis to those polymorphisms and substitutions that are unaffected by BGC – i.e. A<>T and G<>C changes. This reduces our dataset to about 20% of its previous size. We find that there is still a positive correlation although this is only marginally significant (r = 0.102, p = 0.093) (Appendix figure, C2).

Hence we can conclude that ω_a is positively correlated to RR, and negatively correlated to gene age. For the gene length and expression analyses, we are unable to convincingly disentangle the effects of these factors from one another and so cannot draw any conclusion about their individual effects on the rate of adaptive evolution.

4.4.4 Non-adaptive evolution

We repeated the analysis above for the rate of non-adaptive evolution. We find that ω_{na} is highly significantly negatively correlated to RR, gene age, length and expression (Table 4.2; Figure 4.1). All of these correlations remain significant when controlling for potentially confounding factors, and the observed correlation is significantly greater in magnitude than the predicted correlation (Table 4.2). Hence, we can conclude that all four factors have significant independent effects on ω_{na}. As with the analysis of ω_a it is possible that these correlations are due to BGC. However, if we control for RR in our analyses we find that all the negative correlations persist (gene age: $r = -0.886$, $p < 0.001$; gene length: $r = -0.910$ $p < 0.001$; gene expression: $r = 0.989$, $p < 0.001$). In the case of the correlation between ω_{na} and RR, if we restrict the analysis to G<>C and A<>T mutations we find that the correlation persists ($r=-0.648$, $p<0.001$).

4.4.5 Gene function

In the second part of our analysis, we consider the effect of gene function on the rate of adaptive and non-adaptive evolution. It has previously been demonstrated that genes whose products interact with viruses – viral interacting proteins (VIPs) – have higher rates of adaptive evolution than other genes in primates (Enard et al. 2016). We confirm this pattern. In our analysis, in which we have used a different method and statistic to estimate the rate of adaptive evolution, we find that the rate of adaptive evolution amongst VIPs is approximately 40% greater than in non-VIPs ($\omega_a = 0.052$ versus 0.032), a difference that is highly significant ($p<0.001$). This pattern is consistent across almost all GO categories that have at least 100 genes, supporting the results of Enard et al. (2016) (figure 4.2).

It is evident however, that there is substantial variation between GO categories for non-VIP genes, and this variation is significant, taking into account that individual genes can contribute to multiple GO categories (p=0.0012). This pattern is replicated if we include GO categories which do not include VIP proteins (p=0.0010). The GO categories which have the highest rate of adaptive evolution are ubiquitin protein ligase binding, and protein kinase binding (table 4.3).

GO category	ω_a	$\omega_{a\ low}$	$\omega_{a\ high}$
ubiquitin protein ligase binding	0.0843	0.0702	0.0995
protein kinase binding	0.0804	0.0698	0.0914
sequence-specific DNA binding	0.0735	0.0633	0.0842
DNA-binding transcription factor activity	0.0719	0.0628	0.0812
transcription factor complex	0.0682	0.0496	0.0883
transcription by RNA polymerase II	0.0673	0.0518	0.0836
negative regulation of apoptotic process	0.0671	0.0552	0.0796
chromatin organization	0.0669	0.0567	0.0775
DNA-binding transcription activator activity	0.0649	0.0524	0.078
transcription coactivator activity	0.0648	0.0519	0.0786

Table 4.3: Top 10 GO categories, ranked by rate of adaptive substitution.

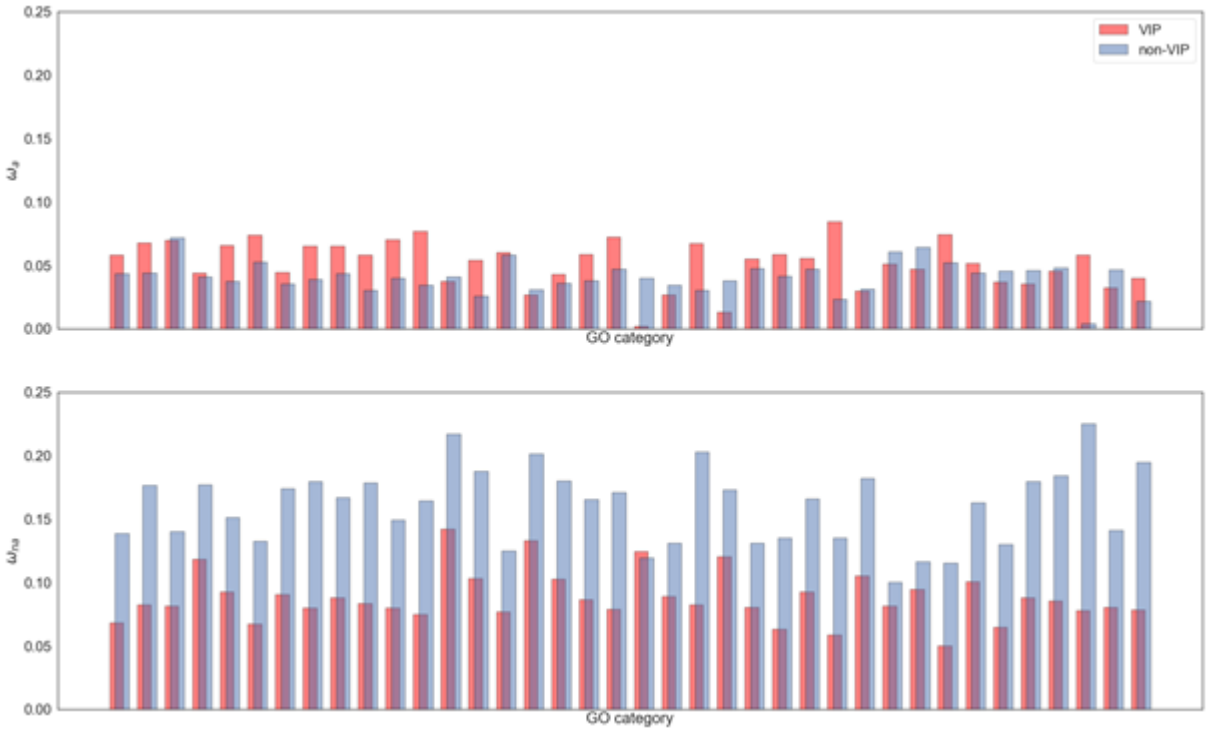


Figure 4.2: Estimates of ω_a (top) and ω_{na} (bottom) for GO categories that contain >100 viable VIP and non-VIP genes.

What are the relative contributions of GO category and VIP status to the variation in the rate of adaptive evolution – i.e. is most of the variation in the rate of adaptive evolution due to whether the gene encodes a VIP or not, or is most of the variation due to other functional considerations? To investigate this, we performed a two-way analysis of variance on ω_a and estimated the variance components. We find that the distinction between VIP and non-VIP contributes approximately 5x the variance in ω_a as the variation between GO categories, suggesting that whether a gene encodes a VIP has a major effect on its rate of adaptation (Appendix table, C2).

But what of non-adaptive evolution? If we divide our data into genes that interact with viruses and those that do not, we find that rates of non-adaptive evolution are substantially higher in

non-VIP genes (ω_{na} = 0.198 vs 0.101), as Enard et al. (2018) found, a pattern that is replicated across GO categories (Figure 4.2). There is substantial and significant variation in ω_{na} across GO categories excluding VIP genes ($p<0.001$). If we partition the variance between VIP/non-VIP and GO categories we find that the distinction between VIP and non-VIP contributes over 8x the variance in ω_{na} as the variation between GO categories, suggesting that whether a gene encodes a VIP has a major effect on its rate of non-adaptive evolution (Appendix table, C3) as well as its rate of adaptation.

There is substantial variation in the rate of non-adaptive evolution between GO categories for non-VIP genes, and this variation is significant, taking into account that individual genes can contribute to multiple GO categories ($p > 0.001$). This pattern is replicated if we include GO categories which do not include VIP proteins ($p > 0.001$). The GO categories that have the highest non-VIP rates of non-adaptive evolution are both related to immune system response (table 4.4).

GO category	ωn_a	$\omega n_{a\ low}$	$\omega n_{a\ high}$
immune system process	0.297	0.283	0.310
innate immune response	0.264	0.248	0.279
chromosome	0.262	0.249	0.274
protein C-terminus binding	0.246	0.228	0.264
centrosome	0.243	0.232	0.253
DNA repair	0.236	0.223	0.249
signal transduction	0.225	0.219	0.231
neutrophil degranulation	0.218	0.206	0.229
extracellular region	0.217	0.211	0.223
proteolysis	0.204	0.195	0.214

Table 4.4: Top 10 GO categories, ranked by rate of non-adaptive substitution

4.5 Discussion

It has been previously shown that the rate of evolution correlates to a number of factors including RR (Presgraves, 2005; Betancourt et al. 2009; Arguello et al. 2010; Mackay et al 2012; Campos et al. 2014; Castellano et al. 2016; Moutinho et al. 2019), gene age (Thornton and Long, 2002; Domazet-Loso and Tautz, 2003; Krylov et al. 2003; Daubin and Ochman, 2004; Alba and Castresena, 2005; Wang, et al., 2005; Cai, et al., 2006; Wolf, et al., 2009; Cai and Petrov, 2010; Zhang et al. 2010; Vishnoi et al. 2010; Tautz and Domazet-Loso, 2011; Cui, et al., 2015), expression level (Pal et al. 2001; Rocha and Danchin, 2004; Subramanian and Kumar, 2004; Wright et al. 2004; Lemos et al. 2005; Moutinho et al. 2019) and protein length (Zhang, 2000;

Lipman et al. 2002; Liao et al. 2006; Moutinho et al. 2019). In addition, the rate of evolution has been shown to vary with gene function (Clark et al. 2003; Nielsen et al. 2005; Chimpanzee Sequencing and Analysis Consortium, 2005). In this study we have correlated each of these factors to ω_a and ω_{na} in hominids, allowing us to disentangle the effects of adaptive and non-adaptive evolution. We find that ω_a is correlated to all four factors, but that when we control for each factor in turn, there is evidence for an independent influence of RR, gene age and probably gene expression. These correlations remain when controlling for the effects of biased gene conversion as well. However, the correlation with gene age could be an artefact of fast evolving genes having higher rates of adaptive evolution and being more difficult to identify in older taxa; when we control for the rate at which a protein evolves the negative correlation between ω_a and gene age becomes non-significant suggesting that this pattern might be an artefact.

In contrast, we find that all four factors have significant independent effects on ω_{na}, and that all of these remain significant when we control for each in turn. Several studies on both Eukaryotes (Pal et al. 2001; Subramanian and Kumar 2004; Wright et al. 2004; Lemos et al. 2005; Moutinho et al. 2019) and Prokaryotes (Rocha and Danchin 2004) have demonstrated that more highly expressed genes have lower rates of protein sequence evolution. Our results support these previous findings, with the negative correlation between ω_{na} and gene expression suggesting that more highly expressed genes are under greater constraint in hominids. Drummond et al. (2005) suggest a general hypothesis that more highly expressed genes evolve slowly (i.e. are under higher selective constraint) because of the selection against the expression level cost of protein misfolding, wherein selection acts by favoring protein sequences that accumulate less translational missense errors. We also find a significant negative correlation between ω_{na} and gene length. This supports former studies that have

shown that smaller genes evolve more rapidly (Zhang 2000; Lipman et al. 2002; Liao et al. 2006; Moutinho et al. 2019), suggesting that smaller protein-coding regions are under more relaxed purifying selection.

4.5.1 Gene function analyses

Our analyses of VIP and non-VIP genes show that a high proportion of the variance in protein evolution in hominids is accounted for by whether or not a gene interacts with viruses, a result that corroborates Enard et al.'s (2016) findings. By disentangling the rates of adaptive and non-adaptive evolution, we find that VIP genes are under greater constraint than nonVIPs, and that despite this greater level of constraint, VIPs exhibit a higher rate of adaptive evolution. We also estimate the variance components using two-way analyses of variance, finding that the distinction between VIP and non-VIP contributes about 5x the variance in ω_a, and 8x the variance in ω_{na} as the variation between GO categories, suggesting that whether a gene encodes a VIP has a major effect on its rate of adaptation and non-adaptation (Appendix table, C2). These results could explain why there appears to be little variation in the rate of adaptive evolution across biological functions categorised using Gene Ontology (Bierne and Eyre-Walker, 2004), with viruses acting across a range of biological functions likely to be a key factor in these estimates.

Our study is likely to underestimate the amount of adaptive evolution attributable to viruses, for reasons outlined by Enard et al (2016). Briefly, we used the categorisation of VIPs and non-VIPs provided by Enard et al (2016). However new VIPs are being discovered regularly, suggesting there are many VIPs that were not included in our analysis. Secondly, the categorisation of VIP and non-VIP necessarily cannot account for proteins that adapt to viruses

but do not physically interact with them (e.g. in proteins that are upstream or downstream of VIPs in signalling cascades).

4.5.2 No asymptote in the correlation between ω_a and RR

Both Campos et al. (2014) and Castellano et al. (2016) found that there is a positive relationship between the rate of adaptive evolution and RR in *Drosophila*. However, Castellano et al. (2016) showed that the positive correlation between RR and ω_a asymptotes in *Drosophila,* suggesting that above a certain level of recombination Hill-Robertson interference has little effect. In this study we find no evidence of this asymptote in hominids for either the rate of adaptive or non-adaptive evolution, suggesting that most coding sequences may experience some level of HRi. This is perhaps not unexpected. The level of HRi will depend on several factors - the effectiveness of recombination in breaking down associations, the density of selected sites and the mutation rate to alleles that are subject to selection; if weakly selected mutations are responsible for HRi then the effective population size and the level of nearly neutral genetic diversity will also be important. Recombination is a considerably more effective force in Drosophila; linkage disequilibrium (LD) decays over a scale of 10s of base pairs (Mackay et al. 2012) rather than the 10,000s that we observe in humans (The 1000 Genomes Project Consortium, 2015). This 1000-fold difference in the effectiveness of recombination is likely to more than compensate for the fact that humans have ~20-fold greater genome size, and a higher rate of deleterious mutation (2.1 in humans (Lesecque et al., 2012) to 1.2 in Drosophila (Haag-Liautard et al., 2007) respectively).

4.5.3 Gene age

Cai and Petrov (2010) found that older genes exhibit a lower rate of protein evolution (as measured by the Ka/Ks ratio) than younger genes. The authors demonstrated that this was at least in part due to stronger purifying selection acting on older genes than on younger ones, by showing that levels of non-synonymous to synonymous polymorphism were lower in older genes. Our findings corroborate these results, with the strong negative correlation between ω_{na} and gene age showing that older genes are under a lower rate of protein evolution than younger genes. However, we also find a significant negative correlation between gene age and the rate of adaptive evolution, ω_a, whilst Cai and Petrov found no such correlation. There are two potential causes of this discrepancy. Firstly, for this analysis Cai and Petrov group genes by their age based on lineage specificity (LS), that is, how specifically a gene and orthologs of a gene are distributed on a given phylogeny (Cai et al. 2006), whilst we group our genes by phylostratigraphic category (PL), that is, where genes are ranked by phylostratigraphic category based on their earliest ortholog (Domazet-Loso et al. 2007). Each method has its limitations. Because the LS method relies on the phylogenetic profiles of individual genes, Cai and Petrov removed genes with patchy distributions (Cai et al. 2006), resulting in 10,032 of 20,150 genes being removed from the dataset for having irregular phylogenetic profiles. The PL method relies on parsimony and assumes that a gene family can be lost, but cannot re-evolve in different lineages (Domazet-Loso et al. 2007), meaning that those genes that would be removed using the LS method are maintained in the PL method. By using the PL method, our dataset contained 15,439 grouped into 19 phylostratigraphic bins. Secondly, Cai and Petrov obtained divergence and polymorphism data from the compiled Applera dataset (Bustamante et al. 2005; Lohmueller, et al., 2008) of 39 humans (19 African Americans and 20 European Americans), whilst we have used data from the 661 African samples within the 1000 genomes dataset (The 1000 Genomes Project Consortium, 2015). Notably, the African population has undergone a more stable demographic history than Europeans, who carry

proportionally more deleterious genetic variation, which Lohmueller et al (2008) ascribe to the bottleneck encountered by the Eurasian population at the time of the migration out of Africa. This higher proportion of segregating deleterious alleles will inevitably affect estimates of the rate of adaptive evolution, but not the Ka/Ks ratio (the latter of which yields a strong correlation with gene age using both the PL and LS methods in Cai and Petrov's study).

4.5.4 The effect of population contraction

It has been shown previously that the MK test can generate artifactual evidence of adaptive evolution if some nonsynonymous mutations are slightly deleterious and the population in question has undergone recent expansion, because selection is more effective during the polymorphism phase than during the divergence phase (McDonald and Kreitman, 1991; Eyre-Walker, 2002). Although, the effective population size in humans has increased recently, the effective population size is considerably reduced from that in the human-chimpanzee ancestor (Hobolth et al. 2007; Burgess and Yang 2008; Prado-Martinez et al. 2013; Schrago, 2014). This population contraction can depress the signal of adaptive evolution in humans. Furthermore, we show elsewhere (unpublished ref) that if a factor, for example gene age, is correlated to the mean strength of selection against deleterious mutations, population size change will generate an artifactual correlation between that factor and the rate of adaptive evolution. The direction of this correlation depends on the direction of the correlation between the mean strength of selection acting against deleterious mutations and the factor in question and whether the population has expanded or contracted; for example, if factor X is positively correlated to the mean strength of selection (i.e. selection is stronger against genes with large values of X), then population contraction will induce an artifactual positive correlation between ω_a and X.

Figure 4.3 shows that all four factors are positively correlated to the mean strength of selection against deleterious mutations, estimated from the site frequency spectrum (gene age: r=0.916, p<0.001; RR: r=0.828, p<0.001; gene length: r=0.818, p<0.001; gene expression: r=0.948, p<0.001). Population contraction undergone by humans should therefore tend to induce a positive correlation between ω_a, gene age and RR. This artifactual positive correlation is contrary to the negative correlation that we observe (Figure 4.1). This may be one reason why we observe a weaker correlation between gene age and the rate of adaptive evolution in hominids compared with *Drosophila* and *Arabidopsis* species (Moutinho et al. unpublished). However, population contraction might also be responsible for the positive correlation between ω_a and RR.

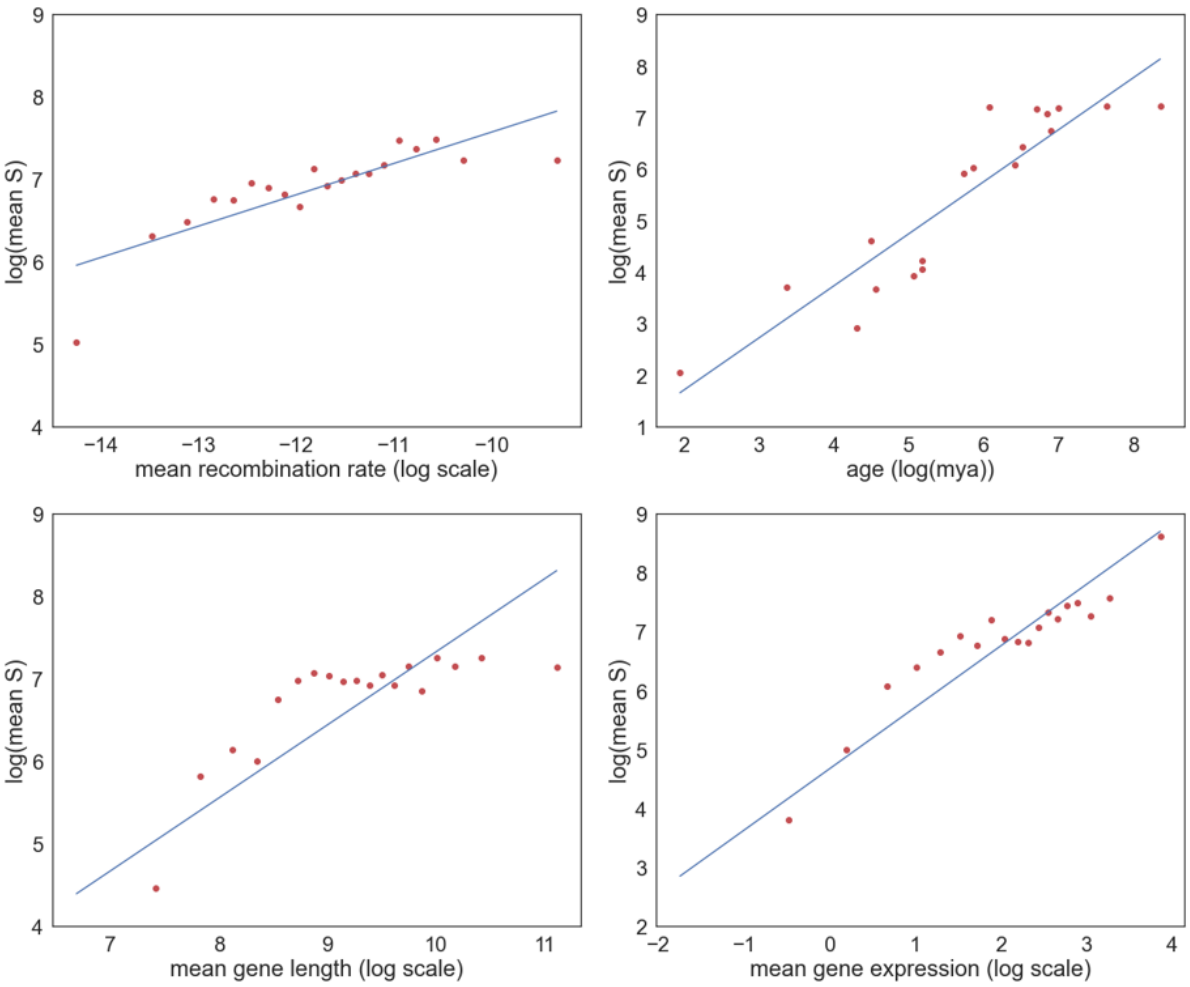


Figure 4.3: Correlation between the log of the mean strength of selection against deleterious mutations and gene age (top left), RR (top right), gene length (bottom left), gene expression (bottom right). A linear regression has been fitted to each dataset.

Because ω_{na} is estimated exclusively from polymorphism phase data, we do not expect the correlations between ω_{na} and our four factors to be attenuated by the population contraction.

In summary, we observe a significant correlation between the rate of adaptive evolution, RR, gene age, protein length and gene expression. However, we can only be confident that there is a genuine correlation between ω_a and expression; the correlation between ω_a and RR might be

due to an artifactual effect of population size contraction, and the correlation between ω_a and age might be due to the problems of identifying rapidly evolving genes, with high values of ω_a, in more distant taxa. The correlation between ω_a and length could be due to the fact that both are correlated to gene expression. In contrast, the rate of non-adaptive evolution is independently negatively correlated to all factors. We quantify the proportion of variance in the rate of adaptive and non-adaptive evolution that is due to whether a gene interacts with viruses or not, and show that this single factor explains 5-8 more variance, than any other categorisation of genes encapsulated in GO categories.

5. Why does genetic variation vary so little across the human genome?

5.1 Abstract

Genetic diversity varies across the human genome, however the reasons for this variation are not fully understood. We examine variation in the level of genetic diversity in non-overlapping 10KB windows masking those regions of the genome that are thought to be directly affected by natural selection. We show that diversity varies by ~3-fold across the genome and that this variation in genetic diversity is correlated to rate of mutation, the density of selected sites and the rate of recombination. These correlations suggest that the level of diversity is affected by both the mutation rate and effects of selection at linked sites. We estimate the distribution of mutation rates using de novo mutation (DNM) data, and find that the distribution is broader than the distribution of SNP densities. We also find that the slope of the relationship between SNP and DNM densities is shallower than might be expected. We demonstrate that both of these observations are consistent with a model in which the effects of linked selection in reducing diversity, are dependent upon the mutation rate. However, a simple model implies that the effects of linked selection are widespread, reducing genetic diversity by an average of

50%; the effects of linked selection are also highly variable, generating some regions with minimal effects of linked selection and others in which diversity is reduced by 80%.

5.2 Introduction

Genetic diversity is known to vary across the genomes of many species. The primary evidence for this variation comes from the observation that the level of genetic variation is correlated to the rate of recombination, which was first described in *Drosophila melanogaster* in the landmark paper by Begun and Aquadro (1992); such correlations have been observed in diverse multicellular animals, plants and fungi (Cutter and Payseur 2013) and even some bacteria (Vigue and Eyre-Walker, 2019).

This variation in diversity might arise from several sources. It is known that the mutation rate can vary across the genome. This has been particularly well studied in humans and other primates (Matassi et al. 1999; Webster et al. 2004; Tyekucheva et al. 2008; Terekhanova et al. 2017). The analysis of *de novo* mutation (DNM) data from human trios (an offspring and their parents) provides particularly compelling evidence for this variation (Michaelson et al. 2012; Francioli et al. 2015; Smith et al. 2018). In some species, in which there is a correlation between diversity and RR, there is also a correlation between divergence in putatively neutral sequences and RR suggesting that RR is mutagenic and the mutation rate varies across the genome (Cutter and Payseur, 2013). This was first observed in humans (Lercher and Hurst, 2002; Hellmann et al. 2003; Hellmann et al. 2005), however it has also been observed in a small number of other species (Cutter and Payseur, 2013).

The variation in diversity could also be due to the direct or indirect effects of natural selection. The density of selected sites varies across genomes, however, in many species the density of selected sites is so low that it is thought that this not a major source of variation in diversity across the genome; for example, only 5-10% of the human genome is thought to be subject to selection (Chiaromonte et al. 2003; Smith et al. 2004; Cooper et al. 2005; Asthana et al. 2007; Garber et al. 2009; Davydov et al. 2010; Meader et al. 2010; Pollard et al. 2010; Ponting and Hardison, 2010; Linblad-Toh et al. 2011; Ward and Kellis, 2012; Rands et al. 2014). Furthermore, most analyses attempt to control for these effects by focussing on diversity that is likely to be neutral. However, selection can have indirect effects on diversity through the processes of genetic hitch-hiking (Maynard Smith and Haigh, 1976) and background selection (Charlesworth et al. 1993); these are often characterized as causing variation in the effective population size across the genome. Evidence for the variation in the effective population size across the genome comes from three sources. First, in many species there is a correlation between diversity and the RR (Cutter and Payseur, 2013). This potentially could be due to recombination being mutagenic, but there is no correlation between a measure of the mutation rate, the divergence between species and the RR, in most species in which there is a correlation between diversity and the RR – in Cutter and Payseur's (2013) compilation studies only 3 out of 14 studies that show a correlation between diversity and RR, also show a correlation between divergence and RR. However, some caution should be exercised with these results, because it is known from studies in humans that the mutation rate evolves relatively rapidly at both a regional (Terekhanova et al. 2017; Smith et al. 2018) and a site level (Harris, 2015). Variation in divergence can also be a consequence of variation in the depth of the genealogy in the ancestor of the two species, something that is well documented in humans (McVicker et al. 2009). The divergence between species may therefore not give a good estimate of the variation in the mutation rate. The second line of evidence for variation in the effective population size comes from a negative correlation between diversity and the density

of linked sites, that has been observed in a number of species (Cutter and Payseur 2013; Castellano et al. 2019); this is expected because the effects of linked selection are expected to depend on both the RR and the number of selected sites. The third line of evidence for variation in the effective population size comes from a study in which Gossman et al. (2011) estimated the distribution of N_e by simultaneously considering diversity and divergence at putatively neutral sites; they found significant evidence of variation in N_e across the genomes of 6 out 10 species, but the estimates of the variation in N_e were modest. Finally, it has been observed that a measure of the efficiency of natural selection, the ratio of the number of non-synonymous to synonymous or non-coding diversity, is correlated to diversity across genomes (Gossman et al. 2011; Murray et al. 2017; Castellano et al. 2018; Castellano et al. 2020; Chen et al. 2020), consistent with variation in the effective population size across a genome.

In many species biased gene conversion is thought to act (Duret and Galtier 2009), and this can potentially affect diversity across a genome, although this has not been extensively studied, despite the fact that biased gene conversion is widespread (Eyre-Walker 1993; Montoya-Burgos et al. 2003; Meunier and Duret 2004; Webster et al. 2004; Webster et al. 2005; Spencer et al. 2006; Mancera et al. 2008; Escobar et al. 2011; Pessia et al. 2012; Lesecque et al. 2013; Williams et al. 2015; Halldorsson et al. 2016; Smeds et al. 2016; Keith et al. 2016; Long et al. 2018; Galtier et al. 2018; Smith et al. 2018) and there is a correlation between diversity and the RR. Finally, there is an expectation that diversity will vary simply because of variation in the coalescent process; i.e. we expect variation in diversity even for neutral loci with the same mutation rate, no linked selection and no biased gene conversion, because the genealogy will vary between regions. In estimating the variation in N_e, Gossman et al. (2011) considered models in which there is no intra-locus recombination and

hence substantial variation in the genealogy; they still find significant variation in diversity that could not be attributed to variation in the genealogy or the mutation rate.

Humans were one of the first species in which a correlation between diversity and RR was described (Nachman et al. 1998; Nachman et al. 2001). However, in contrast to Drosophila, subsequent studies also showed that the divergence between humans and other species was correlated to the RR, which suggested that the correlation might be due to a mutagenic effect of recombination (Lercher and Hurst, 2002; Hellmann et al. 2003; Hellmann et al. 2005), a conjecture which is supported by more recent investigations of recombination and mutation (Pratto et al. 2014; Arbeithuber et al. 2015). However, despite the mutagenic effects of RR, Hellmann et al. (2005) concluded that diversity was correlated to the RR, even controlling for this mutational effect, suggesting a role for linked selection in determining diversity levels across the human genome. Recently, Castellano et al. (2019) have revisited this question in humans and other hominids and shown that levels of putatively neutral diversity at the 50KB scale are correlated equally to the RR, the density of selected sites and a measure of the mutation rate, the divergence in putatively non-coding sequences. However, both the studies of Hellman et al. (2005) and Castellano et al. (2019) used divergence between species as a measure of the mutation rate, and given the speed at which the mutation rate evolves (Harris 2015; Tekehanova et al. 2017; Smith et al. 2018), this might be a relatively poor measure of the mutation rate that pertains to extant genetic diversity. Using DNM data, Smith et al. (2018) show that more than 70% of the variation in diversity at the 100KB and 1MB scale can be explained in terms of variation in the mutation rate. This then might suggest that there is relatively little variation in N_e across the genome, although this variation can be detected. In their analysis, Castellano et al. (2019) show that a measure of the effectiveness of selection,

the ratio of non-synonymous to non-coding SNPs, is correlated to both RR and the level of non-coding diversity in several hominid species.

Here we revisit the question of what factors determine the level of neutral diversity across the human genome using SNP and DNM data. We confirm previous results – that the level of diversity is correlated to measures of the mutation rate, the density of selected sites and the rate of recombination. However, we demonstrate that the inferred distribution of mutation rates is broader than the distribution of SNPs. This leads us to explore a model in which the effects of linked selection are dependent upon the mutation rate.

5.3 Materials and methods

5.3.1 Data

Human variation data was obtained from 1000 genomes Grch37.p13 vcf files (The 1000 Genomes Project Consortium, 2015)**.** Variants were annotated using Annovar's hg19 database (Wang and Li, 2010). We considered SNPs from European populations since the *de novo* mutation (DNM) data had been obtained from European individuals (Wong et al. 2016, Jonsson 2017).

DNM data from the studies of Wong et al. (2016) and Jonsson et al. (2017) were obtained from the supplementary materials of the papers. Wong et al. do not specify the nucleotide change associated with the DNM.

Recombination rate maps were obtained from Spence and Song (2019).

5.3.2 Statistical analysis

Variation in the number of DNMs and SNPs per window is composed of two factors, systematic variation in the underlying process, for example the mutation rate, and sampling error; this is particularly pertinent to the distribution of DNMs because we have less than halve a DNM on average per window. To estimate the underlying mutation rate, or in the case of SNPs, the product of the mutation rate, effective population size and mean genealogy length, we assume that the rate of the underlying process is gamma distributed, and that the observed number of DNMs or SNPs is Poisson distributed around the expectation (although it is not certain that SNPs will be Poisson distributed). For example, the mutation rate per site for window *I* might be u_i; we therefore expect to observe on average lku_i DNMs where *l* is the length of the window and *k* is a constant which reflects the sampling scheme (the number of chromosomes sampled and the demography of the population). The observed number of DNMs is assumed to be a Poisson variate with a mean of lku_i. The resultant distribution is a negative binomial and we can estimate the underlying gamma distribution by maximising the likelihood.

It is helpful to approximate the gamma distribution using a lognormal distribution since the product or ratio of two lognormally distributed variates is itself lognormal; for gamma distributions with shape parameters > 6 the fit is good (appendix figure D3). To infer the relationship between the gamma distribution and best fitting lognormal, we generated 100,000 random samples from gamma distributions with shape parameters between 2 and 1024 and a mean of one. To this data we fit a log normal distribution using maximum likelihood. The relationship between the shape parameter of the lognormal distribution and the log of the shape parameter of the gamma distribution is well approximated by a 4th order polynomial (appendix figure D12): $\sigma = 1 - 0.4534 \ln(\beta) + 0.08222 \ln(\beta)^2 - 0.006194 \ln(\beta)^3 +$

$0.0001141\ \ln(\beta)^4$ where σ is the shape parameter of the lognormal distribution and β is the shape parameter of the gamma distribution. Both distributions are assumed to have a mean of one.

To investigate whether the slope of the regression between SNP density and DNM density is what we would expect in a model in which the effective population size of a genomic region is independent of the mutation rate we simulated data as follows; for the *i*th window we sampled a relative mutation rate, u_i, (scaled such that the mean is one) from a lognormal distribution with a shape parameter estimated from the DNM data. This mutation rate was multiplied by the number of sites present in the genome build and with Phastcons score <0.5, w_i, and the mean number of DNMs per site, $\overline{D}$, to generate the expected number of DNMs for the window.

$$\widehat{D_i} = u_i w_i \overline{D}$$

We then generated a random Poisson variate with this expectation. To simulate the number of SNPs in the window we followed a similar process, multiplying the mutation rate by the number of bases in the window, the mean number SNPs per base pair, $\bar{S}$, and another random variate drawn from a lognormal distribution, h_i, representing the variation in the effective population size and the mean genealogy length.

$$\widehat{S_i} = u_i w_i h_i \bar{S}$$

Again, we generated a random Poisson variate with this expected value. For each window we calculated the number of DNMs and SNPs per site, and we divided this by the mean number of DNMs and SNPs per site across windows so the normalised mean number of DNMs and SNPs per window was unity.

We investigated the fit of models in which the effective population size was a function of the mutation rate. We followed a simulation scheme similar to that described above, but in generating the expected number of SNPs in each window we assumed that the effective population size was a function of the mutation rate $f(u_i)$ and we multiplied this by a random variate drawn from a lognormal distribution which represents the variation in N_e not due to variation in the mutation rate and the variation due to variance in the mean genealogy length, k_i.

$$\widehat{S_i} = u_i w_i k_i \bar{S} f(u_i)$$

We simulated 250,000 windows of 10,000bp each. As in the simulation above we calculated the number of DNMs and SNPs per site for each window and then normalised these so the mean across windows in each case was equal to one. We estimated the slope of the regression between the number SNPs and DNMs per site, and fit a lognormal to the distribution of the number of SNPs per site. Since we have two unknown parameters in our model, the parameter governing the relationship between the effective population size and the mutation rate, and the parameter associated with the distribution of residual variation in N_e and the mean genealogy length, and two observations, it should be possible to fit the model perfectly to the data under many circumstances. However, fitting the model is not straightforward since the values from the model are not known without error, because they are simulated; most maximisation algorithms would struggle. We therefore took the following approach. We initially generated simulated datasets in which the residual variation was assumed to be zero, since this residual variation does not affect the slope of the relationship between the number of SNPs and DNMs per site. We varied the parameter governing the relationship between N_e and the mutation rate over parameter values close to that value that would fit the data; this was found by trial and error. We then regressed the slope from the simulation against the

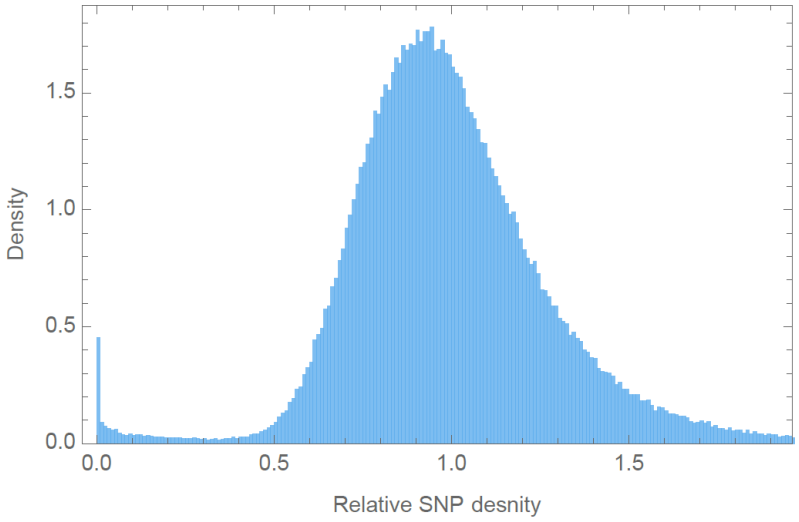
parameter value and used this to estimate the parameter value that best fit the data (see appendix figure D6 for an example). Having obtained this value, we simulated datasets using this value and varying the parameter associated with the residual variation. Again, we regressed the shape parameter of the distribution fitted to the simulated data against the parameter value and used linear regression to infer the parameter value that best fit the data. The observed value of the slope of the regression between the numbers of SNPs and DNMs per site and the shape parameter of the lognormal fit to the distribution of SNPs per site are known with little error; however, the shape parameter of the mutation rate distribution is estimated with some error; we take into account this error by estimating the parameters of the best fitting model for the 95% CIs for the shape parameter.

5.4 Results

Genetic diversity is known to vary across the human genome. To quantify this variation, we divided the genome up into 10KB windows; this is the smallest window size for which we can reliably estimate the distribution of mutation rates from *de novo* mutation data, something that is important for understanding the variation (see below). The distribution of SNPs per bp varies substantially from windows that contain no SNPs to those that have more than 3-times the average diversity (Figure 5.1a). The regions of the genome with very little diversity may be due to problems in calling SNPs in certain regions of the genome; on average there are 85 SNPs per window in the dataset that we are using and many of these windows have 10KB in the human genome build that these SNPs have been mapped to, but there are some windows with no SNPs. We therefore trimmed the data excluding the regions of the genome with the 1.5% of the lowest diversity values. After this trimming there is still substantial variation with some regions having half the diversity of the mean and others with almost twice as much variation (Figure 5.1b). This distribution of SNPs per window is reasonably well described by a log-

normal distribution with a shape parameter of 0.26 (SE = 0.00) (with the mean normalised to unity) (Figure 5.1b). Fitting a distribution is useful for subsequent modelling work, and the log-normal is convenient because the product or ratio of two log-normally distributed variates, is itself log-normally distributed.

A)



B)

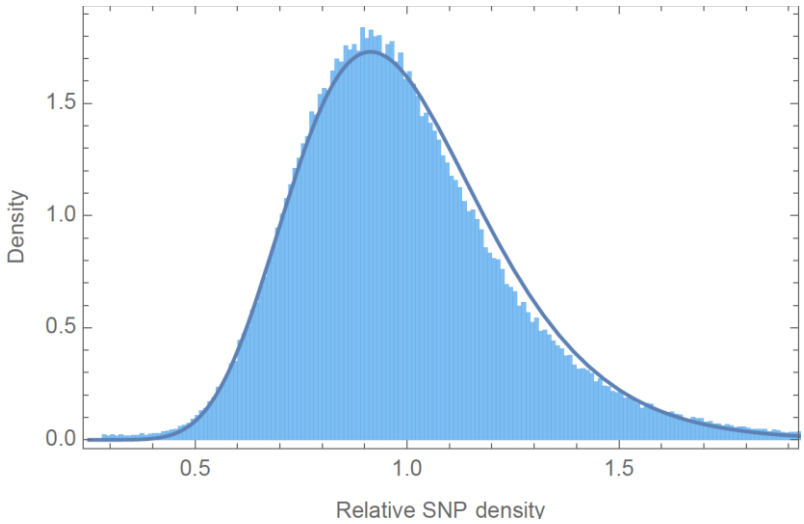
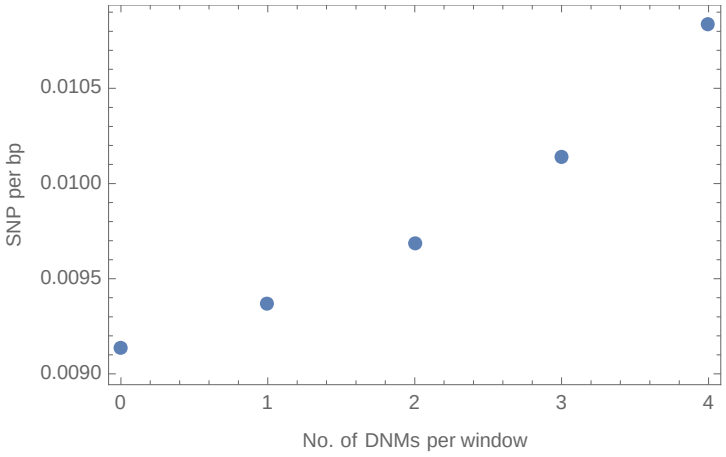
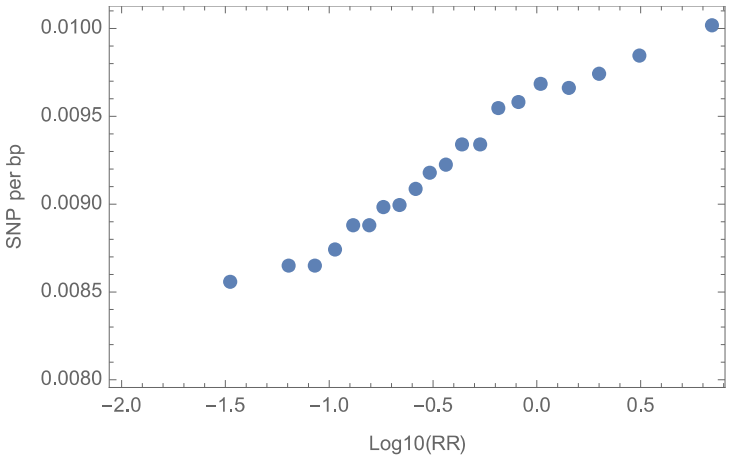


Figure 5.1: The distribution of the number of SNPs per bp, normalised so the mean is one. A) All windows, B) excluding the 1.5% of windows with the lowest levels of diversity. A log-normal distribution is fitted to this data.

A)


SNP per bp
0.0105
0.0100
0.0095
0.0090
0
1
2
3
4
No. of DNMs per window

B)


SNP per bp
0.0100
0.0095
0.0090
0.0085
0.0080
−2.0
−1.5
−1.0
−0.5
0.0
0.5
Log10(RR)

C)

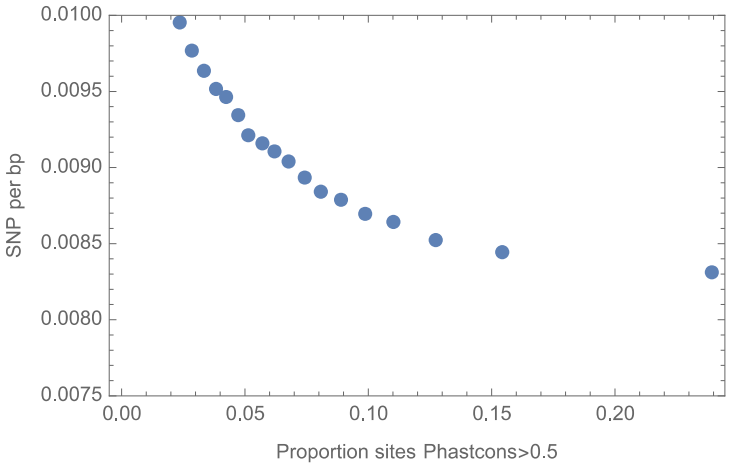

SNP per bp
0.0100
0.0095
0.0090
0.0085
0.0080
0.0075
0.00
0.05
0.10
0.15
0.20
Proportion sites Phastcons>0.5

Figure 5.2: The number of SNPs per bp in 10KB windows as a function of (A) DNM density, (B) log(RR) and (C) the density of selected sites. The data have been binned into 20 equal sized bins in panels (B) and (C).

There are a number of potential reasons for this variation in diversity across the human genome: it could be due to variation in the mutation rate, the direct or indirect effects of selection, or biased gene conversion. To investigate the role that mutation rate variation might have on genetic diversity we used two sets of *de novo* mutation (DNM) data; ~98,000 autosomal DNMs from Jonsson *et al.* (2017) and ~25,000 autosomal DNMs from Wong et al. (2016). We find there is a highly significant correlation between SNP and DNM density (Jonsson: $r = 0.056$, $p < 0.001$; Wong: $r = 0.015$, $p < 0.001$) (Figure 5.2A); the correlations are low in part because there is substantial sampling error associated with the DNM data, since there are < 0.5 DNMs per window in each dataset. We can estimate how strong the correlation might be if all the variation in SNP density was due to mutation rate variation by following the method of Francioli et al. (2015); we assume that the SNP density yields an error free estimate of the mutation rate, and then simulate DNMs according to this distribution of mutation rates. In this simulation we observe the expected correlation between SNP and DNM density is 0.18 and 0.093 for Jonsson and Wong datasets respectively; i.e. the correlations are 31 and 16% as strong as they could be if all the variation in diversity was due to variation in the mutation rate. This is in sharp contrast to a similar analysis at the 100KB and 1MB scales in which the observed correlation is about 70% the expected value (Smith et al. 2018).

Direct selection probably generates relatively little variation in diversity across the human genome, because only ~8% of the human genome is estimated to be under the effects of selection (Rands et al. 2014). Furthermore, we have excluded all sites with Phastcons scores >

0.5. In contrast, the indirect effects of selection are thought to widespread in the human genome and decrease diversity by an average of 15-20% (McVicker et al. 2009; Murphy et al. 2021). Consistent with this, we find highly significant correlations between diversity and log(RR) ($r = -0.16$, $p<0.001$) (Figure 5.2B), and diversity and the density of selected sites, as inferred from those with Phastcons scores > 0.5 ($r = -0.19$, $p < 0.001$) (Figure 5.2C). In a multiple regression these two factors and DNM density all remain highly significant ($p < 0.001$). Perhaps surprisingly the effects associated with RR and the density of selected sites are fairly similar as judged by standardised regression coefficients (RR: $b_s = 0.15$; density of selected sites: $b_s = -0.19$; DNM per bp: $b_s = 0.058$). However, it should be appreciated that there are unknown levels of sampling error associated with each of these variables (Phastcons score is only a proxy for the sites under selection) and hence the multiple regression should be interpreted with caution.

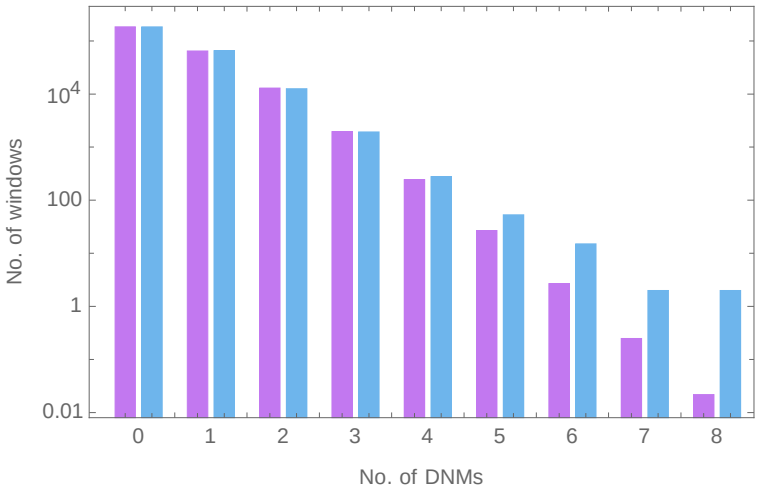
The effects of biased gene conversion on the levels of genetic diversity have not been extensively studied. If we focus our analysis on SNPs that are not subject to BGC, we find the distribution of SNP density is very similar to the distribution including all SNPs (appendix figure D1); and the fitted lognormal distribution has a shape = 0.30, not very different to shape estimated using all SNPs of 0.26. This suggests that BGC is not a major factor generating variation in diversity across the genome; if anything, it decreases the variance.

5.4.1 Distribution of mutation rates

We have shown that diversity across the genome is correlated to the mutation rate and that it is also likely affected by linked selection. To further investigate the role that mutation rate variation plays in the distribution of diversity across the genome we estimated the distribution of mutation rates across the genome from the distribution of DNMs per window by fitting a

gamma distribution of rates across windows, taking into account the sampling error associated with having so few SNPs. The two DNM mutation datasets give similar estimates of the distribution of mutation rates (Jonsson gamma shape parameter $\beta = 7.13$; Wong $\beta = 5.34$) (appendix figure D2). A gamma distribution fits the distribution of DNMs for the Jonsson and Wong DNM datasets, although in the case of the Jonsson data, a goodness-of-fit rejects the gamma distribution ($p < 0.0001$); this is primarily driven by 17 windows that have 6 or more DNMs (Figure 5.3).

A)



B)

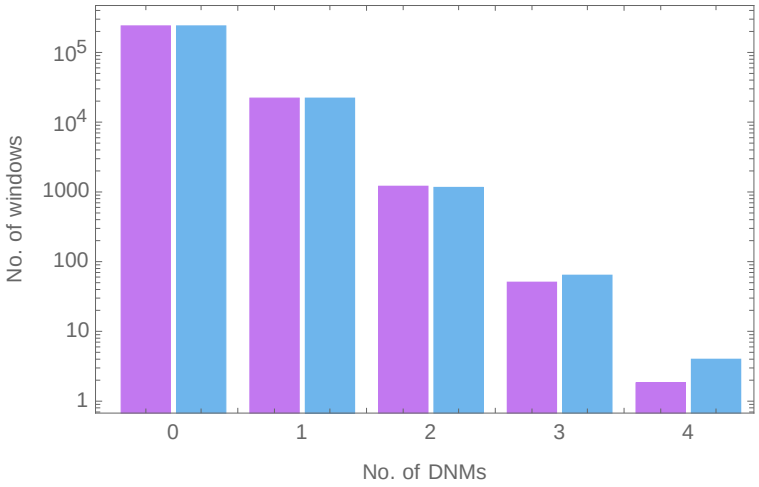


Figure 5.3: The observed (blue) and expected (magenta) number of windows with a certain number of DNMs per window for A) Jonsson and B) Wong DNMs. The expected number is from the fitted gamma distribution. Note the Y-axis is on a log scale.

If we approximate the gamma distribution with a log-normal distribution so that the distribution of mutation rates and SNP density are comparable we find the shape parameter of the mutation rate distribution (Jonsson shape = 0.39 (SE = 0.00); Wong shape = 0.45 (0.01)) is

significantly broader than the distribution of SNPs (0.26 (0.00)). The log-normal distribution approximates the gamma well for these parameter values (appendix figures D3, D4).

It is surprising that the mutation rate distribution is so broad, since we might naively have expected the distribution of SNPs to have a larger variance than the distribution of mutation rates since the number of SNPs in a window is a product of the mutation rate, the effective population size and the average genealogy length; each of these factors should generate variance in addition to that associated with the variance in the mutation rate. There are a number of potential reasons why the mutation rate distribution might be broader than the SNP distribution. First, the method to estimate the distribution of the mutation rate might be biased upwards given the very low number of DNMs per window in the two datasets – 0.36 and 0.092 DNMs per 10KB in Jonsson and Wong respectively. To investigate this we simulated data; we find that the method is slightly biased upwards when the shape parameter is greater than 0.3, but not sufficiently to explain the discrepancy (Table 5.1).

As a further test of whether we are overestimating the variation in the mutation rate we estimated the number of substitutions that have occurred along the human lineage since the split between humans and chimpanzees in our 10KB windows. We find that a log-normal distribution with a shape parameter of 0.32 (0.00) fits the distribution well (appendix figure D5). As expected, the distribution is narrower than the distribution of mutation rates, but it is still significantly broader than the distribution of SNPs.

Number of DNMs per window	Shape	Mean simulated shape (SE)
0.092	0.2	0.15 (0.02)
0.092	0.3	0.31 (0.02)
0.092	0.4	0.42 (0.01)
0.092	0.5	0.56 (0.01)
0.36	0.2	0.21 (0.01)
0.36	0.3	0.30 (0.01)
0.36	0.4	0.43 (0.00)
0.36	0.5	0.57 (0.00)

Table 5.1. Performance of the method to infer the shape parameter of the mutation rate distribution. Ten simulated datasets of 250,000 windows were generated and the shape parameter estimated as in the data analysis; note that the data were simulated assuming the mutation rate was lognormally distributed; we then estimated the rate assuming the rate was gamma distributed before approximating this gamma with a log-normal distribution.

The second potential explanation for why the mutation rate distribution is broader than the SNP distribution, is that the mutation rate has evolved such that the mutation rate that pertains to the polymorphism data is not the same as that measured in pedigree studies. The mutation rate at the 100KB and 1MB scale is known to evolve (Terekhanova et al. 2017; Smith et al. 2018) but the speed at which it would have to evolve to explain the discrepancy between the SNP and mutation rate distributions would have to be very rapid. If we imagine that the mutation rate for a region changes instantaneously to a new value drawn from some distribution then the variance in the average mutation rate after t episodes is V/t. The average shape parameter of the mutation rate distribution is 0.39 and 0.45 for the Jonsson and Wong datasets, and this translates into variances of 0.16 and 0.21; the shape parameter for the SNP

distribution is 0.26 and this translates into a variance of 0.070. Hence the mutation rate in region would have had to change completely 2-3-times over the average age of polymorphisms in the population to explain the discrepancy between the inferred variance in the mutation rate and variance in the distribution of SNPs. Given that the distribution of substitution density along the human lineage is only slightly narrower than the distribution of mutation rates, this explanation is clearly not correct.

A third possible explanation for the discrepancy between the SNP and mutation rate distributions is biased gene conversion; if the strength of BGC is correlated to the mutation rate, then it might reduce the variance in SNP density. DNM density is correlated to log(RR) but the relationship is very weak ($r = 0.026$, $p < 0.001$); the mutation rate is also very weakly negatively correlation to the proportion of sites with Phastcons scores > 0.5 ($r = -0.0076$, $p < 0.001$). However, as we have shown above, the variance in SNP density is largely unaffected by restricting the analysis to mutations that are unaffected by BGC – G<>C and A<>T mutations (appendix figure D1).

The most likely explanation is that the effective population size depends on the mutation rate. This is not unexpected; it is believed that the effective population varies across a genome through the action of linked selection, either in the form of background selection or genetic hitch-hiking. The strength of background selection is expected to depend on the mutation rate (Charlesworth et al. 1994; Nordborg et al. 1996; Hudson and Kaplan 1995). In the case of hitch-hiking, the effective population size is expected to depend on the mutation rate if adaptation is limited by the supply of mutations, for which there is some evidence (Gossmann et al. 2012; paper by Besenbacher et al. 2019; Rousselle et al. 2020; though see Galtier 2016).

To test whether the effective population size is correlated to the mutation rate we considered the slope of the regression between the number of SNPs per site and the number of DNMs per site. Under a model in which the effective population size and mutation rate are uncorrelated, SNP density should be linearly related to the mutation rate with an intercept of zero and a slope of one, if we normalise the SNP and DNM densities such that they are unity. Unfortunately, this simple prediction only applies if we know the mutation rate without error, which we do not. Sampling error in the DNMs per window will reduce the predicted slope. To investigate whether the observed slope is less than we might predict under a model in which the mutation rate and N_e are uncorrelated, we simulated data under a model in which we randomly sampled mutations rates, mean genealogy lengths and effective population sizes from lognormal distributions. The number of DNMs was generated as a random Poisson variate of the mutation rate. For each simulated dataset we estimated the slope of the regression between the number of SNPs per window, and the mean DNMs per window (both normalised so their mean was unity). We investigated the effects of varying the shape parameter of each distribution, but as expected only the shape parameter of the mutation rate distribution affected the slope; the shape parameters of the mean genealogy length and N_e distributions have no effect on the expected slope because they only affect the dependent variable.

The observed slope between SNP and DNM density using the DNMs from Jonsson is 0.0096 (0.0003) and this is significantly lower than the mean simulated slope of 0.054 (0.000); for the Wong DNMs the corresponding slopes are 0.0013 (0.0002) and 0.0080 (0.000), and these are also significantly different; hence the observed slope is less than expected under a model in

which the mutation rate and effective population are uncorrelated to each other; the slopes are consistent with the effective population being negatively correlated to the mutation rate.

To investigate the matter further we attempted to fit a model in which N_e is a function of the mutation rate. We have two observations that the model needs to explain: the slope of the relationship between SNP and DNM density, and the variation in the distribution of the SNP density. In our model we randomly sampled mutation rates (u) from a log-normal distribution with the shape parameter estimated from the DNM data, and from this we generated a simulated number of DNMs using a Poisson distribution. The expected N_e was related to the mutation rate assuming a model of exponential decay: $f(u) = e^{-\beta u}$, where β is a constant which describes how fast N_e declines as a function the mutation rate, u. This is the expected relationship under a model of background selection. We also randomly sampled a variate from a lognormal distribution representing the product of the residual variation in effective population size (i.e. variation in the effective population size not explained by variation in the mutation rate) and the mean genealogy length - k_i. The expected number of SNPs in the ith window is therefore equal to $\widehat{S_i} = u_i w_i k_i \bar{S} f(u_i)$ where w_i is the number of nucleotides in the window, $\bar{S}$ is the average number of SNPs per site and $f(u_i)$ is the function that relates the N_e to the mutation rate. The realised number of SNPs was generated as a random Poisson deviate with this expectation.

Our simple model fits the data and the parameter estimates are similar for the Jonsson and Wong datasets of DNMs (Table 5.2) (Figure 5.3). Note that we have two observations, which are estimated with very little error, and two parameters in the model, so in principle the model can fit perfectly. However, our estimate of the shape parameter of the mutation rate distribution is subject to some level of uncertainty so we fit our model using the maximum

likelihood estimate of the mutation rate shape parameter, and its 95% confidence intervals, inferred from the likelihood surface. The model predicts that diversity is substantially reduced by linked selection (Figure 5.3); with a mean reduction of ~50%. Under this model there is also substantial variation in N_e across the genome due to variation in the mutation rate (Figure 5.4).

DNM data	S_u	Linked selection parameter	S_{gn}	Mean diversity relative to no linked selection
Jonsson lower 95%	0.36	0.72	0.23	0.50
Jonsson ML	0.38	0.74	0.23	0.50
Jonsson upper 95%	0.40	0.75	0.22	0.49
Wong lower 95%	0.37	0.83	0.24	0.46
Wong ML	0.44	0.84	0.22	0.46
Wong upper 95%	0.51	0.84	0.19	0.47
Jonsson	0.30	0.66	0.24	0.53
Jonsson	0.25	0.55	0.24	0.58
Jonsson	0.20	0.34	0.22	0.71
Wong	0.30	0.81	0.25	0.46
Wong	0.25	0.75	0.25	0.48
Wong	0.20	0.64	0.25	0.53

Table 5.2. Parameter estimates under a simple exponential model in which the N_e is a function of the mutation rate ($N_e(u) = e^{-\beta u}$). Parameters are estimated for the ML estimates of the shape parameter of the mutation rate distribution, S_u, and its 95% confidence intervals. S_{gn} is the shape parameter of the distribution for the residual variation.

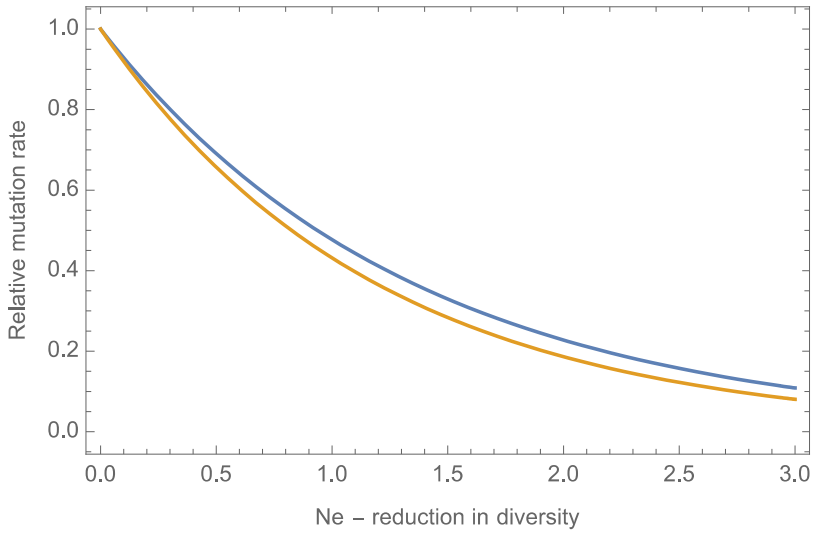


Figure 5.3. The relationship between N_e and the relative mutation rate, normalised such that the mean is one, estimated assuming $N_e(u) = e^{-\beta u}$ using estimates of the mutation rate distribution from the Jonsson (blue) and Wong (yellow) DNM datasets.

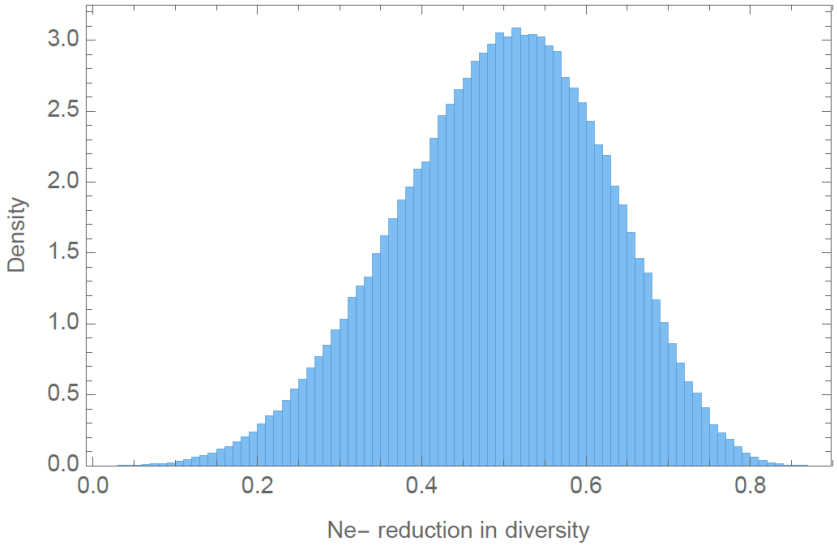


Figure 5.4. The distribution in effective population size, the degree to which diversity is reduced by linked selection, caused by variation in the mutation rate under the model in which $N_e(u) = e^{-\beta u}$.

5.5 Discussion

There is substantial variation in the level of genetic diversity across the human genome. We have investigated the factors that are correlated to this variation in diversity and find, as others have done, that both the mutation rate (Castellano et al. 2019) and linked selection (McVicker et al. 2009; Hernandez et al. 2011; Murphy et al. 2021) seem to play a role. However, we estimate the distribution of mutation rates from *de novo* mutation data, and find that the distribution of mutation rates is substantially broader than the distribution of SNP density. This is surprising since the density of SNPs should depend upon the mutation rate, the effective population size and the mean genealogy length, so we might naively expect the total variance in SNP density to be greater than the variance in the mutation rate. It seems likely that the variance in SNP density is smaller than the variance in the mutation rate because the effects of linked selection depend on the mutation rate and hence as the mutation rate increases, so diversity tends to increase, but at the same time the effective population size is decreased. We show that a simple model, in which the effective population size is correlated to the mutation rate, can explain two salient observations – why the distribution of SNP density is so narrow, and why the slope of the relationship between SNP and DNM density is so shallow. However, the model implies that background selection is extremely prevalent in the human genome, reducing diversity by an average of ~50%. The model also suggests that the effects of linked selection vary across the genome substantially, with some regions experiencing a decrease in diversity of ~20% and others ~80% simply due to variation in the mutation rate; this is not factoring in variation in N_e due to variation in the density of selected sites or recombination rate.

Our model is attempting to explain two observations; why the variance in the mutation rate distribution is greater than the variance in the density of SNPs, and why the slope of the

relationship between SNP and DNM density is shallow. We should be cautious about our estimate of the mutation rate distribution since this is inferred from DNM data, and it is evident that there are biases in these datasets; for example, the density of DNMs depends on GC-content, but this relationship differs between different DNM datasets, including the two used here, at the 100KB and 1MB scale (Smith et al. 2018). It is therefore possible that the variance in the mutation rate distribution is being over-estimated due to variance associated with these biases; although, the two datasets give similar estimates for the amount of variation there is at the 10KB scale. To investigate how sensitive our analyses are to the estimate of the mutation rate distribution we simulated data assuming the mutation rate distribution had less variance. As expected, we find that as we reduce the variance in the mutation rate distribution, the estimated relationship between the effective population size and mutation rate becomes less steep (Table 5.2). As a consequence, the inferred level of background selection is reduced, but even halving the shape parameter of the mutation rate distribution yields a mean reduction in diversity of 29%.

Is our model credible? The model predicts that on average linked selection reduces diversity by an average of 50% in the human genome. This is far higher than two previous estimates of 15-20% (McVicker et al. 2009; Murphy et al. 2021). These two analyses and ours take very different approaches to inferring the level of linked selection over very different scales. McVicker et al. (2009) and Murphy et al. (2021) identify sites that might be subject to selection from conservation across species. They then estimate the distribution of fitness effects that under a background selection model would generate the observed levels of diversity. The model of Murphy et al. (2021) explains an impressive 60% of the variance in diversity at the 1MB scale, although it only explains 15% of the variance at the 10KB scale; the relatively poor performance of the method at smaller scales might be due to variance in the mutation rate or

genealogy length, variance that is averaged out at larger scales or it might be due to the inability to identify those sites that are under selection and contributing to background selection. In contrast, we are inferring the presence of linked selection and its scale indirectly. Potentially these two approaches could be combined.

How much of the genome needs to be subject to selection to reduce diversity by 50% on average? In their analysis, Murphy et al. (2021) find that the 6% of sites with the highest CADD scores, an evolutionary measure of the selection acting at a site, reduce diversity by approximately 17% across the human genome. Hence, a reduction of 50% in our model might be compatible with ~18% of the human genome being subject to selection; this is much greater than previous estimates – for example, Rands et al. (2014) estimate that only 8% of the human genome is constrained by natural selection. However, although Rands et al. (2014) take into account the turnover of constrained DNA they assume that all sites within a functional category turn-over at the same rate; if there is substantial variation in turnover rate then the amount of DNA subject to selection might be much greater than previously thought. In Drosophila, Charlesworth (2012) estimated that background selection reduces diversity by about 50%; however, Drosophila has a far shorter genetic map length than humans (290cM (Catcheside, 1977) versus 3600cM (Ott, 1999), and higher density of selected sites.

The model allows there to be residual variation that affects the density of SNPs across the genome, and we estimate the shape parameter of this distribution to be 0.23 and 0.22 using the DNM data from Jonsson et al. (2017) and Wong et al. (2016) respectively. This residual variation comes from two sources; variation in the mean genealogy length and variation in the effective population size that is not associated with variation in the mutation rate; this arises through variation in the density of selected sites and the recombination rate. We have found

that whilst diversity is correlated to both recombination rate and the density of selected sites, neither explains very much of the variance in diversity. How much variation in the genealogy length do we expect? To investigate this, we simulated a 10kb locus subject to a level of recombination, that was homogeneous across the locus, for a sample size of 1000 chromosomes – the European sample from the 1000 genome project contains 1006 chromosomes; for each simulation we measured the average genealogy length across the locus and then we fit a log-normal distribution to the distribution of mean genealogy lengths. The distribution of mean genealogy lengths is well approximated by a lognormal distribution (appendix figure D6), and the relationship between the shape parameter of the lognormal and the log of $N_e r$ is well approximated by a logistic equation (appendix figure D7). The mean $N_e r$ value in humans is approximately 0.0002 – there is approximately 1 cM per MB (Dumont and Payseur 2007 Evolution) and the effective population size of humans is ~20,000 (Wall and Przeworski, 2000; Voight et al. 2005), which means the distribution of mean genealogy lengths has a predicted shape parameter of 0.014. This implies that there is little variation in the mean genealogy length between windows with an average rate of recombination. This is an underestimate because most recombination in humans is concentrated in hotspots (McVean et al. 2004; Myers et al. 2005; Coop and Przeworski, 2007; Pratto et al. 2014). However, even when there is no recombination the shape parameter of the distribution is 0.17, which is substantially less the residual variation included in our model. Thus, under our model there is some residual variation that can be attributed to variation in the effective population size that is not due to variation in the mutation rate. Estimating this source of variation is not straightforward because it is difficult to identify all sites under selection.

As an aside we find that both the lognormal (appendix figures D6, D8) and the gamma distributions (appendix figure D10) fit the distribution of mean genealogy lengths.

Furthermore, for the gamma distribution the shape parameter is a simple linear function of the log of the sample size (appendix figure D11). For the lognormal distribution the relationship between the shape parameter and the log sample size is slightly curvilinear. These relationships may prove useful in future work.

We have investigated patterns of diversity across the human genome. We find, as others have before, that diversity is correlated to rates of mutation, the density of selected sites and the rate of recombination. However, we show for the first time that the distribution of mutation rates is broader than the distribution of SNPs. A model in which the effective population size of a genomic region is correlated to the mutation rate fits the data. However, our simple model implies that diversity is decreased by ~50% across the human genome by the action of linked selection, and that the effects of linked selection vary substantially across the genome.

6. General Discussion

The main focus of this thesis has been to increase our understanding of the patterns of natural selection in humans, and the factors that affect these patterns and rates of evolution at the molecular level. Each of the projects described in chapters 2 to 5 have touched on different aspects that contribute to our understanding of natural selection in humans. Here I will briefly summarise each project in turn, the consequences of my results for our understanding of human evolution, and the further work they necessitate.

6.1 Chapter summary

In Chapter 2 we developed and applied a novel method for detecting balancing selection using polymorphism data. Previous methods have used the presence of non-synonymous polymorphisms shared between two populations as evidence of balancing selection. These methods require that a long enough divergence time has passed to ensure that all shared neutral genetic variation to have reached fixation or loss in at least one of the two populations,

making these methods most effective at detecting ancient or long-term balancing selection. By comparing proportion of shared non-synonymous to synonymous polymorphism at shared and neutral sites, we are able to gain greater power for detecting balancing selection on shorter timescales. Through extensive simulations we have shown that although our method is robust to demography, estimates can still be depressed by demographic change, and therefore must be accounted for. We applied our method to human continental populations, finding large numbers of balanced polymorphisms being maintained between all populations.

In chapters 3 and 4 I investigated numerous factors that affect the rate of evolution in humans, both at the gene level and at the site level. In chapter 3, I looked at gene-level factors that affect the rates of adaptive and non-adaptive evolution between humans and chimpanzees. It has been shown that the rate of evolution correlates to recombination rate (RR), gene age, protein length, and gene expression in multiple species. We correlated each factor with the rates of adaptive and non-adaptive evolution, controlling for each other factor individually. By disentangling the rate of adaptive and non-adaptive evolution, we show that the rate of adaptive evolution, ω_a is correlated to all four factors in hominids, but that when we control for each factor in turn, there is evidence for an independent influence of RR, gene age and probably gene expression. These correlations remain when controlling for the effects of biased gene conversion. We also find that all four factors have significant independent effects on the rate of non-adaptive evolution, ω_{na}, and that all of these remain significant when we control for each in turn. We also considered the effect of gene function on the rate of adaptive and non-adaptive evolution. By splitting genes into GO categories and splitting the genes in each GO category by whether they code for viral interacting proteins (VIPs) or not, we confirm Enard et al.'s findings that VIPs have higher rates of adaptive evolution than other genes in primates. We also show that VIPs have a lower rate of non-adaptive evolution than other

genes in primates. Using a two-way analysis of variance on the rates of adaptive and non-adaptive evolution and the estimated variance components, we show that the distinction between VIP and non-VIP contributes several times more to the variance in ω_a and ω_{na} than the variation between GO categories, suggesting that whether a gene encodes a VIP has a major effect on its rate of adaptive and non-adaptive evolution.

In chapter 4, I looked at site-level factors that affect the rates of adaptive and non-adaptive evolution between humans and chimpanzees. Previous studies have observed strong correlations between the rate of adaptive evolution and amino acid dissimilarity (as measured by the difference in polarity, volume or Pn/Ps between amino acids), and the rate of adaptive evolution and relative solvent accessibility in *Drosophila melanogaster*. We found similar correlations in hominids for each of these factors except pN/pS, where the correlation is much weaker than in *Drosophila*. We suggest that this pattern can be explained by the population contraction in humans since the human-chimpanzee split, which tends to reduce genuine correlations between the rate of adaptive evolution and amino acid dissimilarity. We show that population size increases can artifactually generate negative correlations between an estimate of the rate of adaptive evolution and the mean strength of selection against deleterious mutations, even if there is no adaptive evolution, and that the reverse is true in cases of population contraction. In this case, we find that pN/pS is strongly correlated to the mean strength of selection against deleterious mutations, and therefore the correlation with the rate of adaptive evolution is greatly attenuated.

In chapter 5, I attempted to estimate the variation in effective population size across the human genome. Genetic diversity is known to vary across the genomes of many species. For neutral diversity this variation is due to variation in the rate of mutation, the genealogy length

and the effective population size, with variation in the effective population size being due to linked selection. We sought to understand the extent to which these two factors contribute to the variation in genetic diversity in humans. We divided the human genome up into non-overlapping 10KB windows, and quantified the variation in SNP density, and the variation in the mutation rate using de novo mutation data. We find greater variation in the mutation rate than variation in diversity. We explored a number of explanations for this, and conclude that it is most likely due to a negative relationship between the effects of linked selection and the mutation rate; this is expected since the power of linked selection can depend on the mutation rate. However, our models suggest that linked selection is extremely prevalent in the human genome, reducing diversity by more than 40% on average.

Taken together these projects contribute to our understanding of the prevalence of natural selection in humans. In the following section I will highlight some of the limitations of the methods developed and applied in this thesis.

6.2 Limitations

There are typically two approaches to modelling a system: abstraction involves leaving elements out whilst maintaining a literal description of the system being modelled, whilst idealisation treats elements within a system as having features they clearly do not have in reality, in order to produce a description that fictionalises in the service of simplification (Godfrey-Smith, 2009). Population genetic data describes complex systems and therefore modelling tends to use the approach of idealisation via a number of assumptions ascribed to the model. Here I briefly describe some of these assumptions and the limitations they impose on the research presented in this thesis.

6.2.1 Limitations of MK-type tests

Chapters 2, 3 and 4 all utilise methods that have a basis in the MK test (McDonald and Kreitman, 1991). In each analysis our dataset has consisted solely of coding regions of the human genome, despite evidence that selection also acts in noncoding regions in humans (Kryukov et al. 2005; Drake et al. 2006; Pollard et al. 2006; Asthana et al. 2007; Katzman et al. 2007).

MK-type methods require information on the amount of variation and divergence occurring at neutral and non-neutral sites. MK-type methods assume that certain classes of sites within the genome can be categorised as neutrally evolving. For simplicity, it is common to classify all synonymous sites as neutrally evolving, though it is likely that this is not the case for this entire class of sites (Hershberg and Petrov, 2008; Kudia et al. 2009). Little work has been done thus far in understanding to what extent this biases estimates of evolution for MK-type tests, though it has been shown that weak purifying selection does act on synonymous sites, biasing estimates of positive selection upwards (Eyre-Walker et al. 2002; Andolfatto, 2005). Halligan and Keightley (2006) have proposed using fast evolving sites of short introns as an alternative neutral site class, but it has been shown in *Drosophila melanogaster* that these sites tend to have similar levels of polymorphism and divergence as synonymous sites (Parsch et al. 2010).

6.2.2 Biased gene conversion

Patterns of codon usage can be influenced by forces such as GC-biased gene conversion (gBGC), a segregation bias that favours G and C alleles over A and T alleles in regions of high recombination (Duret and Galtier, 2009; Mugal et al. 2015). The occurrence of gBGC has been experimentally demonstrated in a wide range of organisms (Eyre-Walker 1993; Montoya-Burgos et al. 2003; Meunier and Duret 2004; Webster et al. 2004; Webster et al. 2005; Spencer

et al. 2006; Mancera et al. 2008; Escobar et al. 2011; Pessia et al. 2012; Lesecque et al. 2013; Williams et al. 2015; Halldorsson et al. 2016; Smeds et al. 2016; Keith et al. 2016; Long et al. 2018; Galtier et al. 2018; Smith et al. 2018). It can both mimic positive selection by increasing the fixation probability of G or C (S) over A or T (W) neutral alleles (Galtier and Duret 2007; Berglund et al. 2009; Ratnakumar et al. 2010), and promotes the fixation of slightly deleterious GC alleles (Duret and Galtier 2009; Glémin 2010; Necşulea et al. 2011; Lachance and Tishkoff 2014). Although Corcoran et al. (2018) showed that failing to control for the effects of gBGC can lead to an overestimation of α, it remains unclear as to how gBGC affects estimates of ω_a and ω_{na}. There are two methods for controlling for BGC, either by restricting the analysis to those polymorphisms and substitutions that are unaffected by BGC – i.e. A<>T and G<>C changes, or by controlling for recombination rate (as BGC is effective in highly recombining regions (Duret and Galtier, 2009; Mugal et al. 2015). Across the four projects we applied the former strategy where viable, and the latter in other cases. The former directly controls for BGC and is therefore preferable, but inevitably a great deal of data is lost. For instance, in chapter 3 applying this method reduced our dataset to about 20% of its previous size, resulting in a loss of significance in our results. In chapter 2 we applied both methods and found in both cases that we lost significance in our results. Despite the considerably larger datasets available to population geneticists today, we will need more still before we can control for BGC and still estimate results meaningfully.

6.2.3 Demographic models

In chapter 2 it became apparent that the model of human demography used in our simulations (Gravel et al. 2011), fit our 1000 genomes African data (The 1000 Genomes Project Consortium, 2015) poorly. We showed that in the African population there are far too many singleton SNPs even amongst the putative neutral synonymous mutations. The lack of fit is

perhaps not surprising; Gravel et al. inferred their model using 80 chromosomes per population, whereas the 1000 genome data contains >1000 chromosomes per population. Furthermore, the inference of a demographic model should take into account the influence of biased gene conversion and background selection, which appear to be pervasive factors in the human genome (Pouyet et al. 2018), so these simulations will be complex. However, Beichman et al. (2017) showed that the Gutenkunst et al. (2009) model of human demography (which is similar to the Gravel et al. (2011) model) actually fits the 1000 genomes data remarkably well when sampling 10 random unrelated individuals from the 1000 genomes YRI population. We attempted to replicate Beichman et al.'s findings by sampling the same 10 individuals from the YRI population and filtering out sites that do not pass the 1000 genomes "strict mask" filter (The 1000 Genomes Project Consortium, 2015), and simulating the Gravel et al. (2011) demographic model using neutral genetic variation only (i.e. there was no selection involved in our models). We sampled 10 individuals from our simulated data and compared the site frequency spectra for the observed and simulated data. We then repeated this analysis, this time sampling all 661 African individuals in the 1000 genomes data (The 1000 Genomes Project Consortium, 2015), and sampling 661 individuals from our simulations. Figure 6.1 shows the comparisons of these site frequency spectra.

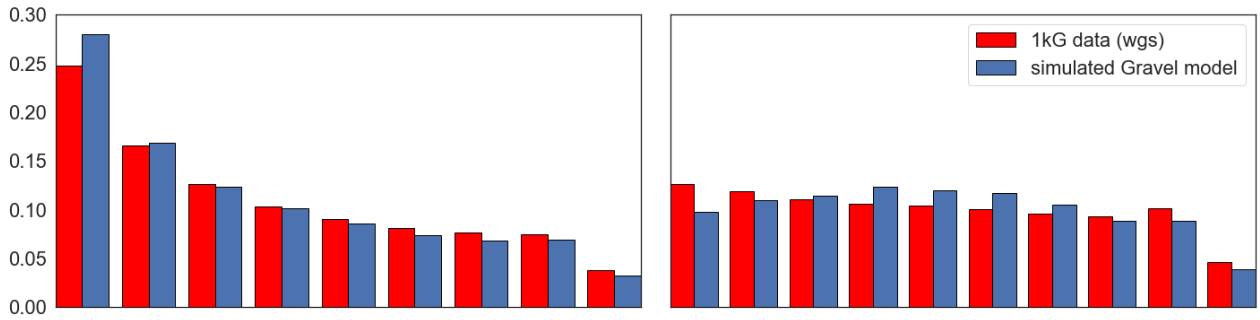


Figure 6.1: Folded SFS comparisons between observed and simulated data for the Beichman et al. (2017) samples (left) and for all African samples (right). For the comparison on the right, frequency categories are combined using a scheme of 1, 2-3, 4-7, 8-15, 16-31 etc).

We find the simulated SFS fits the observed SFS well for the Beichman et al. (2017) samples (figure 5.1 left), but poorly for the 661 African samples, with a tendency for the Gravel model to underestimate the number of low frequency variants and overestimate the number of variants at intermediate frequency (figure 5.1 right). These results suggest that the increase in the number of sampled chromosomes affects how well the Gravel et al. (2011) demographic model fits the 1000 genomes data (The 1000 Genomes Project Consortium, 2015). It is unsurprising that the Beichman et al. (2017) samples fit the data well, as fewer chromosomes are sampled than were used to estimate the Gravel et al. (2011) model. To understand how the fit changes with increasing numbers of sampled individuals, we sampled a varying number of individuals and compared the observed and simulated SFS. To summarise the difference in fit we estimated the mean square error for each different sample size, and plotted the results (figure 6.2).

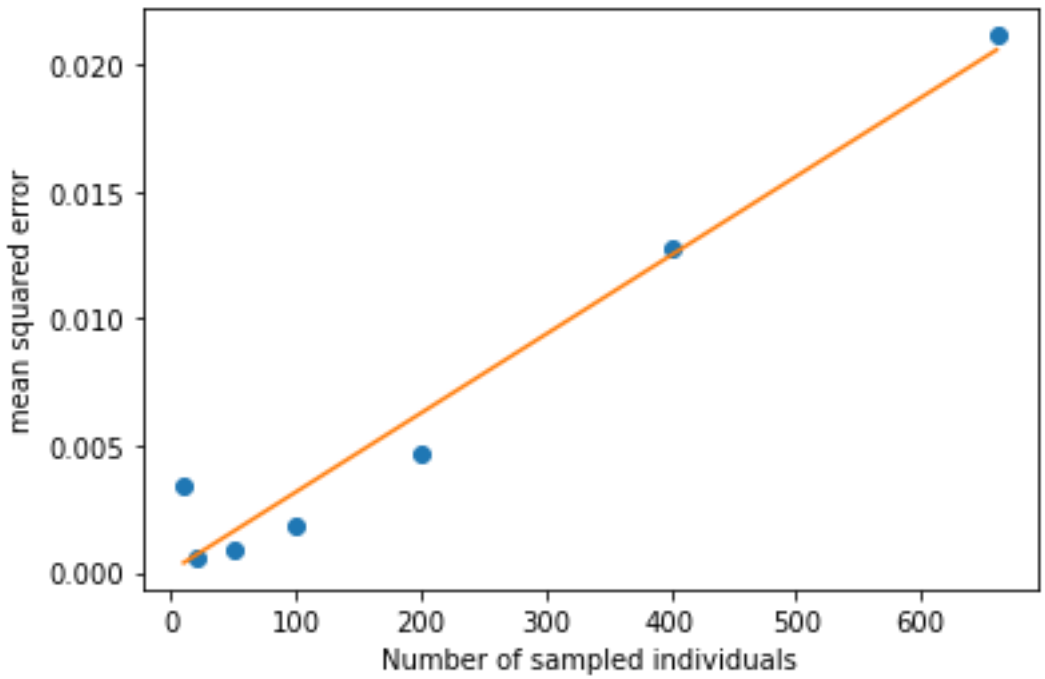


Figure 6.2: Mean squared error of comparisons between observed and simulated data for varying numbers of sampled individuals. A linear regression has been fitted to the estimates.

There is a strong correlation between the number of sampled individuals and the mean squared error of the fit between the observed and simulated SFS, suggesting that inferring a human demographic model from the larger datasets that are now being generated and used in research is likely to be pivotal to ensuring the reliability of neutrality tests. Joint estimation of demographic models is discussed further in section 6.3.2.

6.2.4 Population size change

A more suitable human demographic model is necessary to model the polymorphism phase of human evolution. However demographic models alone do not provide us with information about the divergence phase (meaning the population history of the human-chimpanzee ancestor), which is required for estimating rates of evolution using MK-type methods that compare polymorphism and divergence data (McDonald and Kreitman, 1990). It has been shown that population size change between the divergence phase and polymorphism phase of population history leads to either an underestimate (in the case of population contraction) or an overestimate (in the case of expansion) of the rate of adaptive evolution (McDonald and Kreitman, 1990; Eyre-Walker, 2002). This is due to the effect of *Ne* on the efficacy of selection against SDMs (this is explored in more detail in section 1.3.1 of this thesis). For example, if a large past population undergoes population contraction, selection will have been more effective at removing SDMs in the large past population than after the contraction. This will result in fewer SDMs fixing in the past population, which will then be segregating in the contracted population resulting in an underestimate of ω_a. In chapter 3 we showed that population size change can also attenuate the correlation between a variable and the rate of adaptive evolution if that variable is also correlated to the mean strength of selection against

deleterious mutations (mean s). The stronger the correlation between that variable and mean s, the greater the signal is attenuated.

These effects of population size are relevant to human population genetics because although *Ne* in humans has increased since the human-chimpanzee split (Gutenkunst et al. 2009; Gravel et al. 2011), overall estimates of *Ne* in humans are still considerably lower than that of the human-chimpanzee ancestor (Holboth et al. 2007; Burgess and Yang 2008; Prado-Martinez et al. 2013; Schrago, 2014). This population contraction will result in underestimates of ω_a between humans and chimpanzees. It will also attenuate the correlation between any variable and ω_a, if that variable is correlated to mean S. In chapter 3 we showed that our measure of evolutionary dissimilarity between amino acids (p_N/p_S) was strongly negatively correlated to mean s, resulting in a much weaker correlation between ω_a and p_N/p_S in hominids than *Drosophila*.

6.3 Moving forward

The results obtained in this thesis open up several avenues of further research, in terms of estimating natural selection in humans, the limitations of the methods applied, and the application of novel methods to other species.

6.3.1 Looking at other species

By virtue of an understandable curiosity to understand our own species, human population genetic datasets are some of the largest and most comprehensively annotated available. Throughout this thesis we have made use of human data from the 1000 genomes project (The 1000 Genomes Project Consortium, 2015), containing data for 2,504 individuals across 26

populations. Other animal species for which extremely comprehensive datasets exist include *Drosophila* and mice. One particularly useful avenue of further research would be to apply our novel method for estimating the frequency of balancing selection across the genome (see chapter 2) to *Drosophila* species. Like other tests for balancing selection that compare the number of polymorphisms at selected and neutral sites, our method gains power as tMRCA increases, as shared neutral genetic variation is lost or fixed within at least one of the two populations being compared. Of course we must also consider that a balanced polymorphism can maintain neutral variation in linkage disequilibrium (LD) that may also be shared between populations. Recombination is the force that can break up linkage, and it is notable that recombination is a considerably more effective force in *Drosophila* (where LD decays over a scale of 10s of base pairs (Mackay et al. 2012)) than in humans (where LD decay is in the order of 10,000 base pairs (The 1000 Genomes Project Consortium, 2015)). We would therefore expect our method for detecting balancing selection to have more power in *Drosophila* than in humans, though there are many other factors to consider (including demographic history – discussed in chapter 2).

6.3.2 Joint inference of demographic models

As discussed in section 6.2.3, there is a pressing need for a human demographic model inferred from the 1000s of chromosomes that are now available to researchers, that also accounts for BGS and BGC. Because demographic models assume sites evolve neutrally, it is necessary to parameterise them from regions of the genome that are free from the effects of linked selection (Pouyet et al. 2018). One approach is to identify functional elements in the genome and filter out regions that are in close proximity, leaving only regions that are free from the effects of linked selection. However, methods used to identify conserved elements

(e.g. Siepel et al. 2005) can be susceptible to the failure of identifying rapidly evolving and/or weakly selected regions.

In recent years computationally demanding approaches such as approximate Bayesian computation and machine learning have become viable for demographic inference. Both methods have a similar underlying logic: Simulate data under a chosen model, sampling the parameters of interest from plausible ranges (using the literature to determine these ranges), and compare summary statistics from the observed and simulated data. The parameter estimates that give the best fit are chosen. This is the basic idea behind ABC and machine learning approaches.

Johri et al. (2020) have utilised a novel statistical framework to develop an appropriate null model in *Drosophila melanogaster*. The authors jointly estimated the effects of population history and the DFE using an ABC framework, whilst accounting for the effects of background selection (BGS). A similar approach would be appropriate for inferring a null model for humans.

6.3.3 Conclusion

The research in this thesis furthers our understanding of patterns of natural selection in humans, whilst highlighting numerous limitations that point the way forward for further research.

Bibliography

Aguade, M., Miyashita, N., & Langley, C. H. (1989). Reduced variation in the yellow-achaete-scute region in natural populations of Drosophila melanogaster. *Genetics*, *122*(3), 607–615.

Aguilar, A., Roemer, G., Debenham, S., Binns, M., Garcelon, D., & Wayne, R. K. (2004). High MHC diversity maintained by balancing selection in an otherwise genetically monomorphic mammal. *Proceedings of the National Academy of Sciences*, *101*(10), 3490–3494. https://doi.org/10.1073/pnas.0306582101

Akey, J. M. (2009). Constructing genomic maps of positive selection in humans: Where do we go from here? *Genome Research*, *19*(5), 711–722. https://doi.org/10.1101/gr.086652.108

Albà, M. M., & Castresana, J. (2005). Inverse Relationship Between Evolutionary Rate and Age of Mammalian Genes. *Molecular Biology and Evolution*, *22*(3), 598–606. https://doi.org/10.1093/molbev/msi045

Allison, A. C. (1956). The sickle-cell and haemoglobin C genes in some African populations. *Annals of Human Genetics*, *21*(1), 67–89.

Andolfatto, P. (2005). Adaptive evolution of non-coding DNA in Drosophila. *Nature*, *437*(7062), 1149–1152. https://doi.org/10.1038/nature04107

Andolfatto, P., & Przeworski, M. (2000). A genome-wide departure from the standard neutral model in natural populations of Drosophila. *Genetics*, *156*(1), 257–268.

Andolfatto, P., Wong, K. M., & Bachtrog, D. (2011). Effective population size and the efficacy of selection on the X chromosomes of two closely related Drosophila species. *Genome Biology and Evolution*, *3*, 114–128. https://doi.org/10.1093/gbe/evq086

Andrés, A. M., Hubisz, M. J., Indap, A., Torgerson, D. G., Degenhardt, J. D., Boyko, A. R., Gutenkunst, R. N., White, T. J., Green, E. D., Bustamante, C. D., Clark, A. G., & Nielsen, R. (2009). Targets of balancing selection in the human genome. *Molecular Biology and Evolution*, *26*(12), 2755–2764. https://doi.org/10.1093/molbev/msp190

Arbeithuber, B., Betancourt, A. J., Ebner, T., & Tiemann-Boege, I. (2015). Crossovers are associated with mutation and biased gene conversion at recombination hotspots. *Proceedings of the National Academy of Sciences*, *112*(7), 2109–2114. https://doi.org/10.1073/pnas.1416622112

Arguello, J. R., Zhang, Y., Kado, T., Fan, C., Zhao, R., Innan, H., Wang, W., & Long, M. (2010). Recombination yet inefficient selection along the Drosophila melanogaster subgroup's fourth chromosome. *Molecular Biology and Evolution*, *27*(4), 848–861. https://doi.org/10.1093/molbev/msp291

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., Harris, M. A., Hill, D. P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J. C., Richardson, J. E., Ringwald, M., Rubin, G. M., & Sherlock, G. (2000). Gene ontology: Tool for the unification of biology. The Gene Ontology Consortium. *Nature Genetics*, *25*(1), 25–29. https://doi.org/10.1038/75556

Asthana, S., Noble, W. S., Kryukov, G., Grant, C. E., Sunyaev, S., & Stamatoyannopoulos, J. A. (2007). Widely distributed noncoding purifying selection in the human genome. *Proceedings of the*

National Academy of Sciences, *104*(30), 12410–12415. https://doi.org/10.1073/pnas.0705140104

Asthana, S., Schmidt, S., & Sunyaev, S. (2005). A limited role for balancing selection. *Trends in Genetics*, *21*(1), 30–32. https://doi.org/10.1016/j.tig.2004.11.001

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

Barton, N. H. (1998). The effect of hitch-hiking on neutral genealogies. *Genetical Research*, *72*(2), 123–133. https://doi.org/10.1017/S0016672398003462

Bataillon, T., Duan, J., Hvilsom, C., Jin, X., Li, Y., Skov, L., Glemin, S., Munch, K., Jiang, T., Qian, Y., Hobolth, A., Wang, J., Mailund, T., Siegismund, H. R., & Schierup, M. H. (2015). Inference of Purifying and Positive Selection in Three Subspecies of Chimpanzees (Pan troglodytes) from Exome Sequencing. *Genome Biology and Evolution*, *7*(4), 1122–1132. https://doi.org/10.1093/gbe/evv058

Becher, H., Jackson, B. C., & Charlesworth, B. (2020). Patterns of Genetic Variability in Genomic Regions with Low Rates of Recombination. *Current Biology*, *30*(1), 94-100.e3. https://doi.org/10.1016/j.cub.2019.10.047

Begun, D. J., & Aquadro, C. F. (1991). Molecular population genetics of the distal portion of the X chromosome in Drosophila: Evidence for genetic hitchhiking of the yellow-achaete region. *Genetics*, *129*(4), 1147–1158.

Begun, D. J., & Aquadro, C. F. (1992). Levels of naturally occurring DNA polymorphism correlate with recombination rates in D. melanogaster. *Nature*, *356*(6369), 519–520. https://doi.org/10.1038/356519a0

Beichman, A. C., Phung, T. N., & Lohmueller, K. E. (2017). Comparison of Single Genome and Allele Frequency Data Reveals Discordant Demographic Histories. *G3 Genes|Genomes|Genetics*, *7*(11), 3605–3620. https://doi.org/10.1534/g3.117.300259

Berglund, J., Pollard, K. S., & Webster, M. T. (2009). Hotspots of Biased Nucleotide Substitutions in Human Genes. *PLoS Biology*, *7*(1), e1000026. https://doi.org/10.1371/journal.pbio.1000026

Bergman, J., & Eyre-Walker, A. (2019). Does Adaptive Protein Evolution Proceed by Large or Small Steps at the Amino Acid Level? *Molecular Biology and Evolution*, *36*(5), 990–998. https://doi.org/10.1093/molbev/msz033

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

Besenbacher, S., Hvilsom, C., Marques-Bonet, T., Mailund, T., & Schierup, M. H. (2019). Direct estimation of mutations in great apes reconciles phylogenetic dating. *Nature Ecology & Evolution*, *3*(2), 286–292. https://doi.org/10.1038/s41559-018-0778-x

Betancourt, A. J., & Presgraves, D. C. (2002). Linkage limits the power of natural selection in Drosophila. *Proceedings of the National Academy of Sciences*, *99*(21), 13616–13620. https://doi.org/10.1073/pnas.212277199

Betancourt, A. J., Welch, J. J., & Charlesworth, B. (2009). Reduced effectiveness of selection caused by a lack of recombination. *Current Biology: CB*, *19*(8), 655–660. https://doi.org/10.1016/j.cub.2009.02.039

Bhaskar, A., Wang, Y. X. R., & Song, Y. S. (2015). Efficient inference of population size histories and locus-specific mutation rates from large-sample genomic variation data. *Genome Research*, *25*(2), 268–279. https://doi.org/10.1101/gr.178756.114

Bierne, N., & Eyre-Walker, A. (2004). The genomic rate of adaptive amino acid substitution in Drosophila. *Molecular Biology and Evolution*, *21*(7), 1350–1360. https://doi.org/10.1093/molbev/msh134

Bitarello, B. D., de Filippo, C., Teixeira, J. C., Schmidt, J. M., Kleinert, P., Meyer, D., & Andrés, A. M. (2018). Signatures of Long-Term Balancing Selection in Human Genomes. *Genome Biology and Evolution*, *10*(3), 939–955. https://doi.org/10.1093/gbe/evy054

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., & Bustamante, C. D. (2008). Assessing the Evolutionary Impact of Amino Acid Mutations in the Human Genome. *PLoS Genetics*, *4*(5), e1000083. https://doi.org/10.1371/journal.pgen.1000083

Braverman, J. M., Hudson, R. R., Kaplan, N. L., Langley, C. H., & Stephan, W. (1995). The hitchhiking effect on the site frequency spectrum of DNA polymorphisms. *Genetics*, *140*(2), 783–796. https://doi.org/10.1093/genetics/140.2.783

Broad Institute Sequencing Platform and Whole Genome Assembly Team, Baylor College of Medicine Human Genome Sequencing Center Sequencing Team, Genome Institute at Washington University, Lindblad-Toh, K., Garber, M., Zuk, O., Lin, M. F., Parker, B. J., Washietl, S., Kheradpour, P., Ernst, J., Jordan, G., Mauceli, E., Ward, L. D., Lowe, C. B., Holloway, A. K., Clamp, M., Gnerre, S., Alföldi, J., … Kellis, M. (2011). A high-resolution map of human evolutionary constraint using 29 mammals. *Nature*, *478*(7370), 476–482. https://doi.org/10.1038/nature10530

Bubb, K. L., Bovee, D., Buckley, D., Haugen, E., Kibukawa, M., Paddock, M., Palmieri, A., Subramanian, S., Zhou, Y., Kaul, R., Green, P., & Olson, M. V. (2006). Scan of Human Genome

Reveals No New Loci Under Ancient Balancing Selection. *Genetics*, *173*(4), 2165–2177. https://doi.org/10.1534/genetics.106.055715

Burgess, R., & Yang, Z. (2008). Estimation of hominoid ancestral population sizes under bayesian coalescent models incorporating mutation rate variation and sequencing errors. *Molecular Biology and Evolution*, *25*(9), 1979–1994. https://doi.org/10.1093/molbev/msn148

Bustamante, C. D., Fledel-Alon, A., Williamson, S., Nielsen, R., Todd Hubisz, M., Glanowski, S., Tanenbaum, D. M., White, T. J., Sninsky, J. J., Hernandez, R. D., Civello, D., Adams, M. D., Cargill, M., & Clark, A. G. (2005). Natural selection on protein-coding genes in the human genome. *Nature*, *437*(7062), 1153–1157. https://doi.org/10.1038/nature04240

Bustamante, C. D., Nielsen, R., Sawyer, S. A., Olsen, K. M., Purugganan, M. D., & Hartl, D. L. (2002). The cost of inbreeding in Arabidopsis. *Nature*, *416*(6880), 531–534. https://doi.org/10.1038/416531a

Bustamante, C. D., Townsend, J. P., & Hartl, D. L. (2000). Solvent Accessibility and Purifying Selection Within Proteins of Escherichia coli and Salmonella enterica. *Molecular Biology and Evolution*, *17*(2), 301–308. https://doi.org/10.1093/oxfordjournals.molbev.a026310

Cai, J. J., & Petrov, D. A. (2010). Relaxed Purifying Selection and Possibly High Rate of Adaptation in Primate Lineage-Specific Genes. *Genome Biology and Evolution*, *2*, 393–409. https://doi.org/10.1093/gbe/evq019

Cai, J. J., Woo, P. C. Y., Lau, S. K. P., Smith, D. K., & Yuen, K.-Y. (2006). Accelerated evolutionary rate may be responsible for the emergence of lineage-specific genes in ascomycota. *Journal of Molecular Evolution*, *63*(1), 1–11. https://doi.org/10.1007/s00239-004-0372-5

Campos, J. L., Halligan, D. L., Haddrill, P. R., & Charlesworth, B. (2014). The Relation between Recombination Rate and Patterns of Molecular Evolution and Variation in Drosophila

melanogaster. *Molecular Biology and Evolution*, *31*(4), 1010–1028. https://doi.org/10.1093/molbev/msu056

Cargill, M., Altshuler, D., Ireland, J., Sklar, P., Ardlie, K., Patil, N., Lane, C. R., Lim, E. P., Kalyanaraman, N., Nemesh, J., Ziaugra, L., Friedland, L., Rolfe, A., Warrington, J., Lipshutz, R., Daley, G. Q., & Lander, E. S. (1999). Characterization of single-nucleotide polymorphisms in coding regions of human genes. *Nature Genetics*, *22*(3), 231–238. https://doi.org/10.1038/10290

Castellano, D., Coronado-Zamora, M., Campos, J. L., Barbadilla, A., & Eyre-Walker, A. (2016). Adaptive Evolution Is Substantially Impeded by Hill-Robertson Interference in Drosophila. *Molecular Biology and Evolution*, *33*(2), 442–455. https://doi.org/10.1093/molbev/msv236

Castellano, D., Eyre-Walker, A., & Munch, K. (2020). Impact of Mutation Rate and Selection at Linked Sites on DNA Variation across the Genomes of Humans and Other Homininae. *Genome Biology and Evolution*, *12*(1), 3550–3561. https://doi.org/10.1093/gbe/evz215

Castellano, D., James, J., & Eyre-Walker, A. (2018). Nearly Neutral Evolution across the Drosophila melanogaster Genome. *Molecular Biology and Evolution*. https://doi.org/10.1093/molbev/msy164

Castellano, D., Macià, M. C., Tataru, P., Bataillon, T., & Munch, K. (2019). Comparison of the Full Distribution of Fitness Effects of New Amino Acid Mutations Across Great Apes. *Genetics*, *213*(3), 953–966. https://doi.org/10.1534/genetics.119.302494

Catcheside D. W. (1977). *Genetics of Recombination.* University Park Press.

Charlesworth, B. (1994). *Evolution in Age-Structured Populations* (2nd ed.). Cambridge University Press. https://doi.org/10.1017/CBO9780511525711

Charlesworth, B. (2009). Effective population size and patterns of molecular evolution and variation. *Nature Reviews Genetics*, *10*(3), 195–205. https://doi.org/10.1038/nrg2526

Charlesworth, B. (2012). The Effects of Deleterious Mutations on Evolution at Linked Sites. *Genetics*, *190*(1), 5–22. https://doi.org/10.1534/genetics.111.134288

Charlesworth, B., & Charlesworth, D. (2017). Population genetics from 1966 to 2016. *Heredity*, *118*(1), 2–9. https://doi.org/10.1038/hdy.2016.55

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

Charlesworth, J. (2006). The Rate of Adaptive Evolution in Enteric Bacteria. *Molecular Biology and Evolution*, *23*(7), 1348–1356. https://doi.org/10.1093/molbev/msk025

Charlesworth, J., & Eyre-Walker, A. (2008). The McDonald-Kreitman Test and Slightly Deleterious Mutations. *Molecular Biology and Evolution*, *25*(6), 1007–1015. https://doi.org/10.1093/molbev/msn005

Chen, J., Glémin, S., & Lascoux, M. (2020). From Drift to Draft: How Much Do Beneficial Mutations Actually Contribute to Predictions of Ohta's Slightly Deleterious Model of Molecular Evolution? *Genetics*, *214*(4), 1005–1018. https://doi.org/10.1534/genetics.119.302869

Chen, X., Zhao, Y., Zeng, C., Li, Y., Zhu, L., Wu, J., Chen, J., & Wei, Z. (2019). Assessment contributions of physicochemical properties and bacterial community to mitigate the bioavailability of heavy metals during composting based on structural equation models. *Bioresource Technology*, *289*, 121657. https://doi.org/10.1016/j.biortech.2019.121657

Cheng, X., & DeGiorgio, M. (2019). Detection of Shared Balancing Selection in the Absence of Trans-Species Polymorphism. *Molecular Biology and Evolution*, *36*(1), 177–199. https://doi.org/10.1093/molbev/msy202

Chiaromonte, F., Weber, R. J., Roskin, K. M., Diekhans, M., Kent, W. J., & Haussler, D. (2003). The share of human genomic DNA under selection estimated from human-mouse genomic

alignments. *Cold Spring Harbor Symposia on Quantitative Biology*, *68*, 245–254. https://doi.org/10.1101/sqb.2003.68.245

Chikhi, L., Sousa, V. C., Luisi, P., Goossens, B., & Beaumont, M. A. (2010). The Confounding Effects of Population Structure, Genetic Diversity and the Sampling Scheme on the Detection and Quantification of Population Size Changes. *Genetics*, *186*(3), 983–995. https://doi.org/10.1534/genetics.110.118661

Choi, S. S., Vallender, E. J., & Lahn, B. T. (2006). Systematically assessing the influence of 3-dimensional structural context on the molecular evolution of mammalian proteomes. *Molecular Biology and Evolution*, *23*(11), 2131–2133. https://doi.org/10.1093/molbev/msl086

Clark, A. G. (2003). Inferring Nonneutral Evolution from Human-Chimp-Mouse Orthologous Gene Trios. *Science*, *302*(5652), 1960–1963. https://doi.org/10.1126/science.1088821

Clarkson, C., Guerrero, C., & Soni, V. (n.d.). *Development of a pipeline to identify HERV-K insertional polymorphisms in a Neanderthal genome*. 28.

Comeron, J. M. (2014). Background Selection as Baseline for Nucleotide Variation across the Drosophila Genome. *PLoS Genetics*, *10*(6), e1004434. https://doi.org/10.1371/journal.pgen.1004434

Conant, G. C., & Stadler, P. F. (2009). Solvent Exposure Imparts Similar Selective Pressures across a Range of Yeast Proteins. *Molecular Biology and Evolution*, *26*(5), 1155–1161. https://doi.org/10.1093/molbev/msp031

Coop, G., Pickrell, J. K., Novembre, J., Kudaravalli, S., Li, J., Absher, D., Myers, R. M., Cavalli-Sforza, L. L., Feldman, M. W., & Pritchard, J. K. (2009). The Role of Geography in Human Adaptation. *PLoS Genetics*, *5*(6), e1000500. https://doi.org/10.1371/journal.pgen.1000500

Coop, G., & Przeworski, M. (2007). An evolutionary view of human recombination. *Nature Reviews. Genetics*, *8*(1), 23–34. https://doi.org/10.1038/nrg1947

Cooper, G. M., Stone, E. A., Asimenos, G., NISC Comparative Sequencing Program, Green, E. D., Batzoglou, S., & Sidow, A. (2005). Distribution and intensity of constraint in mammalian genomic sequence. *Genome Research*, *15*(7), 901–913. https://doi.org/10.1101/gr.3577405

Corbett-Detig, R. B., Hartl, D. L., & Sackton, T. B. (2015). Natural Selection Constrains Neutral Diversity across A Wide Range of Species. *PLOS Biology*, *13*(4), e1002112. https://doi.org/10.1371/journal.pbio.1002112

Cui, X., Lv, Y., Chen, M., Nikoloski, Z., Twell, D., & Zhang, D. (2015). Young Genes out of the Male: An Insight from Evolutionary Age Analysis of the Pollen Transcriptome. *Molecular Plant*, *8*(6), 935–945. https://doi.org/10.1016/j.molp.2014.12.008

Cutter, A. D., & Payseur, B. A. (2013). Genomic signatures of selection at linked sites: Unifying the disparity among species. *Nature Reviews. Genetics*, *14*(4), 262–274. https://doi.org/10.1038/nrg3425

Darwin, C. R. & Wallace, A. R. 1858. On the tendency of species to form varieties; and on the perpetuation of varieties and species by natural means of selection. *Journal of the Proceedings of the Linnean Society of London*, *Zoology*, **3**(9): 45-62.

Daubin, V., & Ochman, H. (2004). Bacterial genomes as new gene homes: The genealogy of ORFans in E. coli. *Genome Research*, *14*(6), 1036–1042. https://doi.org/10.1101/gr.2231904

Davidson, A. E., Sergouniotis, P. I., Mackay, D. S., Wright, G. A., Waseem, N. H., Michaelides, M., Holder, G. E., Robson, A. G., Moore, A. T., Plagnol, V., & Webster, A. R. (2013). *RP1L1* Variants are Associated with a Spectrum of Inherited Retinal Diseases Including Retinitis Pigmentosa and Occult Macular Dystrophy. *Human Mutation*, *34*(3), 506–514. https://doi.org/10.1002/humu.22264

Davydov, E. V., Goode, D. L., Sirota, M., Cooper, G. M., Sidow, A., & Batzoglou, S. (2010). Identifying a High Fraction of the Human Genome to be under Selective Constraint Using GERP++. *PLoS Computational Biology*, *6*(12), e1001025. https://doi.org/10.1371/journal.pcbi.1001025

Dean, A. M., Neuhauser, C., Grenier, E., & Golding, G. B. (2002). The Pattern of Amino Acid Replacements in α/β-Barrels. *Molecular Biology and Evolution*, *19*(11), 1846–1864. https://doi.org/10.1093/oxfordjournals.molbev.a004009

DeGiorgio, M., Lohmueller, K. E., & Nielsen, R. (2014). A Model-Based Approach for Identifying Signatures of Ancient Balancing Selection in Genetic Data. *PLoS Genetics*, *10*(8), e1004561. https://doi.org/10.1371/journal.pgen.1004561

Delihas, N. (2018). A family of long intergenic non-coding RNA genes in human chromosomal region 22q11.2 carry a DNA translocation breakpoint/AT-rich sequence. *PLOS ONE*, *13*(4), e0195702. https://doi.org/10.1371/journal.pone.0195702

Denamur, E., & Matic, I. (2006). Evolution of mutation rates in bacteria. *Molecular Microbiology*, *60*(4), 820–827. https://doi.org/10.1111/j.1365-2958.2006.05150.x

Domazet-Loso, T., Brajković, J., & Tautz, D. (2007). A phylostratigraphy approach to uncover the genomic history of major adaptations in metazoan lineages. *Trends in Genetics: TIG*, *23*(11), 533–539. https://doi.org/10.1016/j.tig.2007.08.014

Drake, J. A., Bird, C., Nemesh, J., Thomas, D. J., Newton-Cheh, C., Reymond, A., Excoffier, L., Attar, H., Antonarakis, S. E., Dermitzakis, E. T., & Hirschhorn, J. N. (2006). Conserved noncoding sequences are selectively constrained and not mutation cold spots. *Nature Genetics*, *38*(2), 223–227. https://doi.org/10.1038/ng1710

Dumont, B. L., & Payseur, B. A. (2008). EVOLUTION OF THE GENOMIC RATE OF RECOMBINATION IN MAMMALS. *Evolution*, *62*(2), 276–294. https://doi.org/10.1111/j.1558-5646.2007.00278.x

Duret, L., & Galtier, N. (2009). Biased gene conversion and the evolution of mammalian genomic landscapes. *Annual Review of Genomics and Human Genetics*, *10*, 285–311. https://doi.org/10.1146/annurev-genom-082908-150001

Edgar, R. C. (2004). MUSCLE: Multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research*, *32*(5), 1792–1797. https://doi.org/10.1093/nar/gkh340

Elhaik, E., Sabath, N., & Graur, D. (2006). The "Inverse Relationship Between Evolutionary Rate and Age of Mammalian Genes" Is an Artifact of Increased Genetic Distance with Rate of Evolution and Time of Divergence. *Molecular Biology and Evolution*, *23*(1), 1–3. https://doi.org/10.1093/molbev/msj006

Enard, D., Cai, L., Gwennap, C., & Petrov, D. A. (2016). Viruses are a dominant driver of protein adaptation in mammals. *ELife*, *5*, e12469. https://doi.org/10.7554/eLife.12469

Escobar, J. S., Glémin, S., & Galtier, N. (2011). GC-Biased Gene Conversion Impacts Ribosomal DNA Evolution in Vertebrates, Angiosperms, and Other Eukaryotes. *Molecular Biology and Evolution*, *28*(9), 2561–2575. https://doi.org/10.1093/molbev/msr079

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 in Ecology & Evolution*, *21*(10), 569–575. https://doi.org/10.1016/j.tree.2006.06.015

Eyre-Walker, A., & Keightley, P. D. (2007). The distribution of fitness effects of new mutations. *Nature Reviews Genetics*, *8*(8), 610–618. https://doi.org/10.1038/nrg2146

Eyre-Walker, A., & Keightley, P. D. (2009). Estimating the Rate of Adaptive Molecular Evolution in the Presence of Slightly Deleterious Mutations and Population Size Change. *Molecular Biology and Evolution*, *26*(9), 2097–2108. https://doi.org/10.1093/molbev/msp119

Eyre-Walker, A., Keightley, P. D., Smith, N. G. C., & Gaffney, D. (2002). Quantifying the Slightly Deleterious Mutation Model of Molecular Evolution. *Molecular Biology and Evolution*, *19*(12), 2142–2149. https://doi.org/10.1093/oxfordjournals.molbev.a004039

Fagerberg, L., Hallström, B. M., Oksvold, P., Kampf, C., Djureinovic, D., Odeberg, J., Habuka, M., Tahmasebpoor, S., Danielsson, A., Edlund, K., Asplund, A., Sjöstedt, E., Lundberg, E., Szigyarto, C. A.-K., Skogs, M., Takanen, J. O., Berling, H., Tegel, H., Mulder, J., … Uhlén, M. (2014). Analysis of the Human Tissue-specific Expression by Genome-wide Integration of Transcriptomics and Antibody-based Proteomics. *Molecular & Cellular Proteomics*, *13*(2), 397–406. https://doi.org/10.1074/mcp.M113.035600

Fay, J. C., Wyckoff, G. J., & Wu, C.-I. (n.d.). *Positive and Negative Selection on the Human Genome*. 8.

Ferrer-Admetlla, A., Liang, M., Korneliussen, T., & Nielsen, R. (2014). On Detecting Incomplete Soft or Hard Selective Sweeps Using Haplotype Structure. *Molecular Biology and Evolution*, *31*(5), 1275–1291. https://doi.org/10.1093/molbev/msu077

Field, Y., Boyle, E. A., Telis, N., Gao, Z., Gaulton, K. J., Golan, D., Yengo, L., Rocheleau, G., Froguel, P., McCarthy, M. I., & Pritchard, J. K. (n.d.). *Detection of human adaptation during the past 2000 years*. 5.

Fijarczyk, A., & Babik, W. (2015). Detecting balancing selection in genomes: Limits and prospects. *Molecular Ecology*, *24*(14), 3529–3545. https://doi.org/10.1111/mec.13226

Fisher, R. (1922). On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, *222*(594–604), 309–368. https://doi.org/10.1098/rsta.1922.0009

Francioli, L. C., Polak, P. P., Koren, A., Menelaou, A., Chun, S., Renkens, I., Genome of the Netherlands Consortium, van Duijn, C. M., Swertz, M., Wijmenga, C., van Ommen, G., Slagboom, P. E., Boomsma, D. I., Ye, K., Guryev, V., Arndt, P. F., Kloosterman, W. P., de Bakker,

P. I. W., & Sunyaev, S. R. (2015). Genome-wide patterns and properties of de novo mutations in humans. *Nature Genetics*, *47*(7), 822–826. https://doi.org/10.1038/ng.3292

Franzosa, E. A., & Xia, Y. (2009). Structural determinants of protein evolution are context-sensitive at the residue level. *Molecular Biology and Evolution*, *26*(10), 2387–2395. https://doi.org/10.1093/molbev/msp146

Fu, W., & Akey, J. M. (2013). Selection and Adaptation in the Human Genome. *Annual Review of Genomics and Human Genetics*, *14*(1), 467–489. https://doi.org/10.1146/annurev-genom-091212-153509

Fu, Y. X., & Li, W. H. (1993). Statistical tests of neutrality of mutations. *Genetics*, *133*(3), 693–709.

Galtier, N. (2016). Adaptive Protein Evolution in Animals and the Effective Population Size Hypothesis. *PLOS Genetics*, *12*(1), e1005774. https://doi.org/10.1371/journal.pgen.1005774

Galtier, N., & Duret, L. (2007). Adaptation or biased gene conversion? Extending the null hypothesis of molecular evolution. *Trends in Genetics: TIG*, *23*(6), 273–277. https://doi.org/10.1016/j.tig.2007.03.011

Galtier, N., & Rousselle, M. (2020). *How Much Does Ne Vary Among Species?* 14.

Galtier, N., Roux, C., Rousselle, M., Romiguier, J., Figuet, E., Glémin, S., Bierne, N., & Duret, L. (2018). Codon Usage Bias in Animals: Disentangling the Effects of Natural Selection, Effective Population Size, and GC-Biased Gene Conversion. *Molecular Biology and Evolution*, *35*(5), 1092–1103. https://doi.org/10.1093/molbev/msy015

Gao, F., & Keinan, A. (2016). Explosive genetic evidence for explosive human population growth. *Current Opinion in Genetics & Development*, *41*, 130–139. https://doi.org/10.1016/j.gde.2016.09.002

Gao, Z., Przeworski, M., & Sella, G. (2015). Footprints of ancient-balanced polymorphisms in genetic variation data from closely related species. *Evolution; International Journal of Organic Evolution*, *69*(2), 431–446. https://doi.org/10.1111/evo.12567

Garber, M., Guttman, M., Clamp, M., Zody, M. C., Friedman, N., & Xie, X. (2009). Identifying novel constrained elements by exploiting biased substitution patterns. *Bioinformatics*, *25*(12), i54–i62. https://doi.org/10.1093/bioinformatics/btp190

Gasper, J., & Swanson, W. J. (2006). Molecular Population Genetics of the Gene Encoding the Human Fertilization Protein Zonadhesin Reveals Rapid Adaptive Evolution. *The American Journal of Human Genetics*, *79*(5), 820–830. https://doi.org/10.1086/508473

Gattepaille, L., Günther, T., & Jakobsson, M. (2016). Inferring Past Effective Population Size from Distributions of Coalescent Times. *Genetics*, *204*(3), 1191–1206. https://doi.org/10.1534/genetics.115.185058

Gazave, E., Ma, L., Chang, D., Coventry, A., Gao, F., Muzny, D., Boerwinkle, E., Gibbs, R. A., Sing, C. F., Clark, A. G., & Keinan, A. (2014). Neutral genomic regions refine models of recent rapid human population growth. *Proceedings of the National Academy of Sciences*, *111*(2), 757–762. https://doi.org/10.1073/pnas.1310398110

Giraud, A. (2001). Costs and Benefits of High Mutation Rates: Adaptive Evolution of Bacteria in the Mouse Gut. *Science*, *291*(5513), 2606–2608. https://doi.org/10.1126/science.1056421

Glémin, S. (2010). Surprising Fitness Consequences of GC-Biased Gene Conversion: I. Mutation Load and Inbreeding Depression. *Genetics*, *185*(3), 939–959. https://doi.org/10.1534/genetics.110.116368

Godfrey-Smith, P. (2009). *Darwinian populations and natural selection*. Oxford University Press.

Gojobori, J., Tang, H., Akey, J. M., & Wu, C.-I. (2007). Adaptive evolution in humans revealed by the negative correlation between the polymorphism and fixation phases of evolution. *Proceedings*

of the National Academy of Sciences of the United States of America, *104*(10), 3907–3912. https://doi.org/10.1073/pnas.0605565104

Goldman, N., Thorne, J. L., & Jones, D. T. (1998). Assessing the impact of secondary structure and solvent accessibility on protein evolution. *Genetics*, *149*(1), 445–458.

Gossmann, T. I., Keightley, P. D., & Eyre-Walker, A. (2012). The Effect of Variation in the Effective Population Size on the Rate of Adaptive Molecular Evolution in Eukaryotes. *Genome Biology and Evolution*, *4*(5), 658–667. https://doi.org/10.1093/gbe/evs027

Gossmann, T. I., Song, B.-H., Windsor, A. J., Mitchell-Olds, T., Dixon, C. J., Kapralov, M. V., Filatov, D. A., & Eyre-Walker, A. (2010). Genome wide analyses reveal little evidence for adaptive evolution in many plant species. *Molecular Biology and Evolution*, *27*(8), 1822–1832. https://doi.org/10.1093/molbev/msq079

Gossmann, T. I., Woolfit, M., & Eyre-Walker, A. (2011). Quantifying the Variation in the Effective Population Size Within a Genome. *Genetics*, *189*(4), 1389–1402. https://doi.org/10.1534/genetics.111.132654

Grant, B. J., Rodrigues, A. P. C., ElSawy, K. M., McCammon, J. A., & Caves, L. S. D. (2006). Bio3d: An R package for the comparative analysis of protein structures. *Bioinformatics*, *22*(21), 2695–2696. https://doi.org/10.1093/bioinformatics/btl461

Grantham, R. (1974). Amino Acid Difference Formula to Help Explain Protein Evolution. *Science*, *185*(4154), 862–864. https://doi.org/10.1126/science.185.4154.862

Gravel, S., Henn, B. M., Gutenkunst, R. N., Indap, A. R., Marth, G. T., Clark, A. G., Yu, F., Gibbs, R. A., The 1000 Genomes Project, Bustamante, C. D., Altshuler, D. L., Durbin, R. M., Abecasis, G. R., Bentley, D. R., Chakravarti, A., Clark, A. G., Collins, F. S., De La Vega, F. M., Donnelly, P., … McVean, G. A. (2011). Demographic history and rare allele sharing among human populations.

Proceedings of the National Academy of Sciences, *108*(29), 11983–11988. https://doi.org/10.1073/pnas.1019276108

Guo, H. H., Choe, J., & Loeb, L. A. (2004). Protein tolerance to random amino acid change. *Proceedings of the National Academy of Sciences*, *101*(25), 9205–9210. https://doi.org/10.1073/pnas.0403255101

Gutenkunst, R. N., Hernandez, R. D., Williamson, S. H., & Bustamante, C. D. (2009). Inferring the Joint Demographic History of Multiple Populations from Multidimensional SNP Frequency Data. *PLoS Genetics*, *5*(10), e1000695. https://doi.org/10.1371/journal.pgen.1000695

Haag-Liautard, C., Dorris, M., Maside, X., Macaskill, S., Halligan, D. L., Houle, D., Charlesworth, B., & Keightley, P. D. (2007). Direct estimation of per nucleotide and genomic deleterious mutation rates in Drosophila. *Nature*, *445*(7123), 82–85. https://doi.org/10.1038/nature05388

Haddrill, P. R., Loewe, L., & Charlesworth, B. (2010). Estimating the parameters of selection on nonsynonymous mutations in Drosophila pseudoobscura and D. miranda. *Genetics*, *185*(4), 1381–1396. https://doi.org/10.1534/genetics.110.117614

Haerty, W., Jagadeeshan, S., Kulathinal, R. J., Wong, A., Ravi Ram, K., Sirot, L. K., Levesque, L., Artieri, C. G., Wolfner, M. F., Civetta, A., & Singh, R. S. (2007). Evolution in the fast lane: Rapidly evolving sex-related genes in Drosophila. *Genetics*, *177*(3), 1321–1335. https://doi.org/10.1534/genetics.107.078865

Haldane, J. B. S. (1957). The cost of natural selection. *Journal of Genetics*, *55*(3), 511–524. https://doi.org/10.1007/BF02984069

Halldorsson, B. V., Hardarson, M. T., Kehr, B., Styrkarsdottir, U., Gylfason, A., Thorleifsson, G., Zink, F., Jonasdottir, A., Jonasdottir, A., Sulem, P., Masson, G., Thorsteinsdottir, U., Helgason, A., Kong, A., Gudbjartsson, D. F., & Stefansson, K. (2016). The rate of meiotic gene conversion varies by sex and age. *Nature Genetics*, *48*(11), 1377–1384. https://doi.org/10.1038/ng.3669

Haller, B. C., & Messer, P. W. (2019). SLiM 3: Forward Genetic Simulations Beyond the Wright–Fisher Model. *Molecular Biology and Evolution*, *36*(3), 632–637. https://doi.org/10.1093/molbev/msy228

Halligan, D. L., & Keightley, P. D. (2006). Ubiquitous selective constraints in the Drosophila genome revealed by a genome-wide interspecies comparison. *Genome Research*, *16*(7), 875–884. https://doi.org/10.1101/gr.5022906

Halligan, D. L., Oliver, F., Eyre-Walker, A., Harr, B., & Keightley, P. D. (2010). Evidence for Pervasive Adaptive Protein Evolution in Wild Mice. *PLoS Genetics*, *6*(1), e1000825. https://doi.org/10.1371/journal.pgen.1000825

Harpak, A., Bhaskar, A., & Pritchard, J. K. (2016). Mutation Rate Variation is a Primary Determinant of the Distribution of Allele Frequencies in Humans. *PLOS Genetics*, *12*(12), e1006489. https://doi.org/10.1371/journal.pgen.1006489

Harris, K. (2015). Evidence for recent, population-specific evolution of the human mutation rate. *Proceedings of the National Academy of Sciences*, *112*(11), 3439–3444. https://doi.org/10.1073/pnas.1418652112

Hedrick, P. W. (1998). [No title found]. *Genetica*, *104*(3), 207–214. https://doi.org/10.1023/A:1026494212540

Hedrick, P. W. (2002). PATHOGEN RESISTANCE AND GENETIC VARIATION AT MHC LOCI. *Evolution*, *56*(10), 1902–1908. https://doi.org/10.1111/j.0014-3820.2002.tb00116.x

Hedrick, P. W. (2012). What is the evidence for heterozygote advantage selection? *Trends in Ecology & Evolution*, *27*(12), 698–704. https://doi.org/10.1016/j.tree.2012.08.012

Heller, R., Chikhi, L., & Siegismund, H. R. (2013). The Confounding Effect of Population Structure on Bayesian Skyline Plot Inferences of Demographic History. *PLoS ONE*, *8*(5), e62992. https://doi.org/10.1371/journal.pone.0062992

Hellmann, I., Ebersberger, I., Ptak, S. E., Pääbo, S., & Przeworski, M. (2003). A Neutral Explanation for the Correlation of Diversity with Recombination Rates in Humans. *The American Journal of Human Genetics*, *72*(6), 1527–1535. https://doi.org/10.1086/375657

Hellmann, I., Prüfer, K., Ji, H., Zody, M. C., Pääbo, S., & Ptak, S. E. (2005). Why do human diversity levels vary at a megabase scale? *Genome Research*, *15*(9), 1222–1231. https://doi.org/10.1101/gr.3461105

Henn, B. M., Botigué, L. R., Peischl, S., Dupanloup, I., Lipatov, M., Maples, B. K., Martin, A. R., Musharoff, S., Cann, H., Snyder, M. P., Excoffier, L., Kidd, J. M., & Bustamante, C. D. (2016). Distance from sub-Saharan Africa predicts mutational load in diverse human genomes. *Proceedings of the National Academy of Sciences*, *113*(4), E440–E449. https://doi.org/10.1073/pnas.1510805112

Hernandez, R. D., Kelley, J. L., Elyashiv, E., Melton, S. C., Auton, A., McVean, G., 1000 Genomes Project, Sella, G., & Przeworski, M. (2011). Classic Selective Sweeps Were Rare in Recent Human Evolution. *Science*, *331*(6019), 920–924. https://doi.org/10.1126/science.1198878

Hershberg, R., & Petrov, D. A. (2008). Selection on codon bias. *Annual Review of Genetics*, *42*, 287–299. https://doi.org/10.1146/annurev.genet.42.110807.091442

Hill, W. G., & Robertson, A. (1966). The effect of linkage on limits to artificial selection. *Genetical Research*, *8*(3), 269–294.

Hobolth, A., Christensen, O. F., Mailund, T., & Schierup, M. H. (2007). Genomic Relationships and Speciation Times of Human, Chimpanzee, and Gorilla Inferred from a Coalescent Hidden Markov Model. *PLoS Genetics*, *3*(2), e7. https://doi.org/10.1371/journal.pgen.0030007

Hodgkinson, A., Ladoukakis, E., & Eyre-Walker, A. (2009). Cryptic Variation in the Human Mutation Rate. *PLoS Biology*, *7*(2), e1000027. https://doi.org/10.1371/journal.pbio.1000027

Huber, C. D., DeGiorgio, M., Hellmann, I., & Nielsen, R. (2016). Detecting recent selective sweeps while controlling for mutation rate and background selection. *Molecular Ecology*, *25*(1), 142–156. https://doi.org/10.1111/mec.13351

Huber, C. D., Durvasula, A., Hancock, A. M., & Lohmueller, K. E. (2018). Gene expression drives the evolution of dominance. *Nature Communications*, *9*(1), 2750. https://doi.org/10.1038/s41467-018-05281-7

Hudson, R. R., & Kaplan, N. L. (1995). Deleterious background selection with recombination. *Genetics*, *141*(4), 1605.

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

Hughes, A. L. (2005). Evidence for Abundant Slightly Deleterious Polymorphisms in Bacterial Populations. *Genetics*, *169*(2), 533–538. https://doi.org/10.1534/genetics.104.036939

Hughes, A. L., & Nei, M. (1988). Pattern of nucleotide substitution at major histocompatibility complex class I loci reveals overdominant selection. *Nature*, *335*(6186), 167–170. https://doi.org/10.1038/335167a0

Hughes, A. L., Packer, B., Welch, R., Bergen, A. W., Chanock, S. J., & Yeager, M. (2003). Widespread purifying selection at polymorphic sites in human protein-coding loci. *Proceedings of the National Academy of Sciences*, *100*(26), 15754–15757. https://doi.org/10.1073/pnas.2536718100

Huh, G.-Y., Glantz, S. B., Je, S., Morrow, J. S., & Kim, J. H. (2001). Calpain proteolysis of αII-spectrin in the normal adult human brain. *Neuroscience Letters*, *316*(1), 41–44. https://doi.org/10.1016/S0304-3940(01)02371-0

Hunter-Zinck, H., & Clark, A. G. (2015). Aberrant Time to Most Recent Common Ancestor as a Signature of Natural Selection. *Molecular Biology and Evolution*, *32*(10), 2784–2797. https://doi.org/10.1093/molbev/msv142

Hurst, L. D., & Smith, N. G. (1999). Do essential genes evolve slowly? *Current Biology: CB*, *9*(14), 747–750. https://doi.org/10.1016/s0960-9822(99)80334-0

Ingvarsson, P. K. (2004). Population subdivision and the Hudson–Kreitman–Aguade test: Testing for deviations from the neutral model in organelle genomes. *Genetical Research*, *83*(1), 31–39. https://doi.org/10.1017/S0016672303006529

Ingvarsson, P. K. (2010). Natural Selection on Synonymous and Nonsynonymous Mutations Shapes Patterns of Polymorphism in Populus tremula. *Molecular Biology and Evolution*, *27*(3), 650–660. https://doi.org/10.1093/molbev/msp255

Innan, H. (2006). The Effect of Gene Flow on the Coalescent Time in the Human-Chimpanzee Ancestral Population. *Molecular Biology and Evolution*, *23*(5), 1040–1047. https://doi.org/10.1093/molbev/msj109

Jensen, J. D., Kim, Y., DuMont, V. B., Aquadro, C. F., & Bustamante, C. D. (2005). Distinguishing Between Selective Sweeps and Demography Using DNA Polymorphism Data. *Genetics*, *170*(3), 1401–1410. https://doi.org/10.1534/genetics.104.038224

Jensen, J. D., Payseur, B. A., Stephan, W., Aquadro, C. F., Lynch, M., Charlesworth, D., & Charlesworth, B. (2019). The importance of the Neutral Theory in 1968 and 50 years on: A response to Kern and Hahn 2018: COMMENTARY. *Evolution*, *73*(1), 111–114. https://doi.org/10.1111/evo.13650

Johnson, P. L. F., & Hellmann, I. (2011). Mutation Rate Distribution Inferred from Coincident SNPs and Coincident Substitutions. *Genome Biology and Evolution*, *3*, 842–850. https://doi.org/10.1093/gbe/evr044

Johri, P., Charlesworth, B., & Jensen, J. D. (2020). Toward an Evolutionarily Appropriate Null Model: Jointly Inferring Demography and Purifying Selection. *Genetics*, *215*(1), 173–192. https://doi.org/10.1534/genetics.119.303002

Jónsson, H., Sulem, P., Kehr, B., Kristmundsdottir, S., Zink, F., Hjartarson, E., Hardarson, M. T., Hjorleifsson, K. E., Eggertsson, H. P., Gudjonsson, S. A., Ward, L. D., Arnadottir, G. A., Helgason, E. A., Helgason, H., Gylfason, A., Jonasdottir, A., Jonasdottir, A., Rafnar, T., Besenbacher, S., … Stefansson, K. (2017). Whole genome characterization of sequence diversity of 15,220 Icelanders. *Scientific Data*, *4*(1), 170115. https://doi.org/10.1038/sdata.2017.115

Katzman, S., Kern, A. D., Bejerano, G., Fewell, G., Fulton, L., Wilson, R. K., Salama, S. R., & Haussler, D. (2007). Human genome ultraconserved elements are ultraselected. *Science (New York, N.Y.)*, *317*(5840), 915. https://doi.org/10.1126/science.1142430

Keightley, P. D., Lercher, M. J., & Eyre-Walker, A. (2005). Evidence for Widespread Degradation of Gene Control Regions in Hominid Genomes. *PLoS Biology*, *3*(2), e42. https://doi.org/10.1371/journal.pbio.0030042

Keinan, A., & Clark, A. G. (2012). Recent explosive human population growth has resulted in an excess of rare genetic variants. *Science (New York, N.Y.)*, *336*(6082), 740–743. https://doi.org/10.1126/science.1217283

Keith, N., Tucker, A. E., Jackson, C. E., Sung, W., Lucas Lledó, J. I., Schrider, D. R., Schaack, S., Dudycha, J. L., Ackerman, M., Younge, A. J., Shaw, J. R., & Lynch, M. (2016). High mutational rates of large-scale duplication and deletion in Daphnia pulex. *Genome Research*, *26*(1), 60–69. https://doi.org/10.1101/gr.191338.115

Kern, A. D., & Hahn, M. W. (2018). The Neutral Theory in Light of Natural Selection. *Molecular Biology and Evolution*, *35*(6), 1366–1371. https://doi.org/10.1093/molbev/msy092

Kidd, J. M., Gravel, S., Byrnes, J., Moreno-Estrada, A., Musharoff, S., Bryc, K., Degenhardt, J. D., Brisbin, A., Sheth, V., Chen, R., McLaughlin, S. F., Peckham, H. E., Omberg, L., Bormann Chung, C. A., Stanley, S., Pearlstein, K., Levandowsky, E., Acevedo-Acevedo, S., Auton, A., … Bustamante, C. D. (2012). Population Genetic Inference from Personal Genome Data: Impact of Ancestry and Admixture on Human Genomic Variation. *The American Journal of Human Genetics*, *91*(4), 660–671. https://doi.org/10.1016/j.ajhg.2012.08.025

Kim, Y., & Stephan, W. (2002). Detecting a Local Signature of Genetic Hitchhiking Along a Recombining Chromosome. *Genetics*, *160*(2), 765.

Kimura, M. (1968). Evolutionary Rate at the Molecular Level. *Nature*, *217*(5129), 624–626. https://doi.org/10.1038/217624a0

Kimura, M. (1983). *The Neutral Theory of Molecular Evolution* (1st ed.). Cambridge University Press. https://doi.org/10.1017/CBO9780511623486

Kimura, M. (1991). The neutral theory of molecular evolution: A review of recent evidence. *Idengaku Zasshi*, *66*(4), 367–386. https://doi.org/10.1266/jjg.66.367

King, J. L., & Jukes, T. H. (1969). Non-Darwinian Evolution. *Science*, *164*(3881), 788–798. https://doi.org/10.1126/science.164.3881.788

Kingman, J. F. C. (1982). The coalescent. *Stochastic Processes and Their Applications*, *13*(3), 235–248. https://doi.org/10.1016/0304-4149(82)90011-4

Krylov, D. M., Wolf, Y. I., Rogozin, I. B., & Koonin, E. V. (2003). Gene loss, protein sequence divergence, gene dispensability, expression level, and interactivity are correlated in eukaryotic evolution. *Genome Research*, *13*(10), 2229–2235. https://doi.org/10.1101/gr.1589103

Kryukov, G. V., Schmidt, S., & Sunyaev, S. (2005). Small fitness effect of mutations in highly conserved non-coding regions. *Human Molecular Genetics*, *14*(15), 2221–2229. https://doi.org/10.1093/hmg/ddi226

Kudla, G., Murray, A. W., Tollervey, D., & Plotkin, J. B. (2009). Coding-sequence determinants of gene expression in Escherichia coli. *Science (New York, N.Y.)*, *324*(5924), 255–258. https://doi.org/10.1126/science.1170160

Lachance, J., & Tishkoff, S. A. (2014). Biased gene conversion skews allele frequencies in human populations, increasing the disease burden of recessive alleles. *American Journal of Human Genetics*, *95*(4), 408–420. https://doi.org/10.1016/j.ajhg.2014.09.008

Langley, C. H., MacDonald, J., Miyashita, N., & Aguade, M. (1993). Lack of correlation between interspecific divergence and intraspecific polymorphism at the suppressor of forked region in Drosophila melanogaster and Drosophila simulans. *Proceedings of the National Academy of Sciences*, *90*(5), 1800–1803. https://doi.org/10.1073/pnas.90.5.1800

Langley, C. H., Stevens, K., Cardeno, C., Lee, Y. C. G., Schrider, D. R., Pool, J. E., Langley, S. A., Suarez, C., Corbett-Detig, R. B., Kolaczkowski, B., Fang, S., Nista, P. M., Holloway, A. K., Kern, A. D., Dewey, C. N., Song, Y. S., Hahn, M. W., & Begun, D. J. (2012). Genomic variation in natural populations of Drosophila melanogaster. *Genetics*, *192*(2), 533–598. https://doi.org/10.1534/genetics.112.142018

Leffler, E. M., Gao, Z., Pfeifer, S., Segurel, L., Auton, A., Venn, O., Bowden, R., Bontrop, R., Wall, J. D., Sella, G., Donnelly, P., McVean, G., & Przeworski, M. (2013). Multiple Instances of Ancient Balancing Selection Shared Between Humans and Chimpanzees. *Science*, *339*(6127), 1578–1582. https://doi.org/10.1126/science.1234070

Lemos, B., Bettencourt, B. R., Meiklejohn, C. D., & Hartl, D. L. (2005). Evolution of proteins and gene expression levels are coupled in Drosophila and are independently associated with mRNA abundance, protein length, and number of protein-protein interactions. *Molecular Biology and Evolution*, *22*(5), 1345–1354. https://doi.org/10.1093/molbev/msi122

Lercher, M. J., & Hurst, L. D. (2002). Human SNP variability and mutation rate are higher in regions of high recombination. *Trends in Genetics: TIG*, *18*(7), 337–340. https://doi.org/10.1016/s0168-9525(02)02669-0

Lesecque, Y., Keightley, P. D., & Eyre-Walker, A. (2012). A resolution of the mutation load paradox in humans. *Genetics*, *191*(4), 1321–1330. https://doi.org/10.1534/genetics.112.140343

Lesecque, Y., Mouchiroud, D., & Duret, L. (2013). GC-biased gene conversion in yeast is specifically associated with crossovers: Molecular mechanisms and evolutionary significance. *Molecular Biology and Evolution*, *30*(6), 1409–1419. https://doi.org/10.1093/molbev/mst056

Levins, R., & Lewontin, R. (1977). *The Dialectical Biologist*. Harvard University Press.

Li, H., & Durbin, R. (2011). Inference of human population history from individual whole-genome sequences. *Nature*, *475*(7357), 493–496. https://doi.org/10.1038/nature10231

Li, W. H., Wu, C. I., & Luo, C. C. (1985). A new method for estimating synonymous and nonsynonymous rates of nucleotide substitution considering the relative likelihood of nucleotide and codon changes. *Molecular Biology and Evolution*, *2*(2), 150–174. https://doi.org/10.1093/oxfordjournals.molbev.a040343

Liao, B.-Y., Scott, N. M., & Zhang, J. (2006). Impacts of gene essentiality, expression pattern, and gene compactness on the evolutionary rate of mammalian proteins. *Molecular Biology and Evolution*, *23*(11), 2072–2080. https://doi.org/10.1093/molbev/msl076

Liberles, D. A., Teichmann, S. A., Bahar, I., Bastolla, U., Bloom, J., Bornberg-Bauer, E., Colwell, L. J., de Koning, A. P. J., Dokholyan, N. V., Echave, J., Elofsson, A., Gerloff, D. L., Goldstein, R. A., Grahnen, J. A., Holder, M. T., Lakner, C., Lartillot, N., Lovell, S. C., Naylor, G., … Whelan, S. (2012). The interface of protein structure, protein biophysics, and molecular evolution. *Protein Science*, *21*(6), 769–785. https://doi.org/10.1002/pro.2071

Lin, Y.-S., Hsu, W.-L., Hwang, J.-K., & Li, W.-H. (2007). Proportion of solvent-exposed amino acids in a protein and rate of protein evolution. *Molecular Biology and Evolution*, *24*(4), 1005–1011. https://doi.org/10.1093/molbev/msm019

Lipman, D. J., Souvorov, A., Koonin, E. V., Panchenko, A. R., & Tatusova, T. A. (2002). The relationship of protein conservation and sequence length. *BMC Evolutionary Biology*, *2*, 20. https://doi.org/10.1186/1471-2148-2-20

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., … Louis, E. J. (2009). Population genomics of domestic and wild yeasts. *Nature*, *458*(7236), 337–341. https://doi.org/10.1038/nature07743

Litman, T., & Stein, W. D. (2019a). Obtaining estimates for the ages of all the protein-coding genes and most of the ontology-identified noncoding genes of the human genome, assigned to 19 phylostrata. *Seminars in Oncology*, *46*(1), 3–9. https://doi.org/10.1053/j.seminoncol.2018.11.002

Litman, T., & Stein, W. D. (2019b). Obtaining estimates for the ages of all the protein-coding genes and most of the ontology-identified noncoding genes of the human genome, assigned to 19 phylostrata. *Seminars in Oncology*, *46*(1), 3–9. https://doi.org/10.1053/j.seminoncol.2018.11.002

Lohmueller, K. E., Indap, A. R., Schmidt, S., Boyko, A. R., Hernandez, R. D., Hubisz, M. J., Sninsky, J. J., White, T. J., Sunyaev, S. R., Nielsen, R., Clark, A. G., & Bustamante, C. D. (2008). Proportionally more deleterious genetic variation in European than in African populations. *Nature*, *451*(7181), 994–997. https://doi.org/10.1038/nature06611

Long, H., Sung, W., Kucukyildirim, S., Williams, E., Miller, S. F., Guo, W., Patterson, C., Gregory, C., Strauss, C., Stone, C., Berne, C., Kysela, D., Shoemaker, W. R., Muscarella, M. E., Luo, H., Lennon, J. T., Brun, Y. V., & Lynch, M. (2018). Evolutionary determinants of genome-wide nucleotide composition. *Nature Ecology & Evolution*, *2*(2), 237–240. https://doi.org/10.1038/s41559-017-0425-y

Long, M., & Langley, C. (1993). Natural selection and the origin of jingwei, a chimeric processed functional gene in Drosophila. *Science*, *260*(5104), 91–95. https://doi.org/10.1126/science.7682012

Lourenço, J. M., Glémin, S., & Galtier, N. (2013). The Rate of Molecular Adaptation in a Changing Environment. *Molecular Biology and Evolution*, *30*(6), 1292–1301. https://doi.org/10.1093/molbev/mst026

Lynch, M. (2002). GENOMICS: Gene Duplication and Evolution. *Science*, *297*(5583), 945–947. https://doi.org/10.1126/science.1075472

Lynch, M., Ackerman, M. S., Gout, J.-F., Long, H., Sung, W., Thomas, W. K., & Foster, P. L. (2016). Genetic drift, selection and the evolution of the mutation rate. *Nature Reviews Genetics*, *17*(11), 704–714. https://doi.org/10.1038/nrg.2016.104

Machado, C. A., Kliman, R. M., Markert, J. A., & Hey, J. (2002). Inferring the History of Speciation from Multilocus DNA Sequence Data: The Case of Drosophila pseudoobscura and Close Relatives. *Molecular Biology and Evolution*, *19*(4), 472–488. https://doi.org/10.1093/oxfordjournals.molbev.a004103

Mackay, T. F. C., Richards, S., Stone, E. A., Barbadilla, A., Ayroles, J. F., Zhu, D., Casillas, S., Han, Y., Magwire, M. M., Cridland, J. M., Richardson, M. F., Anholt, R. R. H., Barrón, M., Bess, C., Blankenburg, K. P., Carbone, M. A., Castellano, D., Chaboub, L., Duncan, L., … Gibbs, R. A.

(2012). The Drosophila melanogaster Genetic Reference Panel. *Nature*, *482*(7384), 173–178. https://doi.org/10.1038/nature10811

Malaspinas, A.-S., Westaway, M. C., Muller, C., Sousa, V. C., Lao, O., Alves, I., Bergström, A., Athanasiadis, G., Cheng, J. Y., Crawford, J. E., Heupink, T. H., Macholdt, E., Peischl, S., Rasmussen, S., Schiffels, S., Subramanian, S., Wright, J. L., Albrechtsen, A., Barbieri, C., … Willerslev, E. (2016). A genomic history of Aboriginal Australia. *Nature*, *538*(7624), 207–214. https://doi.org/10.1038/nature18299

Mallick, S., Li, H., Lipson, M., Mathieson, I., Gymrek, M., Racimo, F., Zhao, M., Chennagiri, N., Nordenfelt, S., Tandon, A., Skoglund, P., Lazaridis, I., Sankararaman, S., Fu, Q., Rohland, N., Renaud, G., Erlich, Y., Willems, T., Gallo, C., … Reich, D. (2016). The Simons Genome Diversity Project: 300 genomes from 142 diverse populations. *Nature*, *538*(7624), 201–206. https://doi.org/10.1038/nature18964

Mancera, E., Bourgon, R., Brozzi, A., Huber, W., & Steinmetz, L. M. (2008). High-resolution mapping of meiotic crossovers and non-crossovers in yeast. *Nature*, *454*(7203), 479–485. https://doi.org/10.1038/nature07135

Marais, G., & Charlesworth, B. (2003). Genome evolution: Recombination speeds up adaptive evolution. *Current Biology: CB*, *13*(2), R68-70. https://doi.org/10.1016/s0960-9822(02)01432-x

Martín-Campos, J. M., Comerón, J. M., Miyashita, N., & Aguadé, M. (1992). Intraspecific and interspecific variation at the y-ac-sc region of Drosophila simulans and Drosophila melanogaster. *Genetics*, *130*(4), 805–816.

Matassi, G., Sharp, P. M., & Gautier, C. (1999). Chromosomal location effects on gene sequence evolution in mammals. *Current Biology*, *9*(15), 786–791. https://doi.org/10.1016/S0960-9822(99)80361-3

Mazet, O., Rodríguez, W., Grusea, S., Boitard, S., & Chikhi, L. (2016). On the importance of being structured: Instantaneous coalescence rates and human evolution—lessons for ancestral population size inference? *Heredity*, *116*(4), 362–371. https://doi.org/10.1038/hdy.2015.104

McDonald, J. H., & Kreitman, M. (1991). Adaptive protein evolution at the Adh locus in Drosophila. *Nature*, *351*(6328), 652–654. https://doi.org/10.1038/351652a0

McVean, G. A. T., & Cardin, N. J. (2005). Approximating the coalescent with recombination. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, *360*(1459), 1387–1393. https://doi.org/10.1098/rstb.2005.1673

McVean, G. A. T., Myers, S. R., Hunt, S., Deloukas, P., Bentley, D. R., & Donnelly, P. (2004). The fine-scale structure of recombination rate variation in the human genome. *Science (New York, N.Y.)*, *304*(5670), 581–584. https://doi.org/10.1126/science.1092500

McVicker, G., Gordon, D., Davis, C., & Green, P. (2009). Widespread Genomic Signatures of Natural Selection in Hominid Evolution. *PLoS Genetics*, *5*(5), e1000471. https://doi.org/10.1371/journal.pgen.1000471

Meader, S., Ponting, C. P., & Lunter, G. (2010). Massive turnover of functional sequence in human and other mammalian genomes. *Genome Research*, *20*(10), 1335–1343. https://doi.org/10.1101/gr.108795.110

Messer, P. W., & Petrov, D. A. (2013). Frequent adaptation and the McDonald-Kreitman test. *Proceedings of the National Academy of Sciences of the United States of America*, *110*(21), 8615–8620. https://doi.org/10.1073/pnas.1220835110

Meunier, J., & Duret, L. (2004). Recombination drives the evolution of GC-content in the human genome. *Molecular Biology and Evolution*, *21*(6), 984–990. https://doi.org/10.1093/molbev/msh070

Michaelson, J. J., Shi, Y., Gujral, M., Zheng, H., Malhotra, D., Jin, X., Jian, M., Liu, G., Greer, D., Bhandari, A., Wu, W., Corominas, R., Peoples, A., Koren, A., Gore, A., Kang, S., Lin, G. N., Estabillo, J., Gadomski, T., … Sebat, J. (2012). Whole-genome sequencing in autism identifies hot spots for de novo germline mutation. *Cell*, *151*(7), 1431–1442. https://doi.org/10.1016/j.cell.2012.11.019

Miyashita, N. T. (1990). Molecular and phenotypic variation of the Zw locus region in Drosophila melanogaster. *Genetics*, *125*(2), 407–419.

Montoya-Burgos, J. I., Boursot, P., & Galtier, N. (2003). Recombination explains isochores in mammalian genomes. *Trends in Genetics: TIG*, *19*(3), 128–130. https://doi.org/10.1016/S0168-9525(03)00021-0

Moriyama, E. N., & Powell, J. R. (1996). Intraspecific nuclear DNA variation in Drosophila. *Molecular Biology and Evolution*, *13*(1), 261–277. https://doi.org/10.1093/oxfordjournals.molbev.a025563

Moutinho, A. F., Trancoso, F. F., & Dutheil, J. Y. (2019). The Impact of Protein Architecture on Adaptive Evolution. *Molecular Biology and Evolution*, *36*(9), 2013–2028. https://doi.org/10.1093/molbev/msz134

Mugal, C. F., Weber, C. C., & Ellegren, H. (2015). GC-biased gene conversion links the recombination landscape and demography to genomic base composition: GC-biased gene conversion drives genomic base composition across a wide range of species. *BioEssays: News and Reviews in Molecular, Cellular and Developmental Biology*, *37*(12), 1317–1326. https://doi.org/10.1002/bies.201500058

Murphy, D., Elyashiv, E., Amster, G., & Sella, G. (2021). Broad-scale variation in human genetic diversity levels is predicted by purifying selection on coding and non-coding elements. *BioRxiv*, 2021.07.02.450762. https://doi.org/10.1101/2021.07.02.450762

Murray, G. G. R., Soares, A. E. R., Novak, B. J., Schaefer, N. K., Cahill, J. A., Baker, A. J., Demboski, J. R., Doll, A., Da Fonseca, R. R., Fulton, T. L., Gilbert, M. T. P., Heintzman, P. D., Letts, B., McIntosh, G., O'Connell, B. L., Peck, M., Pipes, M.-L., Rice, E. S., Santos, K. M., … Shapiro, B. (2017). Natural selection shaped the rise and fall of passenger pigeon genomic diversity. *Science*, *358*(6365), 951–954. https://doi.org/10.1126/science.aao0960

Myers, S., Bottolo, L., Freeman, C., McVean, G., & Donnelly, P. (2005). A fine-scale map of recombination rates and hotspots across the human genome. *Science (New York, N.Y.)*, *310*(5746), 321–324. https://doi.org/10.1126/science.1117196

Nachman, M. W. (2001). Single nucleotide polymorphisms and recombination rate in humans. *Trends in Genetics: TIG*, *17*(9), 481–485. https://doi.org/10.1016/s0168-9525(01)02409-x

Nachman, M. W., Bauer, V. L., Crowell, S. L., & Aquadro, C. F. (1998). DNA variability and recombination rates at X-linked loci in humans. *Genetics*, *150*(3), 1133–1141.

Nachman, M. W., & Crowell, S. L. (2000). Estimate of the mutation rate per nucleotide in humans. *Genetics*, *156*(1), 297–304.

Necşulea, A., Popa, A., Cooper, D. N., Stenson, P. D., Mouchiroud, D., Gautier, C., & Duret, L. (2011). Meiotic recombination favors the spreading of deleterious mutations in human populations. *Human Mutation*, *32*(2), 198–206. https://doi.org/10.1002/humu.21407

Nei, M., & Gojobori, T. (1986). Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. *Molecular Biology and Evolution*, *3*(5), 418–426. https://doi.org/10.1093/oxfordjournals.molbev.a040410

Nelson, M. R., Wegmann, D., Ehm, M. G., Kessner, D., St Jean, P., Verzilli, C., Shen, J., Tang, Z., Bacanu, S.-A., Fraser, D., Warren, L., Aponte, J., Zawistowski, M., Liu, X., Zhang, H., Zhang, Y., Li, J., Li, Y., Li, L., … Mooser, V. (2012). An abundance of rare functional variants in 202 drug target

genes sequenced in 14,002 people. *Science (New York, N.Y.)*, *337*(6090), 100–104. https://doi.org/10.1126/science.1217876

Neme, R., & Tautz, D. (2013). Phylogenetic patterns of emergence of new genes support a model of frequent de novo evolution. *BMC Genomics*, *14*, 117. https://doi.org/10.1186/1471-2164-14-117

Nielsen, R. (2000). Estimation of Population Parameters and Recombination Rates From Single Nucleotide Polymorphisms. *Genetics*, *154*(2), 931–942. https://doi.org/10.1093/genetics/154.2.931

Nielsen, R. (2001). Statistical tests of selective neutrality in the age of genomics. *Heredity*, *86*(6), 641–647. https://doi.org/10.1046/j.1365-2540.2001.00895.x

Nielsen, R. (2005). Molecular Signatures of Natural Selection. *Annual Review of Genetics*, *39*(1), 197–218. https://doi.org/10.1146/annurev.genet.39.073003.112420

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., J. Sninsky, J., Adams, M. D., & Cargill, M. (2005). A Scan for Positively Selected Genes in the Genomes of Humans and Chimpanzees. *PLoS Biology*, *3*(6), e170. https://doi.org/10.1371/journal.pbio.0030170

Nielsen, R., Hellmann, I., Hubisz, M., Bustamante, C., & Clark, A. G. (2007). Recent and ongoing selection in the human genome. *Nature Reviews Genetics*, *8*(11), 857–868. https://doi.org/10.1038/nrg2187

Nielsen, R., Williamson, S., Kim, Y., Hubisz, M. J., Clark, A. G., & Bustamante, C. (2005). Genomic scans for selective sweeps using SNP data. *Genome Research*, *15*(11), 1566–1575. https://doi.org/10.1101/gr.4252305

Nordborg, M., Charlesworth, B., & Charlesworth, D. (1996). The effect of recombination on background selection. *Genetical Research*, *67*(2), 159–174. https://doi.org/10.1017/s0016672300033619

Nosil, P., Villoutreix, R., de Carvalho, C. F., Farkas, T. E., Soria-Carrasco, V., Feder, J. L., Crespi, B. J., & Gompert, Z. (2018). Natural selection and the predictability of evolution in *Timema* stick insects. *Science*, *359*(6377), 765–770. https://doi.org/10.1126/science.aap9125

Obbard, D. J., Welch, J. J., Kim, K.-W., & Jiggins, F. M. (2009). Quantifying Adaptive Evolution in the Drosophila Immune System. *PLoS Genetics*, *5*(10), e1000698. https://doi.org/10.1371/journal.pgen.1000698

Ohta, T. (1974). Mutational pressure as the main cause of molecular evolution and polymorphism. *Nature*, *252*(5482), 351–354. https://doi.org/10.1038/252351a0

Orozco-terWengel, P. (2016). The devil is in the details: The effect of population structure on demographic inference. *Heredity*, *116*(4), 349–350. https://doi.org/10.1038/hdy.2016.9

Ott, J. (1999). *Analysis of human genetic linkage*. Johns Hopkins University Press.

Pagani, L., Lawson, D. J., Jagoda, E., Mörseburg, A., Eriksson, A., Mitt, M., Clemente, F., Hudjashov, G., DeGiorgio, M., Saag, L., Wall, J. D., Cardona, A., Mägi, R., Sayres, M. A. W., Kaewert, S., Inchley, C., Scheib, C. L., Järve, M., Karmin, M., … Metspalu, M. (2016). Genomic analyses inform on migration events during the peopling of Eurasia. *Nature*, *538*(7624), 238–242. https://doi.org/10.1038/nature19792

Pál, C., Papp, B., & Hurst, L. D. (2001). Highly expressed genes in yeast evolve slowly. *Genetics*, *158*(2), 927–931.

Papatheodorou, I., Moreno, P., Manning, J., Fuentes, A. M.-P., George, N., Fexova, S., Fonseca, N. A., Füllgrabe, A., Green, M., Huang, N., Huerta, L., Iqbal, H., Jianu, M., Mohammed, S., Zhao, L., Jarnuczak, A. F., Jupp, S., Marioni, J., Meyer, K., … Brazma, A. (2019). Expression Atlas update:

From tissues to single cells. *Nucleic Acids Research*, gkz947. https://doi.org/10.1093/nar/gkz947

Parsch, J., Novozhilov, S., Saminadin-Peter, S. S., Wong, K. M., & Andolfatto, P. (2010). On the utility of short intron sequences as a reference for the detection of positive and negative selection in Drosophila. *Molecular Biology and Evolution*, *27*(6), 1226–1234. https://doi.org/10.1093/molbev/msq046

Paterson, S. (1998). Evidence for balancing selection at the major histocompatibility complex in a free-living ruminant. *Journal of Heredity*, *89*(4), 289–294. https://doi.org/10.1093/jhered/89.4.289

Perutz, M. F., Kendrew, J. C., & Watson, H. C. (1965). Structure and function of haemoglobin. *Journal of Molecular Biology*, *13*(3), 669–678. https://doi.org/10.1016/S0022-2836(65)80134-6

Pessia, E., Popa, A., Mousset, S., Rezvoy, C., Duret, L., & Marais, G. A. B. (2012). Evidence for widespread GC-biased gene conversion in eukaryotes. *Genome Biology and Evolution*, *4*(7), 675–682. https://doi.org/10.1093/gbe/evs052

Peter, B. M., Wegmann, D., & Excoffier, L. (2010). Distinguishing between population bottleneck and population subdivision by a Bayesian model choice procedure. *Molecular Ecology*, *19*(21), 4648–4660. https://doi.org/10.1111/j.1365-294X.2010.04783.x

Piganeau, G., & Eyre-Walker, A. (2009). Evidence for Variation in the Effective Population Size of Animal Mitochondrial DNA. *PLoS ONE*, *4*(2), e4396. https://doi.org/10.1371/journal.pone.0004396

Pollard, K. S., Hubisz, M. J., Rosenbloom, K. R., & Siepel, A. (2010). Detection of nonneutral substitution rates on mammalian phylogenies. *Genome Research*, *20*(1), 110–121. https://doi.org/10.1101/gr.097857.109

Pollard, K. S., Salama, S. R., King, B., Kern, A. D., Dreszer, T., Katzman, S., Siepel, A., Pedersen, J. S., Bejerano, G., Baertsch, R., Rosenbloom, K. R., Kent, J., & Haussler, D. (2006). Forces Shaping the Fastest Evolving Regions in the Human Genome. *PLoS Genetics*, *2*(10), e168. https://doi.org/10.1371/journal.pgen.0020168

Ponting, C. P., & Hardison, R. C. (2011). What fraction of the human genome is functional? *Genome Research*, *21*(11), 1769–1776. https://doi.org/10.1101/gr.116814.110

Popadin, K., Polishchuk, L. V., Mamirova, L., Knorre, D., & Gunbin, K. (2007). Accumulation of slightly deleterious mutations in mitochondrial protein-coding genes of large versus small mammals. *Proceedings of the National Academy of Sciences*, *104*(33), 13390–13395. https://doi.org/10.1073/pnas.0701256104

Pouyet, F., Aeschbacher, S., Thiéry, A., & Excoffier, L. (2018). Background selection and biased gene conversion affect more than 95% of the human genome and bias demographic inferences. *ELife*, *7*, e36317. https://doi.org/10.7554/eLife.36317

Prado-Martinez, J., Sudmant, P. H., Kidd, J. M., Li, H., Kelley, J. L., Lorente-Galdos, B., Veeramah, K. R., Woerner, A. E., O'Connor, T. D., Santpere, G., Cagan, A., Theunert, C., Casals, F., Laayouni, H., Munch, K., Hobolth, A., Halager, A. E., Malig, M., Hernandez-Rodriguez, J., … Marques-Bonet, T. (2013). Great ape genetic diversity and population history. *Nature*, *499*(7459), 471–475. https://doi.org/10.1038/nature12228

Pratto, F., Brick, K., Khil, P., Smagulova, F., Petukhova, G. V., & Camerini-Otero, R. D. (2014a). DNA recombination. Recombination initiation maps of individual human genomes. *Science (New York, N.Y.)*, *346*(6211), 1256442. https://doi.org/10.1126/science.1256442

Pratto, F., Brick, K., Khil, P., Smagulova, F., Petukhova, G. V., & Camerini-Otero, R. D. (2014b). Recombination initiation maps of individual human genomes. *Science*, *346*(6211), 1256442–1256442. https://doi.org/10.1126/science.1256442

Presgraves, D. C. (2005). Recombination enhances protein adaptation in Drosophila melanogaster. *Current Biology: CB*, *15*(18), 1651–1656. https://doi.org/10.1016/j.cub.2005.07.065

Pröschel, M., Zhang, Z., & Parsch, J. (2006). Widespread adaptive evolution of Drosophila genes with sex-biased expression. *Genetics*, *174*(2), 893–900. https://doi.org/10.1534/genetics.106.058008

Przeworski, M., Hudson, R. R., & Di Rienzo, A. (2000). Adjusting the focus on human variation. *Trends in Genetics*, *16*(7), 296–302. https://doi.org/10.1016/S0168-9525(00)02030-8

Ptak, S. E., & Przeworski, M. (2002). Evidence for population growth in humans is confounded by fine-scale population structure. *Trends in Genetics*, *18*(11), 559–563. https://doi.org/10.1016/S0168-9525(02)02781-6

R Core Team, (2021). A language and environment for statistical computing. *R Foundation for statistical computing*.

Ramsey, D. C., Scherrer, M. P., Zhou, T., & Wilke, C. O. (2011). The Relationship Between Relative Solvent Accessibility and Evolutionary Rate in Protein Evolution. *Genetics*, *188*(2), 479–488. https://doi.org/10.1534/genetics.111.128025

Ratnakumar, A., Mousset, S., Glémin, S., Berglund, J., Galtier, N., Duret, L., & Webster, M. T. (2010). Detecting positive selection within genomes: The problem of biased gene conversion. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, *365*(1552), 2571–2580. https://doi.org/10.1098/rstb.2010.0007

Rocha, E. P. C., & Danchin, A. (2004). An analysis of determinants of amino acids substitution rates in bacterial proteins. *Molecular Biology and Evolution*, *21*(1), 108–116. https://doi.org/10.1093/molbev/msh004

Roselius, K., Stephan, W., & Städler, T. (2005). The Relationship of Nucleotide Polymorphism, Recombination Rate and Selection in Wild Tomato Species. *Genetics*, *171*(2), 753–763. https://doi.org/10.1534/genetics.105.043877

Rousselle, M., Mollion, M., Nabholz, B., Bataillon, T., & Galtier, N. (2018). Overestimation of the adaptive substitution rate in fluctuating populations. *Biology Letters*, *14*(5), 20180055. https://doi.org/10.1098/rsbl.2018.0055

Rousselle, M., Simion, P., Tilak, M.-K., Figuet, E., Nabholz, B., & Galtier, N. (2020). Is adaptation limited by mutation? A timescale-dependent effect of genetic diversity on the adaptive substitution rate in animals. *PLOS Genetics*, *16*(4), e1008668. https://doi.org/10.1371/journal.pgen.1008668

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., & Lander, E. S. (2002). Detecting recent positive selection in the human genome from haplotype structure. *Nature*, *419*(6909), 832–837. https://doi.org/10.1038/nature01140

Sackton, T. B., Lazzaro, B. P., Schlenke, T. A., Evans, J. D., Hultmark, D., & Clark, A. G. (2007). Dynamic evolution of the innate immune system in Drosophila. *Nature Genetics*, *39*(12), 1461–1468. https://doi.org/10.1038/ng.2007.60

Sawyer, S. A., Kulathinal, R. J., Bustamante, C. D., & Hartl, D. L. (2003). Bayesian Analysis Suggests that Most Amino Acid Replacements in Drosophila Are Driven by Positive Selection. *Journal of Molecular Evolution*, *57*(0), S154–S164. https://doi.org/10.1007/s00239-003-0022-3

Schäffer, A. A., Aravind, L., Madden, T. L., Shavirin, S., Spouge, J. L., Wolf, Y. I., Koonin, E. V., & Altschul, S. F. (2001). Improving the accuracy of PSI-BLAST protein database searches with

composition-based statistics and other refinements. *Nucleic Acids Research*, *29*(14), 2994–3005. https://doi.org/10.1093/nar/29.14.2994

Schiffels, S., & Durbin, R. (2014). Inferring human population size and separation history from multiple genome sequences. *Nature Genetics*, *46*(8), 919–925. https://doi.org/10.1038/ng.3015

Schmid, K. J., Ramos-Onsins, S., Ringys-Beckstein, H., Weisshaar, B., & 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. https://doi.org/10.1534/genetics.104.033795

Schmid, M., Davison, T. S., Henz, S. R., Pape, U. J., Demar, M., Vingron, M., Schölkopf, B., Weigel, D., & Lohmann, J. U. (2005). A gene expression map of Arabidopsis thaliana development. *Nature Genetics*, *37*(5), 501–506. https://doi.org/10.1038/ng1543

Schneider, A., Charlesworth, B., Eyre-Walker, A., & Keightley, P. D. (2011). A Method for Inferring the Rate of Occurrence and Fitness Effects of Advantageous Mutations. *Genetics*, *189*(4), 1427–1437. https://doi.org/10.1534/genetics.111.131730

Schrago, C. G. (2014). The Effective Population Sizes of the Anthropoid Ancestors of the Human-Chimpanzee Lineage Provide Insights on the Historical Biogeography of the Great Apes. *Molecular Biology and Evolution*, *31*(1), 37–47. https://doi.org/10.1093/molbev/mst191

Sellis, D., Callahan, B. J., Petrov, D. A., & Messer, P. W. (2011). Heterozygote advantage as a natural consequence of adaptation in diploids. *Proceedings of the National Academy of Sciences*, *108*(51), 20666–20671. https://doi.org/10.1073/pnas.1114573108

Sheehan, S., & Song, Y. S. (2016). Deep Learning for Population Genetic Inference. *PLOS Computational Biology*, *12*(3), e1004845. https://doi.org/10.1371/journal.pcbi.1004845

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., & Haussler, D. (2005). Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Research*, *15*(8), 1034–1050. https://doi.org/10.1101/gr.3715005

Siewert, K. M., & Voight, B. F. (2017). Detecting Long-Term Balancing Selection Using Allele Frequency Correlation. *Molecular Biology and Evolution*, *34*(11), 2996–3005. https://doi.org/10.1093/molbev/msx209

Simonsen, K. L., Churchill, G. A., & Aquadro, C. F. (1995). Properties of statistical tests of neutrality for DNA polymorphism data. *Genetics*, *141*(1), 413–429.

Slatkin, M., & Hudson, R. R. (1991). Pairwise comparisons of mitochondrial DNA sequences in stable and exponentially growing populations. *Genetics*, *129*(2), 555–562.

Slotte, T., Foxe, J. P., Hazzouri, K. M., & 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. *Molecular Biology and Evolution*, *27*(8), 1813–1821. https://doi.org/10.1093/molbev/msq062

Smeds, L., Mugal, C. F., Qvarnström, A., & Ellegren, H. (2016). High-Resolution Mapping of Crossover and Non-crossover Recombination Events by Whole-Genome Re-sequencing of an Avian Pedigree. *PLOS Genetics*, *12*(5), e1006044. https://doi.org/10.1371/journal.pgen.1006044

Smith, J. M., & Haigh, J. (1974). The hitch-hiking effect of a favourable gene. *Genetical Research*, *23*(1), 23–35. https://doi.org/10.1017/S0016672300014634

Smith, N. G. C., Brandström, M., & Ellegren, H. (2004). Evidence for turnover of functional noncoding DNA in mammalian genome evolution. *Genomics*, *84*(5), 806–813. https://doi.org/10.1016/j.ygeno.2004.07.012

Smith, N. G. C., & Eyre-Walker, A. (2002). Adaptive protein evolution in Drosophila. *Nature*, *415*(6875), 1022–1024. https://doi.org/10.1038/4151022a

Smith, T. C. A., Arndt, P. F., & Eyre-Walker, A. (2018). Large scale variation in the rate of germ-line de novo mutation, base composition, divergence and diversity in humans. *PLOS Genetics*, *14*(3), e1007254. https://doi.org/10.1371/journal.pgen.1007254

Soni, V., & Eyre-Walker, A. (2021). Factors that affect the rates of adaptive and non-adaptive evolution at the gene level in humans and chimpanzees. *BioRxiv*, 2021.05.05.442740. https://doi.org/10.1101/2021.05.05.442740

Soni, V., Moutinho, A. F., & Eyre-Walker, A. (2021). Site level factors that affect the rate of adaptive evolution in humans and chimpanzees; the effect of contracting population size. *BioRxiv*, 2021.05.28.446098. https://doi.org/10.1101/2021.05.28.446098

Soni, V., Vos, M., & Eyre-Walker, A. (2021). A new test suggests that balancing selection maintains hundreds of non-synonymous polymorphisms in the human genome. *BioRxiv*, 2021.02.08.430226. https://doi.org/10.1101/2021.02.08.430226

Spence, J. P., & Song, Y. S. (2019). Inference and analysis of population-specific fine-scale recombination maps across 26 diverse human populations. *Science Advances*, *5*(10), eaaw9206. https://doi.org/10.1126/sciadv.aaw9206

Spencer, C. C. A., Deloukas, P., Hunt, S., Mullikin, J., Myers, S., Silverman, B., Donnelly, P., Bentley, D., & McVean, G. (2006). The Influence of Recombination on Human Genetic Diversity. *PLoS Genetics*, *2*(9), e148. https://doi.org/10.1371/journal.pgen.0020148

Stajich, J. E. (2004). Disentangling the Effects of Demography and Selection in Human History. *Molecular Biology and Evolution*, *22*(1), 63–73. https://doi.org/10.1093/molbev/msh252

Stephan, W., & Langley, C. H. (1989). Molecular genetic variation in the centromeric region of the X chromosome in three Drosophila ananassae populations. I. Contrasts between the vermilion and forked loci. *Genetics*, *121*(1), 89–99.

Stephan, W., & Mitchell, S. J. (1992). Reduced levels of DNA polymorphism and fixed between-population differences in the centromeric region of Drosophila ananassae. *Genetics*, *132*(4), 1039–1045.

Stoletzki, N., & Eyre-Walker, A. (2011). Estimation of the neutrality index. *Molecular Biology and Evolution*, *28*(1), 63–70. https://doi.org/10.1093/molbev/msq249

Stoltzfus, A., & Norris, R. W. (2016). On the Causes of Evolutionary Transition:Transversion Bias. *Molecular Biology and Evolution*, *33*(3), 595–602. https://doi.org/10.1093/molbev/msv274

STOP-HCV Consortium, Ansari, M. A., Pedergnana, V., L C Ip, C., Magri, A., Von Delft, A., Bonsall, D., Chaturvedi, N., Bartha, I., Smith, D., Nicholson, G., McVean, G., Trebes, A., Piazza, P., Fellay, J., Cooke, G., Foster, G. R., Hudson, E., McLauchlan, J., … Spencer, C. C. A. (2017). Genome-to-genome analysis highlights the effect of the human innate and adaptive immune systems on the hepatitis C virus. *Nature Genetics*, *49*(5), 666–673. https://doi.org/10.1038/ng.3835

Strasburg, J. L., Kane, N. C., Raduski, A. R., Bonin, A., Michelmore, R., & Rieseberg, L. H. (2011). Effective Population Size Is Positively Correlated with Levels of Adaptive Divergence among Annual Sunflowers. *Molecular Biology and Evolution*, *28*(5), 1569–1580. https://doi.org/10.1093/molbev/msq270

Strasburg, J. L., & 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; International Journal of Organic Evolution*, *62*(8), 1936–1950. https://doi.org/10.1111/j.1558-5646.2008.00415.x

Strasburg, J. L., Scotti-Saintagne, C., Scotti, I., Lai, Z., & Rieseberg, L. H. (2009). Genomic Patterns of Adaptive Divergence between Chromosomally Differentiated Sunflower Species. *Molecular Biology and Evolution*, *26*(6), 1341–1355. https://doi.org/10.1093/molbev/msp043

Subramanian, S., & Kumar, S. (2004). Gene expression intensity shapes evolutionary rates of the proteins encoded by the vertebrate genome. *Genetics*, *168*(1), 373–381. https://doi.org/10.1534/genetics.104.028944

Sundström, H., Webster, M. T., & Ellegren, H. (2004). Reduced Variation on the Chicken Z Chromosome. *Genetics*, *167*(1), 377–385. https://doi.org/10.1534/genetics.167.1.377

Taddei, F., Radman, M., Maynard-Smith, J., Toupance, B., Gouyon, P. H., & Godelle, B. (1997). Role of mutator alleles in adaptive evolution. *Nature*, *387*(6634), 700–702. https://doi.org/10.1038/42696

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

Tataru, P., Mollion, M., Glémin, S., & Bataillon, T. (2017). Inference of Distribution of Fitness Effects and Proportion of Adaptive Substitutions from Polymorphism Data. *Genetics*, *207*(3), 1103–1119. https://doi.org/10.1534/genetics.117.300323

Tautz, D., & Domazet-Lošo, T. (2011). The evolutionary origin of orphan genes. *Nature Reviews. Genetics*, *12*(10), 692–702. https://doi.org/10.1038/nrg3053

Tenaillon, M. I. (2004). Selection Versus Demography: A Multilocus Investigation of the Domestication Process in Maize. *Molecular Biology and Evolution*, *21*(7), 1214–1225. https://doi.org/10.1093/molbev/msh102

Tenaillon, O., Toupance, B., Le Nagard, H., Taddei, F., & Godelle, B. (1999). Mutators, population size, adaptive landscape and the adaptation of asexual populations of bacteria. *Genetics*, *152*(2), 485–493.

Tennessen, J. A., & Akey, J. M. (2011). Parallel Adaptive Divergence among Geographically Diverse Human Populations. *PLoS Genetics*, *7*(6), e1002127. https://doi.org/10.1371/journal.pgen.1002127

Tennessen, J. A., Bigham, A. W., O'Connor, T. D., Fu, W., Kenny, E. E., Gravel, S., McGee, S., Do, R., Liu, X., Jun, G., Kang, H. M., Jordan, D., Leal, S. M., Gabriel, S., Rieder, M. J., Abecasis, G., Altshuler, D., Nickerson, D. A., Boerwinkle, E., … NHLBI Exome Sequencing Project. (2012). Evolution and functional impact of rare coding variation from deep sequencing of human exomes. *Science (New York, N.Y.)*, *337*(6090), 64–69. https://doi.org/10.1126/science.1219240

Terekhanova, N. V., Seplyarskiy, V. B., Soldatov, R. A., & Bazykin, G. A. (2017). Evolution of local mutation rate and its determinants. *Molecular Biology and Evolution*, msx060. https://doi.org/10.1093/molbev/msx060

The 1000 Genomes Project Consortium. (2015). A global reference for human genetic variation. *Nature*, *526*(7571), 68–74. https://doi.org/10.1038/nature15393

The Chimpanzee Sequencing and Analysis Consortium. (2005). Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature*, *437*(7055), 69–87. https://doi.org/10.1038/nature04072

The Gene Ontology Consortium, Carbon, S., Douglass, E., Good, B. M., Unni, D. R., Harris, N. L., Mungall, C. J., Basu, S., Chisholm, R. L., Dodson, R. J., Hartline, E., Fey, P., Thomas, P. D., Albou, L.-P., Ebert, D., Kesling, M. J., Mi, H., Muruganujan, A., Huang, X., … Elser, J. (2021). The Gene Ontology resource: Enriching a GOld mine. *Nucleic Acids Research*, *49*(D1), D325–D334. https://doi.org/10.1093/nar/gkaa1113

The Genome of the Netherlands Consortium. (2014). Whole-genome sequence variation, population structure and demographic history of the Dutch population. *Nature Genetics*, *46*(8), 818–825. https://doi.org/10.1038/ng.3021

The GTEx Consortium, Ardlie, K. G., Deluca, D. S., Segre, A. V., Sullivan, T. J., Young, T. R., Gelfand, E. T., Trowbridge, C. A., Maller, J. B., Tukiainen, T., Lek, M., Ward, L. D., Kheradpour, P., Iriarte, B., Meng, Y., Palmer, C. D., Esko, T., Winckler, W., Hirschhorn, J. N., … Dermitzakis, E. T. (2015). The Genotype-Tissue Expression (GTEx) pilot analysis: Multitissue gene regulation in humans. *Science*, *348*(6235), 648–660. https://doi.org/10.1126/science.1262110

The International HapMap Consortium. (2007). A second generation human haplotype map of over 3.1 million SNPs. *Nature*, *449*(7164), 851–861. https://doi.org/10.1038/nature06258

Theisen, D. J., Davidson, J. T., Briseño, C. G., Gargaro, M., Lauron, E. J., Wang, Q., Desai, P., Durai, V., Bagadia, P., Brickner, J. R., Beatty, W. L., Virgin, H. W., Gillanders, W. E., Mossamaparast, N., Diamond, M. S., Sibley, L. D., Yokoyama, W., Schreiber, R. D., Murphy, T. L., & Murphy, K. M. (2019). Wdfy4-deficiency reveals a critical role for cross-presentation in anti-viral and anti-tumor responses. *The Journal of Immunology*, *202*(1 Supplement), 177.23.

Thornton, K., & Long, M. (2002). Rapid divergence of gene duplicates on the Drosophila melanogaster X chromosome. *Molecular Biology and Evolution*, *19*(6), 918–925. https://doi.org/10.1093/oxfordjournals.molbev.a004149

Tien, M. Z., Meyer, A. G., Sydykova, D. K., Spielman, S. J., & Wilke, C. O. (2013). Maximum allowed solvent accessibilites of residues in proteins. *PloS One*, *8*(11), e80635. https://doi.org/10.1371/journal.pone.0080635

Tyekucheva, S., Makova, K. D., Karro, J. E., Hardison, R. C., Miller, W., & Chiaromonte, F. (2008). Human-macaque comparisons illuminate variation in neutral substitution rates. *Genome Biology*, *9*(4), R76. https://doi.org/10.1186/gb-2008-9-4-r76

Vigué, L., & Eyre-Walker, A. (2019). The comparative population genetics of Neisseria meningitidis and Neisseria gonorrhoeae. *PeerJ*, *7*, e7216. https://doi.org/10.7717/peerj.7216

Vishnoi, A., Kryazhimskiy, S., Bazykin, G. A., Hannenhalli, S., & Plotkin, J. B. (2010). Young proteins experience more variable selection pressures than old proteins. *Genome Research*, *20*(11), 1574–1581. https://doi.org/10.1101/gr.109595.110

Voight, B. F., Adams, A. M., Frisse, L. A., Qian, Y., Hudson, R. R., & Di Rienzo, A. (2005). Interrogating multiple aspects of variation in a full resequencing data set to infer human population size changes. *Proceedings of the National Academy of Sciences of the United States of America*, *102*(51), 18508–18513. https://doi.org/10.1073/pnas.0507325102

Voight, B. F., Kudaravalli, S., Wen, X., & Pritchard, J. K. (2006). A Map of Recent Positive Selection in the Human Genome. *PLoS Biology*, *4*(3), e72. https://doi.org/10.1371/journal.pbio.0040072

Wakeley, J. (2009). *Coalescent Theory: An Introduction*. Roberts and Company Publishers.

Wall, J. D., Andolfatto, P., & Przeworski, M. (2002). Testing models of selection and demography in Drosophila simulans. *Genetics*, *162*(1), 203–216.

Wall, J. D., & Przeworski, M. (2000). When did the human population size start increasing? *Genetics*, *155*(4), 1865–1874.

Wang, J., Santiago, E., & Caballero, A. (2016). Prediction and estimation of effective population size. *Heredity*, *117*(4), 193–206. https://doi.org/10.1038/hdy.2016.43

Wang, K., Li, M., & Hakonarson, H. (2010). ANNOVAR: Functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Research*, *38*(16), e164–e164. https://doi.org/10.1093/nar/gkq603

Wang, W., Zheng, H., Yang, S., Yu, H., Li, J., Jiang, H., Su, J., Yang, L., Zhang, J., McDermott, J., Samudrala, R., Wang, J., Yang, H., Yu, J., Kristiansen, K., Wong, G. K.-S., & Wang, J. (2005). Origin and evolution of new exons in rodents. *Genome Research*, *15*(9), 1258–1264. https://doi.org/10.1101/gr.3929705

Ward, L. D., & Kellis, M. (2012). Evidence of abundant purifying selection in humans for recently acquired regulatory functions. *Science (New York, N.Y.)*, *337*(6102), 1675–1678. https://doi.org/10.1126/science.1225057

Webster, M. T. (2004). Gene Expression, Synteny, and Local Similarity in Human Noncoding Mutation Rates. *Molecular Biology and Evolution*, *21*(10), 1820–1830. https://doi.org/10.1093/molbev/msh181

Webster, M. T., Smith, N. G. C., Hultin-Rosenberg, L., Arndt, P. F., & Ellegren, H. (2005). Male-driven biased gene conversion governs the evolution of base composition in human alu repeats. *Molecular Biology and Evolution*, *22*(6), 1468–1474. https://doi.org/10.1093/molbev/msi136

Webster, M. T., Smith, N. G. C., Lercher, M. J., & Ellegren, H. (2004). Gene expression, synteny, and local similarity in human noncoding mutation rates. *Molecular Biology and Evolution*, *21*(10), 1820–1830. https://doi.org/10.1093/molbev/msh181

Weedall, G. D., & Conway, D. J. (2010). Detecting signatures of balancing selection to identify targets of anti-parasite immunity. *Trends in Parasitology*, *26*(7), 363–369. https://doi.org/10.1016/j.pt.2010.04.002

Welch, J. J., Eyre-Walker, A., & Waxman, D. (2008). Divergence and Polymorphism Under the Nearly Neutral Theory of Molecular Evolution. *Journal of Molecular Evolution*, *67*(4), 418–426. https://doi.org/10.1007/s00239-008-9146-9

Werner, B., Case, J., Williams, M. J., Chkhaidze, K., Temko, D., Fernández-Mateos, J., Cresswell, G. D., Nichol, D., Cross, W., Spiteri, I., Huang, W., Tomlinson, I. P. M., Barnes, C. P., Graham, T. A., & Sottoriva, A. (2020). Measuring single cell divisions in human tissues from multi-region sequencing data. *Nature Communications*, *11*(1), 1035. https://doi.org/10.1038/s41467-020-14844-6

Wickham, H. (2016). *Ggplot2: Elegant graphics for Data Analysis.* Springer-Verlag New Work.

Williams, A. L., Genovese, G., Dyer, T., Altemose, N., Truax, K., Jun, G., Patterson, N., Myers, S. R., Curran, J. E., Duggirala, R., Blangero, J., Reich, D., Przeworski, M., & on behalf of the T2D-GENES Consortium. (2015). Non-crossover gene conversions show strong GC bias and unexpected clustering in humans. *ELife*, *4*, e04637. https://doi.org/10.7554/eLife.04637

Williamson, S. H., Hubisz, M. J., Clark, A. G., Payseur, B. A., Bustamante, C. D., & Nielsen, R. (2007). Localizing Recent Adaptive Evolution in the Human Genome. *PLoS Genetics*, *3*(6), e90. https://doi.org/10.1371/journal.pgen.0030090

Wolf, Y. I., Novichkov, P. S., Karev, G. P., Koonin, E. V., & Lipman, D. J. (2009). The universal distribution of evolutionary rates of genes and distinct characteristics of eukaryotic genes of different apparent ages. *Proceedings of the National Academy of Sciences of the United States of America*, *106*(18), 7273–7280. https://doi.org/10.1073/pnas.0901808106

Wong, W. S. W., Solomon, B. D., Bodian, D. L., Kothiyal, P., Eley, G., Huddleston, K. C., Baker, R., Thach, D. C., Iyer, R. K., Vockley, J. G., & Niederhuber, J. E. (2016). New observations on maternal age effect on germline de novo mutations. *Nature Communications*, *7*(1), 10486. https://doi.org/10.1038/ncomms10486

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

Wright, S. (1932). The roles of mutation, inbreeding, crossbreeding and selection in evolution. *Proceedings of the Sixth International Congress on Genetics* **1**, 356–366.

Wright, S. I. (2016). Charlesworth et al. On Background Selection and Neutral Diversity. *Genetics*, *204*(3), 829–832. https://doi.org/10.1534/genetics.116.196170

Wright, S. I., & Charlesworth, B. (2004). The HKA Test Revisited. *Genetics*, *168*(2), 1071–1076. https://doi.org/10.1534/genetics.104.026500

Wright, S. I., Yau, C. B. K., Looseley, M., & Meyers, B. C. (2004). Effects of gene expression on molecular evolution in Arabidopsis thaliana and Arabidopsis lyrata. *Molecular Biology and Evolution*, *21*(9), 1719–1726. https://doi.org/10.1093/molbev/msh191

Xie, V. C., Pu, J., Metzger, B. P., Thornton, J. W., & Dickinson, B. C. (2021). Contingency and chance erase necessity in the experimental evolution of ancestral proteins. *ELife*, *10*, e67336. https://doi.org/10.7554/eLife.67336

Yang, C., Yu, T., Han, C., Qin, W., Liao, X., Yu, L., Liu, X., Zhu, G., Su, H., Lu, S., Chen, Z., Liu, Z., Huang, K., Liu, Z., Liang, Y., Huang, J., Mo, Z., Qin, X., Li, L., … Peng, T. (2017). Genome-Wide Association Study of MKI67 Expression and its Clinical Implications in HBV-Related Hepatocellular Carcinoma in Southern China. *Cellular Physiology and Biochemistry*, *42*(4), 1342–1357. https://doi.org/10.1159/000478963

Yang, Z. (2007). PAML 4: Phylogenetic Analysis by Maximum Likelihood. *Molecular Biology and Evolution*, *24*(8), 1586–1591. https://doi.org/10.1093/molbev/msm088

Yates, A. D., Achuthan, P., Akanni, W., Allen, J., Allen, J., Alvarez-Jarreta, J., Amode, M. R., Armean, I. M., Azov, A. G., Bennett, R., Bhai, J., Billis, K., Boddu, S., Marugán, J. C., Cummins, C., Davidson, C., Dodiya, K., Fatima, R., Gall, A., … Flicek, P. (2019). Ensembl 2020. *Nucleic Acids Research*, gkz966. https://doi.org/10.1093/nar/gkz966

Zhang, J. (2000a). Rates of Conservative and Radical Nonsynonymous Nucleotide Substitutions in Mammalian Nuclear Genes. *Journal of Molecular Evolution*, *50*(1), 56–68. https://doi.org/10.1007/s002399910007

Zhang, J. (2000b). Protein-length distributions for the three domains of life. *Trends in Genetics: TIG*, *16*(3), 107–109. https://doi.org/10.1016/s0168-9525(99)01922-8

Zhang, L., & Li, W.-H. (2005). Human SNPs reveal no evidence of frequent positive selection. *Molecular Biology and Evolution*, *22*(12), 2504–2507. https://doi.org/10.1093/molbev/msi240

Zhang, Y. E., Vibranovski, M. D., Krinsky, B. H., & Long, M. (2010). Age-dependent chromosomal distribution of male-biased genes in Drosophila. *Genome Research*, *20*(11), 1526–1533. https://doi.org/10.1101/gr.107334.110

Zhang, Z. (2003). Genomic Background Drives the Divergence of Duplicated Amylase Genes at Synonymous Sites in Drosophila. *Molecular Biology and Evolution*, *21*(2), 222–227. https://doi.org/10.1093/molbev/msg243

Zhao, L., & Charlesworth, B. (2016). Resolving the Conflict Between Associative Overdominance and Background Selection. *Genetics*, *203*(3), 1315–1334. https://doi.org/10.1534/genetics.116.188912

Zhen, Y., Huber, C. D., Davies, R. W., & Lohmueller, K. E. (2021). Greater strength of selection and higher proportion of beneficial amino acid changing mutations in humans compared with mice and Drosophila melanogaster. *Genome Research*, *31*(1), 110–120. https://doi.org/10.1101/gr.256636.119

Zirkle, C. (1941). Natural Selection before the 'Origin of Species.' *Proceedings of the American Philosophical Society*, *84*(1), 71–123.

Appendices

Appendix A: Chapter 2 supplementary material

Target population	Comparative population	all polymorphism data			filtered for BGC		
		Z	Z_{low}	Z_{high}	Z	Z_{low}	Z_{high}
African	non-African	1.13	1.10	1.17	1.12	1.03	1.21
African	East Asian	1.05	1.03	1.07	1.00	0.95	1.05
African	European	1.04	1.01	1.07	1.01	0.93	1.10
African	South Asian	1.04	1.01	1.06	1.04	0.98	1.09
South Asian	East Asian	1.10	1.00	1.23	1.36	1.04	1.73

Table A1: Testing the effects of BGC for population comparisons which show Z>1. Confidence intervals were generated by bootstrapping the data by gene 100 times.

target	comparative	Z	Z_{low}	Z_{high}	α	α_{b_low}	α_{b_low}	*b*	b_{low}	b_{high}
AFR	nonAFR	3.354	0.989	4.533	0.702	-0.011	0.779	299	Na	332
AFR	EAS	1.403	0.995	2.007	0.287	-0.005	0.502	134	Na	234
AFR	EUR	3.714	1.465	5.568	0.731	0.317	0.820	338	147	379
AFR	SAS	2.123	1.443	3.272	0.529	0.307	0.694	247	143	324

Table A2: Estimates of Z for HLA genes only. Confidence intervals were generated by bootstrapping the data by gene 100 times.

target	comparative	Z	Z_{low}	Z_{high}	α	α_{b_low}	α_{b_low}	*b*	b_{low}	b_{high}
AFR	nonAFR	1.112	1.083	1.146	0.101	0.077	0.127	1193	907	1502
AFR	EAS	1.034	1.017	1.052	0.033	0.016	0.049	453	227	678
AFR	EUR	1.025	0.999	1.052	0.024	-0.001	0.049	355	Na	716
AFR	SAS	1.021	1.001	1.043	0.021	0.001	0.041	293	19	587

Table A3: Estimates of Z for non-HLA genes only. Confidence intervals were generated by bootstrapping the data by gene 100 times.

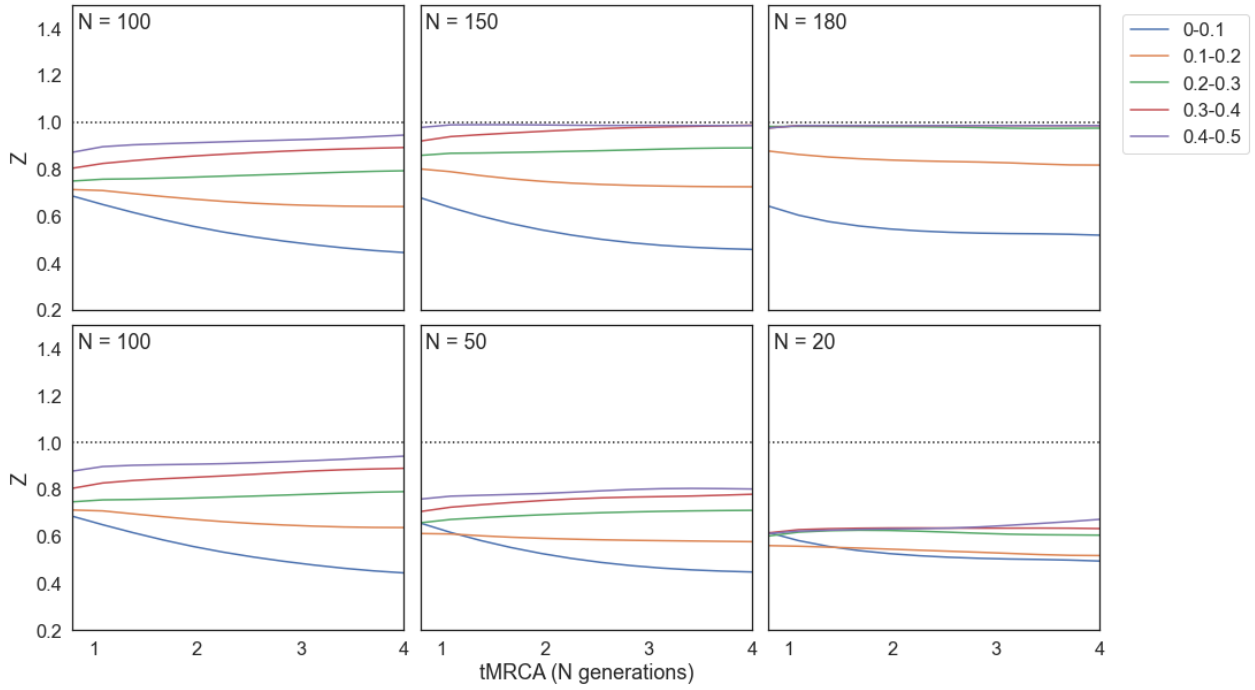


Figure A1: Vicariance simulations in which the ancestral population splits to form two daughter populations of the size specified in the panel. Each column is a separate set of simulations, with the top row plotting Z against tMRCA (measured in N generations, where N is the population size) for the larger daughter population, and the bottom row the smaller. Deleterious mutations are drawn from a gamma DFE with parameters inferred from human population data.

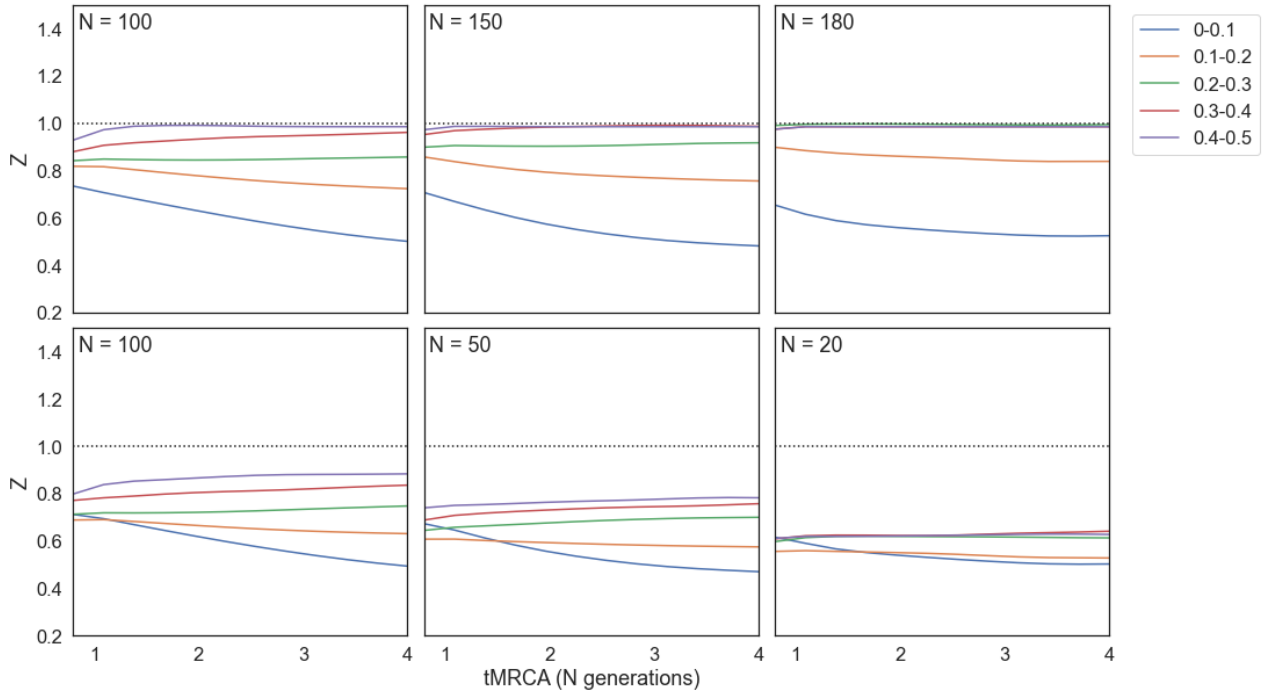


Figure A2: Dispersal simulations in which a single daughter population disperses from the ancestral population. Each column is a separate set of simulations, with the top row plotting Z against tMRCA (measured in N generations, where N is the population size) for the ancestral population, and the bottom row the daughter population. Deleterious mutations are drawn from a gamma DFE with parameters inferred from human population data.

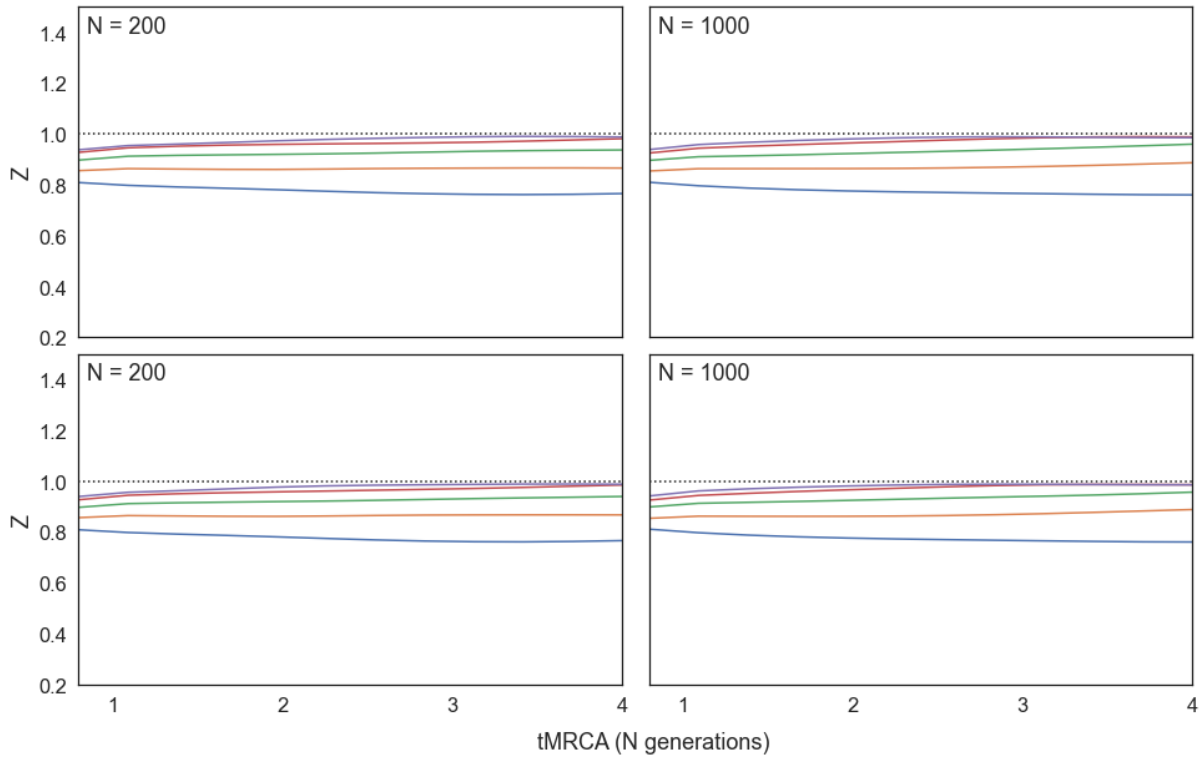


Figure A3: Vicariance expansion simulations in which both daughter populations expand. The ancestral population (of size N=200) splits to form two daughter populations of size N=100. Both daughter populations go on to expand in size. In the left column the daughter populations double in size. In the right panel they reach 10x their initial size. Deleterious mutations are drawn from a gamma DFE with parameters inferred from human population data.

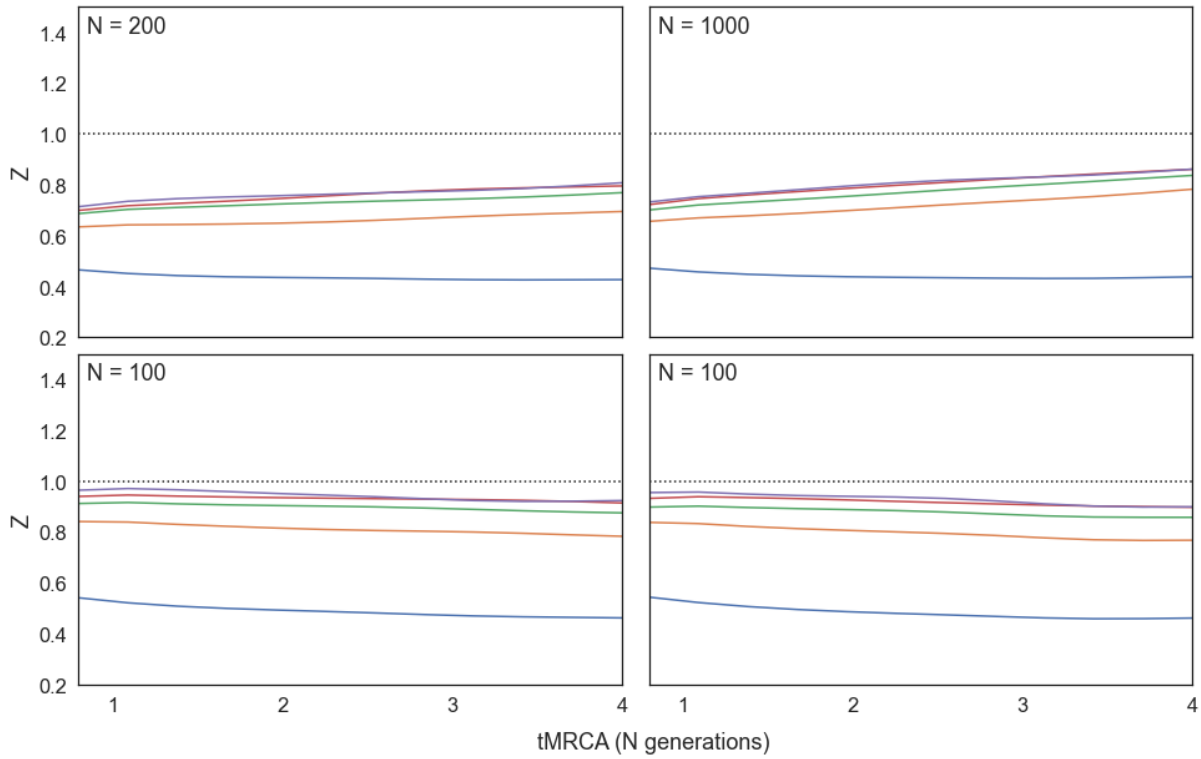


Figure A4: Vicariance expansion simulations in which only one daughter population expands. The ancestral population (of size N=200) splits to form two daughter populations of size N=100. One daughter population (upper panels) goes on to expand in size. In the left column the daughter populations double in size. In the right panel they reach 10x their initial size. Deleterious mutations are drawn from a gamma DFE with parameters inferred from human population data.

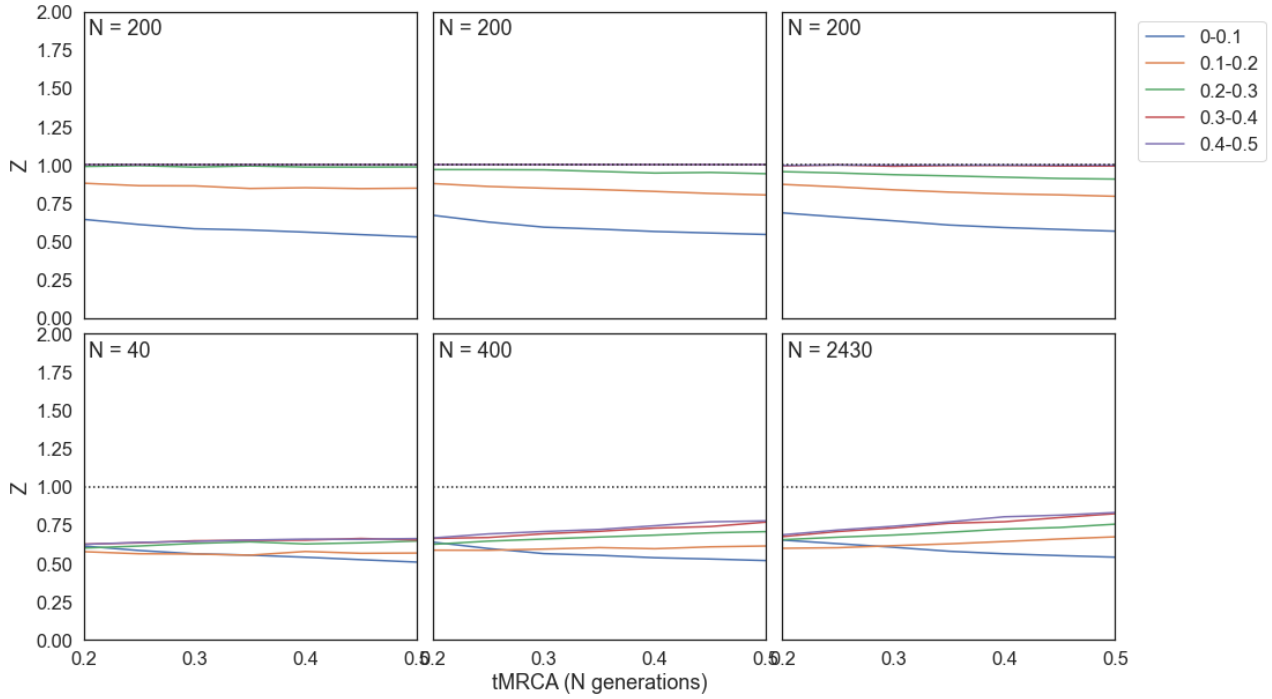


Figure A5: Dispersal expansion simulations in which a single daughter population disperses from the ancestral population and then expands. The ancestral population (of size N=200) splits to form a daughter population of size N=100, which expands to the final population size shown in the panel. Each column is a separate set of simulations, with the top row plotting Z against tMRCA (measured in N generations, where N is the population size) for the ancestral population, and the bottom row the daughter population. Deleterious mutations are drawn from a gamma DFE with parameters inferred from human population data.

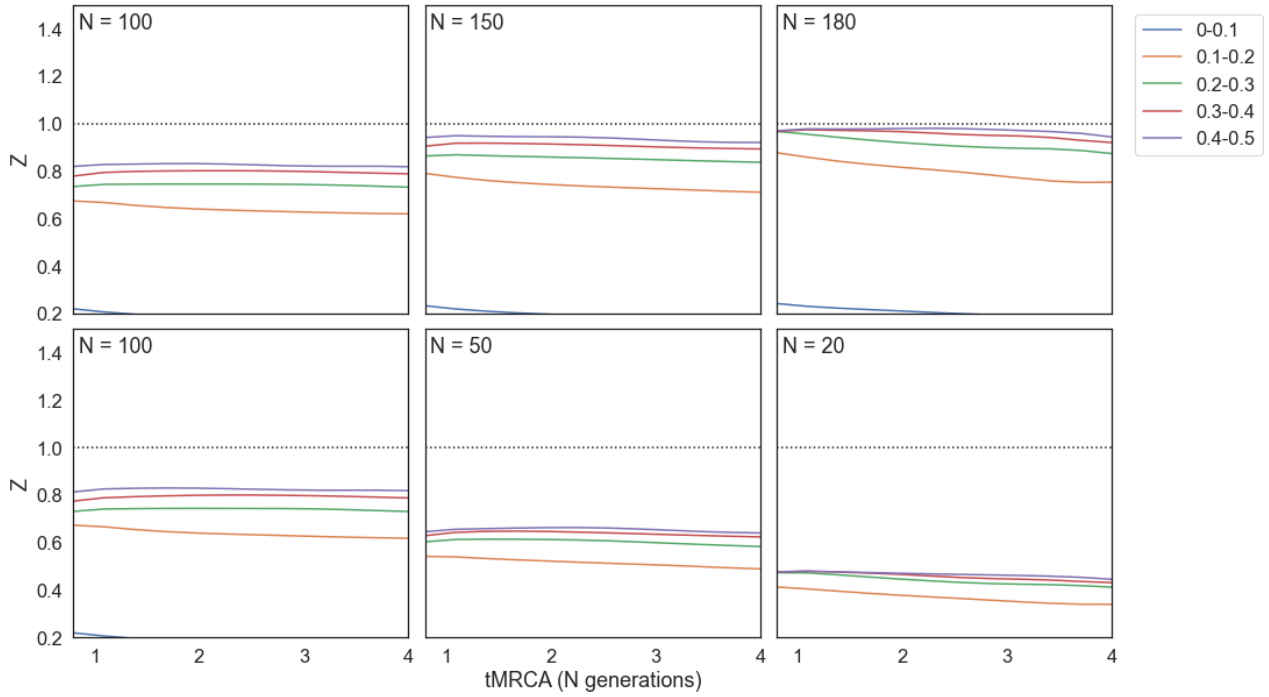


Figure A6: Vicariance simulations in which the ancestral population splits to form two daughter populations of the size specified in the panel. Each column is a separate set of simulations, with the top row plotting Z against tMRCA (measured in N generations, where N is the population size) for the larger daughter population, and the bottom row the smaller. Deleterious mutations are drawn from a gamma DFE with parameters inferred from *Drosophila melanogaster* population data.

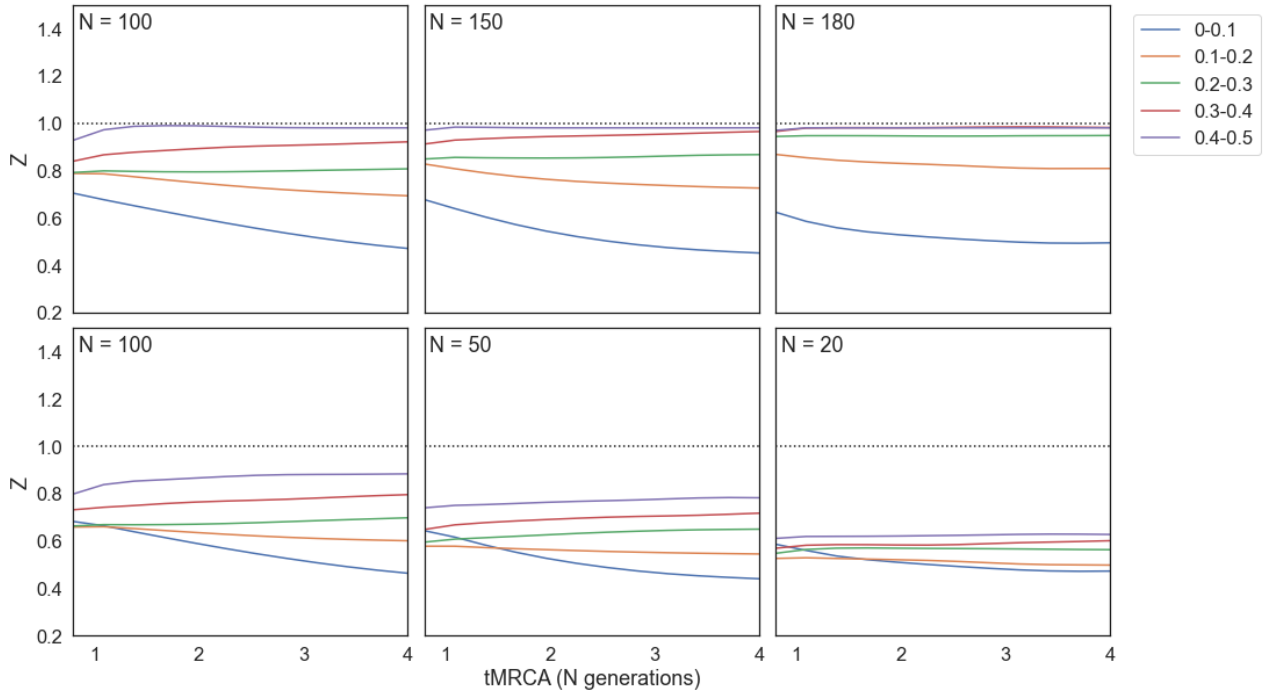


Figure A7: Dispersal simulations in which a single daughter population disperses from the ancestral population. Each column is a separate set of simulations, with the top row plotting Z against tMRCA (measured in N generations, where N is the population size) for the ancestral population, and the bottom row the daughter population. Deleterious mutations are drawn from a gamma DFE with parameters inferred from *Drosophila melanogaster* population data.

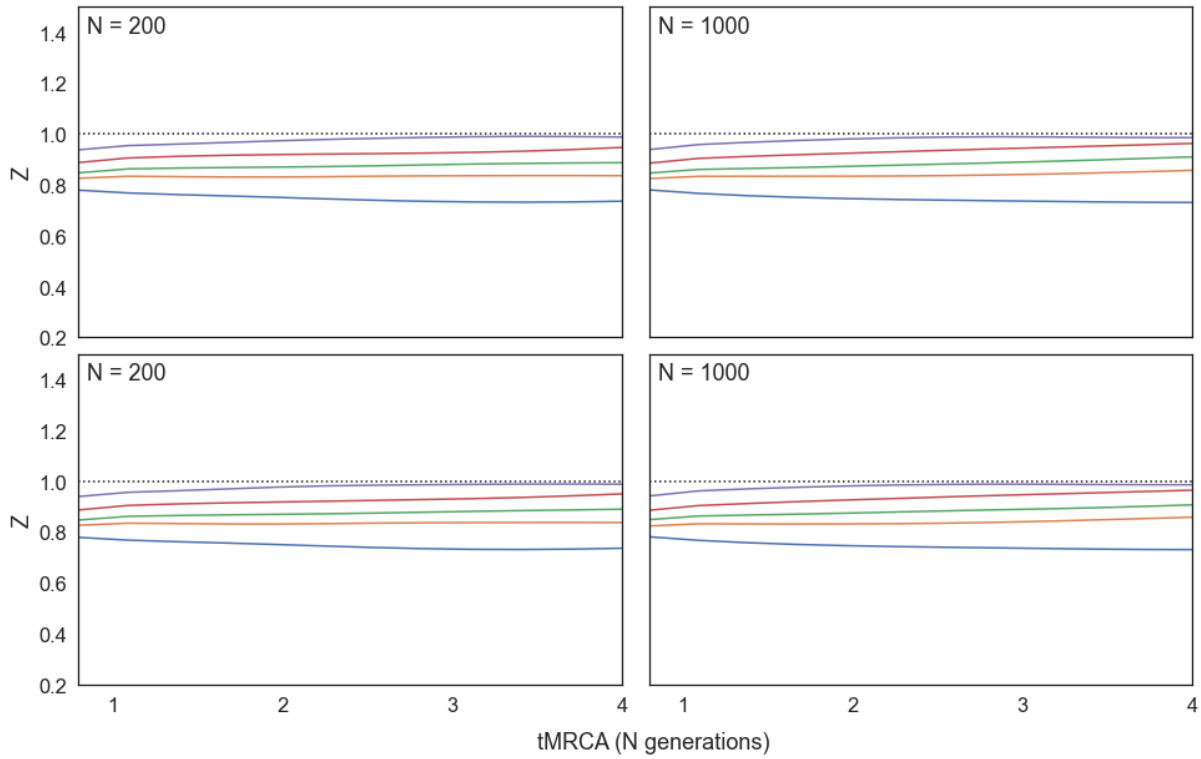


Figure A8: Vicariance expansion simulations in which both daughter populations expand. The ancestral population (of size N=200) splits to form two daughter populations of size N=100. Both daughter populations go on to expand in size. In the left column the daughter populations double in size. In the right panel they reach 10x their initial size. Deleterious mutations are drawn from a gamma DFE with parameters inferred from *Drosophila melanogaster* population data.

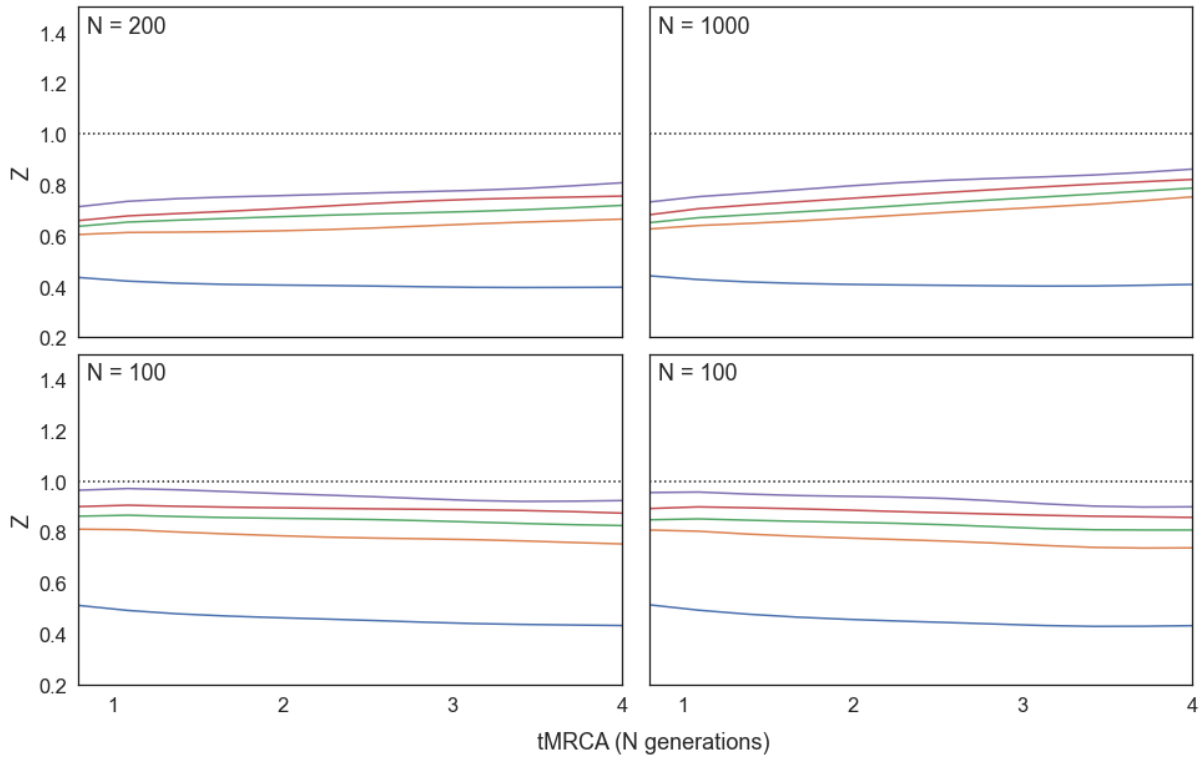


Figure A9: Vicariance expansion simulations in which only one daughter population expands. The ancestral population (of size N=200) splits to form two daughter populations of size N=100. One daughter population (upper panels) goes on to expand in size. In the left column the daughter populations double in size. In the right panel they reach 10x their initial size. Deleterious mutations are drawn from a gamma DFE with parameters inferred from *Drosophila melanogaster* population data.

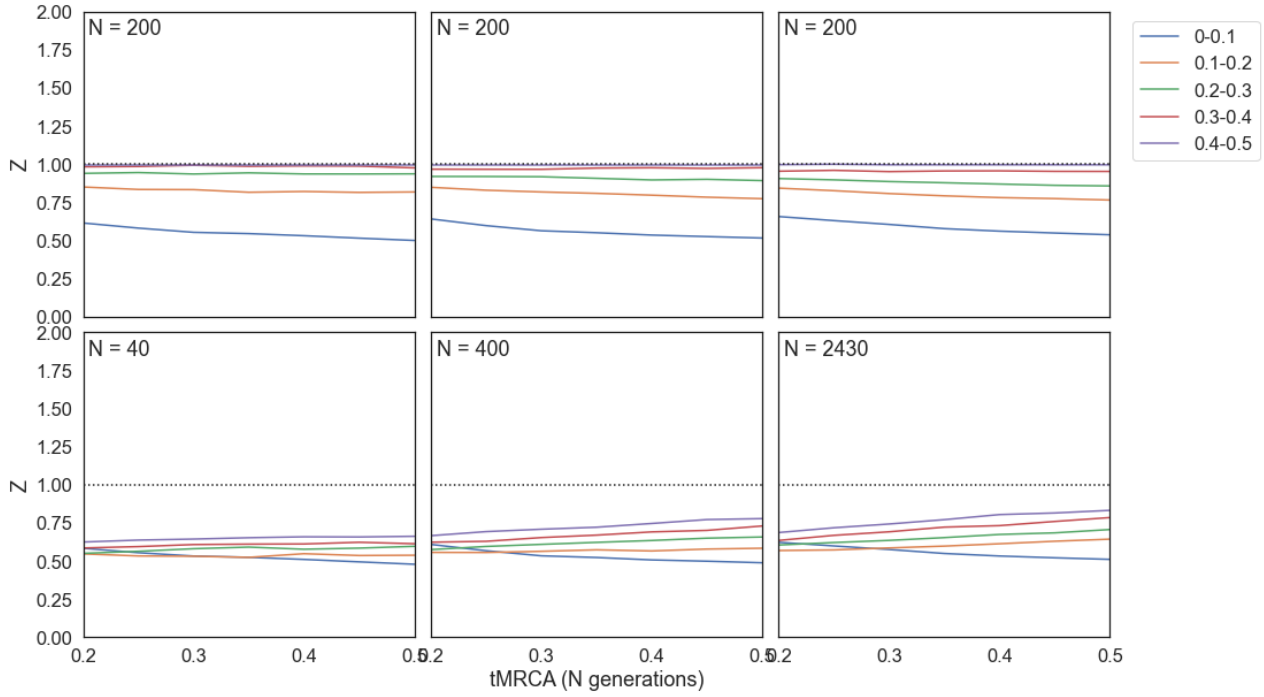


Figure A10: Dispersal expansion simulations in which a single daughter population disperses from the ancestral population and then expands. The ancestral population (of size N=200) splits to form a daughter population of size N=100, which expands to the final population size shown in the panel. Each column is a separate set of simulations, with the top row plotting Z against tMRCA (measured in N generations, where N is the population size) for the ancestral population, and the bottom row the daughter population. Deleterious mutations are drawn from a gamma DFE with parameters inferred from *Drosophila melanogaster* population data.

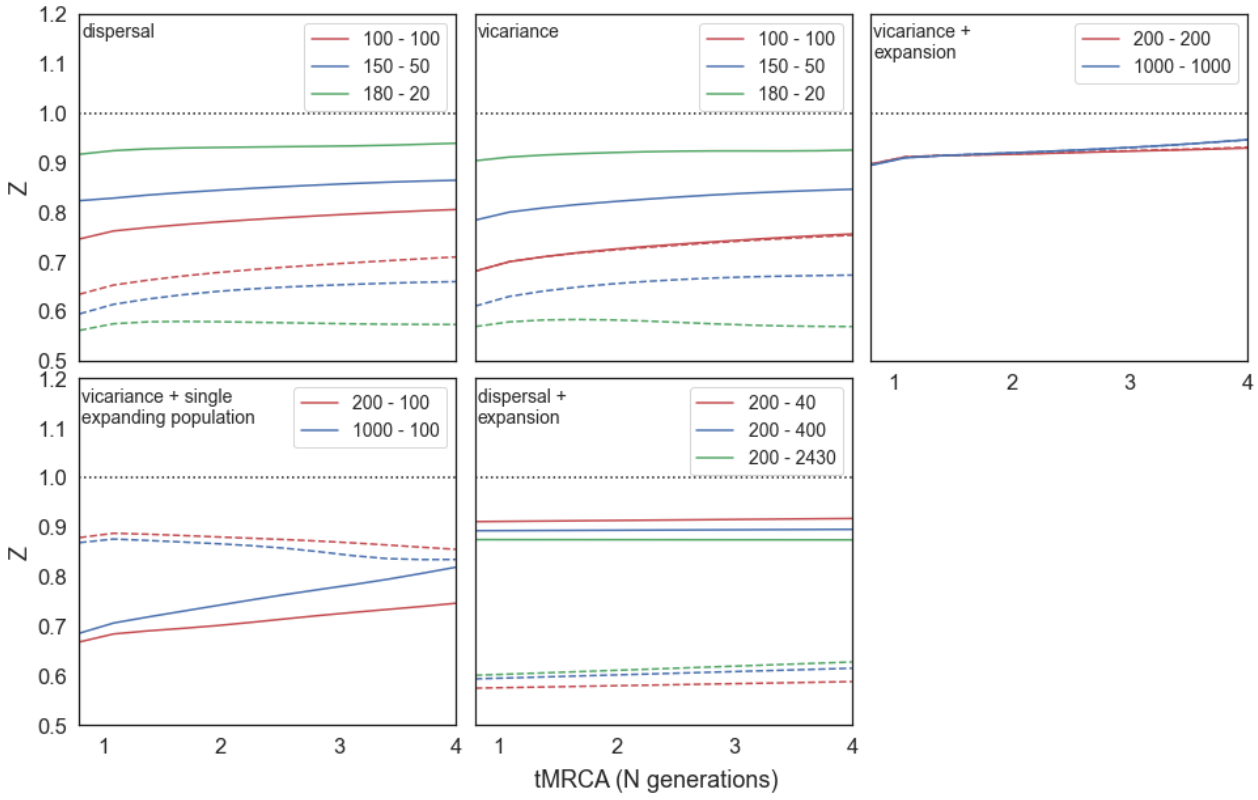


Figure A11: Simulations with for combined 0.1-0.5 minor allele frequencies. Each panel is a separate simulated scenario, with population sizes listed in the panel legend. The first number is for the filled in data lines, denoting the ancestral population in dispersal scenarios, and for the larger population in the vicariance scenarios. The second number is for the dotted data lines, denoting the daughter population in dispersal scenarios, and the smaller population in the vicariance scenarios. For more details on each scenario please see figures A1-10. Deleterious mutations are drawn from a gamma DFE with parameters inferred from human population data.

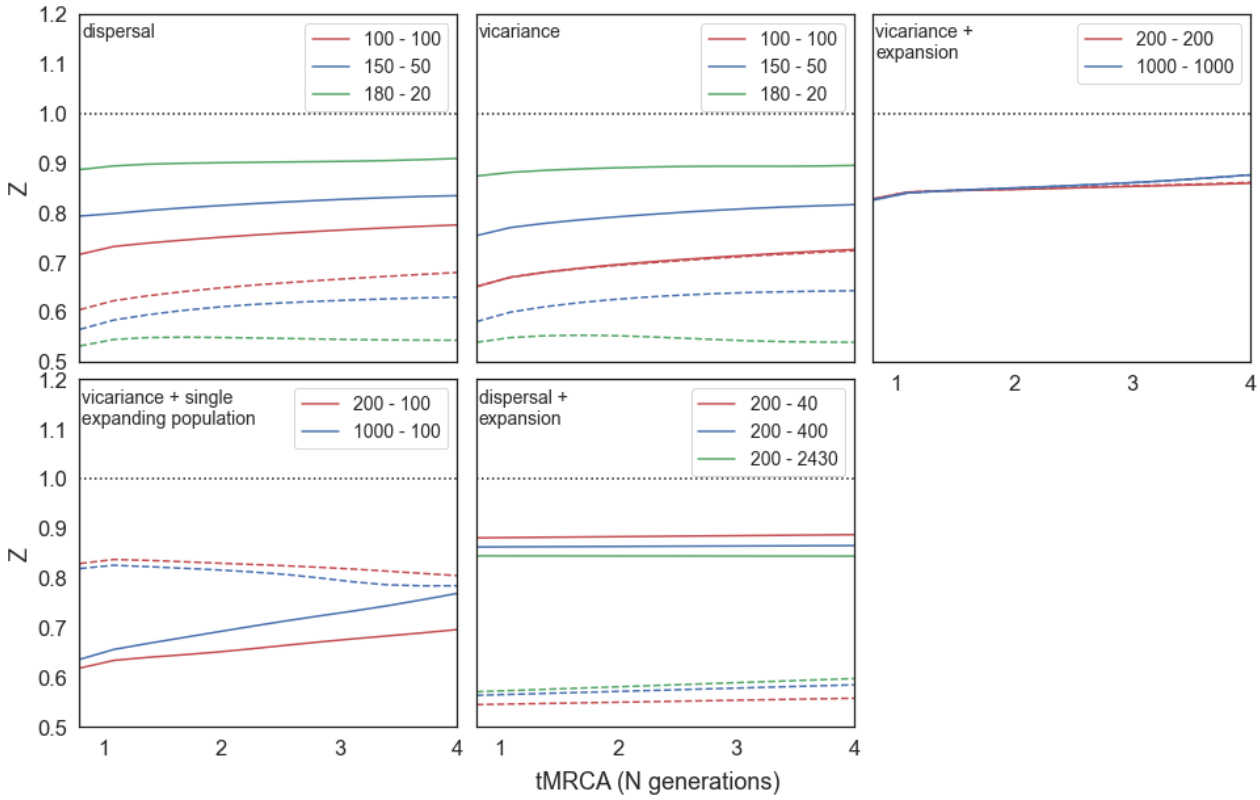


Figure A12: Simulations with for combined 0.1-0.5 minor allele frequencies. Each panel is a separate simulated scenario, with population sizes listed in the panel legend. The first number is for the filled in data lines, denoting the ancestral population in dispersal scenarios, and for the larger population in the vicariance scenarios. The second number is for the dotted data lines, denoting the daughter population in dispersal scenarios, and the smaller population in the vicariance scenarios. For more details on each scenario please see supplementary figures S1-10. Deleterious mutations are drawn from a gamma DFE with parameters inferred from *Drosophila melanogaster* population data.

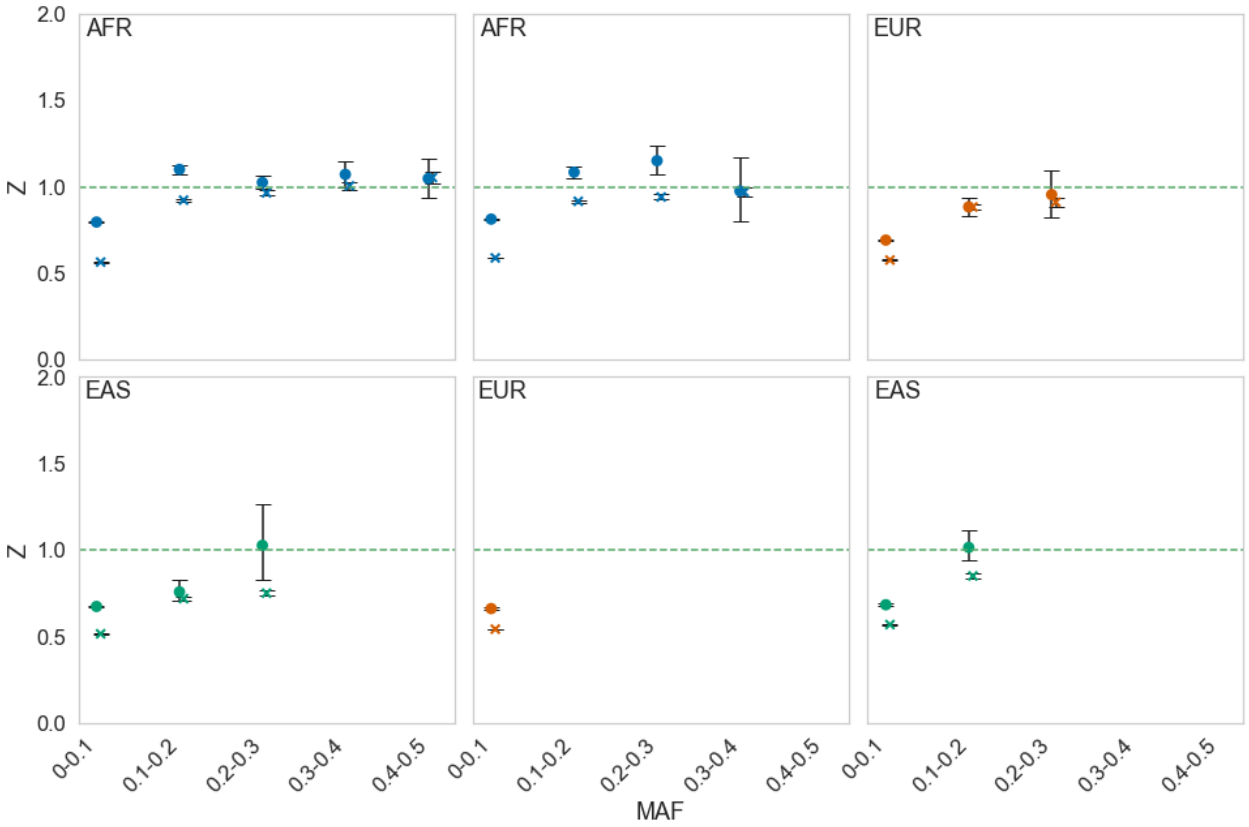


Figure A13: Simulations using the Gravel model of human demography (Gravel et al, 2011). Shown are the observed (filled circles) and simulated (crosses) values of Z. Each column represents a different population comparison. From left to right: Africans (AFR) and East Asians (EAS), Africans and Europeans (EUR), Europeans and East Asians. The population name in the upper left indicates which set of private polymorphisms are used to calculate Z in each population comparison. The x-axis represents private polymorphism minor allele frequency bins. Confidence intervals generated by bootstrapping.

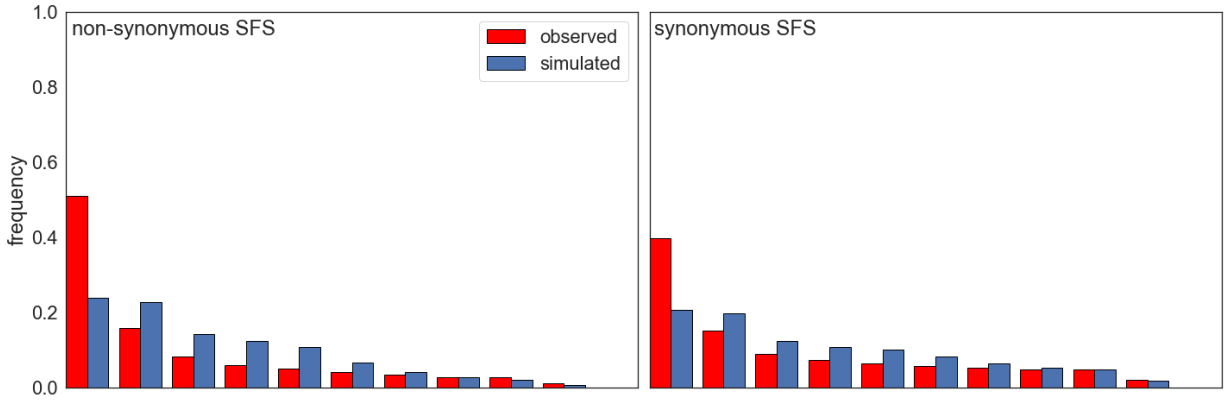


Figure A14: Comparison of simulated (under the Gravel et al. (2011) model of human demography) and observed SFS from the African population.

Appendix B: Chapter 3 supplementary material

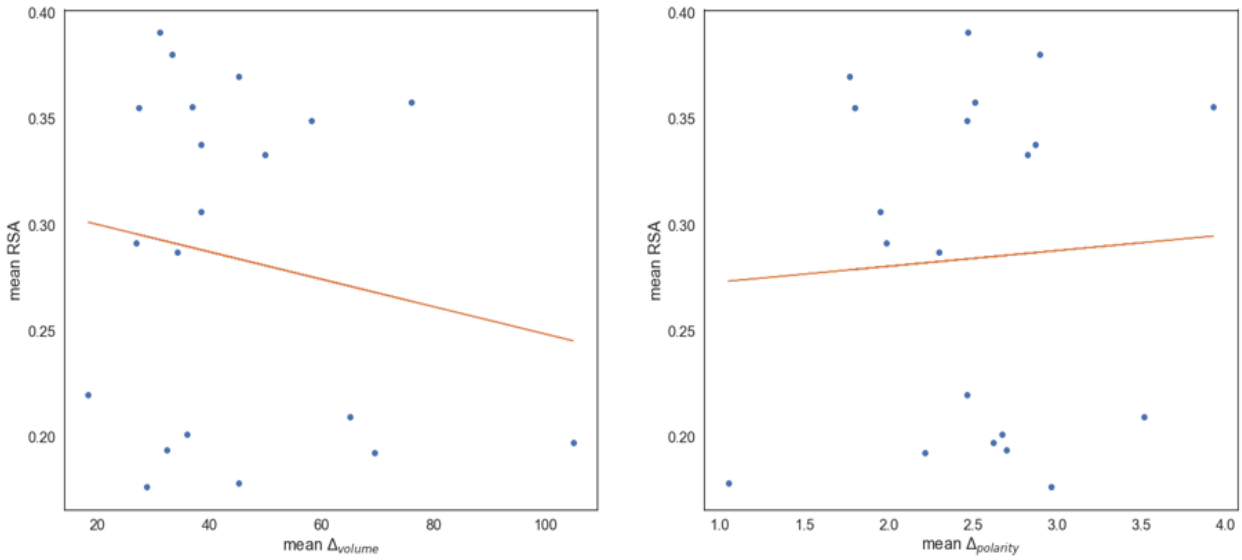


Figure B1: Average RSA of an amino acid and the average difference in volume or polarity to its one mutation step neighbours.

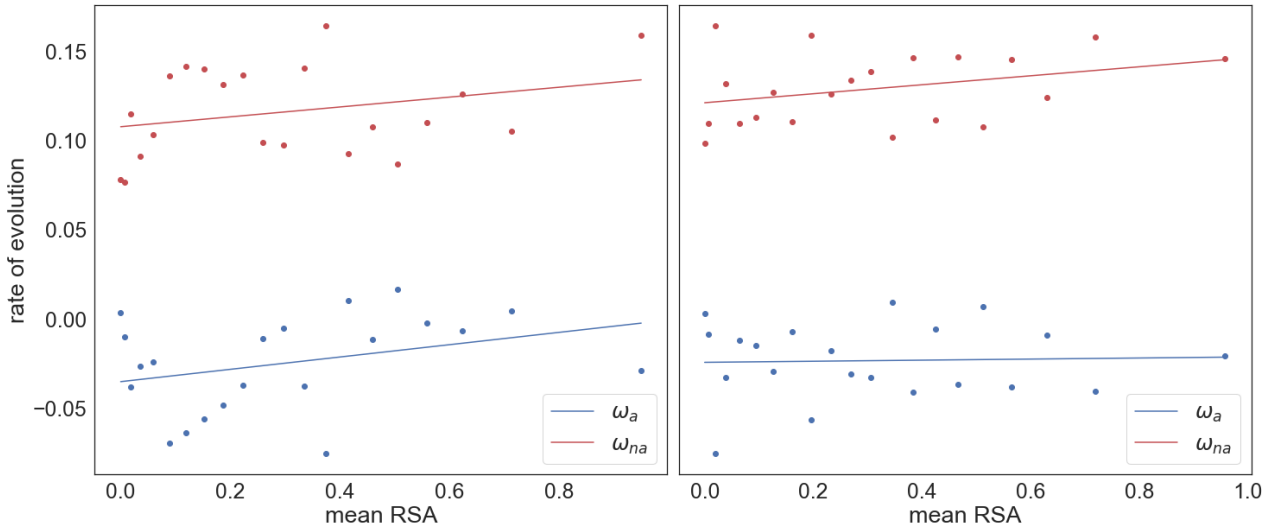


Figure B2: Estimates of ω_a and ω_{na} plotted against mean relative solvent accessibility, controlling for volume difference (left) and polarity difference (right). Data binned into 20 RSA bins of roughly equal size. For each analysis, a weighted linear regression is fitted to the data. The respective significance of each correlation is shown in the plot legend, (*P < 0.05; **P < 0.01; ***P < 0.001; "." 0.05 ≤ P < 0.10) for ω_a and ω_{na}). Regression is weighted by the reciprocal of the variance for each estimate of ω_a and ω_{na}, which were estimated by bootstrapping the data by gene 100 times for each data point.

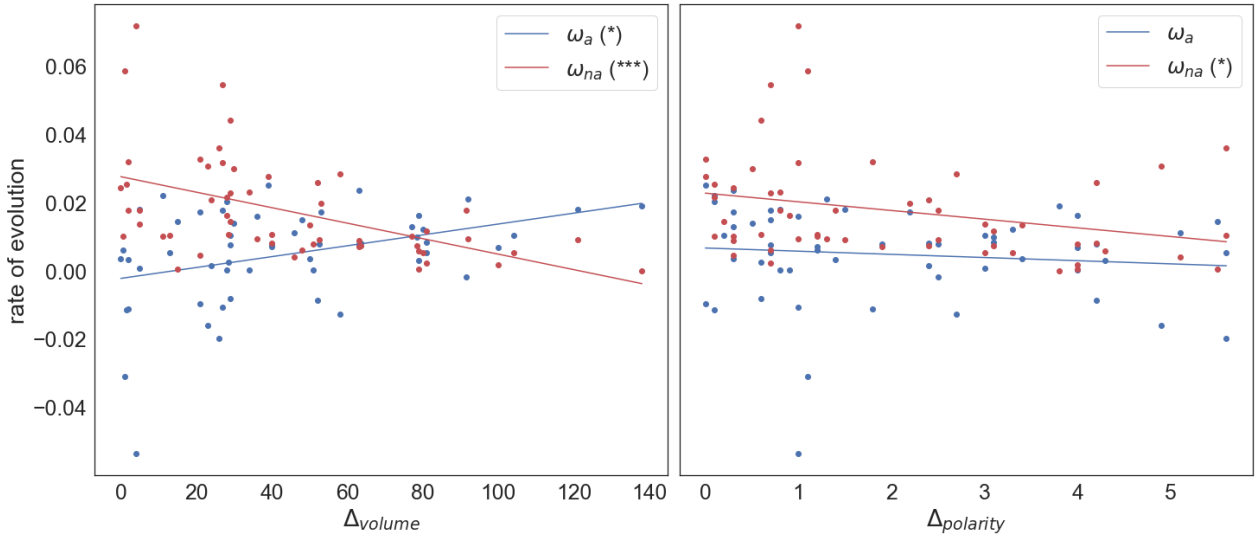


Figure B3: The adaptive and non-adaptive substitution rate plotted against the difference in a) volume, b) polarity, controlling for relative solvent accessibility. A weighted linear regression is fitted to the data, weighted by the variance of each estimate. The respective significance of each correlation is shown in the legend, (*P < 0.05; **P < 0.01; ***P < 0.001; "." 0.05 ≤ P < 0.10).

Appendix C: Chapter 4 supplementary material

	gene expression	gene length	recombination rate
gene age	0.868 (***)	0.860 (***)	-0.621 (**)
gene expression		0.437 (***)	-0.035 (***)
gene length			0.101 (***)

Table S1: Linear regression correlations between gene age, gene expression, gene length and recombination rate. (*P < 0.05; **P < 0.01; ***P < 0.001; "." 0.05 ≤ P < 0.10).

Term	Number	MS	Both fixed Variance	Both random Variance
Residual	2	5.2E-05	0.000052	0.000052
VIP	2	0.02919	0.00112054	0.001120538
GO	13	0.00097	0.000229	0.000229

Table C2: Estimated variance components from two-way analysis of variance on ω_a for GO categories with 200,000 sites or more.

Term	Number	MS	Both fixed Variance	Both random Variance
Residual	2	0.000017	0.000017	0.000017
VIP	2	0.000983	3.7154E-05	3.71538E-05
GO	13	0.000034	4.25E-6	4.25E-06

Table C3: Estimated variance components from two-way analysis of variance on ω_{na} for GO categories with 200,000 sites or more.

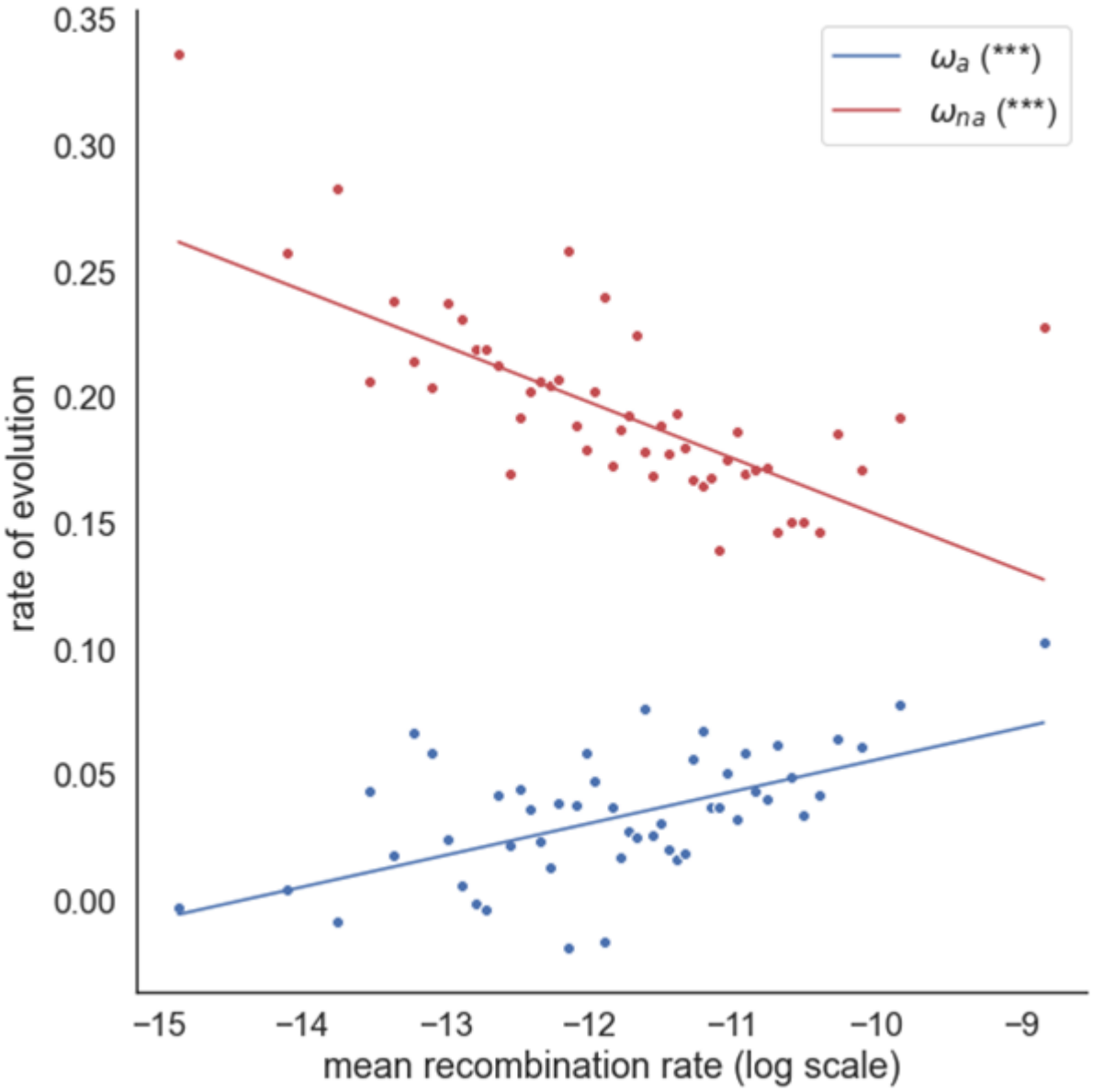


Figure C1: Estimates of ω_a and ω_{na} plotted against the log of the mean recombination rate for genes binned into 50 recombination bins of equal size. An uweighted linear regression is fitted to the data. The respective significance of each correlation is shown in the plot legend, (*P < 0.05; **P < 0.01; ***P < 0.001; "." 0.05 ≤ P < 0.10) for ω_a and ω_{na}).

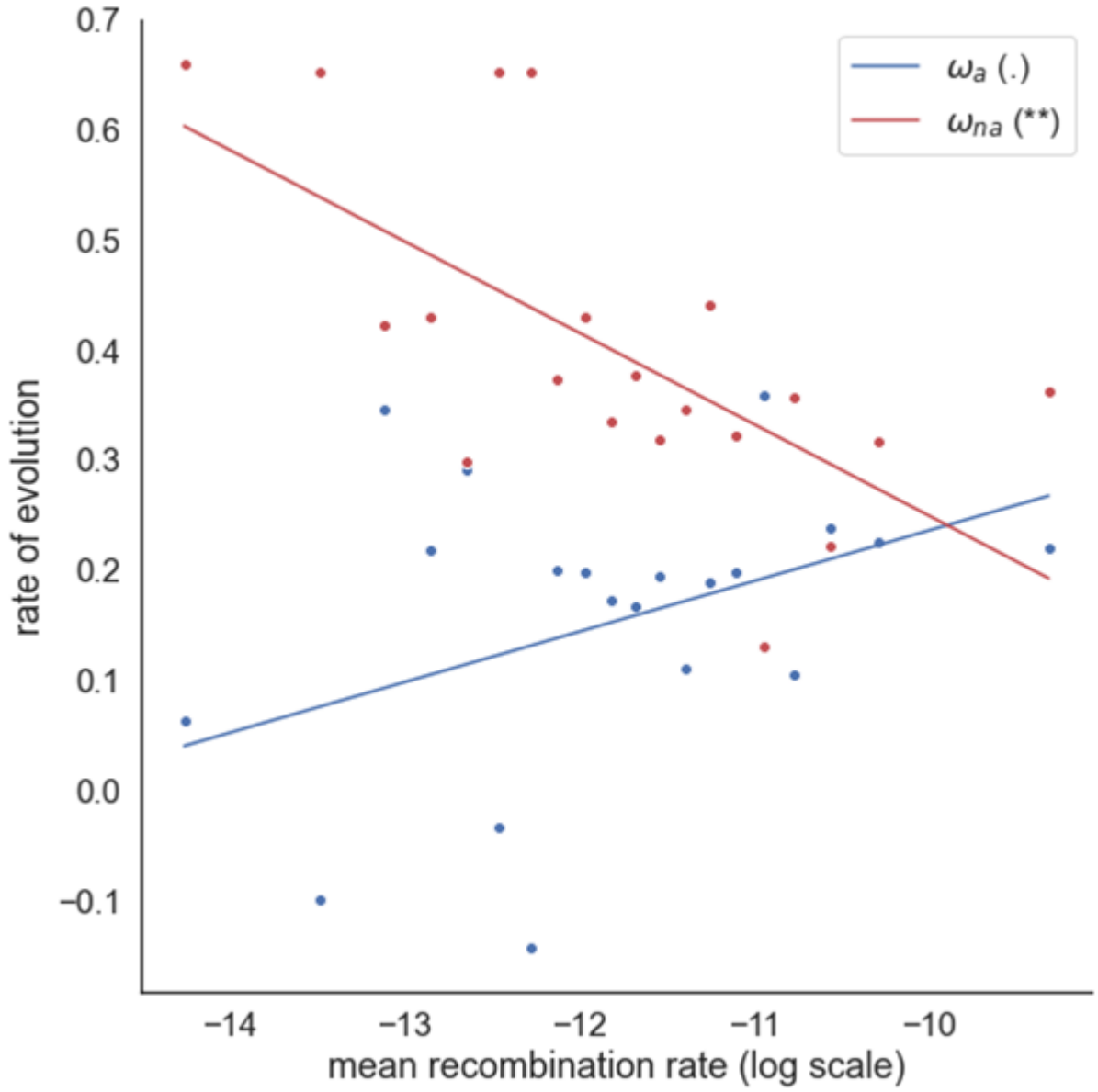


Figure C2: Estimates of ω_a and ω_{na} plotted against the log of the mean recombination rate, controlling for biased gene conversion, for genes binned into 20 recombination bins of equal size. An unweighted linear regression is fitted to the data. The respective significance of each correlation is shown in the plot legend, (*P < 0.05; **P < 0.01; ***P < 0.001; "." 0.05 ≤ P < 0.10) for ω_a and ω_{na}).

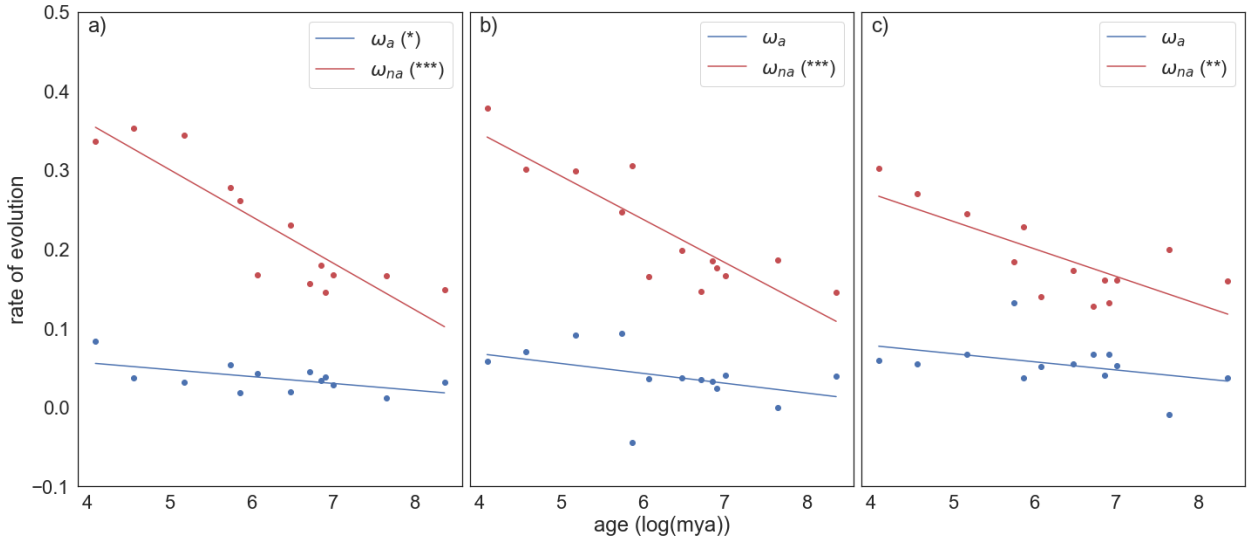


Figure C3: Estimates of ω_a and ω_{na} plotted against log gene age for genes binned into phylostratigraphic age categories, controlling for a) recombination rate; b) gene length; c) gene expression. An unweighted linear regression is fitted to the data. The respective significance of each correlation is shown in the plot legend, (*P < 0.05; **P < 0.01; ***P < 0.001; "." 0.05 ≤ P < 0.10) for ω_a and ω_{na}).

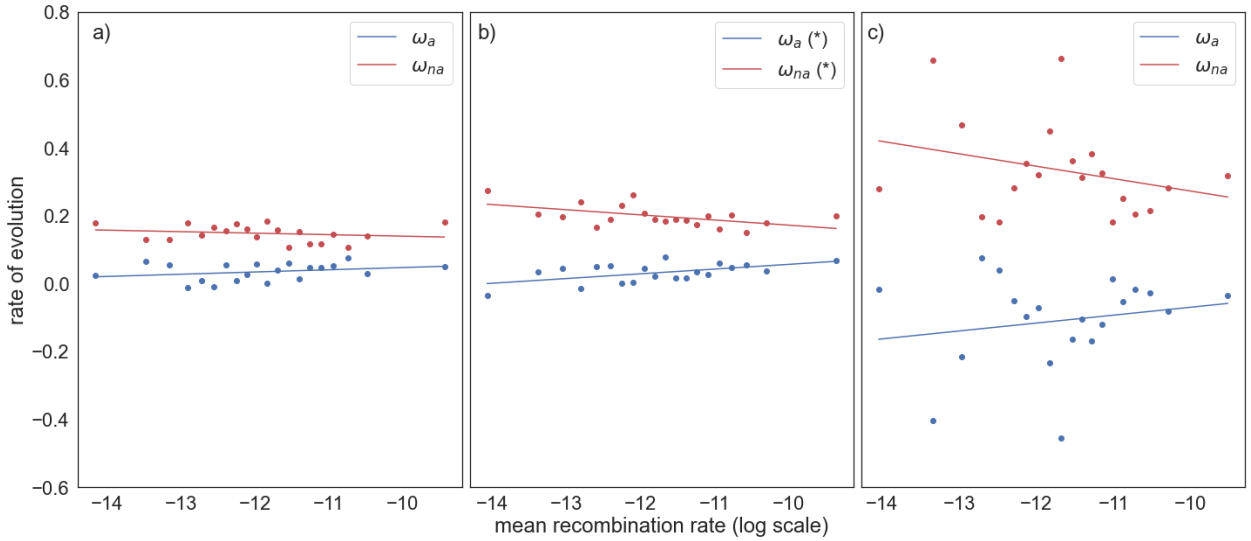


Figure C4: Estimates of ω_a and ω_{na} plotted against the log of the mean recombination rate for genes binned into 20 recombination bins of equal size, controlling for a) gene age; b) gene length; c) gene expression. An unweighted linear regression is fitted to the data. The respective significance of each correlation is shown in the plot legend, (*P < 0.05; **P < 0.01; ***P < 0.001; "." 0.05 ≤ P < 0.10) for ω_a and ω_{na}).

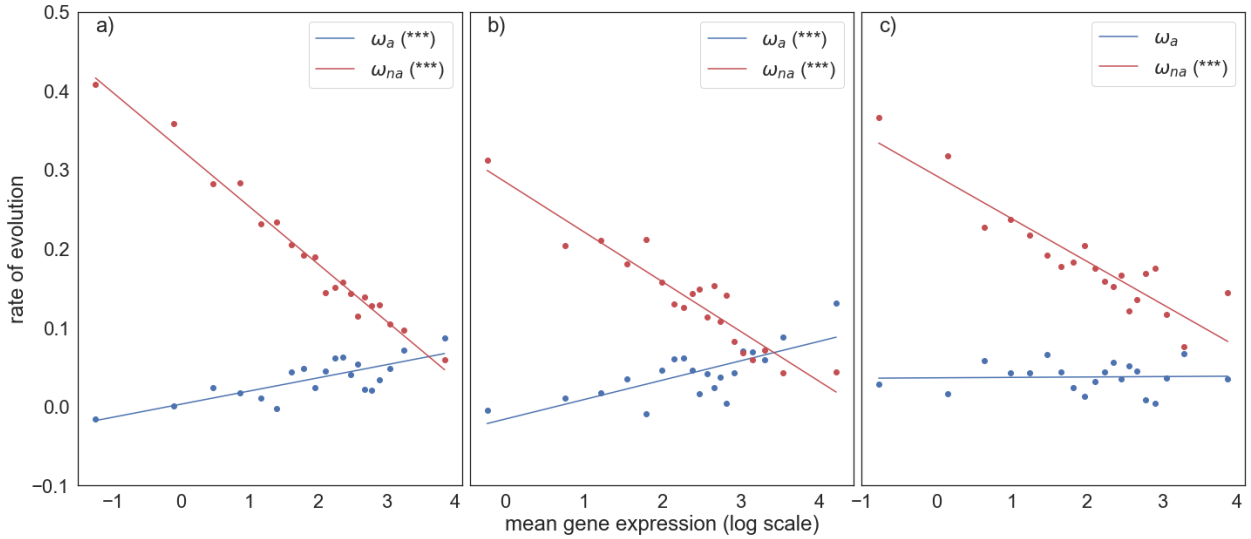


Figure C5: Estimates of ω_a and ω_{na} plotted against the log of the mean gene expression for genes binned into 20 mean expression bins of equal size, controlling for a) recombination rate; b) gene age; c) gene length. An unweighted linear regression is fitted to the data. The respective significance of each correlation is shown in the plot legend, (*P < 0.05; **P < 0.01; ***P < 0.001; "." 0.05 ≤ P < 0.10) for ω_a and ω_{na}).

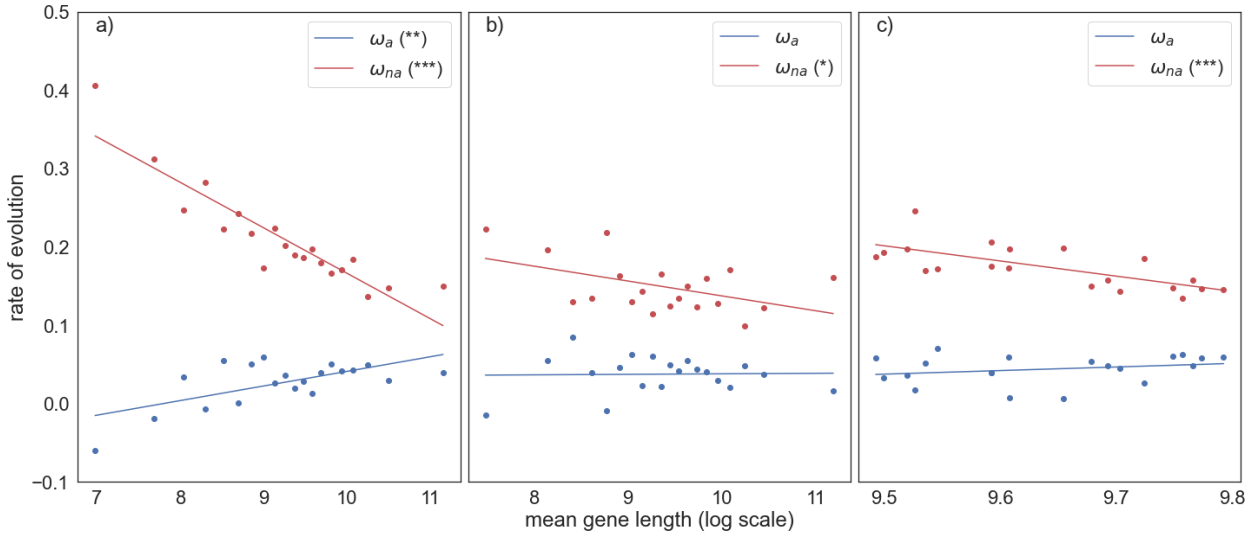


Figure C6: Estimates of ω_a and ω_{na} plotted against the log of the mean gene length for genes binned into 20 mean length bins of equal size, controlling for a) recombination rate; b) gene age; c) gene expression. An unweighted linear regression is fitted to the data. The respective significance of each correlation is shown in the plot legend, (*P < 0.05; **P < 0.01; ***P < 0.001; "." 0.05 ≤ P < 0.10 for ω_a and ω_{na}).

Appendix D: Chapter 5 supplementary material

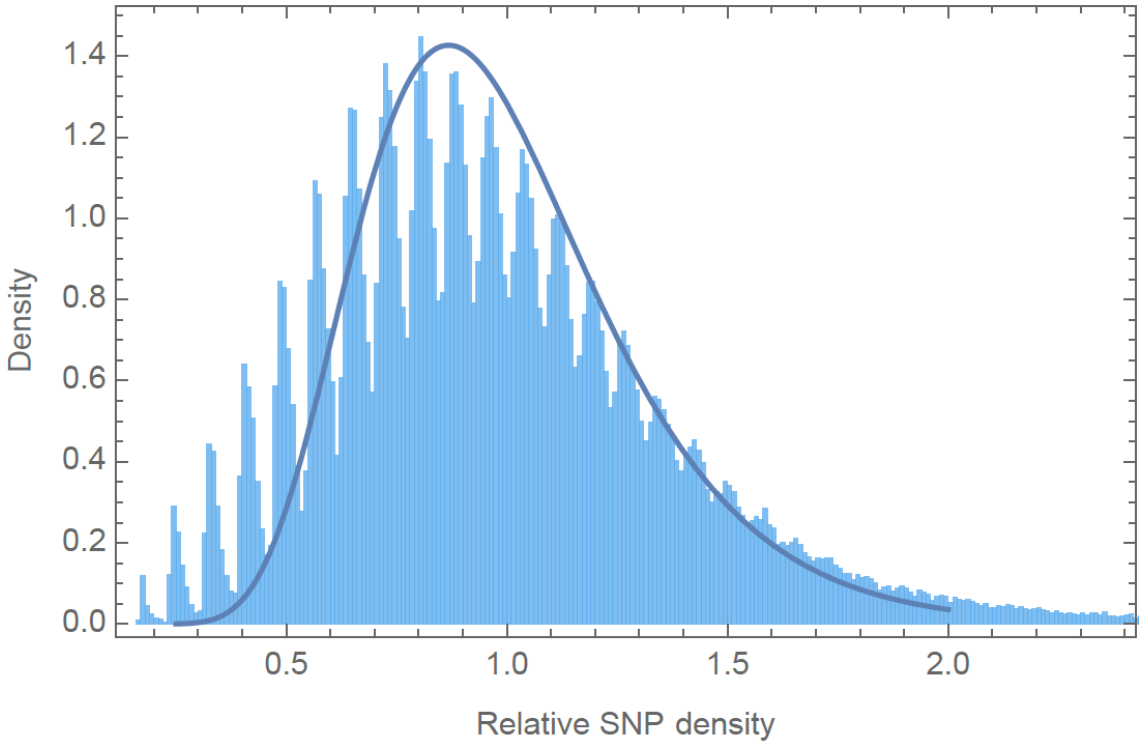


Figure D1. The distribution of SNP density across 10KB windows for SNPs unaffected by BGC. The density has been normalised such that the mean is one, and the 1.5% of windows with the lowest SNP density have been removed. Also shown is the fitted lognormal distribution, which has a shape parameter of 0.31. The graph is ragged because many windows have single digit numbers of SNPs; the variation around each peak is then generated by variation in the number of sites in each window.

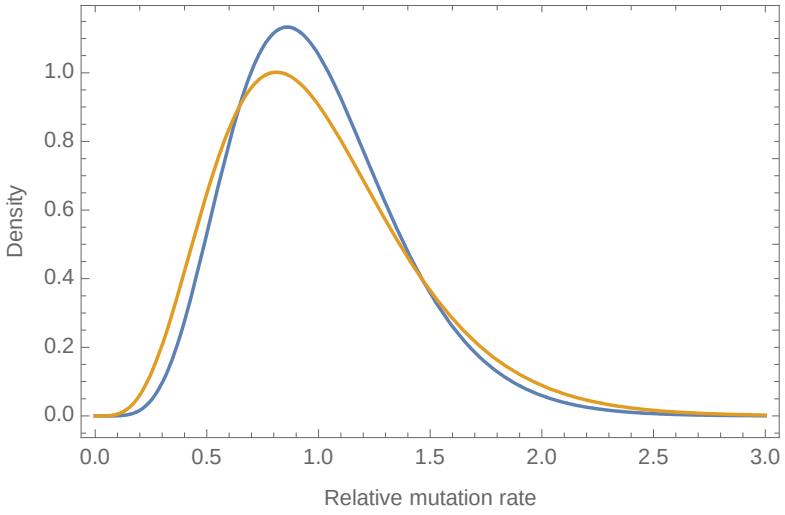


Figure D2. Comparing the gamma distribution fitted to the Jonsson (blue) and Wong (orange) DNM datasets.

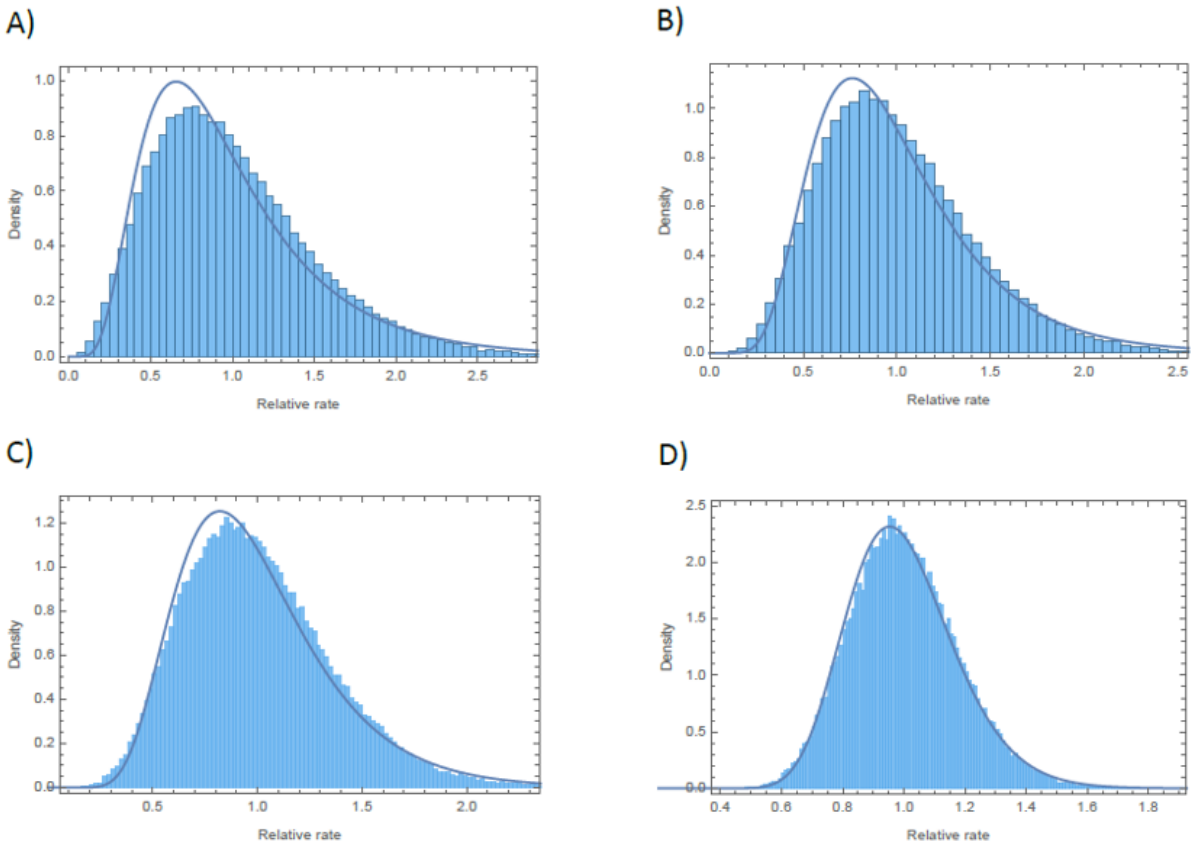
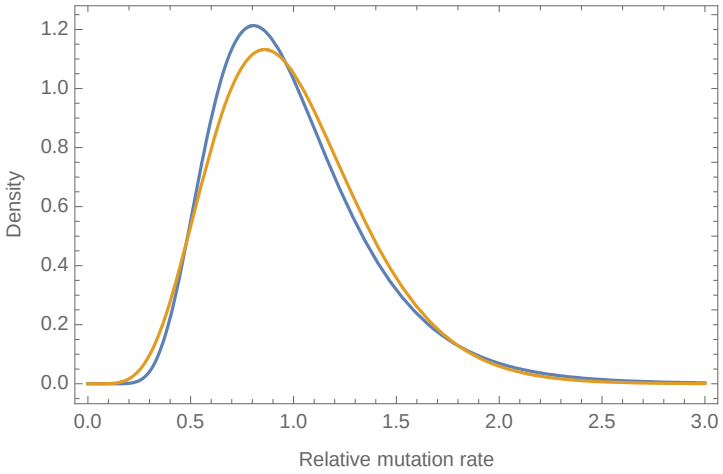


Figure D3. Fit of the lognormal distribution to the gamma. Panels show the fit of the lognormal distribution to the distribution of 100,000 random samples from gamma distributions with shape parameters of (A) 4, (B) 6, (C) 8 and (D) 32. The estimated lognormal shape parameters are 0.53, 0.43, 0.36 and 0.18 respectively.

A)



B)

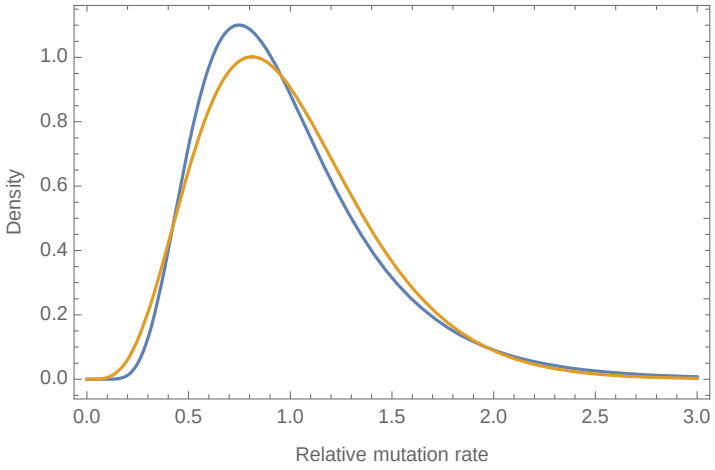


Figure D4. A comparison of the gamma distribution (orange) fitted to the DNM data, and the lognormal distribution (blue) approximating this distribution for the (A) Jonsson and (B) Wong datasets.

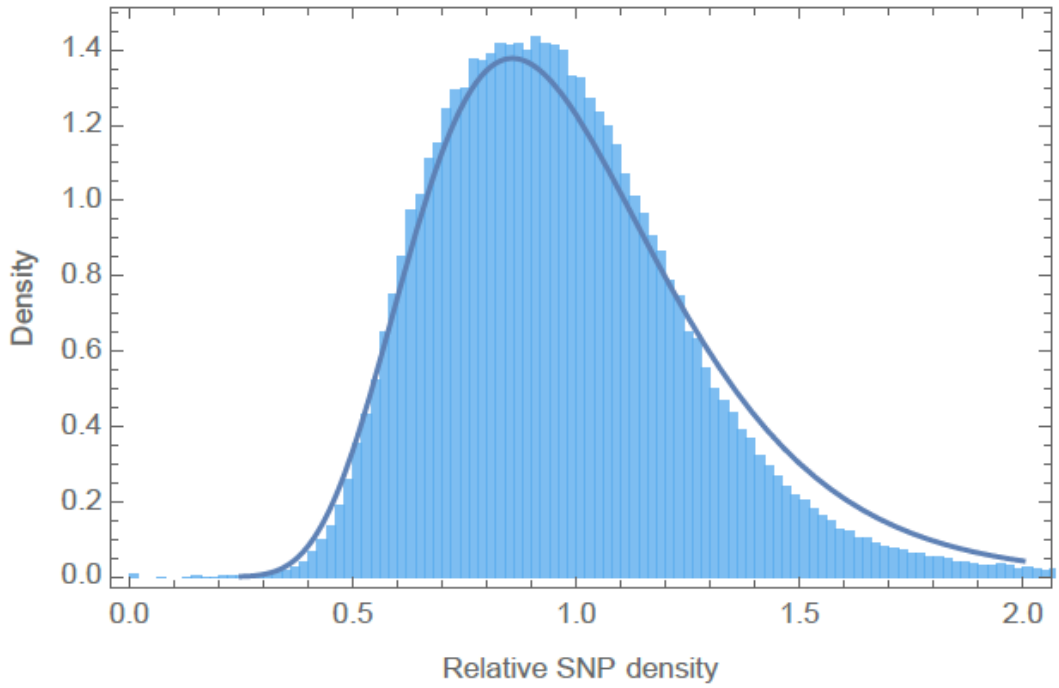


Figure D5. The distribution of substitutions per site estimated to have occurred along the human lineage since humans and chimpanzees split, along with the fitted lognormal distribution which has a shape parameter of 0.32.

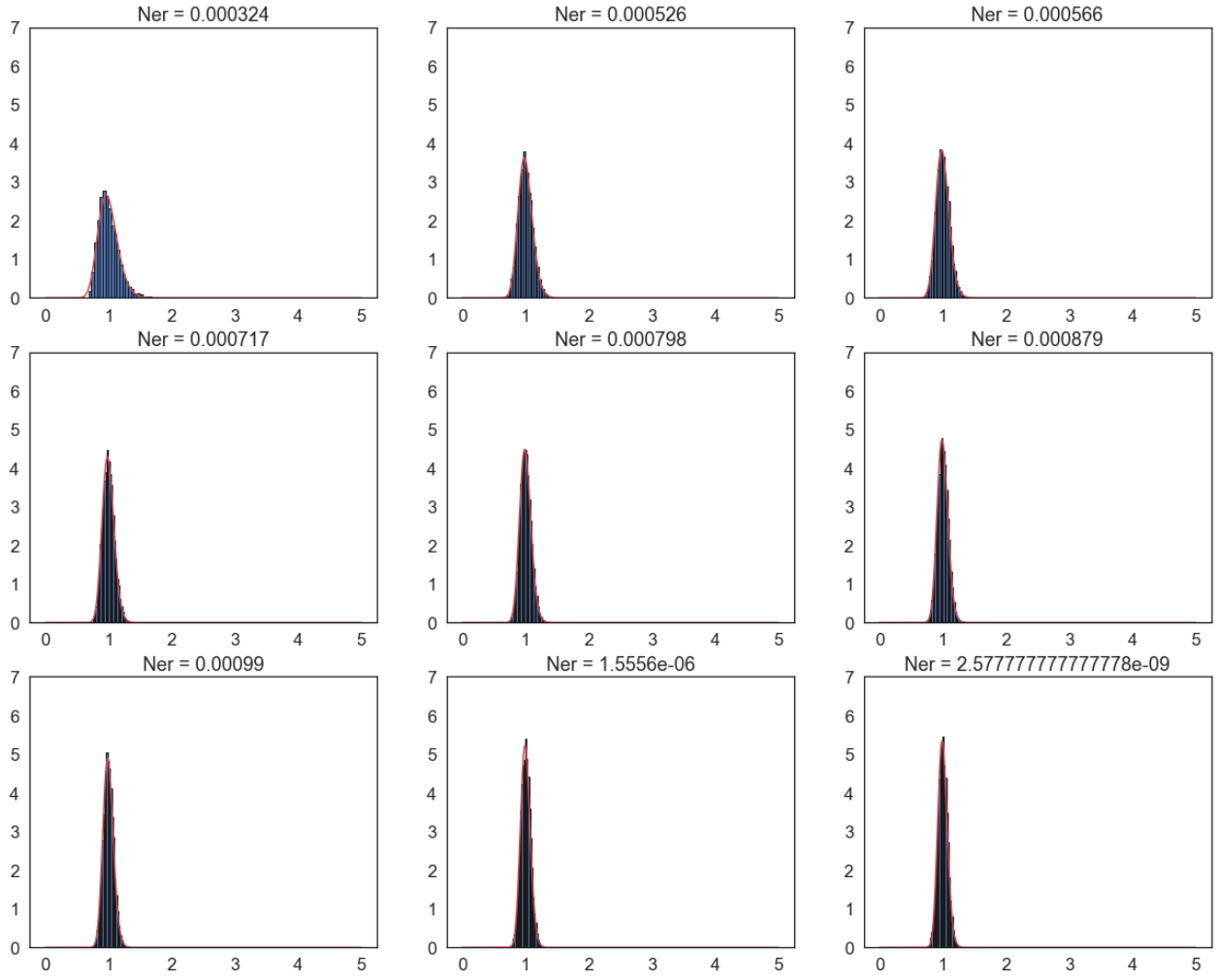


Figure D6. The fit of the lognormal distribution to the distribution of mean genealogy lengths as a function of the product of the effective population size and the rate of recombination, $N_e r$ for a sample of 1000 chromosomes when the population size is stationary.

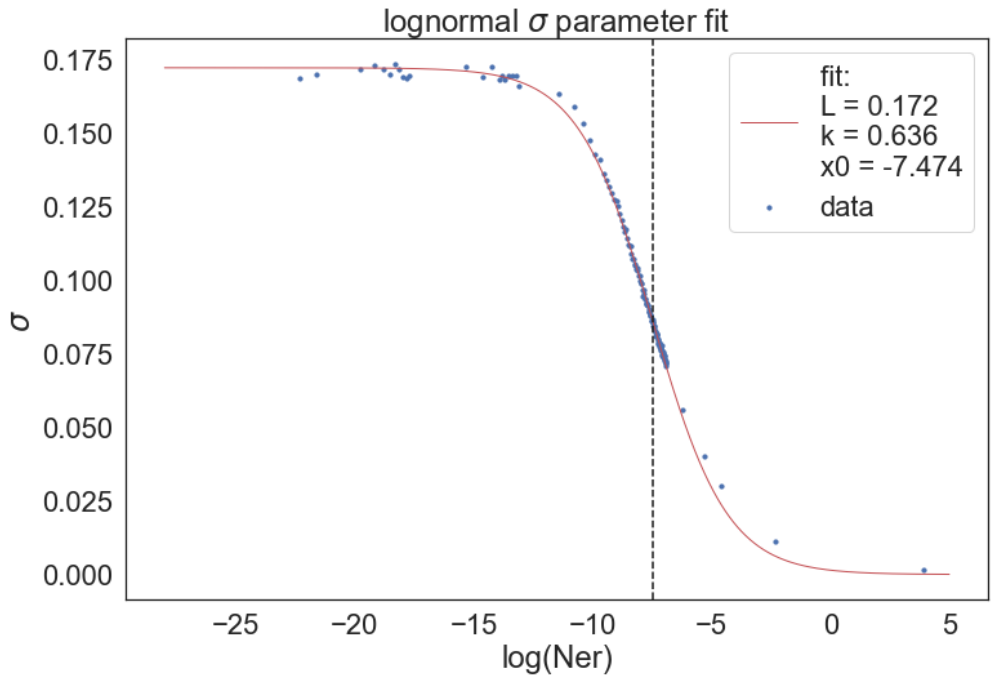


Figure D7. A logistic equation fitted to the relationship between the shape parameter of the mean genealogy length and the log of $N_e r$ for a sample size of 1000 chromosomes. The logistic equation takes the form $S_{\bar{g}} = \frac{0.172}{1+e^{(0.636(Log_{10}(N_e r)+0.747)}}$

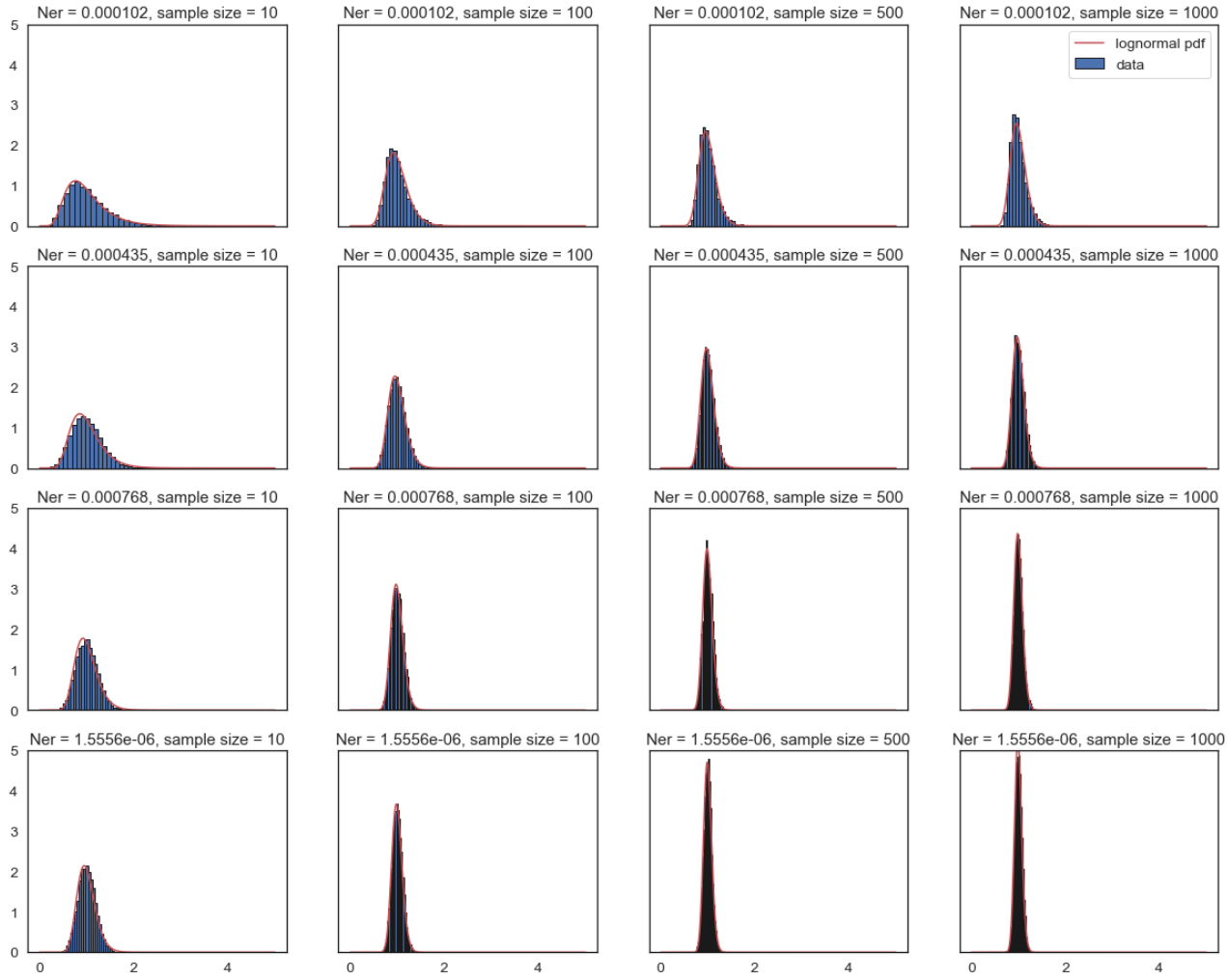


Figure D8. The fit of the lognormal distribution to the distribution of mean genealogy lengths for sample sizes of 10, 100, 500, 1000 and differing levels of recombination.

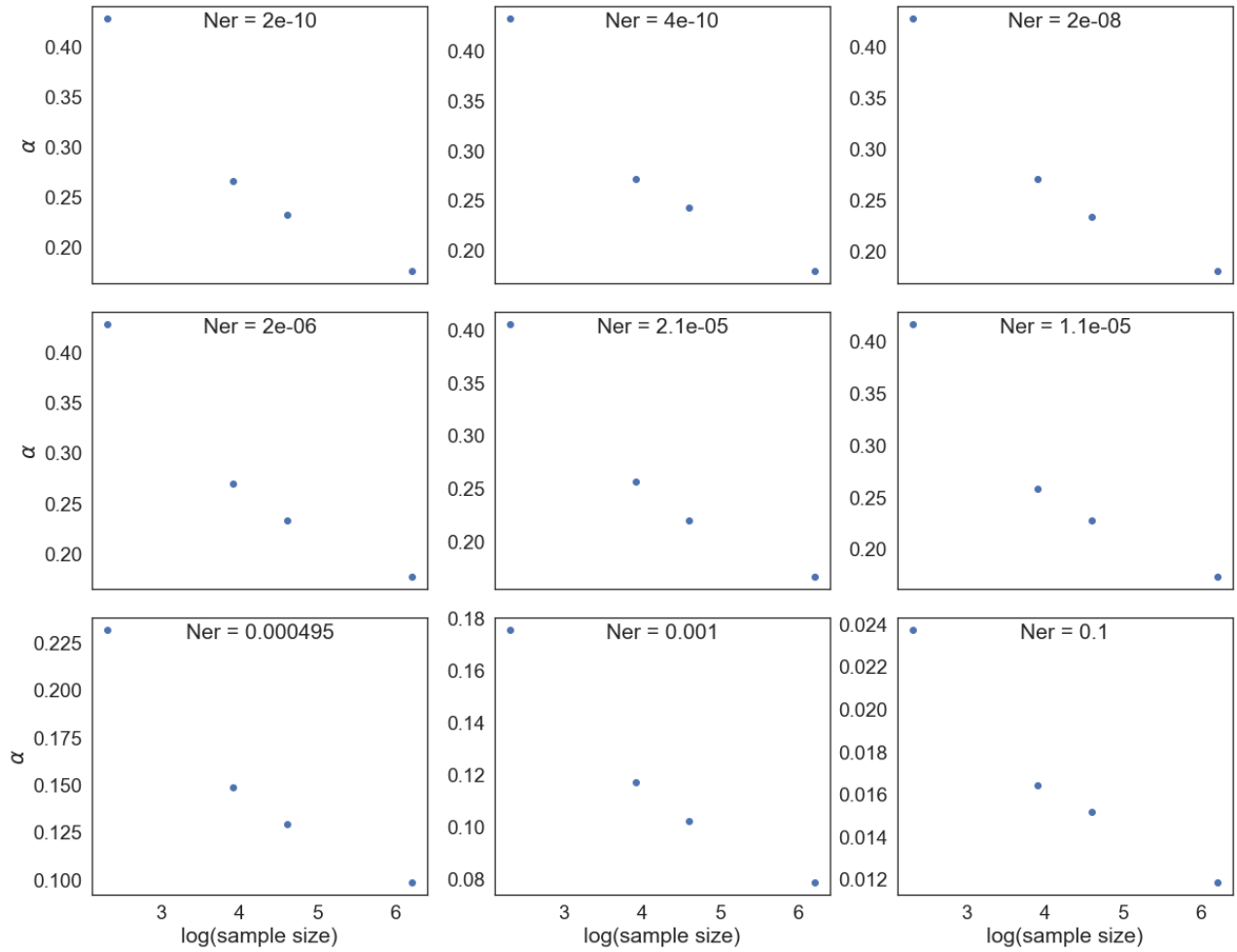


Figure D9. The relationship between the shape parameter of the lognormal distribution fitted to the distribution of mean genealogy lengths as a function of sample size for different values of $N_e r$.

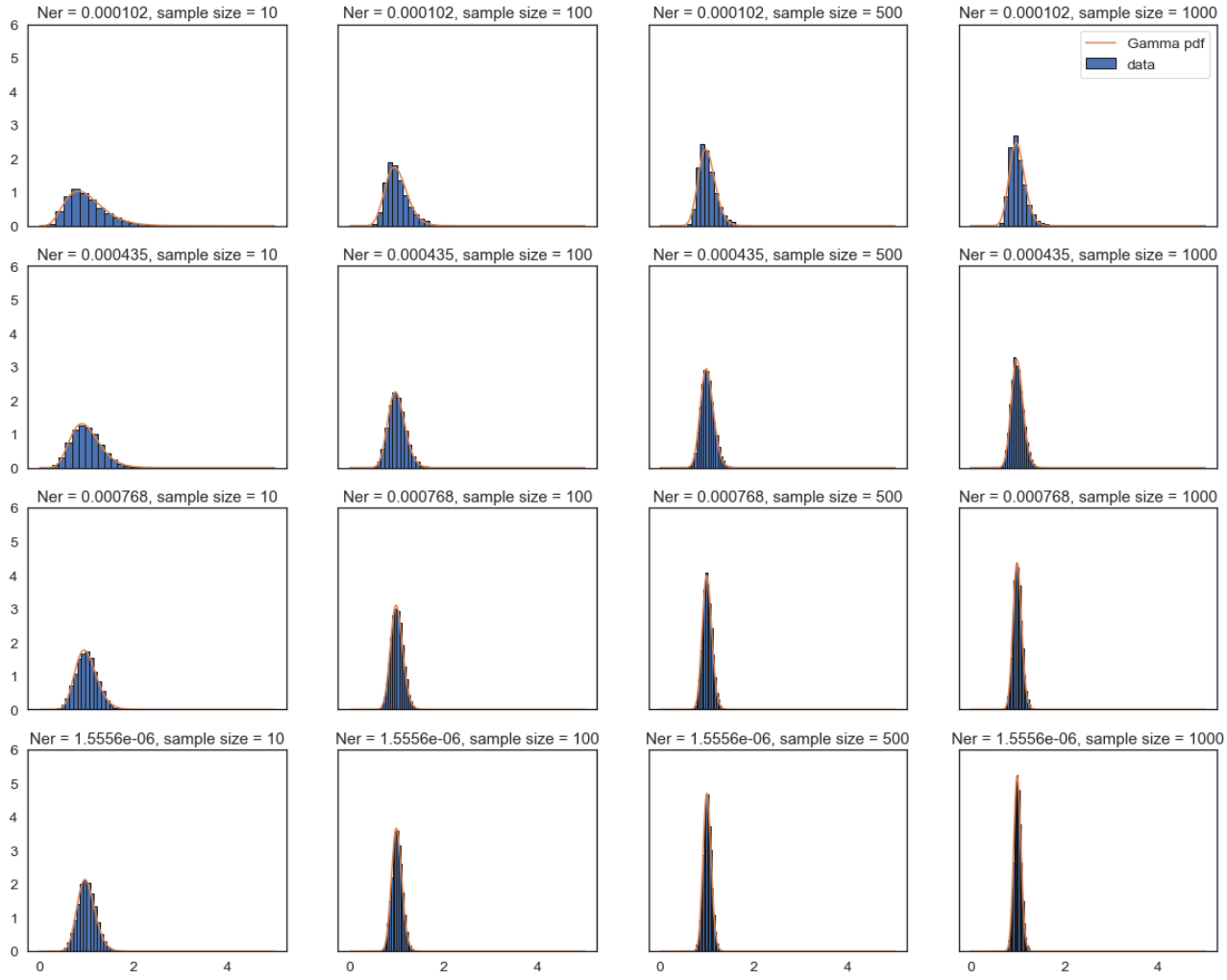


Figure D10. The fit of the gamma distribution to the distribution of mean genealogy lengths for sample sizes of 10, 100, 500, 1000 and differing levels of recombination.

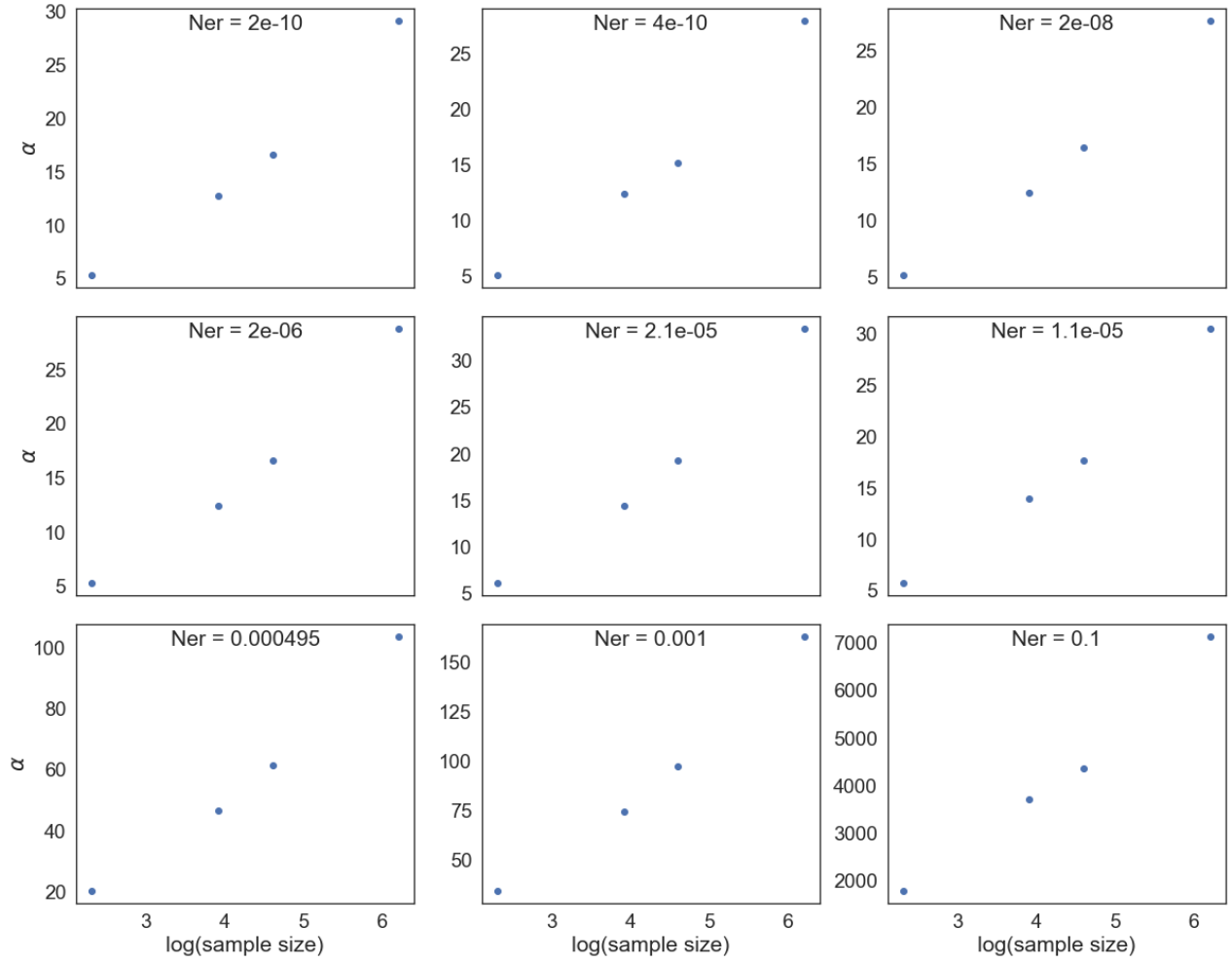


Figure D11. The relationship between the shape parameter of the gamma distribution fitted to the distribution of mean genealogy lengths as a function of sample size for different values of $N_e r$.

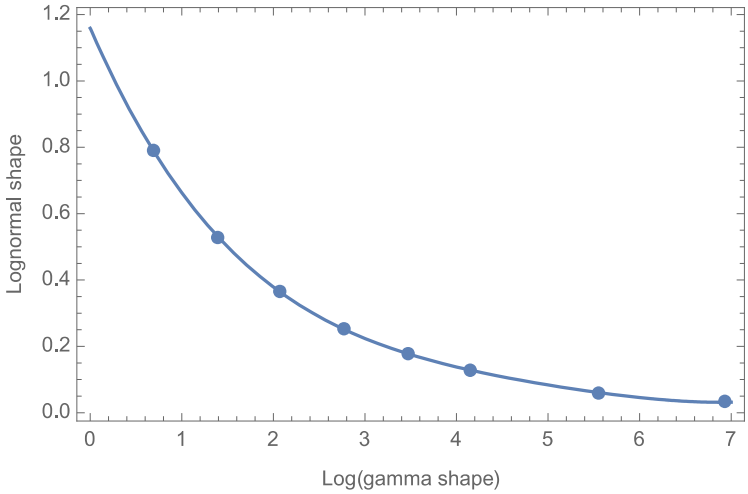


Figure D12. Relationship between estimated lognormal shape parameter and the log of the gamma shape parameter used to simulate the data. The fitted equation has the form $\sigma = 1 - 0.4534 \ln(\beta) + 0.08222 \ln(\beta)^2 - 0.006194 \ln(\beta)^3 + 0.0001141 \ln(\beta)^4$ where σ is the shape parameter of the lognormal distribution and β is the shape parameter of the gamma distribution. Both distributions are assumed to have a mean of one.