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

Vivak Soni[1], Ana Filipa Moutinho[1,2] and Adam Eyre-Walker[1,2,3]

1. School of Life Sciences

University of Sussex

Brighton

BN1 9QG

2. Department for Evolutionary Genetics

Max Planck Institute for Evolutionary Biology

Plon,

Germany

3. Correspondence: a.c.eyre-walker@sussex.ac.uk

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.

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.

Results

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).

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) in hominids (figure 1).

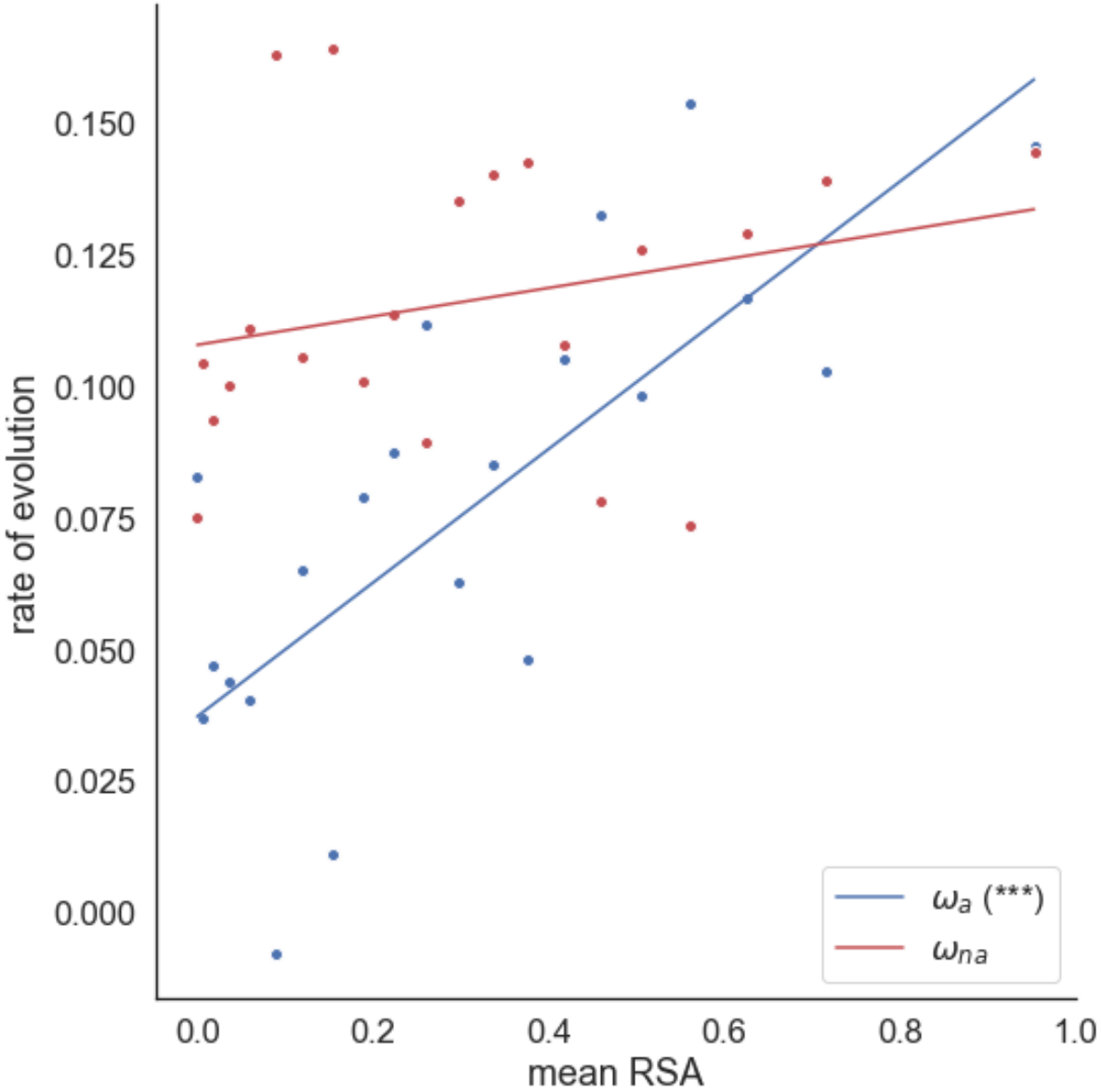

Figure 1: Estimates of ω_a and ω_{na} plotted against mean relative solvent accessibility. Data binned into 20 RSA bins of roughly equal number of sites. For each analysis, a weighted linear regression is fitted to the data. The respective significance of each correlation is shown in the plot legend (*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.

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 the difference in volume ($r = -0.290$, $p = 0.018$) and polarity ($r = -0.269$, $p = 0.027$) (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 (volume difference: $r = -0.545$, $p < 0.001$; polarity difference: $r = -0.170$, $p<0.001$). The slopes are significantly steeper for ω_{na} (Table 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.

		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 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 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 volume and polarity 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.

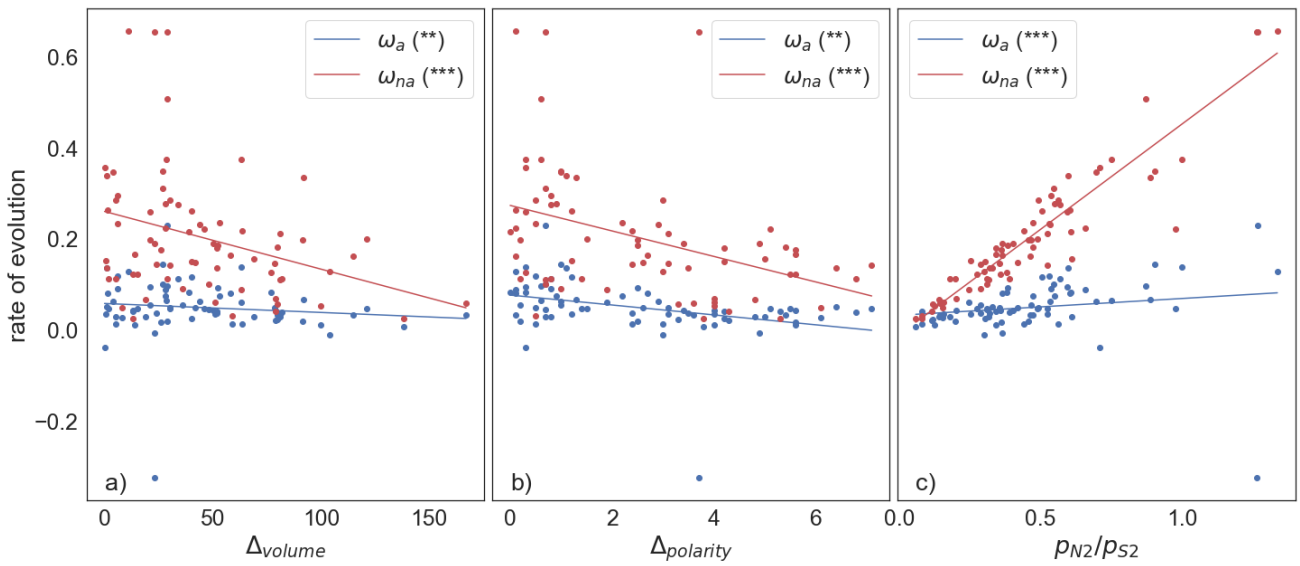

Figure 2: The adaptive and non-adaptive substitution rate plotted against the difference in a) volume, b) polarity and c) the ratio of nonsynonymous to synonymous polymorphisms, p_{N2}/p_{S2}. In c) the polymorphisms are split by sampling from a hypergeometric distribution, with one set used to calculate rates of adaptive and non-adaptive substitution and the other to estimate the two 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 ($r = -0.269$, $p = 0.047$) (supplementary figure S1). 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 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 S1).

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, difference in volume and p_N/p_S all remain significant (RSA: $r = 0.260$, $p < 0.05$; volume difference: $r=-0.576$, $p<0.01$; polarity difference: $r=-0.166$, $p<0.1$; p_{N2}/p_{S2}: $r =0.796$, $p<0.001$); the correlations between ω_{na} and the difference in volume and p_{N2}/p_{S2} remain significant (RSA: $r=0.011$, $p=0.370$; volume difference: $r=0.513$, $p<0.01$; polarity difference: $r=0.115$, $p=0.150$; p_{N2}/p_{S2}: $r=0.804$, $p<0.001$).

Discussion

Our analyses showed that the rate of adaptive evolution is significantly positively correlated to the relative solvent accessibility and measures of amino acid dissimilarity in humans and chimpanzees, being consistent with previous findings in *Drosophila* (Moutinho et al. 2019; Bergman and Eyre-Walker, 2019). We find similar correlations for the rate of non-adaptive evolution, except in the case of RSA. The majority of these correlations are robust to controlling for the effects of biased gene conversion by restricting the analysis to mutational changes that are unaffected by BGC, despite the huge reduction in the dataset.

It is of interest to ask how the slopes of the relationships between ω_a and the site level factors compare to those previously estimated in Drosophila species. We find that the slope is not significantly different for RSA, the difference in volume and the difference in polarity. However, there is a very striking difference in the slope between ω_a and p_N/p_S between hominids and Drosophilids (Table 2). This might be because of population size contraction, which is believed to have occurred since the common ancestor of humans and chimpanzees along both lineages (Holboth et al. 2007; Burgess and Yang 2008; Prado-Martinez et al. 2013; Schrago 2014). It is well known that population size expansion leads to an overestimate, and population size contraction to an underestimation of the rate of adaptive evolution in MK-style approaches (McDonald and Kreitman 1991; Eyre-Walker 2002). This is because the effective population differs between the divergence and polymorphism phases of evolution if the population size changes;. If there has been population size contraction, there will be an excess of slightly deleterious mutations segregating in the polymorphism data because they had no chance of being fixed during the divergence phase with its larger effective population size. How do population size expansion and contraction affect the relationship between ω_a and other variables?

Dataset	Independent variable	Hominids (this analysis) Slope	SE of slope	Drosophila (Bergman and Eyre-Walker 2019) Slope	SE of slope	Sig.
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 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).

Let us assume that synonymous mutations are neutral and non-synonymous mutations are neutral or subject to 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. 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 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

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

This equation demonstrates 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 strength of selection and a positive correlation with population contraction, if there is no adaptive evolution.

If we note that

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

(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

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

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.

The method we have used to estimate ω_a generates an estimate of the mean strength of selection acting against deleterious mutations. We find that log($\bar{s}$) is significantly positively correlated to the difference in volume (r=0.205, p=0.08) and the difference in polarity (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). However, the slope of the relationship is much steeper for p_N/p_S: if we standardise the variables by subtracting the mean and dividing by the standard deviation the slopes are: RSA = -0.101, Volume b = 0.862, Polarity, b= 1.30, p_N/p_S = -3.90. Hence, we expect the effects of population contraction to have more of an effect on the relationship between $\omega_a$and p_N/p_S than any other variable. This is what we observe: the relationship between ω_a and RSA, volume and polarity is similar in hominids and *Drosophila*, but the relationship between ω_a and pN/pS is much weaker in hominids than *Drosophila* (Table 2).

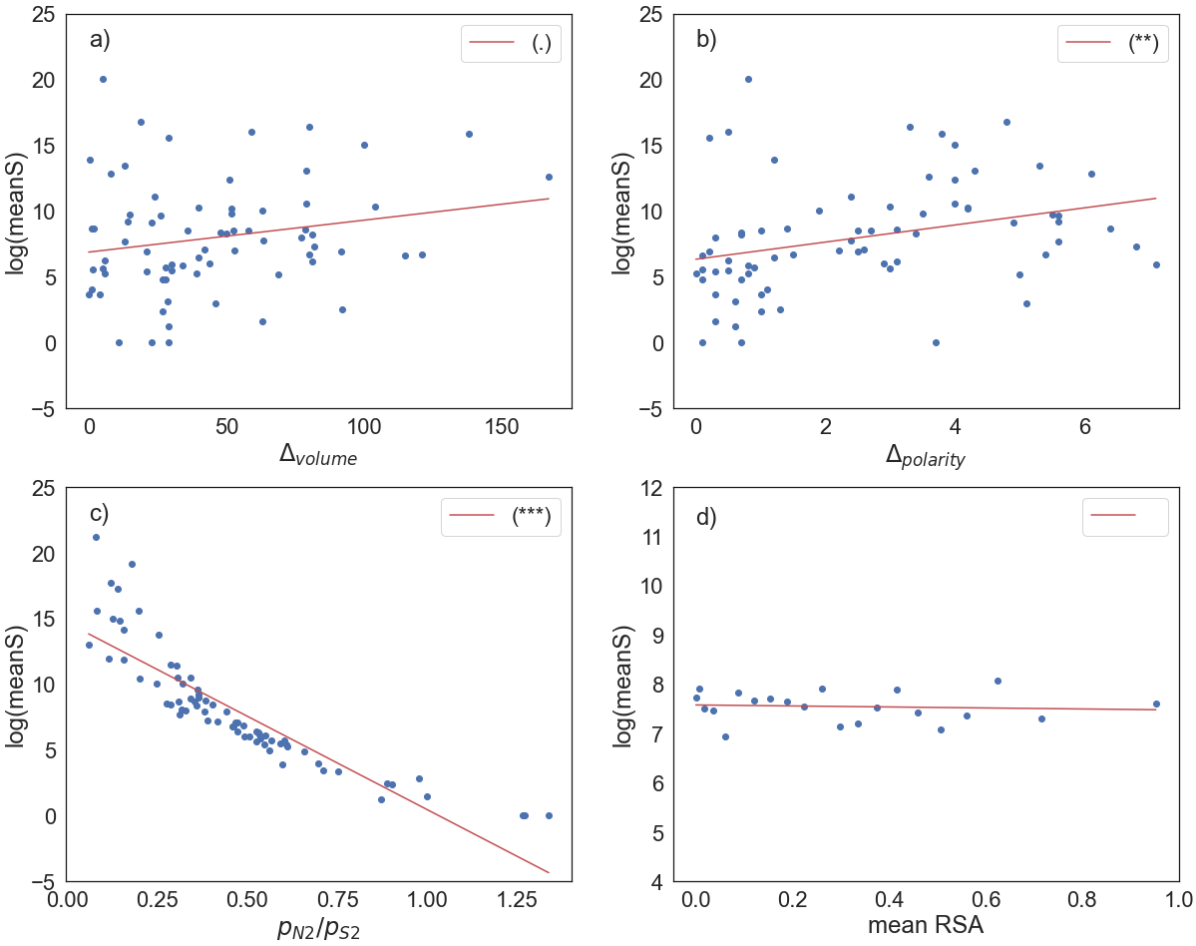

Figure 3: log(meanS) plotted against a) volume difference, b) polarity difference, c) p_{N2}/p_{S2}, d) mean RSA. The respective significance of each correlation is shown in the plot legend, (*P < 0.05; **P < 0.01; ***P < 0.001; "." 0.05 ≤ P < 0.10) based on an unweighted regression fit to the data.

The fact that ω_a is negatively correlated to the difference in volume and polarity, and positively correlated to RSA and p_N/p_S in a pair of species for which we have independent evidence of population contraction (Holboth et al. 2007; Burgess and Yang 2008; Prado-Martinez et al. 2013; Schrago, 2014) suggests that these correlations are due to variation in the rate of adaptive evolution, because population contraction should lead to the opposite patterns if there is no adaptive evolution.

It is well known that population contraction leads to an underestimate of the rate of adaptive evolution when using MK-style methods (McDonald and Kreitman 1990; Eyre-Walker 2002). 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

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

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 (volume: r=-0.012, p=0.707; polarity: r=0.0003, p=0.314).

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. For example, the negative correlation between ω_a and amino acid dissimilarity in *Drosophila* species could be generated by population expansion. It is therefore important to check whether the variable being considered is correlated to the mean strength of selection and/or confirm the correlation in a species which has undergone contraction.

Materials and methods

Data

We obtained gene sequences from Ensembl's biomart (Yates et al. 2020) for the human GRCh38 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.

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, by using the "pdbaa" library and an *E*-value threshold of 10^{-10}. 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.

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.

References

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

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.

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

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

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

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., 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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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. (2000). 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

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