Changing Population Size in McDonald–Kreitman Style Analyses: Artifactual Correlations and Adaptive Evolution between Humans and Chimpanzees

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

[1]School of Life Sciences, University of Sussex, Brighton, United Kingdom

[2]Department for Evolutionary Genetics, Max Planck Institute for Evolutionary Biology, Plon, Germany

*Corresponding author: E-mail: a.c.eyre-walker@sussex.ac.uk.

Accepted: February 2, 2022

Abstract

It is known that methods to estimate the rate of adaptive evolution, which are based on the McDonald–Kreitman test, can be biased by changes in effective population size. Here, we demonstrate theoretically that changes in population size can also generate an artifactual correlation between the rate of adaptive evolution and any factor that is correlated to the strength of selection acting against deleterious mutations. In this context, we have investigated whether several site-level factors influence the rate of adaptive evolution in the divergence of humans and chimpanzees, two species that have been inferred to have undergone population size contraction since they diverged. We 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 nonsynonymous to synonymous polymorphisms (p_N/p_S). All of these correlations are opposite to the artifactual correlations expected under contracting population size. We therefore conclude that these correlations are genuine.

Key words: adaptive evolution, McDonald–Kreitman, human, chimpanzee.

Significance

Understanding the factors that affect the rate of adaptive evolution is a major goal of evolutionary biology. We demonstrate that a commonly used method to estimate the rate of adaptive evolution can generate artifactual correlations between the rate of adaptive evolution and another variable when there has been a change in population size. We investigate a number of factors that might affect the rate of adaptive evolution in humans and chimpanzees, two species which have undergone a contraction in their effective population size since they diverged. We show that the rate of adaptive evolution is correlated to the residue's relative solvent accessibility and the difference between amino acids in their physiochemical properties. We demonstrate that these correlations are not a consequence of contracting population size and are, therefore, genuine.

Introduction

The rate of adaptive evolution in protein coding genes appears to vary 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; Gossmann et al. 2010; also see Strasburg et al. 2009; Ingvarsson 2010; Slotte et al. 2010) and the yeasts of the genus Saccharomyces (Gossmann 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. 2020), 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 mutations more rapidly. There is some evidence that species with large population sizes do undergo significantly faster adaptive evolution (Gossmann et al. 2012; Bataillon et al. 2015; Corbett-Detig et al. 2015; Rousselle et al. 2020; though see Galtier 2016). However, 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 (Lourenço et al. 2013).

At the next level down, there appears to be 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; Chimpanzee Sequencing and Analysis Consortium 2005; Nielsen 2005; Sackton et al. 2007; Obbard et al. 2009), interaction with viruses (Enard et al. 2016), and male reproductive success (Pröschel et al. 2006; Haerty et al. 2007) having higher 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 (Langley et al. 2012; MacKay et al. 2012; Campos et al. 2014). Genes that have lower expression levels (Pál et al. 2001; Rocha and Danchin 2004; Subramanian and Kumar 2004; Wright et al. 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 nonsynonymous with synonymous substitution (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. 2006; 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 nonsynonymous substitution is positively correlated to the 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 substantially 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; 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 nonsynonymous to synonymous polymorphisms, p_N/p_S.

Many of the analyses discussed above used methods based on the McDonald–Kreitman test (McDonald and Kreitman 1991) to infer the rate of adaptive evolution (Boyko et al. 2008; Eyre-Walker and Keightley 2009; Galtier 2016). In these methods, evolution at sites subject to selection is compared with that at putatively neutral sites, using both polymorphism and divergence data; the site frequency spectrum (SFS), derived from the polymorphism data, is used to infer the distribution of fitness effects (DFE), and the DFE is then used to predict how many neutral or deleterious substitutions are expected at the selected sites between the two species. If more divergence is observed than expected, then adaptive evolution is inferred and quantified. It has however, been appreciated for a long time that population size change can lead to biased estimates of the rate of adaptive evolution (McDonald and Kreitman 1991; Eyre-Walker 2002). If the current effective population size, from which the polymorphism data are sampled, is larger than that during divergence, then rates of adaptive evolution will be overestimated (Eyre-Walker 2002). This is because slightly deleterious mutations, which might have been fixed during the divergence phase, no longer segregate because selection is more effective in the current larger population size. If the effective population size during the divergence phase is greater than the current, the rate of adaptive evolution tends to be underestimated.

Population size change might also affect the relationship between the estimated rate of adaptive evolution and a genomic variable. Here, we explore this possibility theoretically and show that population size change induces an apparent correlation between the rate of adaptive evolution and any genomic variable that is correlated to the mean strength of selection acting against deleterious mutations, even when no adaptive evolution has occurred. Hence, some of the correlations that have been observed between the rate of adaptive evolution and another variable could in fact be an artifact of population size change.

Humans and chimpanzees present an interesting case because both ancestral and current effective population sizes have been estimated, and these two species appear to have undergone a substantial decrease in their effective population size since they diverged (Hobolth et al. 2007; Burgess and Yang 2008; Prado-Martinez et al. 2013; Schrago 2014). Here, 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 we investigate whether the apparent correlations could be an artifact of the decrease in effective population size. What we discover is the opposite. The decrease in effective population size is predicted to generate correlations that are contrary to those we observe, suggesting that the rate of adaptive evolution is genuinely correlated to a number of different genomic variables at the site level.

Results

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 artifactually 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 nonsynonymous mutations are neutral or subject to negative selection. The ratio of the nonsynonymous to synonymous substitution rates $\omega = \omega_a + \omega_{na}$, where ω_a and ω_{na} are the rate of adaptive and nonadaptive nonsynonymous 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 nonsynonymous 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, equation 23) 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 SFS at synonymous and nonsynonymous sites, to estimate the DFE at nonsynonymous 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 equations (2) and (3) into (1), we get an expression for the estimated value of ω_a,

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

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 al. 2008, equation 26), where m is a constant which depends on how many chromosomes have been sampled and a scaling factor, 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 population size expansion and negatively correlated if there has been contraction and 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:

$$\omega \approx \frac{k}{(N_d\bar{s})^{\beta}} + A(\bar{s}), \quad (8)$$

which leads to a revision of equations (4) and (6):

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

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

Thus, if the rate of adaptive evolution is independent of the mean strength of selection acting against deleterious mutations, that is, $A(\bar{s}) = a$, then it is evident that our predictions, derived under the assumption of no adaptive evolution, hold—that is, population contraction will induce an artifactual 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.

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 nonadaptive evolution in hominids—RSA, and measures of physicochemical dissimilarity (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 nonadaptive evolution using the statistics ω_a and ω_{na}, which are respectively estimates of the rate of adaptive and nonadaptive evolution relative to the mutation rate. Both statistics were estimated using an extension of the McDonald–Kreitman method (McDonald and Kreitman 1991) 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 nonadaptive evolution across several RSA categories in *Drosophila* and *Arabidopsis*, finding positive correlations between RSA and the rates of both adaptive and nonadaptive 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 nonadaptive evolution. However, the correlation with the rate of nonadaptive evolution is nonsignificant ($r = 0.001$, $P = 0.324$) (fig. 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 randomized data sets did we see a correlation as strong as that observed for ω_a in the real data.

Amino Acid Dissimilarity

To investigate whether the rates of adaptive and nonadaptive evolution are affected by amino acid dissimilarity, we

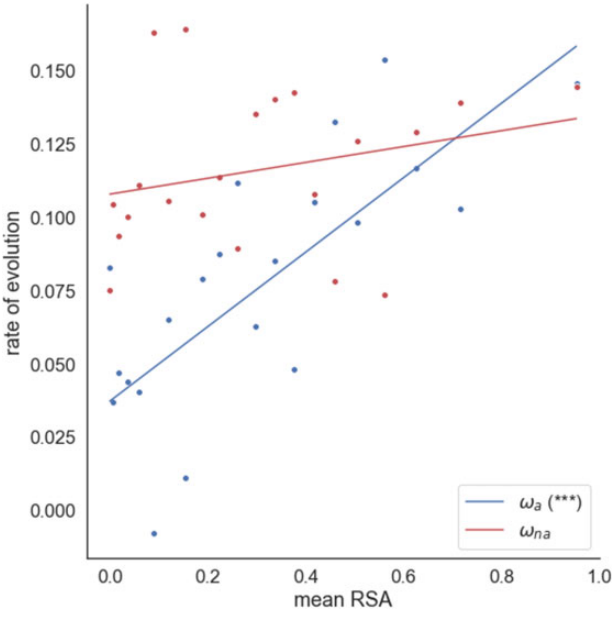

Fig. 1.—Estimates of ω_a and ω_{na} plotted against mean relative solvent accessibility. Data binned into 20 RSA bins of roughly equal number of sites. For each analysis, a weighted linear regression is fitted to the data. The respective significance of each correlation is shown in the plot legend (*$P < 0.05$; **$P < 0.01$; ***$P < 0.001$; "." $0.05 \leq P < 0.10$). 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.

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 ($r=-0.269$, $P=0.027$) (fig. 2*a* and *b*) 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 nonadaptive evolution (Δvolume: $r=-0.545$, $P<0.001$; Δpolarity: $r=-0.170$, $P<0.001$); these correlations remain highly significant ($P<0.001$ in both cases) even if the four datapoints in the top left-hand corner are removed. The slopes are significantly steeper for ω_{na} than ω_a (table 1); however, this appears to be simply because rates of nonadaptive 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 1).

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 nonadaptive evolution. A multiple regression confirms this for ω_{na} with both factors being highly significant and of similar influence, as judged by standardized 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.

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 nonsynonymous 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 are used to estimate both the rates of adaptive and nonadaptive 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 SFS 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_{N1}/p_{S1} ($r=0.419$, $P<0.001$) in hominids, consistent with previous findings in *Drosophila* (Bergman and Eyre-Walker 2019). Consistent with our physico-chemical dissimilarity correlations, ω_{na} also shows a positive correlation with p_{N1}/p_{S1}. The correlation is stronger and the slope steeper than we see for ω_a ($r=0.882$, $P<0.001$) (fig. 2*c* and table 1); however, if we divide ω_a and ω_{na} by their means, we find that the slopes are no longer significantly different (table 1).

It is possible that the correlations between ω_a and ω_{na} and various site level factors are interrelated; for example, the

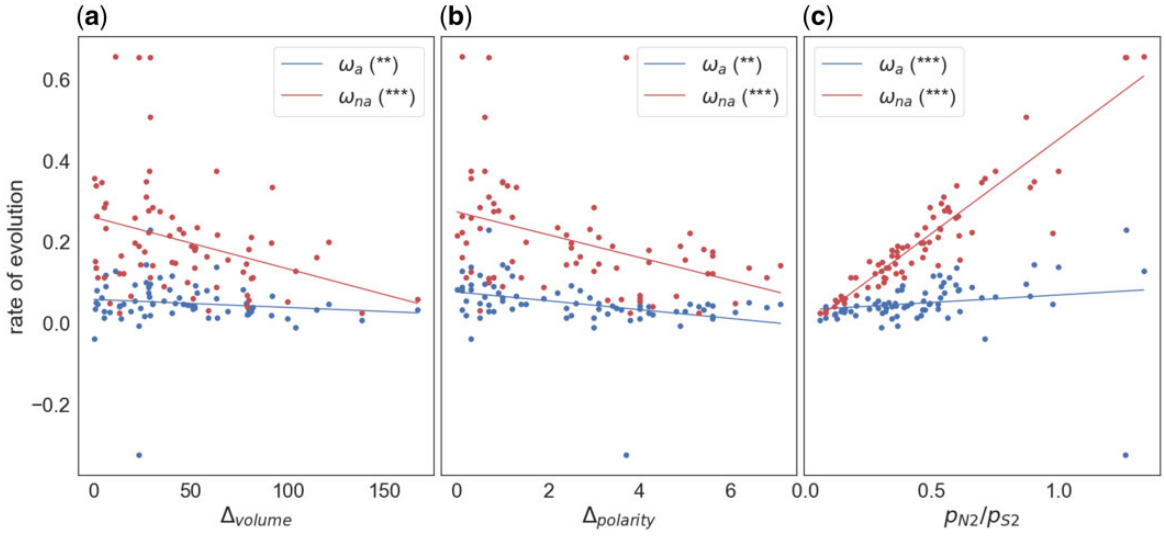

Fig. 2.—The adaptive and nonadaptive substitution rate plotted against the difference in (*a*) volume, (*b*) polarity, and (*c*) the ratio of nonsynonymous to synonymous polymorphisms, p_{N2}/p_{S2}, for 75 pairs of amino acids In (*c*), the polymorphisms are split by sampling from a hypergeometric distribution, with one set used to calculate rates of adaptive and nonadaptive 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 ($^*P<0.05$; $^{**}P<0.01$; $^{***}P<0.001$; "." $0.05\leq P<0.10$).

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

		ω_a		ω_{na}		
Statistic	**Rescaled**	**Slope**	**SE**	**Slope**	**SE**	**Sig.**
ΔVolume	No	−0.00027	0.000098	−0.00009	0.00026	0.012
ΔPolarity	No	−0.0064	0.0020	−0.0020	0.0054	0.010
p_N/p_S	No	0.060	0.020	0.41	0.021	0.000
ΔVolume	Yes	−0.0054	0.0020	−0.0020t	0.0013	ns
ΔPolarity	Yes	−0.13	0.042	−0.042	0.027	ns
p_N/p_S	Yes	1.3	0.41	2.0	0.10	ns

NOTE.—Significance was measured using an analysis of variance.

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 neighbors (RSA-Δvolume: $r = -0.171$, $P = 0.471$; RSA-Δpolarity: $r = 0.059$, $P = 0.803$—supplementary fig. S1, Supplementary Material online).

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 nonsynonymous 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 nonadaptive 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—that is, A<>T and G<>C changes. This reduces our data set substantially by 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 Δpolarity becoming nonsignificant (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 nonadaptive 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}: $r = 0.804$, $P < 0.001$).

Are the Correlations Artifactual?

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 artifact 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 $\log(|\bar{s}|)$ 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 ($r = -0.088$, $P = 0.704$) (fig. 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 artifact of population size contraction, but are genuine.

Comparison to *Drosophila*

It is of interest to ask how the slopes of the relationships between ω_a and each factor compares with 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 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 standardize 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, Δvolume $b = 0.862$, Δpolarity, $b = 1.30$, 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.

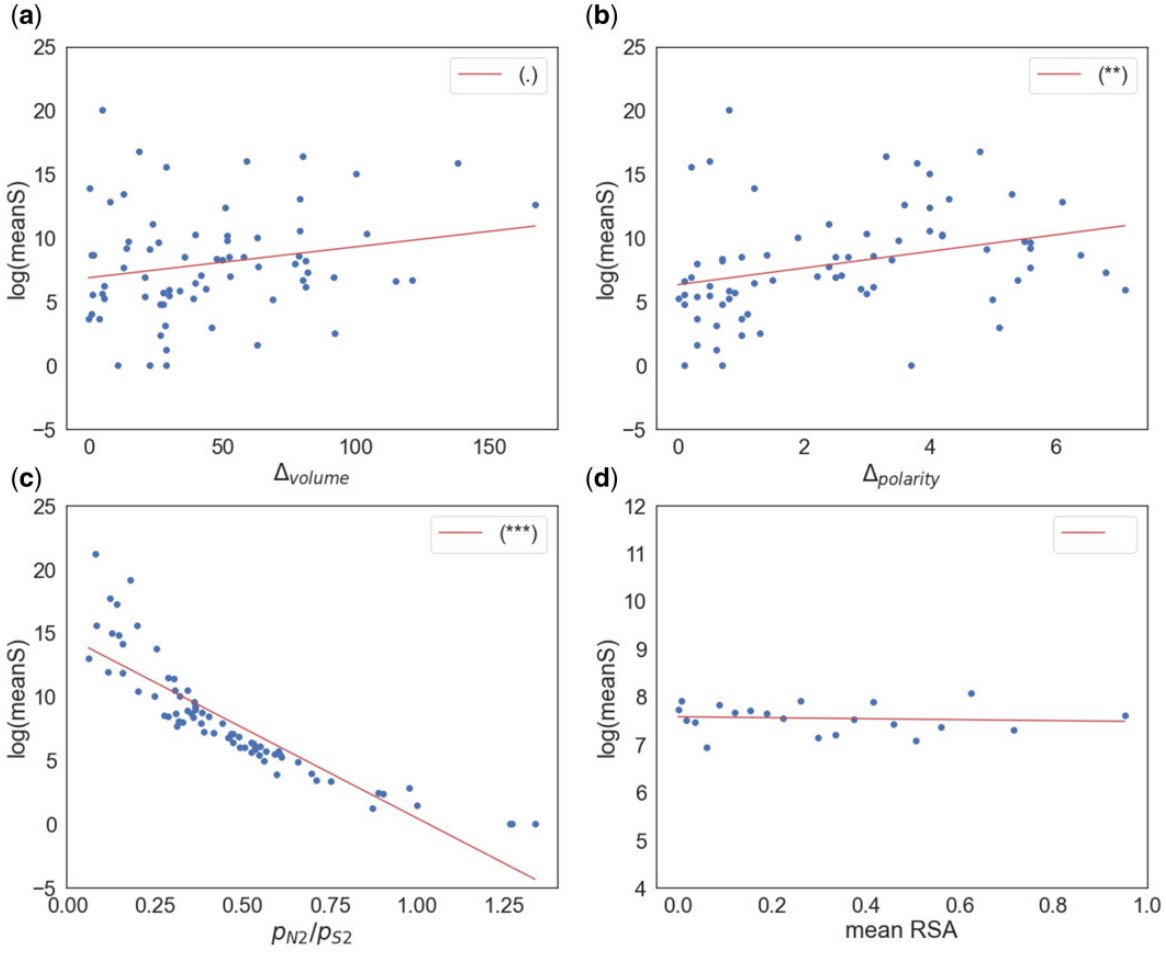

FIG. 3.—Log(mean*S*) plotted against (*a*) volume difference, (*b*) polarity difference, (*c*) p_{N2}/p_{S2}, (*d*) mean RSA. The respective significance of each correlation is shown in the plot legend, ($*P < 0.05$; $**P < 0.01$; $***P < 0.001$; "." $0.05 \leq P < 0.10$) based on an unweighted regression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

Table 2

Slopes of the Regressions between ω_a and Measures of Amino Acid Dissimilarity in Hominid and *Drosophila* Data Sets

		Hominids (This Analysis)		Drosophila (Bergman and Eyre-Walker 2019)		Sig.
Data Set	**Independent Variable**	**Slope**	**SE of Slope**	**Slope**	**SE of Slope**	
Original	RSA	0.13	0.029	0.078	0.0065	ns
Original	ΔVol	−0.00026	0.00010	−0.00027	0.000061	ns
Original	ΔPol	−0.0064	0.0020	−0.0047	0.0011	ns
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	ns
Rescaled	ΔVol	−0.0054	0.0020	−0.011	0.0024	ns
Rescaled	ΔPol	−0.13	0.042	−0.18	0.041	ns
Rescaled	p_N/p_S	1.3	0.40	11	1.1	<0.001

NOTE.—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).

Discussion

One of the main weaknesses of the 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 artifactual 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 artifact 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 (Hobolth et al. 2007; Burgess and Yang 2008; Prado-Martinez et al. 2013; Schrago 2014). Given the observed correlations between each factor 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 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 in between these timepoints. For example, the ancestral population may have contracted after the species diverged and subsequently re-expanded toward the present. Under this scenario, the effective population during the divergence phase could have been lower than that during the polymorphism phase. Furthermore, changes in N_e affect neutral and selected mutations differently (Otto and Whitlock 1997). Since, human population sizes have increased dramatically recently, the effective population size for deleterious genetic variation is greater than for neutral variation, because the deleterious mutations are younger on average. It is therefore possible that the slightly deleterious genetic variation, which can potentially bias MK-style methods, has not experienced a smaller N_e during the polymorphism relative to the divergence phase. However, this seems unlikely; the current N_e for neutral variation is estimated to be between 5- and 10-fold lower than the population size of human–chimpanzee ancestor (Hobolth et al. 2007; Burgess and Yang 2008; Prado-Martinez et al. 2013; Schrago 2014), and the mutations that are most likely to affect the method to estimate the rate of adaptive evolution are weakly selected.

Population contraction leads to an underestimate of the rate of adaptive evolution when using MK-style methods (McDonald and Kreitman 1991; 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 $\omega_a = 0.0488$ [SE $= 0.0072$]; *Drosophila* mean $\omega_a = 0.0258$ [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:

$$\omega_a(\text{obs}) = \omega_a(\text{true}) + \omega_a(\text{predicted}), \qquad (10)$$

where $\omega_a(\text{true})$ is the true value, and $\omega_a(\text{predicted})$ is the value predicted in the absence of adaptive evolution from equations (4) or (6); that is, 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 nonadaptive 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 (2019)—rates of adaptive and nonadaptive 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 focused on a statistic that is related to the proportion of substitutions that are adaptive, and hence conflates the pattern of adaptive and nonadaptive evolution. In fact, consistent with their findings and those of Bergman and Eyre-Walker (2019), 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: $r = 0.0003$, $P = 0.314$); however, the proportion is significantly positively correlated to both p_{N1}/p_{S1} ($r = 0.20$, $P = 0.046$) and RSA ($r = 0.44$, $P = 0.027$).

In summary, we demonstrate that population size change can lead to an artifactual 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.

Materials and Methods

Data

We obtained gene sequences from Ensembl's biomart (Yates et al. 2019) 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 nonsynonymous substitutions per site were obtained using PAML's codeml (Yang 2007) program. We used polymorphism data from the African superpopulation of the 1000 genomes data set (1000 Genomes Project Consortium 2015) to construct our SFS, with rates of adaptive and nonadaptive 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 data set by gene 100 times.

RSA Analysis

In order to obtain structural information for each protein sequence, we ran BlastP (Schäffer et al. 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 SA area (Tien et al. 2013), giving us a final data set 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 nonadaptive 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. 2007). 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 nonadaptive evolution, and one used to estimate p_N/p_S.

Supplementary Material

Supplementary data are available at *Genome Biology and Evolution* online.

Data Availability

The analysis uses publicly available data. Scripts used to process the data are available at https://github.com/vivaksoni/site_level_factors_affecting_rates_of_evolution_in_hominids.

Literature Cited

1000 Genomes Project Consortium. 2015. A global reference for human genetic variation. Nature 526(7571):68–74.

Arguello JR, et al. 2010. Recombination yet inefficient selection along the *Drosophila melanogaster* subgroup's fourth chromosome. Mol Biol Evol. 27(4):848–861.

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

Bataillon T, et al. 2015. Inference of purifying and positive selection in three subspecies of chimpanzees (*Pan troglodytes*) from exome sequencing. Genome Biol Evol. 7(4):1122–1132.

Berglund J, Pollard KS, Webster MT. 2009. Hotspots of biased nucleotide substitutions in human genes. PLoS Biol. 7(1):e1000026.

Bergman J, Eyre-Walker A. 2019. Does adaptive protein evolution proceed by large or small steps at the amino acid level? Mol Biol Evol. 36(5):990–998.

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

Boyko AR, et al. 2008. Assessing the evolutionary impact of amino acid mutations in the human genome. PLoS Genet. 4(5):e1000083.

Burgess R, Yang Z. 2008. Estimation of hominoid ancestral population sizes under Bayesian coalescent models incorporating mutation rate variation and sequencing errors. Mol Biol Evol. 25(9):1979–1994.

Bustamante CD, et al. 2002. The cost of inbreeding in *Arabidopsis*. Nature 416(6880):531–534.

Bustamante CD, Townsend JP, Hartl DL. 2000. Solvent accessibility and purifying selection within proteins of *Escherichia coli* and *Salmonella enterica*. Mol Biol Evol. 17(2):301–308.

Campos JL, Halligan DL, Haddrill PR, Charlesworth B. 2014. The relation between recombination rate and patterns of molecular evolution and variation in *Drosophila melanogaster*. Mol Biol Evol. 31(4):1010–1028.

Castellano D, Coronado-Zamora M, Campos JL, Barbadilla A, Eyre-Walker A. 2016. Adaptive evolution is substantially impeded by Hill-Robertson interference in *Drosophila*. Mol Biol Evol. 33(2):442–455.

Chen X, et al. 2019. Assessment contributions of physicochemical properties and bacterial community to mitigate the bioavailability of heavy metals during composting based on structural equation models. Bioresour Technol. 289:121657.

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

Choi SS, Vallender EJ, Lahn BT. 2006. Systematically assessing the influence of 3-dimensional structural context on the molecular evolution of mammalian proteomes. Mol Biol Evol. 23(11):2131–2133.

Clark AG, et al. 2003. Inferring nonneutral evolution from human-chimp-mouse orthologous gene trios. Science 302(5652):1960–1963.

Conant GC, Stadler PF. 2009. Solvent exposure imparts similar selective pressures across a range of yeast proteins. Mol Biol Evol. 26(5):1155–1161.

Corbett-Detig RB, Hartl DL, Sackton TB. 2015. Natural selection constrains neutral diversity across a wide range of species. PLoS Biol. 13(4):e1002112.

Dean AM, Neuhauser C, Grenier E, Golding GB. 2002. The pattern of amino acid replacements in α/β-barrels. Mol Biol Evol. 19(11):1846–1864.

Edgar RC. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 32(5):1792–1797.

Enard D, Cai L, Gwennap C, Petrov DA. 2016. Viruses are a dominant driver of protein adaptation in mammals. ELife 5:e12469.

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

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

Franzosa EA, Xia Y. 2009. Structural determinants of protein evolution are context-sensitive at the residue level. Mol Biol Evol. 26(10):2387–2395.

Galtier N. 2016. Adaptive protein evolution in animals and the effective population size hypothesis. PLoS Genet. 12(1):e1005774.

Galtier N, Duret L. 2007. Adaptation or biased gene conversion? Extending the null hypothesis of molecular evolution. Trends Genet. 23(6):273–277.

Gojobori J, Tang H, Akey JM, Wu C-I. 2007. Adaptive evolution in humans revealed by the negative correlation between the polymorphism and fixation phases of evolution. Proc Natl Acad Sci U S A. 104(10):3907–3912.

Goldman N, Thorne JL, Jones DT. 1998. Assessing the impact of secondary structure and solvent accessibility on protein evolution. Genetics 149(1):445–458.

Gossmann TI, et al. 2010. Genome wide analyses reveal little evidence for adaptive evolution in many plant species. Mol Biol Evol. 27(8):1822–1832.

Gossmann TI, Keightley PD, Eyre-Walker A. 2012. The effect of variation in the effective population size on the rate of adaptive molecular evolution in eukaryotes. Genome Biol Evol. 4(5):658–667.

Grant BJ, Rodrigues APC, ElSawy KM, McCammon JA, Caves LSD. 2006. Bio3d: an R package for the comparative analysis of protein structures. Bioinformatics 22(21):2695–2696.

Gravel S, et al. 2011. Demographic history and rare allele sharing among human populations. Proc Natl Acad Sci U S A. 108(29):11983–11988.

Guo HH, Choe J, Loeb LA. 2004. Protein tolerance to random amino acid change. Proc Natl Acad Sci U S A. 101(25):9205–9210.

Gutenkunst RN, Hernandez RD, Williamson SH, Bustamante CD. 2009. Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data. PLoS Genet. 5(10):e1000695.

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

Haerty W, et al. 2007. Evolution in the fast lane: rapidly evolving sex-related genes in *Drosophila*. Genetics 177(3):1321–1335.

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

Hobolth A, Christensen OF, Mailund T, Schierup MH. 2007. Genomic relationships and speciation times of human, chimpanzee, and gorilla inferred from a coalescent hidden Markov model. PLoS Genet. 3(2):e7.

Ingvarsson PK. 2010. Natural selection on synonymous and nonsynonymous mutations shapes patterns of polymorphism in *Populus tremula*. Mol Biol Evol. 27(3):650–660.

Kawashima S, et al. 2007. AAindex: amino acid index database, progress report 2008. Nucleic Acids Res. 36:D202–D205.

Langley CH, et al. 2012. Genomic variation in natural populations of *Drosophila melanogaster*. Genetics 192(2):533–598.

Lemos B, Bettencourt BR, Meiklejohn CD, Hartl DL. 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. Mol Biol Evol. 22(5):1345–1354.

Liao B-Y, Scott NM, Zhang J. 2006. Impacts of gene essentiality, expression pattern, and gene compactness on the evolutionary rate of mammalian proteins. Mol Biol Evol. 23(11):2072–2080.

Liberles DA, et al. 2012. The interface of protein structure, protein biophysics, and molecular evolution. Protein Sci. 21(6):769–785.

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. Mol Biol Evol. 24(4):1005–1011.

Lipman DJ, Souvorov A, Koonin EV, Panchenko AR, Tatusova TA. 2002. The relationship of protein conservation and sequence length. BMC Evol Biol. 2:20.

Lourenço JM, Glémin S, Galtier N. 2013. The rate of molecular adaptation in a changing environment. Mol Biol Evol. 30(6):1292–1301.

Mackay TFC, et al. 2012. The *Drosophila melanogaster* genetic reference panel. Nature 482(7384):173–178.
McDonald JH, Kreitman M. 1991. Adaptive protein evolution at the Adh locus in *Drosophila*. Nature 351(6328):652–654.
Moutinho AF, Trancoso FF, Dutheil JY. 2019. The impact of protein architecture on adaptive evolution. Mol Biol Evol. 36(9):2013–2028.
Nielsen R. 2005. Molecular signatures of natural selection. Annu Rev Genet. 39(1):197–218.
Obbard DJ, Welch JJ, Kim K-W, Jiggins FM. 2009. Quantifying adaptive evolution in the *Drosophila* immune system. PLoS Genet. 5(10):e1000698.
Otto SP, Whitlock MC. 1997. The probability of fixation in populations of changing size. Genetics 146(2):723–733.
Overington J, Donnelly D, Johnson MS, Sali A, Blundell TL. 1992. Environment-specific amino acid substitution tables: tertiary templates and prediction of protein folds. Protein Sci. 1(2):216–226.
Pál C, Papp B, Hurst LD. 2001. Highly expressed genes in yeast evolve slowly. Genetics 158(2):927–931.
Perutz MF, Kendrew JC, Watson HC. 1965. Structure and function of haemoglobin. J Mol Biol. 13(3):669–678.
Prado-Martinez J, et al. 2013. Great ape genetic diversity and population history. Nature 499(7459):471–475.
Presgraves DC. 2005. Recombination enhances protein adaptation in *Drosophila melanogaster*. Curr Biol. 15(18):1651–1656.
Pröschel M, Zhang Z, Parsch J. 2006. Widespread adaptive evolution of *Drosophila* genes with sex-biased expression. Genetics 174(2):893–900.
Ramsey DC, Scherrer MP, Zhou T, Wilke CO. 2011. The relationship between relative solvent accessibility and evolutionary rate in protein evolution. Genetics 188(2):479–488.
Ratnakumar A, et al. 2010. Detecting positive selection within genomes: the problem of biased gene conversion. Philos Trans R Soc Lond B Biol Sci. 365(1552):2571–2580.
Rocha EPC, Danchin A. 2004. An analysis of determinants of amino acids substitution rates in bacterial proteins. Mol Biol Evol. 21(1):108–116.
Rousselle M, et al. 2020. Is adaptation limited by mutation? A timescale-dependent effect of genetic diversity on the adaptive substitution rate in animals. PLoS Genet. 16(4):e1008668.
Sackton TB, et al. 2007. Dynamic evolution of the innate immune system in *Drosophila*. Nat Genet. 39(12):1461–1468.
Sawyer SA, Kulathinal RJ, Bustamante CD, Hartl DL. 2003. Bayesian analysis suggests that most amino acid replacements in *Drosophila* are driven by positive selection. J Mol Evol. 57(0):S154–S164.
Schäffer AA, et al. 2001. Improving the accuracy of PSI-BLAST protein database searches with composition-based statistics and other refinements. Nucleic Acids Res. 29(14):2994–3005.
Schmid KJ, 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.
Schrago CG. 2014. The effective population sizes of the anthropoid ancestors of the human-chimpanzee lineage provide insights on the historical biogeography of the great apes. Mol Biol Evol. 31(1):37–47.
Slotte T, Foxe JP, Hazzouri KM, Wright SI. 2010. Genome-wide evidence for efficient positive and purifying selection in Capsella grandiflora, a plant species with a large effective population size. Mol Biol Evol. 27(8):1813–1821.
Smith NGC, Eyre-Walker A. 2002. Adaptive protein evolution in *Drosophila*. Nature 415(6875):1022–1024.
Strasburg JL, Scotti-Saintagne C, Scotti I, Lai Z, Rieseberg LH. 2009. Genomic patterns of adaptive divergence between chromosomally differentiated sunflower species. Mol Biol Evol. 26(6):1341–1355.
Subramanian S, Kumar S. 2004. Gene expression intensity shapes evolutionary rates of the proteins encoded by the vertebrate genome. Genetics 168(1):373–381.
Tien MZ, Meyer AG, Sydykova DK, Spielman SJ, Wilke CO. 2013. Maximum allowed solvent accessibilites of residues in proteins. PLoS One 8(11):e80635.
Welch JJ, Eyre-Walker A, Waxman D. 2008. Divergence and polymorphism under the nearly neutral theory of molecular evolution. J Mol Evol. 67(4):418–426.
Wright SI, Yau CBK, Looseley M, Meyers BC. 2004. Effects of gene expression on molecular evolution in *Arabidopsis thaliana* and *Arabidopsis lyrata*. Mol Biol Evol. 21(9):1719–1726.
Yang Z. 2007. PAML 4: phylogenetic analysis by maximum likelihood. Mol Biol Evol. 24(8):1586–1591.
Yates AD, et al. 2019. Ensembl 2020. Nucleic Acids Res. 8(D1):D682–D688.
Zhang J. 2000. Protein-length distributions for the three domains of life. Trends Genet. 16(3):107–109.
Zhen Y, Huber CD, Davies RW, Lohmueller KE. 2021. Greater strength of selection and higher proportion of beneficial amino acid changing mutations in humans compared with mice and *Drosophila melanogaster*. Genome Res. 31(1):110–120.

Associate editor: David Enard