PLOS | GENETICS

RESEARCH ARTICLE

Adaptive Protein Evolution in Animals and the Effective Population Size Hypothesis

Nicolas Galtier*

Institut des Sciences de l'Evolution UMR5554, Université Montpellier–CNRS–IRD–EPHE, Montpellier, France

* nicolas.galtier@univ-montp2.fr

Citation: Galtier N (2016) Adaptive Protein Evolution in Animals and the Effective Population Size Hypothesis. PLoS Genet 12(1): e1005774. doi:10.1371/journal.pgen.1005774

Editor: Mikkel H. Schierup, Aarhus University, DENMARK

Received: July 6, 2015

Accepted: December 5, 2015

Published: January 11, 2016

Data Availability Statement: Data are available from the NCBI-SRA database (accession numbers: SRX1470185, SRX1470188, SRX1470199, SRX1470201, SRX1470203).

Funding: This study was supported by European Research Center grant 232971 (PopPhyl). The funder had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The author has declared that no competing interests exist.

Abstract

The rate at which genomes adapt to environmental changes and the prevalence of adaptive processes in molecular evolution are two controversial issues in current evolutionary genetics. Previous attempts to quantify the genome-wide rate of adaptation through amino-acid substitution have revealed a surprising diversity of patterns, with some species (*e.g*. *Drosophila*) experiencing a very high adaptive rate, while other (*e.g*. humans) are dominated by nearly-neutral processes. It has been suggested that this discrepancy reflects between-species differences in effective population size. Published studies, however, were mainly focused on model organisms, and relied on disparate data sets and methodologies, so that an overview of the prevalence of adaptive protein evolution in nature is currently lacking. Here we extend existing estimators of the amino-acid adaptive rate by explicitly modelling the effect of favourable mutations on non-synonymous polymorphism patterns, and we apply these methods to a newly-built, homogeneous data set of 44 non-model animal species pairs. Data analysis uncovers a major contribution of adaptive evolution to the amino-acid substitution process across all major metazoan phyla—with the notable exception of humans and primates. The proportion of adaptive amino-acid substitution is found to be positively correlated to species effective population size. This relationship, however, appears to be primarily driven by a decreased rate of nearly-neutral amino-acid substitution because of more efficient purifying selection in large populations. Our results reveal that adaptive processes dominate the evolution of proteins in most animal species, but do not corroborate the hypothesis that adaptive substitutions accumulate at a faster rate in large populations. Implications regarding the factors influencing the rate of adaptive evolution and positive selection detection in humans *vs*. other organisms are discussed.

Author Summary

The rate at which species adapt to environmental changes is a controversial topic. The theory predicts that adaptation is easier in large than in small populations, and the genomic studies of model organisms have revealed a much higher adaptive rate in large population-sized flies than in small population-sized humans and apes. Here we build and analyse a large data set of protein-coding sequences made of thousands of genes in 44 pairs of

species from various groups of animals including insects, molluscs, annelids, echinoderms, reptiles, birds, and mammals. Extending and improving existing data analysis methods, we show that adaptation is a major process in protein evolution across all phyla of animals: the proportion of amino-acid substitutions that occurred adaptively is above 50% in a majority of species, and reaches up to 90%. Our analysis does not confirm that population size, here approached through species genetic diversity and ecological traits, does influence the rate of adaptive molecular evolution, but points to human and apes as a special case, compared to other animals, in terms of adaptive genomic processes.

Introduction

Characterizing and quantifying adaptation at the molecular level is one of the major goals of evolutionary genomics. Recent research in this area has yielded a number of remarkable examples [1–4]. However, in spite of the increased size of molecular data sets and intensive methodological developments, we still lack a quantitative assessment of the relative impact of adaptive *vs*. nearly-neutral evolution at the genomic scale—an issue hotly debated ever since the origins of molecular evolutionary studies [5–7].

Particular attention has been paid to the adaptive evolution of proteins trough amino-acid substitution, following the seminal work of McDonald and Kreitman [8]. Analysing the *Adh* gene in *Drosophila*, these authors compared the ratio of the number of non-synonymous (potentially selected) to synonymous (supposedly neutral) substitutions between species, d_N/d_S, to the ratio of the number of non-synonymous to synonymous polymorphisms within species, p_N/p_S. They showed that the former ratio was higher than the latter, which was interpreted as resulting from adaptive evolution. This is because the rare, positively selected non-synonymous mutations contribute negligibly to the polymorphism and to p_N/p_S, but have a much higher fixation probability than neutral mutations, thus accumulating detectably over time as species diverge, and inflating d_N/d_S. In absence of adaptive evolution, d_N/d_S is expected to be equal to p_N/p_S (neutral model), or lower than p_N/p_S (nearly-neutral model [9]).

The idea that the excess of d_N/d_S over p_N/p_S is a signature of adaptive protein evolution was turned into methods for estimating the proportion of amino-acid substitutions that result from positive selection, a quantity called α [10–15]. These methods differ in the way information is combined across genes, and in the way they cope with the two main confounding factors that have been identified, i.e., (i) slightly deleterious mutations, which tend to inflate p_N/p_S while not affecting d_N/d_S [16], and (ii) variation in time of effective population size (N_e), which tends to decouple d_N/d_S from p_N/p_S irrespective of adaptive processes [17].

To account for the effect of segregating deleterious mutations, the most sophisticated McDonald-Kreitman-based methods model the deleterious effects of non-synonymous mutations by assuming some (typically Gamma or log-normal) distribution of selection coefficients. The model is fitted to polymorphism data, typically depicted by the synonymous and non-synonymous site-frequency spectra (SFS) [18,19]. Slightly deleterious mutations, which confound the McDonald-Kreitman signal, are expected to segregate at lower frequency, on average, than neutral mutations, thus distorting the non-synonymous SFS, compared to the synonymous one. SFS-based estimates of population genetic parameters are used to predict the expected divergence process under near-neutrality, and specifically the expected d_N/d_S ratio, which, compared to the observed one, leads to the estimation of α [15].

These methods have been applied to various species in which large numbers of protein-coding sequences were available from several individuals and a closely related outgroup. Interestingly, the resulting estimates of α varied considerably among species. For instance, the estimated α was close to zero in apes [14,20], giant Galapagos tortoise [21], yeast [22] and nine plant species [23], but above 0.5 in *Drosophila* [11], mouse [24], rabbit [25], sea squirt [26], sunflower [23] and enterobacteria [27]. According to these results, positive selection would be the prevalent mode of protein evolution in some species, but a minor one in others, including humans, in which the vast majority of amino-acid substitutions are predicted to be nearly neutral and fixed through genetic drift. Such a contrasted pattern among species is in itself intriguing, and deserves an explanation.

The main hypothesis that has been considered so far regarding across-species variation in the molecular adaptive rate invokes a population size effect. Several lines of evidence suggest that α is higher in large-N_e than in small-N_e species. The hypothesis first arose from the human *vs*. *Drosophila* comparison [15, 28], and was somewhat corroborated by the report of a low α in typical small-N_e species [21] and a high α in typical large-N_e ones [26]. In sunflower [29] and house mouse [30], the comparison of closely related species revealed a greater α in the most highly polymorphic species, which is consistent with the N_e hypothesis. Finally, a meta-analysis of 13 eukaryotic species revealed a significant—albeit not so strong—positive relationship between N_e and the rate of adaptive substitution [31]. Not all the empirical evidence, however, is consistent with the N_e hypothesis [23, 32], the most obvious counter-examples being provided by the yeast and maize proteomes, which, despite presumably very large N_e, are apparently dominated by non-adaptive processes.

Theoretically, there are mainly two reasons why large-N_e species are expected to experience a higher adaptive rate than small-N_e ones [28]. First, the population rate of appearance of new beneficial mutations is proportional to N_e, i.e., higher in large populations. Secondly, once they have appeared, beneficial mutations have a greater probability of fixation in large populations than in small populations, in which the random effect of genetic drift is stronger. These arguments have been debated [6,7], however, and two recent simulation studies using Fisher's geometric formalism recovered only a weak relationship between N_e and the adaptive rate [33, 34].

The determinants of the variation in α among species and the relevance of the effective population size hypothesis therefore remain contentious, in large part because we still lack a broad overview of the prevalence of adaptive processes in nature. First, the sample of species for which we have an estimate of α, which are mostly model organisms, is limited and disparate. Secondly, the set of analysed genes varied much between studies. For instance, in humans, more than 10,000 protein-coding sequences have been analysed [14], whereas the above mentioned plant study was based on a few dozens to a few hundreds of loci [23]. Finally, estimates of the adaptive rate have been obtained using distinct methods and models across studies, each making specific assumptions about the evolutionary processes at work.

Here, we analyse patterns of coding sequence polymorphism and divergence in the high-expressed genomic fraction of 44 taxonomically and ecologically diverse metazoan species characterized by a substantial variation in N_e [35]. A battery of models for the distribution of fitness effects of mutations (DFE) are introduced and assessed in the maximum likelihood framework, with particular emphasis on the modelling of beneficial mutations. The confounding effects of sample size, polymorphism orientation errors, gene expression level and GC-content are considered and controlled for. The analysis revealed a pervasive impact of adaptive processes on the evolution of coding sequences in the eight sampled animal phyla, but did not detect any influence of N_e on the rate of adaptive molecular evolution.

Results

Simulations

We re-implemented and extended a method of estimation of the adaptive rate introduced by Eyre-Walker and collaborators [15,19], in which the DFE is modelled as a continuous distribution. Two versions of this method have been proposed [15], which differ in the way demographic effects are accounted for. One is based on an explicit demographic model assuming a shift in N_e at some time point in the past [36]. We used the other version, which refrains from modelling the variation of N_e in time, and rather relies on "anonymous" parameters called r_i's [19]. There is one such parameter per allele frequency class, which multiplies both the synonymous and the non-synonymous expected number of single-nucleotide polymorphisms (SNPs). The r_i's are intended to capture a wide range of departures from the mutation/selection/drift equilibrium (see Methods).

We performed a series of simulations in order to assess the reliability of the newly-introduced methods. We first reproduced the conditions of a recent study [37], in which the effects of linked positive and negative selection on estimates of the adaptive rate was examined. Then we used a more complex model accounting for long-term fluctuations in N_e. In both cases, we found that our estimates of α were quite accurate and robust to a wide range of conditions, for simulated data sets of the size of the ones analysed in this study (see S1 Text for detailed results).

Data

Forty four metazoan species pairs were analysed (S1 Table). Each pair consisted of a focal species, in which at least four diploid individuals were available, and a closely-related outgroup species. The data set was based on two recent transcriptome-based population genomic reports [35, 38]. Five species were newly sequenced in this study (one individual each). The sample covered eight metazoan phyla: 18 pairs of vertebrates (of which twelve mammals), eleven arthropods (of which eight insects), five molluscs, four echinoderms, two tunicates, one annelid, one nematode, one nemertian and one cnidarian. Focal species sample size varied from four to eleven individuals, and was seven or more in 26 species. Ten of the twelve mammalian species had a sample size of four [38]. The number of analysed coding sequences varied between 295 (Black and white ruffed lemur *Varecia variegata variegata*) and 7951 (small skipper *Thymelicus sylvestris*), the median being 1900. The total number of SNPs varied between 315 (*Varecia variegata variegata*) and 62,553 (Essex skipper *Thymelicus lineola*), the median being 3241. The number of synonymous SNPs and the number non-synonymous SNPs were both above 100 in all the analysed species. The per site synonymous divergence between focal and outgroup species, d_S, varied between 0.7% (in penguins *Aptenodytes patagonicus* vs. *A. forsteri*) and 36% (in nematodes *Caenorhabditis brenneri* vs. *Caenorhabditis* sp.16).

DFE model comparison

In all pairs of species, a population genetic model was fitted to the synonymous and non-synonymous polymorphism (*i.e.*, SFS) and divergence data. Model parameters included population mutation rate θ, demography *sensu lato* $\mathbf{r}$, neutral divergence T, and the distribution of fitness effects of non-synonymous mutations (DFE). Six distinct parametrizations of the DFE were used. These included the classical Neutral[11] and Gamma [19] models, which only consider negative selection coefficients, and four newly introduced models explicitly accounting for beneficial mutations (see Methods). DFE models were compared using the Akaike Information Criterion (AIC) and likelihood ratio tests.

Among the 44 species pairs, the Neutral model obtained the highest AIC score in four cases, and the Gamma model in 13 cases. In the remaining 27 cases, *i.e.*, 60% of the data sets, a model accounting for slightly advantageous mutations best fitted the data. The GammaExpo model, which assumes a Gamma distribution of negative effects and an exponential distribution of positive effects, obtained the highest AIC score in 18 data sets. The ScaledBeta model, which uses a Beta-shaped distribution of mild negative and positive effects, obtained the highest AIC score in 8 data sets. The DisplacedGamma model [39] essentially converged towards Gamma, the displacement parameter always taking very low values. The BesselK model [40] sometimes yielded a likelihood similar to (but not above) GammaExpo, and sometimes performed badly —perhaps because of optimization problems. These two models were not considered any longer in this study. When we performed likelihood ratio tests between the nested Neutral, Gamma and GammaExpo models, we found that Gamma significantly rejected Neutral in 40 species, and GammaExpo significantly rejected Gamma in 19 species out of 44.

Consistent with published methods [15], the above-described approach implicitly assumes the existence of strongly advantageous non-synonymous mutations that contribute to divergence but negligibly affect polymorphism patterns (parameter A in eq (7), see Methods). Since the GammaExpo and ScaledBeta models account for the beneficial effect of non-synonymous mutations, we implemented a distinct version, the [-A] version, in which no additional, strongly adaptive class of mutation was assumed. This was done for the 26 data sets in which either GammaExpo or ScaledBeta was the model best supported by AIC. In 14 of these, the [-A] version was not significantly rejected by the classical one according to likelihood-ratio test. In these cases the new DFE models seem to appropriately account for both polymorphism patterns and divergence rates. In the remaining twelve data sets, the fit was significantly worse in the [-A] than in the classical version. Estimates of α were strongly correlated across species between the two analyses ($n = 44$, $r^2 = 0.79$, p-val$<10^{-10}$).

Adaptive substitution rate

The estimated population genetic parameters were used to predict the expected rate of non-adaptive non-synonymous substitutions relative to neutral, ω_{na}, deduce the estimated rate of adaptive ones, $\omega_a = d_N/d_S - \omega_{na}$, and the proportion of adaptive non-synonymous substitutions, $\alpha = \omega_a/(d_N/d_S)$ (see Methods). This was achieved using four distinct DFE models. Fig 1 plots the Neutral, GammaExpo and ScaledBeta estimates of α as a function of the Gamma estimate. The Neutral model provided estimates generally lower than, and loosely correlated with, the other three models. This confirms that accounting for the effect of weakly deleterious non-synonymous mutations makes a significant difference, as previously demonstrated [14–16]. The GammaExpo and ScaledBeta estimates of α, in contrast, were strongly correlated with the Gamma estimate, suggesting that the inclusion of positive effects in the assumed DFE does not impact much the estimation of the adaptive rate, even when a better fit to the data is achieved (Fig 1, closed circles). Below we focus on the estimate of α based on the GammaExpo model, but very similar results would be obtained with either the Gamma of ScaledBeta estimates. Maximum likelihood confidence intervals for α are provided in S1 Table.

The distribution of α among animal species was bimodal, with a peak around 0.3 and another one around 0.7, whereas the distribution of ω_a was unimodal (S1 Fig). No significant difference in α was detected between vertebrates (18 species) and invertebrates (26 species, Fig 2, top left). Within invertebrates, no difference was observed between arthropods (eleven species), molluscs (five species) and echinoderms (four species). Within vertebrates, however, a surprising significant difference was detected between mammals (twelve species, of which ten primates) and sauropsids (three bird and two turtle species), α being lower than in the average

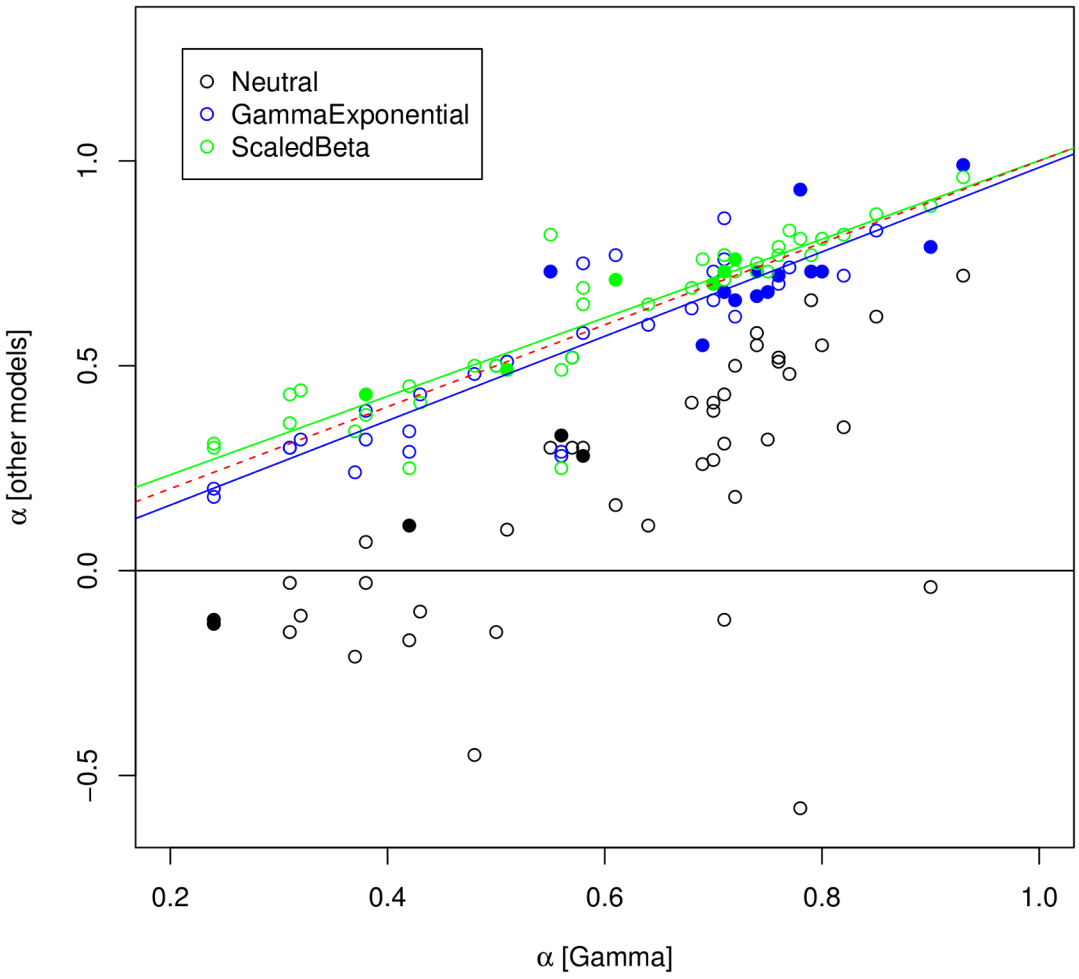

Fig 1. Effect of the underlying DFE model on estimators of α. Each dot is for a species pair. X-axis: estimated α under the Gamma model. Y-axis: estimated α under the Neutral (black), GammaExpo (blue) and ScaledBeta models (red). Dotted red line: (y = x) line. Closed circles indicate species pairs for which the considered model obtained the highest AIC.

doi:10.1371/journal.pgen.1005774.g001

metazoan in the former, but higher in the latter (Fig 2, bottom left). The mammalian pattern is mainly driven by primates: the two non-primates species pairs yielded relatively high estimates of α (Iberian hare *Lepus granatensis*: α = 0.77; common vole *Microtus arvalis*: α = 0.60). A more or less similar pattern was found regarding the taxonomic distribution of ω_a (Fig 2, right).

We correlated d_N/d_S, ω_{na}, ω_a, and α to species genetic diversity, π_S, here taken as a predictor of N_e, the effective population size. A significant negative relationship between ω_{na} and log-transformed π_S was obtained (Fig 3a), which is indicative of a higher efficiency of purifying selection against slightly deleterious mutations in large-N_e species [35]. The relationship between d_N/d_S and π_S was also a significantly negative one (Fig 3b), suggesting that the effect of N_e on the efficiency of purifying selection also affects the rate of non-synonymous substitution. Interestingly, the slopes of the ω_{na} vs. $\log(\pi_S)$ and d_N/d_S vs. $\log(\pi_S)$ relationships were quite similar, as we illustrated by reporting in Fig 3b the regression line of Fig 3a (dotted line). According to the McDonald-Kreitman rationale, the parallel lines indicate that the adaptive component of the non-synonymous rate, $\omega_a = d_N/d_S - \omega_{na}$, is unrelated to N_e, as shown by Fig 3c. The proportion of adaptive amino-acid substitutions, finally, was significantly and positively related to

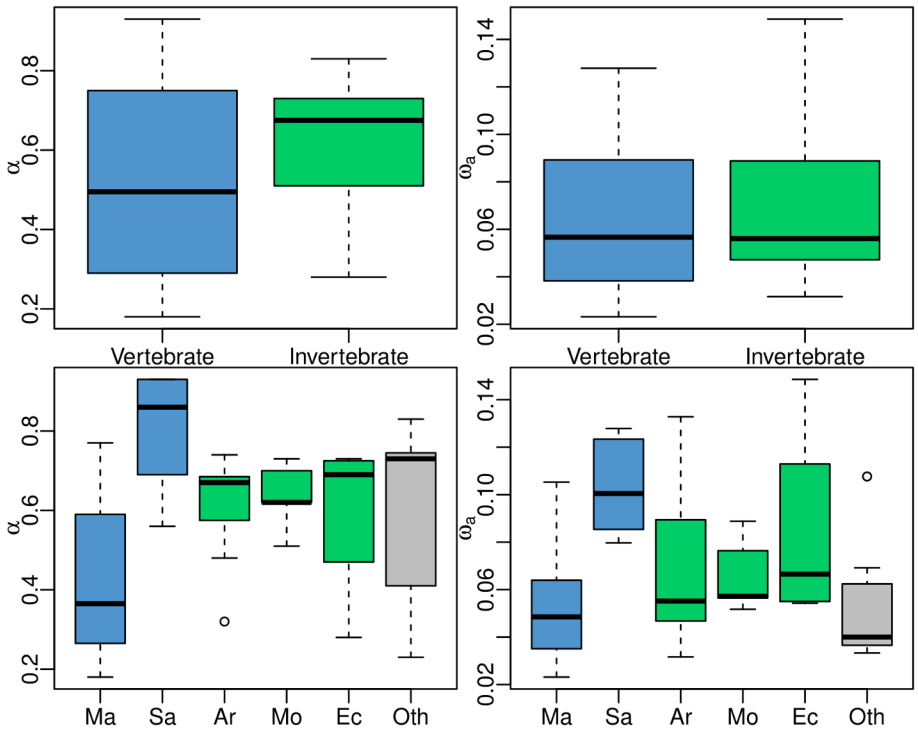

Fig 2. Distributions of α and ω_a split by taxa. Ma: mammals; Sa: sauropsids; Ar: arthropods; Mo: molluscs; Ec: echinoderms; Ot: others

doi:10.1371/journal.pgen.1005774.g002

log-transformed π_S (Fig 3d), consistent with the existing literature. However, Fig 3a–3c suggests that the variation of α among species is primarily driven by the fixation rate of slightly deleterious changes. α tends to be lower in small-N_e than in large-N_e species because non-adaptive substitutions accumulate at a faster rate in the former, thus decreasing the proportion of adaptive ones. The correlation coefficients and *p*-values corresponding to the four regression analyses shown in Fig 3 are provided in Table 1 (analysis A).

In Fig 3a, 3c and 3d the X-axis, π_S, is not independent of the Y-axis—estimators of α, ω_a and ω_{na} depend on the synonymous diversity. We therefore re-conducted the same analysis using other predictors of N_e, namely species maximal longevity, and size of propagule—propagule being defined as the stage that leaves the mother. These two life-history traits were identified as strong negative correlates of N_e in animals [35]. We obtained results very similar to those shown in Fig 3, *i.e.*, a significant correlation with these two traits for d_N/d_S and ω_{na}, but not for ω_a, α being in this case only marginally significant (S2 and S3 Figs). The proportion of adaptive non-synonymous substitutions tends to be higher in short-lived, small propagule-sized species, but this is due in the first place to a faster accumulation of non-adaptive changes in long-lived, large propagule-sized species. None of the other life-history traits considered in ref. [35] (body mass, fecundity, size of geographic range) did provide additional explanation to the across-species variation in α and ω_a.

Effect of sample size

Our dataset is heterogeneous across species in terms of number of analyzed individuals. We investigated the influence of sample size on our results in two ways. We first re-conducted the above analyses using only the 25 species for which at least eight individuals were available.

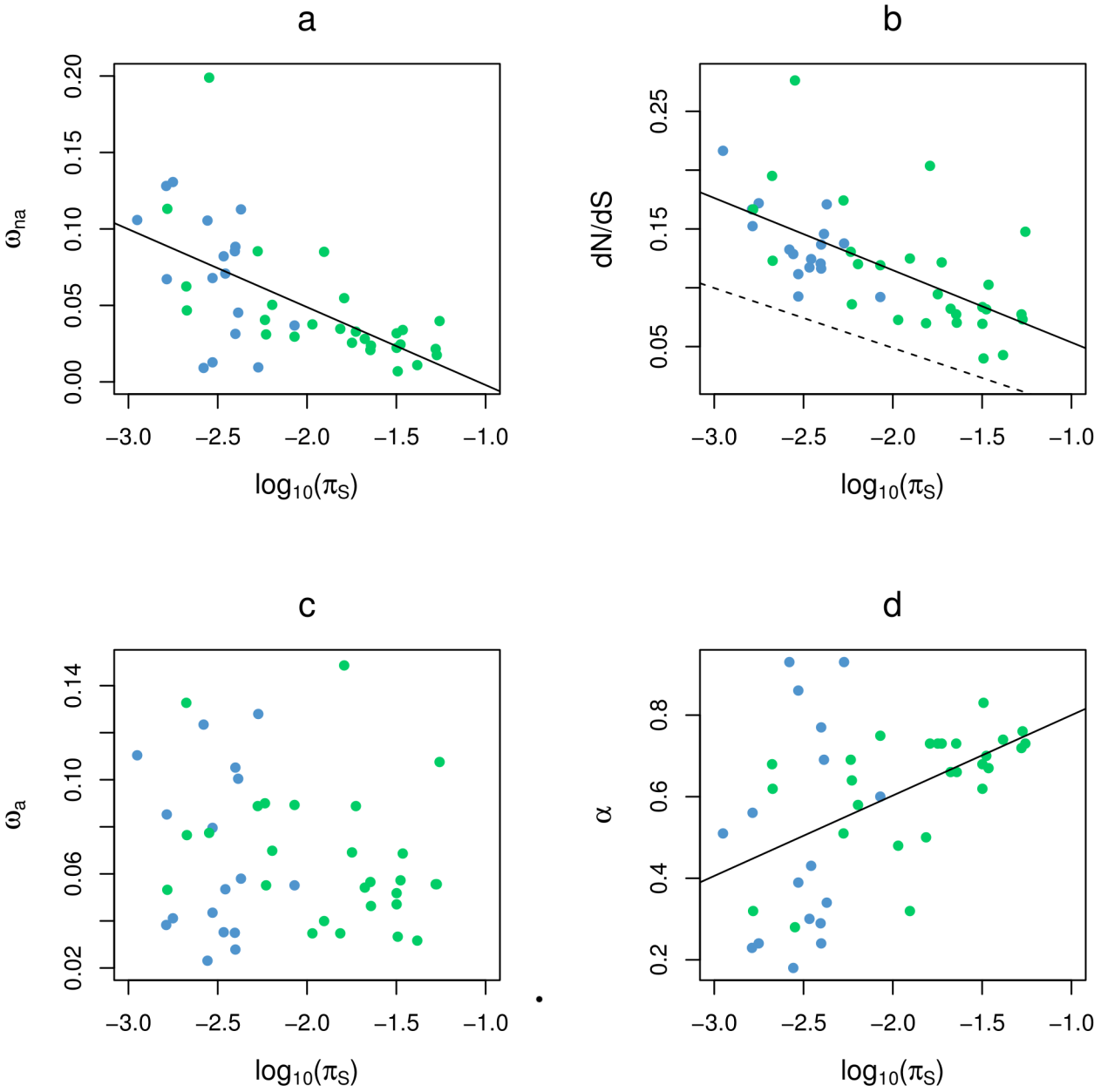

Fig 3. Effect of species neutral genetic diversity on protein divergence parameters. Each dot is for one species pair. Blue: vertebrates; green: invertebrates. a: non-adaptive rate; b: non-synonymous/synonymous ratio; c: adaptive rate; d: proportion of adaptive non-synonymous substitutions. X-axis: focal species synonymous heterozygosity.

doi:10.1371/journal.pgen.1005774.g003

Similar to the full dataset analysis, we found a significantly negative relationship between d_N/d_S and log-transformed π_S, and between ω_{na} and log-transformed π_S. No significant correlation, however, was detected between α and π_S with the large sample-size dataset, and ω_a was negatively related to π_S (Table 1, analysis B)–an intriguing result that contradicts the N_e hypothesis. It should be noted that taxon sampling is more evenly balanced amongst animal phyla in this 25 species data set, which includes only two mammalian species, than in the complete data set, which includes twelve mammals. In a second analysis, we analysed the whole set of species but used sub-samples of exactly four individuals in each species—only the 40 species for which the

Table 1. Genetic diversity vs. protein divergence: control analyses.

analysis	sp.[a]	ind.[b]	α vs. π_S [c]	d_N/d_S vs. π_S	ω_{na} vs. π_S	ω_a vs. π_S
A: main	44	7.2	0.48; $<10^{-3}$	-0.64; $<10^{-5}$	-0.62; $<10^{-5}$	-0.17; NS
B: >6 individuals	26	9.2	0.20; NS	-0.70; $<10^{-4}$	-0.48; 0.012	-0.60; $<10^{-2}$
C: subsampled 4	40	4	0.49; $<10^{-2}$	-0.57; $<10–^{3}$	-0.61; $<10^{-4}$	-0.10; NS
D: folded SFS	26	9.2	0.11; NS	-0.70; 10^{-4}	-0.49; 0.012	-0.61; $<10^{-2}$
E: d_S control	21	7.3	0.76; $<10^{-4}$	-0.63; $<10^{-2}$	-0.80; $<10^{-5}$	0.17; NS
F: GC-poor	18	8.8	0.46; NS	-0.50; 0.03	-0.65; $<10^{-2}$	-0.23; NS
G: GC-medium	21	9.0	0.32; NS	-0.65; $<10^{-2}$	-0.69; $<10^{-3}$	-0.49; 0.024
H: GC-rich	21	9.0	0.32; NS	-0.65; $<10^{-2}$	-0.55; <0.011	-0.51; 0.018
I: low-expressed	23	9.0	0.17; NS	-0.62; $<10^{-2}$	-0.48; <0.021	-0.49; 0.018
J: high-expressed	25	9.1	0.18; NS	-0.69; $<10^{-3}$	-0.50; 0.012	-0.56; $<10^{-2}$
K: whole body	22	8.4	0.57; $<10^{-2}$	-0.69; $<10^{-3}$	-0.79; $<10^{-4}$	-0.38; NS
L: no mirror	36	7.6	0.44; $<10^{-2}$	-0.63; $<10^{-4}$	-0.59; $<10^{-4}$	-0.25; NS

[a] number of analysed species

[b] mean number of individuals per species

[c] [correlation coefficient; *p*-value];

NS: not significant

doi:10.1371/journal.pgen.1005774.t001

number of non-synonymous SNPs was above 100 were retained. The results were again strongly consistent with the main analysis, suggesting that the heterogeneity of sample size is not a major issue in this analysis (Table 1, analysis C).

SNP mis-orientation

In the above analyses we used so-called unfolded SFS, in which the derived and ancestral alleles are defined based on the allelic state of the outgroup. Orientation errors can deeply affect the folded SFS in that mis-oriented low-frequency mutations are mistaken for high-frequency ones. This problem, however, is alleviated in the current approach by the "demographic" r_i's parameters, which are expected to capture any departure from the expected SFS as soon as it is shared by synonymous and non-synonymous sites. Still mis-orientation could be problematic if it did not affect synonymous and non-synonymous sites to the same extent. To control for this potential bias we built folded SFS, in which SNPs in frequency k and $2n$-k are merged in a single bin, n being the diploid sample size. Only the 25 species in which sample size was seven individuals or more were considered here. Estimates of α were significantly correlated between the two analyses ($n = 26$, $r^2 = 0.28$, p-val$<10^{-2}$), but there was a couple of outliers in which the folded and unfolded SFS yielded quite different estimates of α–e.g., 0.78 (folded) *vs*. 0.32 (unfolded) in the yellow gorgonian *Eunicella cavolinii*. Regarding the relationship to π_S, however, the results with folded SFS were essentially similar to those obtained with this set of species and unfolded SFS (Table 1, analysis D). We also restricted the analysis to a subset of 21 species in which d_S was between 0.05 and 0.25 substitution per synonymous site, thus avoiding biases resulting from a too high, or a too low, level of divergence. Again, this control yielded results qualitatively similar to the main analysis (Table 1, analysis E).

GC-content

For each species pair, we split the data set in three bins of equal size according to coding sequence third codon position GC-content (GC3) and re-estimated α separately for GC-poor, GC-median and GC-rich genes. Within each category of genes, the relationships between α,

d_N/d_S, ω_a, ω_{na} and π_S shown in Fig 3 were recovered (Table 1). Considering only the species in which sample size was seven of more and above 100 non-synonymous SNPs were available, we found a slight, positive effect of GC3 on α (Fig 4). This might reflect the effect on the efficacy of positive selection of local recombination rate, which alleviates Hill-Robertson effects by breaking genetic linkage—GC3 is associated to local recombination rate in a number of animal species [41,42]. Alternatively, a higher α in high-GC regions might reflect the confounding effect of GC-biased gene conversion, a recombination-associated molecular drive that mimics the effect of directional selection [43].

Gene expression level

For each species pair, we split the data set in two bins according to gene expression level and re-estimated α separately for low-expressed and high-expressed genes. In this case bins were unequal in size: the low-expressed bin included 90% of the genes, and the high-expressed bin 10% of the genes. This is because there is a strong positive correlation between ORF length, genotyping success and expression level in this transcriptome-based data set—high-expressed genes are more efficiently assembled and genotyped—so that using bins of equal size would result in a strong discrepancy in terms of number of SNPs. Consistent with the existing literature [44], we detected a significant effect of expression level on the ratio of non-synonymous to synonymous changes: the across-species average π_N/π_S of high-expressed genes (0.084) was lower than that (0.096) of low-expressed genes (p-val<0.001, $n = 44$, Student's t-test for paired samples). However, no difference in α was detected between the two categories of gene expression (Fig 4), which, considered separately, yielded results similar to those of Fig 3 (Table 1).

Tissue specificity

The sets of analysed genes varied among species, particularly because not the same tissues were used in distinct species for RNA extraction—e.g., blood in reptiles, liver in mammals, intestine in urchins. This might affect the comparison if the prevalence of positive selection was higher in genes expressed in specific tissues. To control for this effect, we analysed the subset of 22 species in which whole-body RNA extraction had been performed. This includes insects, molluscs, annelids, tunicates, nematodes, cnidarians, some echinoderms and some crustaceans. Vertebrates were excluded from the sub-sample, but the range of π_S was unaffected. The analysis provided results highly similar to the main one (Table 1, analysis K).

"Mirror" species

Our data set included eight pairs of "mirror" species, in which species 1 served as outgroup for species 2, and reciprocally (S1 Table). First, when we randomly selected a single species per mirror species pairs, reducing the data set to 36 species, results were essentially unchanged (Table 1, analysis L). Mirror species, on the other hand, are useful because they offer biological replicates: the two analyses (species 1 outgroup to species 2 *vs*. species 2 outgroup to species 1) concern the same history of divergence, and therefore estimate the same biological quantities.

To get a hint about the robustness of estimates of α (and ω_a), we compared their values between mirror species. In four mirror pairs out of eight the difference in α was below 0.1 between the two mirror analyses, but in three mirror pairs the difference reached values above 0.15 –and up to 0.26 in the *Ciona intestinalis A* / *Ciona intestinalis B* pair. In neither of the eight mirror pairs was the α estimate of species 1 within the confidence interval of α in species 2, and reciprocally. This indicates that the ML confidence intervals (S1 Table), which implicitly trust the assumptions of the model, are unrealistically narrow. Interestingly, in six mirror pairs

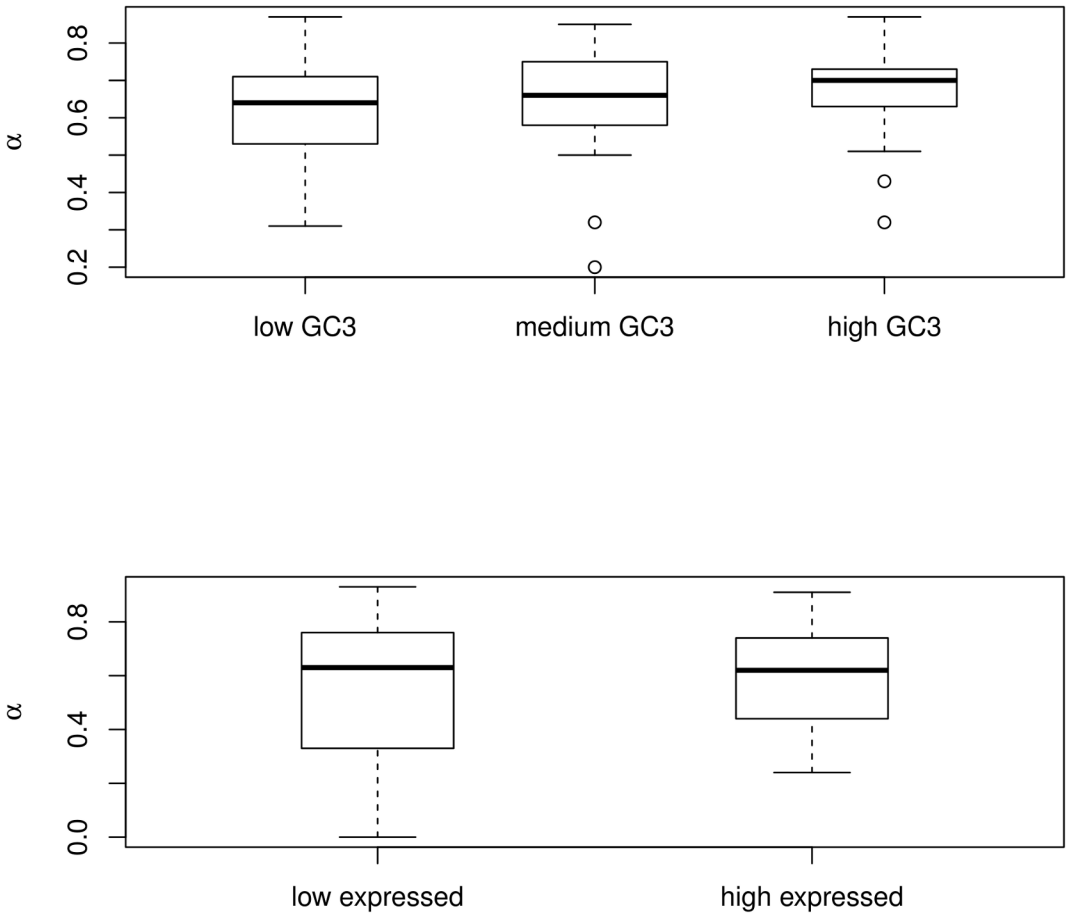

Fig 4. Effect of gene GC-content and expression level on estimates of α. Only the 26 species for which at least seven individuals have been sampled are considered here.

doi:10.1371/journal.pgen.1005774.g004

out of eight, the higher estimate of α was obtained in the more polymorphic of the two mirror species (S4 Fig).

Characterizing the DFE

Besides α and ω_a, our model-fitting procedure also provides estimates of parameters characterizing the DFE. When the Gamma model was used, the shape parameter of the negative Gamma distribution of selection coefficients (call it β) varied greatly among species—from 0.009 to its maximum allowed value 100.0 –and took a median value of 0.23. When the GammaExponential model was used, the median β was higher and equal to 0.46. With Gamma distributed negative selection coefficients the theory predicts a log-linear relationship of slope 1- β between π_N and π_S [45,46]. The linear regression of $\log(\pi_N)$ on $\log(\pi_S)$ across species yielded an estimated slope of 0.62 and an estimated β of 0.38, in reasonable agreement with species-specific, SFS-based estimates. These results indicate that the data examined in this study are broadly consistent with a negative Gamma distribution of shape in the 0.2–0.4 range for selection coefficients applying to non-synonymous mutations, even though single species data can be too noisy to provide reliable estimates of the DFE.

Discussion

The comparative approach offers a unique opportunity to relate species traits to population genomics, and particularly investigate the effect of selection and drift on molecular evolutionary processes [35, 47]. In this study, we extended existing McDonald-Kreitman-based approaches to estimate the rate of adaptive amino-acid substitution from coding sequence polymorphism and divergence analysis. We applied the new methods to the high-expressed genomic fraction of 44 metazoan species pairs, thus offering for the first time a general picture of the distribution of the adaptive rate across taxa in this group.

Enhanced McDonald-Kreitman modelling

Various population genomic models involving distinct shapes for the DFE were implemented in the maximum likelihood framework and assessed through simulations (S1 Text). Including a continuous distribution of positive effects in the model significantly improved the fit in a majority of data sets. In roughly half of these, this addition was sufficient to correctly fit both the polymorphism and divergence data. In the other half, d_N/d_S was still in excess compared to model predictions, thus requiring an extra class of arbitrarily high selection coefficients, similarly to classical methods [15]. It should be noted that the proportion of adaptive non-synonymous mutations and the average strength of adaptive effects could not be efficiently disentangled—the two parameters were strongly, negatively correlated. Similar results were reported in a previous attempt to incorporate adaptive effects in the DFE [48]–in this case, a single, discrete category of positive selection coefficients was assumed. Among the newly introduced DFE models, GammaExpo and ScaledBeta were the ones best fitting the data. The two DFE models that were suggested from theoretical studies based on Fisher's geometric formalism did not perform well in this analysis. Importantly, the new approaches did not make a strong difference with respect to estimates of α and ω_a, compared to existing ones.

In this study, departure of the synonymous SFS from the neutral, constant-N_e expectation was accounted for thanks to the "anonymous" r_i 's parameters [19]. This is biologically less meaningful but statistically more flexible than explicit models of population size changes [36]. The r_i 's parameters are intended to capture the effects of a variety of potential confounding factors, such as demographic changes, population structure, strong selection at linked loci, genotyping error or SNP misorientation—provided these factors affect synonymous and non-synonymous SNPs to the same extent (S1 Text). This methodological choice was motivated by a recent simulation study [37], which showed that linked selection can distort the synonymous and non-synonymous SFS in a way that confounds explicit models of demographic changes.

Pervasive adaptive protein evolution in animals

The analysis of 44 species paired revealed a prominent effect of adaptation on amino-acid sequence evolution in animals. Our estimate of α, the proportion of adaptive amino-acid substitutions, was above 0.5 in a majority of the analysed species and taxa—with the notable exception of primates (Fig 2). Observing a high α therefore seems to be the rule, and a low α the exception, making humans and apes a peculiar, intriguing case. The report of prevalent adaptive protein evolution in animals contrasts with the plant situation, in which eight out of nine analysed species pairs showed little evidence for positive selection [23]. Identifying the environmental and genomic drivers of protein sequence adaptation in animals and understanding the causes of the difference with plants would be two exciting goals to be pursued.

An exciting goal of current population genomics is to uncover the targets of positive selection. Our results suggest that the challenge is even stronger in humans and apes than in most animal species. The confounding effect of neutral and slightly deleterious mutations, which are

prevalent in primates, is likely to affect not only d_N/d_S-based studies, but also analyses of local SNP-density [49] and of Fst-outlying loci [50]. The focus of this study is on point mutations in coding sequences, but the results are expected to be general to a wider range of mutations, including structural and regulatory changes, unless the DFE associated to these mutation types differ substantially from that of amino-acid sequence changes [51].

No effect of N_e on the adaptive rate?

A significant, positive effect of various predictors of N_e on α was detected, in broad agreement with the existing literature. This relationship, however, seems to be primarily explained by deleterious effects. Being more strongly affected by genetic drift, small-N_e species accumulate slightly deleterious non-synonymous substitutions at a higher rate than large-N_e ones, in agreement with the nearly neutral theory of molecular evolution ([52,53]). This mechanically decreases the proportion of adaptive non-synonymous substitutions. No positive relationship was found, however, between ω_a, the (relative to neutral) rate of adaptive amino-acid substitution, and N_e. This result, which was robust to controls for DFE modelling, sample size, divergence time, SNP orientation errors, GC-content, and gene expression level, does not corroborate a recent report of a weak but significant effect of N_e on ω_a across 13 species from four eukaryotic groups [31], but confirms earlier suggestions that the long-term adaptive rate does not always respond to changes in N_e [32].

Our results are consistent with theoretical predictions recently obtained by simulations under Fisher's geometric model [34], in which α was only weakly positively related to N_e, and ω_a independent from N_e. A plausible reason for these perhaps counter-intuitive results is that proteins in small-N_e species tend to be less adapted (*i.e.*, further away from their fitness optimum), and therefore more prone to adaptation, than in large-N_e species. The accumulation of slightly deleterious amino-acid substitutions in small-N_e species would open the opportunity for compensatory, adaptive ones—at least under the hypothesis of a single-peaked fitness landscape [34], and assuming that mutation does not limit adaptation [54]. A positive relationship between ω_a and N_e would be expected if the DFE was independent of N_e, *i.e.*, if the adaptive mutation rate was similar in small and large populations. Our study, if confirmed, would suggest that this hypothesis does not hold, at least in animals: adaptive mutations could be more common in small-N_e species, and this would compensate for their lower fixation probability [55,56].

The lack of a strong positive relationship between ω_a and N_e does not contradict the prediction that adaptive walks are more efficient, in terms of rate of fitness increase, in large than in small populations [57]. This is because McDonald-Kreitman-related methods estimate the number of adaptive steps, not the associated fitness effect. To our best knowledge, no analytical result is available regarding the relationship between the number of steps in adaptive walks and N_e. Simulations under Fisher's geometric model and a fluctuating environment, however, seem to indicate that the length of adaptive walks is essentially independent of N_e, at least for the range of parameters considered in references [33] and [34]. In these simulations large populations adapt more efficiently than small populations by making larger steps (on average), not more steps.

Not detecting a positive relationship between ω_a and N_e does not demonstrate the absence of such a relationship. The McDonald-Kreitman method is known to be sensitive to various confounding effects, and our data could be just too noisy (see below methodological discussion). However, what our analysis reveals is a significantly negative relationship between d_N/d_S and N_e (Fig 3b). This indicates that N_e primarily affects the rate of amino-acid sequence

evolution through its effect on slightly deleterious mutations. The effect of N_e on the adaptive rate, if any, must be of second order—or d_N/d_S would correlate positively with N_e.

In this study, like in previous ones [23, 29, 30], the synonymous diversity, π_S, was used as a proxy for N_e. This can be problematic for several reasons. First, π_S is influenced not only by N_e but also by the mutation rate. This can in principle be addressed by incorporating direct or indirect estimates of the per generation mutation rate in the analysis [31]. Unfortunately, data on generation time and divergence time are lacking for many of the species analysed here. The strong relationships between π_S and ω_{na} and between π_S and d_N/d_S, however, suggest that N_e is the major determinant of π_S, since ω_{na} and are d_N/d_S independent of the mutation rate (and see [35]). Secondly, some of the species analysed here are highly polymorphic (*e.g. C. brenneri*), and the standard formula for π_S does not hold in such extreme cases [58]. Fig 3, however, does not reveal any particular effect of the most highly polymorphic species on the main results of this analysis. Thirdly, synonymous sites are not always neutrally evolving and might be affected by selection on codon usage. The robustness of our results to variation in GC-content, which is typically correlated to codon usage bias, does not suggest that this is a major issue with this analysis. Furthermore, we recall that the selective pressure applying to synonymous positions is likely to affect the non-synonymous ones as well. The McDonald-Kreitman method, and other methods based on the comparison between the non-synonymous and synonymous rates, implicitly focus on the selective effects applying at the protein level, not at the DNA or RNA level.

What determines the between-species variation of ω_a?

The evolutionary genomic literature in animals has long been dominated by studies in humans (and apes) and *Drosophila*. Quite plausibly, the patterns that distinguish these two groups have often been associated to variation in N_e. Our analysis of a taxonomically extended data set suggests that this is not obviously the right explanation as far as the adaptive rate is concerned, reopening the question of the determination of the variation of the adaptive rate across species. We report a wide range of estimated α (minimum: 0.18; maximum: 0.93) and ω_a (minimum: 0.023; maximum: 0.15) among animals, without obvious connection to species biology or taxonomy. Several hypotheses can be considered.

First, it might well be that distinct species do not face the same rate of environmental change. Unfortunately, measuring or predicting the rate at which abiotic and biotic environmental fluctuations affect the fitness landscape of distinct species appears essentially impossible, so that this hypothesis appears difficult to test empirically. Secondly, a part of the variance of α and ω_a might be explained by the fact that different sets and numbers of genes have been analysed among species—even though the data set analysed here is arguably more homogeneous than in previous meta-analyses. Gene expression levels vary substantially among metazoan phyla, and, depending on the species, RNA was extracted from either the whole body or specific tissues [35], so that distinct sets of genes were captured. The prevalence of adaptive evolution is known to vary across gene sets and functions, with a enrichment among adaptive loci of genes associated to the immune response [49, 59]. Results were unchanged, however, when we restricted the analysis to species in which whole-body RNA extraction was performed. Clarifying the influence of gene function on estimates of α and ω_a would be a promising continuation of this work; this should preferably be achieved based on whole proteomes in model organisms, in which functional annotations are available.

Finally, the dispersion of α and ω_a estimates among species could in part result from methodological issues. First, the data are somewhat noisy, as suggested by the wide variation across species in estimated shape of the DFE, and by the higher variance in estimated α among low-

diversity species, compared to high-diversity ones (Fig 3d). Secondly, it might be that not all the sources of biological variation are appropriately modelled in the McDonald-Kreitman framework, as previously suggested [60, 61]. This includes local adaptation, balancing selection and other sources of protected selected polymorphisms, which if common might push π_N/π_S above its nearly-neutral expectation. Another question is whether the prevalence of slightly deleterious mutations, which might pose methodological problems, is related to N_e. Of note, the ratio of slightly deleterious to neutral mutations is expected to be essentially independent of N_e under Gamma-distributed DFE—the Gamma distribution converges to a power law in the neighbourhood of zero, and power laws have the property of scale invariance. However, it is currently uncertain whether the DFE of non-synonymous mutations is truly Gamma-distributed.

Illustrative of shortcomings of the model, the likelihood-based confidence intervals we obtained are implausibly narrow, compared to the variation resulting from slight modifications of the data analysis strategy—for instance, switching from unfolded to folded SFS substantially impacted the estimate of α in some species. One particularly critical assumption is the time-constancy of N_e during the whole period of divergence. Our model accounts from recent demographic changes through the r_i 's parameters, but events older than the coalescence time of the sampled individuals, or having appeared in the outgroup lineage, might have influenced the d_N/d_S ratio while being undetectable from current polymorphism patterns. A former bottleneck, for example, might have inflated d_N/d_S through the accumulation of slightly deleterious non-synonymous substitutions in a way indistinguishable from truly adaptive evolution. Our simulations, however, suggest that the approach is sufficiently robust to detect large differences in α and ω_a between species in case of fluctuating N_e (S1 Text).

Polymorphism data only provide information on current, or very recent, strength of purifying selection, whereas d_N/d_S is influenced by the long-term average. Our analysis of "mirror" species, in which distinctive estimates of α were obtained depending on which species was taken as the focal one, highlights limitations of the McDonald-Kreitman approach when based on polymorphism data from a single species. The advent of genome-wide polymorphism data in multiple, closely related species should help disentangling the effects of demography *vs*. positive selection on the evolution of amino-acid sequences.

Methods

Data sets

The data set was built based on two recent population genomic studies of, respectively, 16 mammalian [38] and 76 metazoan [35] species. In these two contributions, two to eleven individuals per species were sampled from distinct populations and subjected to massive mRNA sequencing. From this sample we identified 35 focal species for which (i) at least four individuals had been sequenced, (ii) a closely-related outgroup was available, and (iii) at least 100 synonymous SNPs and 100 non-synonymous SNPs had been called. We added nine outgroup species (one individual each), of which four were already published [21,62,63] and five were newly sequenced here, increasing the total sample size to 44 species pairs (species list and SRA IDs in S1 Table). Not all these species pairs are fully independent from each other: when four individuals or more were available in each of two closely species, two species pairs were defined, hereafter referred to as "mirror species pairs", each species serving as outgroup to the other one. A restricted data set of 36 species pairs was created by randomly keeping just one mirror species pair out of two.

The newly-sampled individuals were subjected to RNA extraction, library preparation and illumina sequencing as previously described [64]. In each of the analysed species, sequence

reads were assembled to predicted cDNAs and open reading frames (ORFs), then SNPs and diploid genotypes were called the same way as in [35], according to the methods described in previous papers of ours [26,62,65]. For each species pair, orthologous coding sequences between the focal and outgroup species were predicted by reciprocal best-BLAST hit, aligned using MACSE [66] and cleaned as previously described [63].

For each species pair, the numbers of fixed synonymous, D_S, and fixed non-synonymous, D_N, differences were computed. In each focal species, the synonymous, $\mathbf{P_S}$, and the non-synonymous, $\mathbf{P_N}$, unfolded SFS were built by computing the observed frequency of derived alleles, and summing counts across genes. An unfolded SFS is a vector of $2n$-1 entries corresponding to the counts of SNPs at which the absolute frequency of the derived allele is 1, 2, . . ., $2n$-1, respectively, where n is the sample number of diploid individuals. SNP orientation was deduced from the outgroup state. SNPs for which the outgroup was polymorphic or distinct from any of the focal species alleles were discarded, as well as tri-allelic and tetra-allelic SNPs. For each species pair, the data set consisted of $4n$ data points: D_N, D_S, and the $2n$-1 entries of each of $\mathbf{P_N}$ and $\mathbf{P_S}$. Capital letters are used here to denote counts, lower case d_N, d_S, $\mathbf{p_N}$ and $\mathbf{p_S}$ being used for rates–*i.e.*, counts divided by site numbers.

Not every position of every individual was genotyped, so that sample sized varied among SNPs, complicating the calculation of the SFS. To cope with missing data we arbitrarily defined n, the minimal required number of genotyped diploid individuals, and applied a hypergeometric projection of the observed SFS into a subsample of size n, SNPs sampled in less than n individuals being discarded [67]. We used $n = 4$ when total sample size was four or five individuals, $n = 5$ when total sample size was six individuals, $n = 6$ when total sample size was seven or eight individuals, and $n = 7$ when total sample size was above eight individuals.

Population genetic model

We used a modified version of Eyre-Walker's approach [15,19] to fit a population genetic model to the data. Consider a focal species of population size N that diverged t generations ago from its sister species (outgroup). Mutations occur at rate μ per site per generation and drift applies according to the Wright-Fisher model under panmixy. Synonymous mutations are assumed to be neutral, whereas non-synonymous mutations experience protein-level codominant selection with constant in time selection coefficient s against heterozygotes, s being drawn from distribution $\phi(s)$–the DFE. The expected number of synonymous and non-synonymous SNPs at derived allele frequency i,$\widehat{P_S}[i]$ and $\widehat{P_N}[i]$, can be expressed as:

$$\widehat{P_S}[i] = 4L_S N\mu/i \tag{1}$$

$$\widehat{P_N}[i] = 2L_N N\mu \int_{-1}^{1} \phi(s) g(N, s, i, n) ds \tag{2}$$

where L_S and L_N are the sampled numbers of synonymous and non-synonymous sites, respectively.

In eq (2), g(N, s, i, n) is the probability that a mutation of selection coefficient s segregates at observed frequency i in a sample of size $2n$. The latter can be expressed by integrating on allele frequency, x:

$$g(N, s, i, n) = \int_{0}^{1} h(N, s, x) q(i, n, x) dx \tag{3}$$

where h(N, s, x) is the relative sojourn time at frequency x of a mutation of selection coefficient

s, and q(*i*, *n*, *x*) is the binomial probability that an allele at population frequency *x* is observed at frequency *i* in a sample of size 2*n*.

$$h(N,s,x) = \frac{2(1-e^{4Ns(1-x)})}{x(1-x)(1-e^{-4Ns})} \quad (4)$$

$$q(i,n,x) = \frac{2n!}{i!(2n-i)!}x^i(1-x)^{2n-i} \quad (5)$$

Departure from the Wright-Fisher assumption

The shape of observed SFS's typically differ from the expectations of eqs (1) and (2). There are many potential reasons for this, including population sub-structure, variation in time of *N*, and interference between linked selected mutations. Following [19], we used a generalized version of eqs (1) and (2):

$$\widehat{P_S}[i] = 4L_S N_e \mu r_i / i \quad (6)$$

$$\widehat{P_N}[i] = 2L_N N_e \mu r_i \int_{-1}^{1} \phi(s) g(N_e, s, i, n) ds \quad (7)$$

in which $r_1 = 1$ and $\{r_2, r_3, \ldots, r_{2n-1}\}$ are intended to capture any effect, typically demographic, affecting both the synonymous and non-synonymous SFS. Equating all the r_i's coefficients to unity would model the case of a constant-size Wright-Fisher population. The population size parameter in eqs (6) and (7) is now called "effective" and denoted by N_e. It represents the size of a Wright-Fisher population that would confer the same amount of distortion between non-synonymous and synonymous SFS as observed in current data. Accordingly, *N* should be replaced with N_e in eqs (4) and (5) as well.

Divergence model

The expected number of synonymous and non-synonymous substitutions, $\widehat{D_S}$ and $\widehat{D_N}$, were expressed as:

$$\widehat{D_S} = L_S \mu t \quad (8)$$

$$\widehat{D_N} = 2L_N N_e \mu t \int_{-1}^{1} \phi(s) f(N_e, s) ds + A \quad (9)$$

The first term in eq (9) corresponds to the expected number of non-synonymous substitutions given μ, *t*, N_e, *s*, and ϕ. $f(N_e, s) = 2s/(1-\exp(-4 N_e s))$ is the fixation probability of a mutation of selection coefficient *s* in a population of size N_e. The additional parameter *A* accounts for the excess of non-synonymous substitutions under strong positive selection. Introducing parameter *A* here is equivalent to adding to the DFE a class of mutations of arbitrarily high selection coefficient that contribute to divergence but negligibly affect polymorphism patterns.

Not all model parameters are identifiable since N_e, μ, *s* and *t*, which appear as products in the above equations, capture just three degrees of freedom. We used the following re-

parametrization:

$$\theta = 4N_e\mu \tag{10}$$

$$T = \mu t \tag{11}$$

$$S = 4N_e s \text{ and } \Phi(S) = \phi(s) \tag{12}$$

DFE models

Several functions were used to model the distribution of fitness effects of non-synonymous mutations. The Neutral model considers a discrete DFE, with a class of neutral mutations and a class of strongly deleterious mutations, which do not contribute to polymorphism or divergence [8,11]. The Gamma model, in contrast, assumes the existence of weakly deleterious non-synonymous mutations, modelled as a continuous, negative Gamma distribution [19].

Besides these classical DFE, we investigated models assuming a continuous distribution of positive effects. The GammaExpo model builds upon Gamma by adding a proportion of weakly advantageous mutations, assumed to be exponentially distributed. The Displaced-Gamma model assumes a displaced negative Gamma distribution [39], and the BesselK model implements the DFE obtained by Lourenco et al (their eq 8) [40]. These two DFE derive from two distinct versions of Fisher's geometric model. Finally, the ScaledBeta model is similar to the Neutral model in including a class of strongly deleterious mutations, but assumes a Beta-shaped distribution of weak-effect mutations:

$$\Phi(S) = Beta\left(\frac{S}{2S_{max}} + 0.5\right),\ S \in [-S_{max}, S_{max}] \tag{13}$$

Here $S_{\max}$ was arbitrarily set to 25. The Beta distribution has two parameters and can take many different shapes—monotonously increasing, monotonously decreasing, flat, unimodal, U-like.

Maximum likelihood estimation and the adaptive amino-acid substitution rate

Parameters θ, Φ, **r**, T and *A* were fit to the data using the maximum likelihood method. The likelihood was calculated assuming a Poisson distribution of polymorphism and substitution counts:

$$Prob(P_N, P_S, D_N, D_S/\theta, \Phi, r, T, A) = \gamma(D_N, \widehat{D_N})\gamma(D_S, \widehat{D_S})\prod_{i=1}^{2n-1}\gamma(P_N[i], \widehat{P_N}[i])\gamma(P_S[i], \widehat{P_S}[i]) \tag{14}$$

where $\gamma(a, b) = \frac{e^{-b}b^a}{a!}$.

Based on these parameter estimates we calculated D_N^{na}, the expected number of non-adaptive (*i.e.*, nearly-neutral) non-synonymous substitutions. Non-synonymous substitutions were said to be non-adaptive when corresponding to a mutation of population selection coefficient below threshold S_{adv}:

$$\widehat{D_N^{na}} = 2L_N N_e \mu t \int_{-\infty}^{S_{adv}} \Phi(S) f(N_e, s) dS \tag{15}$$

S_{adv} was here arbitrarily set to $S_{adv} = 5$. For the Neutral and Gamma DFE models, this is equivalent to the existing methods [11, 15].

Comparing this expectation to the actual number of non-synonymous substitutions and dividing by the observed number of synonymous substitutions, we calculated ω_a, the rate of non-adaptive amino-acid substitution relative to neutral divergence, ω_a, the rate of adaptive amino-acid substitution relative to neutral divergence, and α, the proportion of adaptive non-synonymous substitutions:

$$\omega_{na} = (\widehat{D_N^{na}}/L_N)/(D_S/L_S) = \widehat{d_N^{na}}/d_s \quad (16)$$

$$\omega_a = (d_N - \widehat{d_N^{na}})/d_S \quad (17)$$

$$\alpha = (d_N - \widehat{d_N^{na}})/d_N \quad (18)$$

Alternatively, for DFE models explicitly accounting for beneficial mutations, we used a version in which A was set to zero, so that no additional class of strong selection coefficient was assumed. This version, which was called [-A], has one less degree of freedom than the classical. Comparing the two versions is a way to test whether the divergence pattern is appropriately predicted by polymorphism data and the model, or if an extra parameter is needed. Models were compared using likelihood-ratio tests (LRT) when nested, and Akaike's Information Criterion (AIC) otherwise.

Integrals in eqs (3), (7), (9) and (15) were calculated numerically. The likelihood was maximized using the Newton-Raphson method. Confidence intervals around estimates of α were defined as values of α for which the log-likelihood was within two units of its maximum. A C++ program implementing these methods was written based on the Bio++ libraries [68]. Source code and Linux executable files are available upon request. The reliability of α and ω_a estimates obtained under the Gamma, GammaExponential and ScaledBeta models were assessed from simulations (S1 Text). Linear regressions were performed in R using Pearson's parametric method.

Supporting Information

S1 Table. Species pairs, polymorphism and divergence data, α and ω_a estimates, confidence intervals. Pn, Ps, Dn, Ds refer to the total number of non-synonymous and synonymous SNPs and substitutions, respectively. Lpn, Lps, Ldn, Lds refer to the total number of non-synonymous and synonymous sites available in the within-species and between-species alignments, respectively. α_down and α_up refer to the confidence interval boundaries for α (and similarly for ω_a). Estimates were obtained under the GammaExpo model.
(XLS)

S1 Fig. Distribution of α and ω_a estimates across 44 animal species pairs. Top: all species; Bottom: 26 species for which at least seven individuals have been sampled.
(PDF)

S2 Fig. Effect of species longevity on protein divergence parameters. $n = 41$ species; blue: vertebrates; green: invertebrates; a: $r^2 = 0.25$, p-val$<10^{-3}$; b: $r^2 = 0.28$, p-val$<10^{-3}$; c: $r^2 = 0.02$, not significant; d: $r^2 = 0.15$, p-val = 0.01.
(PDF)

S3 Fig. Effect of species propagule size on protein divergence parameters. $n = 43$ species; blue: vertebrates; green: invertebrates; a: $r^2 = 0.13$, p-val = 0.02; b: $r^2 = 0.12$, p-val = 0.02;

c: $r^2 = 0.005$, not significant; d: $r^2 = 0.12$, p-val = 0.02.
(PDF)

S4 Fig. Differences in π_S and α between "mirror" species. Segments connect pairs of mirror species. Blue: increasing α with π_S; red decreasing α with π_S; A: (*Homo sapiens*, *Pan troglodytes*); B: (*Nycticebus coucang*, *Galago senegalensis*); C: (*Eulemur mongoz*, *Eulemur coronatus*); D: (*Chlorocebus aethiops*, *Macaca mulatta*); E: (*Varecia variegata variegata*, *Propithecus coquereli*); F: (*Ciona intestinalis* A, *Ciona intestinalis* B); G: (*Echinocardium mediterraneum*, *Echinocardium cordatum* B2); H: (*Thymelicus sylvestris*, *Thymelicus lineola*)
(PDF)

S1 Text. Simulation protocol and results.
(PDF)

Acknowledgments

The author is grateful to Thomas Bataillon and Sylvain Glémin for help with the theoretical part, Philip Messer for help with his simulation program, the Montpellier Bioinformatics & Biodiversity platform for computing facilities, Dmitri Petrov and four anonymous reviewers for thoughtful comments.

Author Contributions

Conceived and designed the experiments: NG. Performed the experiments: NG. Analyzed the data: NG. Contributed reagents/materials/analysis tools: NG. Wrote the paper: NG.

References

1. Zhang J. Parallel adaptive origins of digestive RNases in Asian and African leaf monkeys. Nat Genet. 2006 Jul; 38(7):819–23. PMID: 16767103
2. Labbé P, Berticat C, Berthomieu A, Unal S, Bernard C, Weill M, et al. Forty years of erratic insecticide resistance evolution in the mosquito *Culex pipiens*. PLoS Genet. 2007 Nov; 3(11):e205. PMID: 18020711
3. Liu Y, Cotton JA, Shen B, Han X, Rossiter SJ, Zhang S. Convergent sequence evolution between echolocating bats and dolphins. Curr Biol. 2010 Jan 26; 20(2):R53–4. doi: 10.1016/j.cub.2009.11.058 PMID: 20129036
4. Huerta-Sánchez E, Jin X, Asan, Bianba Z, Peter BM, Vinckenbosch N, et al. Altitude adaptation in Tibetans caused by introgression of Denisovan-like DNA. Nature. 2014 Aug 14; 512(7513):194–7. doi: 10.1038/nature13408 PMID: 25043035
5. Kimura M. The neutral theory of molecular evolution. Cambridge University Press; 1983.
6. Gillespie J.H. The causes of molecular evolution. Oxford University Press; 1991.
7. Maynard Smith J. What determines the rate of evolution. Am Nat. 1976; 110:331–8.
8. McDonald JH, Kreitman M. Adaptive protein evolution at the *Adh* locus in *Drosophila*. Nature. 1991 Jun 20; 351(6328):652–4. PMID: 1904993
9. Rand DM, Kann LM. Excess amino acid polymorphism in mitochondrial DNA: contrasts among genes from *Drosophila*, mice, and humans. Mol Biol Evol. 1996 Jul; 13(6):735–48. PMID: 8754210
10. Fay JC, Wyckoff GJ, Wu CI. Positive and negative selection on the human genome. Genetics. 2001 Jul; 158(3):1227–34. PMID: 11454770
11. Smith NGC, Eyre-Walker A. Adaptive protein evolution in *Drosophila*. Nature. 2002 Feb 28; 415 (6875):1022–4. PMID: 11875568
12. Bierne N, Eyre-Walker A. The genomic rate of adaptive amino acid substitution in *Drosophila*. Mol Biol Evol. 2004 Jul; 21(7):1350–60. PMID: 15044594
13. Welch JJ. Estimating the genomewide rate of adaptive protein evolution in *Drosophila*. Genetics. 2006 Jun; 173(2):821–37. PMID: 16582427

14. Boyko AR, Williamson SH, Indap AR, Degenhardt JD, Hernandez RD, Lohmueller KE, et al. Assessing the evolutionary impact of amino acid mutations in the human genome. PLoS Genet. 2008 May 30; 4 (5):e1000083. doi: 10.1371/journal.pgen.1000083 PMID: 18516229

15. Eyre-Walker A, Keightley PD. Estimating the rate of adaptive molecular evolution in the presence of slightly deleterious mutations and population size change. Mol Biol Evol. 2009 Sep; 26(9):2097–108. doi: 10.1093/molbev/msp119 PMID: 19535738

16. Charlesworth J, Eyre-Walker A. The McDonald-Kreitman test and slightly deleterious mutations. Mol Biol Evol. 2008 Jun; 25(6):1007–15. doi: 10.1093/molbev/msn005 PMID: 18195052

17. Eyre-Walker A. Changing effective population size and the McDonald-Kreitman test. Genetics. 2002 Dec; 162(4):2017–24. PMID: 12524367

18. Loewe L, Charlesworth B, Bartolomé C, Nöel V. Estimating selection on nonsynonymous mutations. Genetics. 2006 Feb; 172(2):1079–92. PMID: 16299397

19. Eyre-Walker A, Woolfit M, Phelps T. The distribution of fitness effects of new deleterious amino acid mutations in humans. Genetics. 2006 Jun; 173(2):891–900. PMID: 16547091

20. Hvilsom C, Qian Y, Bataillon T, Li Y, Mailund T, Sallé B, et al. Extensive X-linked adaptive evolution in central chimpanzees. Proc Natl Acad Sci U S A. 2012 Feb 7; 109(6):2054–9. doi: 10.1073/pnas.1106877109 PMID: 22308321

21. Loire E, Chiari Y, Bernard A, Cahais V, Romiguier J, Nabholz B, et al. Population genomics of the endangered giant Galápagos tortoise. Genome Biol. 2013 Dec 16; 14(12):R136. doi: 10.1186/gb-2013-14-12-r136 PMID: 24342523

22. Liti G, Carter DM, Moses AM, Warringer J, Parts L, James SA, et al. Population genomics of domestic and wild yeasts. Nature. 2009 Mar 19; 458(7236):337–41. doi: 10.1038/nature07743 PMID: 19212322

23. Gossmann TI, Song B-H, Windsor AJ, Mitchell-Olds T, Dixon CJ, Kapralov MV, et al. Genome wide analyses reveal little evidence for adaptive evolution in many plant species. Mol Biol Evol. 2010 Aug; 27(8):1822–32. doi: 10.1093/molbev/msq079 PMID: 20299543

24. Halligan DL, Oliver F, Eyre-Walker A, Harr B, Keightley PD. Evidence for pervasive adaptive protein evolution in wild mice. PLoS Genet. 2010 Jan 22; 6(1):e1000825. doi: 10.1371/journal.pgen.1000825 PMID: 20107605

25. Carneiro M, Albert FW, Melo-Ferreira J, Galtier N, Gayral P, Blanco-Aguiar JA, et al. Evidence for widespread positive and purifying selection across the european rabbit (*Oryctolagus cuniculus*) genome. Mol Biol Evol 29:1837–1849. doi: 10.1093/molbev/mss025 PMID: 22319161

26. Tsagkogeorga G, Cahais V, Galtier N. The population genomics of a fast evolver: high levels of diversity, functional constraint and molecular adaptation in the tunicate *Ciona intestinalis*. Genome Biol Evol. 2012; 4(8):740–9. doi: 10.1093/gbe/evs054 PMID: 22745226

27. Charlesworth J, Eyre-Walker A. The rate of adaptive evolution in enteric bacteria. Mol Biol Evol. 2006 Jul; 23(7):1348–56. PMID: 16621913

28. Eyre-Walker A. The genomic rate of adaptive evolution. Trends Ecol Evol. 2006 Oct; 21(10):569–75. PMID: 16820244

29. Strasburg JL, Kane NC, Raduski AR, Bonin A, Michelmore R, Rieseberg LH. Effective population size is positively correlated with levels of adaptive divergence among annual sunflowers. Mol Biol Evol. 2011 May; 28(5):1569–80. doi: 10.1093/molbev/msq270 PMID: 20952500

30. Phifer-Rixey M, Bonhomme F, Boursot P, Churchill GA, Piálek J, Tucker PK et al. Adaptive evolution and effective population size in wild house mice. Mol Biol Evol. 2012 Oct; 29(10):2949–55. PMID: 22490822

31. Gossmann TI, Keightley PD, Eyre-Walker A. The effect of variation in the effective population size on the rate of adaptive molecular evolution in eukaryotes. Genome Biol Evol. 2012; 4(5):658–67. doi: 10.1093/gbe/evs027 PMID: 22436998

32. Jensen JD, Bachtrog D. Characterizing the influence of effective population size on the rate of adaptation: Gillespie's Darwin domain. Genome Biol Evol. 2011; 3:687–701. doi: 10.1093/gbe/evr063 PMID: 21705473

33. Razeto-Barry P, Díaz J, Vásquez RA. The nearly neutral and selection theories of molecular evolution under the fisher geometrical framework: substitution rate, population size, and complexity. Genetics. 2012 Jun; 191(2):523–34. doi: 10.1534/genetics.112.138628 PMID: 22426879

34. Lourenço JM, Glémin S, Galtier N. The rate of molecular adaptation in a changing environment. Mol Biol Evol. 2013 Jun; 30(6):1292–301. doi: 10.1093/molbev/mst026 PMID: 23412912

35. Romiguier J, Gayral P, Ballenghien M, Bernard A, Cahais V, Chenuil A, et al. Comparative population genomics in animals uncovers the determinants of genetic diversity. Nature. 2014 Nov 13; 515 (7526):261–3. doi: 10.1038/nature13685 PMID: 25141177

36. Keightley PD, Eyre-Walker A. Joint inference of the distribution of fitness effects of deleterious mutations and population demography based on nucleotide polymorphism frequencies. Genetics. 2007 Dec; 177(4):2251–61. PMID: 18073430

37. Messer PW, Petrov DA. Frequent adaptation and the McDonald-Kreitman test. Proc Natl Acad Sci U S A. 2013 May 21; 110(21):8615–20. doi: 10.1073/pnas.1220835110 PMID: 23650353

38. Perry GH, Melsted P, Marioni JC, Wang Y, Bainer R, Pickrell JK, et al. Comparative RNA sequencing reveals substantial genetic variation in endangered primates. Genome Res. 2012 Apr; 22(4):602–10. doi: 10.1101/gr.130468.111 PMID: 22207615

39. Martin G, Lenormand T. A general multivariate extension of Fisher's geometrical model and the distribution of mutation fitness effects across species. Evolution. 2006 May; 60(5):893–907. PMID: 16817531

40. Lourenço J, Galtier N, Glémin S. Complexity, pleiotropy, and the fitness effect of mutations. Evolution. 2011 Jun; 65(6):1559–71. doi: 10.1111/j.1558-5646.2011.01237.x PMID: 21644948

41. Duret L, Galtier N. Biased gene conversion and the evolution of mammalian genomic landscapes. Annu Rev Genomics Hum Genet. 2009; 10:285–311. doi: 10.1146/annurev-genom-082908-150001 PMID: 19630562

42. Pessia E, Popa A, Mousset S, Rezvoy C, Duret L, Marais GA. Evidence for widespread GC-biased gene conversion in eukaryotes. Genome Biol Evol. 2012; 4(7):675–82. doi: 10.1093/gbe/evs052 PMID: 22628461

43. Ratnakumar A, Mousset S, Glémin S, Berglund J, Galtier N, Duret L, et al. Detecting positive selection within genomes: the problem of biased gene conversion. Philos Trans R Soc Lond B Biol Sci. 2010 Aug 27; 365(1552):2571–80. doi: 10.1098/rstb.2010.0007 PMID: 20643747

44. Drummond DA, Wilke CO. Mistranslation-induced protein misfolding as a dominant constraint on coding-sequence evolution. Cell. 2008 Jul 25; 134(2):341–52. doi: 10.1016/j.cell.2008.05.042 PMID: 18662548

45. Welch JJ, Eyre-Walker A, Waxman D. Divergence and polymorphism under the nearly neutral theory of molecular evolution. J Mol Evol. 2008 Oct; 67(4):418–26. doi: 10.1007/s00239-008-9146-9 PMID: 18818860

46. Betancourt AJ, Blanco-Martin B, Charlesworth B. The relation between the neutrality index for mitochondrial genes and the distribution of mutational effects on fitness. Evolution. 2012 Aug; 66(8):2427–38. doi: 10.1111/j.1558-5646.2012.01628.x PMID: 22834742

47. Corbett-Detig RB, Hartl DL, Sackton TB. Natural selection constrains neutral diversity across a wide range of species. PLoS Biol. 2015 Apr 10; 13(4):e1002112. doi: 10.1371/journal.pbio.1002112 PMID: 25859758

48. Schneider A, Charlesworth B, Eyre-Walker A, Keightley PD. A method for inferring the rate of occurrence and fitness effects of advantageous mutations. Genetics. 2011 Dec; 189(4):1427–37. doi: 10.1534/genetics.111.131730 PMID: 21954160

49. Enard D, Messer PW, Petrov DA. Genome-wide signals of positive selection in human evolution. Genome Res. 2014 Jun; 24(6):885–95. doi: 10.1101/gr.164822.113 PMID: 24619126

50. Daub JT, Hofer T, Cutivet E, Dupanloup I, Quintana-Murci L, Robinson-Rechavi M, et al. Evidence for polygenic adaptation to pathogens in the human genome. Mol Biol Evol. 2013 Jul; 30(7):1544–58. doi: 10.1093/molbev/mst080 PMID: 23625889

51. Bank C, Ewing GB, Ferrer-Admettla A, Foll M, Jensen JD. Thinking too positive? Revisiting current methods of population genetic selection inference. Trends Genet. 2014 Dec; 30(12):540–6. doi: 10.1016/j.tig.2014.09.010 PMID: 25438719

52. Ohta T. Very slightly deleterious mutations and the molecular clock. J Mol Evol. 1987; 26(1–2):1–6.

53. Akashi H, Osada N, Ohta T. Weak selection and protein evolution. Genetics. 2012 Sep; 192(1):15–31. doi: 10.1534/genetics.112.140178 PMID: 22964835

54. Lesecque Y, Keightley PD, Eyre-Walker A. A resolution of the mutation load paradox in humans. Genetics. 2012 Aug; 191(4):1321–30. doi: 10.1534/genetics.112.140343 PMID: 22661324

55. Cherry JL. Should we expect substitution rate to depend on population size? Genetics. 1998 Oct; 150 (2):911–9. PMID: 9755219

56. Gillespie JH. The role of population size in molecular evolution. Theor Popul Biol. 1999 Apr; 55(2):145–56. PMID: 10329514

57. Orr HA. Adaptation and the cost of complexity. Evolution. 2000 Feb; 54(1):13–20. PMID: 10937178

58. Charlesworth B, Jain K. Purifying selection, drift, and reversible mutation with arbitrarily high mutation rates. Genetics. 2014 Dec; 198(4):1587–602. doi: 10.1534/genetics.114.167973 PMID: 25230951

59. Obbard DJ, Welch JJ, Kim KW, Jiggins FM. Quantifying adaptive evolution in the *Drosophila* immune system. PLoS Genet. 2009 Oct; 5(10):e1000698. doi: 10.1371/journal.pgen.1000698 PMID: 19851448

60. Fay JC. Weighing the evidence for adaptation at the molecular level. Trends Genet. 2011 Sep; 27 (9):343–9. doi: 10.1016/j.tig.2011.06.003 PMID: 21775012

61. Lanfear R, Kokko H, Eyre-Walker A. Population size and the rate of evolution. Trends Ecol Evol. 2014 Jan; 29(1):33–41. doi: 10.1016/j.tree.2013.09.009 PMID: 24148292

62. Gayral P, Melo-Ferreira J, Glémin S, Bierne N, Carneiro M, Nabholz B, et al. Reference-free population genomics from next-generation transcriptome data and the vertebrate-invertebrate gap. PLoS Genet. 2013 Apr; 9(4):e1003457. doi: 10.1371/journal.pgen.1003457 PMID: 23593039

63. Figuet E, Ballenghien M, Romiguier J, Galtier N. Biased gene conversion and GC-content evolution in the coding sequences of reptiles and vertebrates. Genome Biol Evol. 2014 Dec 19; 7(1)240–250. doi: 10.1093/gbe/evu277 PMID: 25527834

64. Gayral P, Weinert L, Chiari Y, Tsagkogeorga G, Ballenghien M, et al. Next-generation sequencing of transcriptomes: a guide to RNA isolation in nonmodel animals. Mol Ecol Resour. 2011 Jul; 11(4):650–61. doi: 10.1111/j.1755-0998.2011.03010.x PMID: 21481219

65. Cahais V, Gayral P, Tsagkogeorga G, Melo-Ferreira J, Ballenghien M, Weinert L, et al. Reference-free transcriptome assembly in non-model animals from next-generation sequencing data. Mol Ecol Resour. 2012 Sep; 12(5):834–45. doi: 10.1111/j.1755-0998.2012.03148.x PMID: 22540679

66. Ranwez V, Harispe S, Delsuc F, Douzery EJ. MACSE: Multiple Alignment of Coding SEquences accounting for frameshifts and stop codons. PLoS One. 2011; 6(9):e22594. doi: 10.1371/journal.pone.0022594 PMID: 21949676

67. Hernandez RD, Williamson SH, Zhu L, Bustamante CD. Context-dependent mutation rates may cause spurious signatures of a fixation bias favoring higher GC-content in humans. Mol Biol Evol. 2007 Oct; 24(10):2196–202. PMID: 17656634

68. Guéguen L, Gaillard S, Boussau B, Gouy M, Groussin M, Rochette NC, et al. Bio++: efficient extensible libraries and tools for computational molecular evolution. Mol Biol Evol. 2013 Aug; 30(8):1745–50. doi: 10.1093/molbev/mst097 PMID: 23699471