RESEARCH ARTICLE

Nonlinear Dynamics of Nonsynonymous (d_N) and Synonymous (d_S) Substitution Rates Affects Inference of Selection

Jochen B. W. Wolf, Axel Künstner, Kiwoong Nam, Mattias Jakobsson, and Hans Ellegren

Department of Evolutionary Biology, Uppsala University, Uppsala, Sweden

Selection modulates gene sequence evolution in different ways by constraining potential changes of amino acid sequences (purifying selection) or by favoring new and adaptive genetic variants (positive selection). The number of nonsynonymous differences in a pair of protein-coding sequences can be used to quantify the mode and strength of selection. To control for regional variation in substitution rates, the proportionate number of nonsynonymous differences (d_N) is divided by the proportionate number of synonymous differences (d_S). The resulting ratio (d_N/d_S) is a widely used indicator for functional divergence to identify particular genes that underwent positive selection. With the ever-growing amount of genome data, summary statistics like mean d_N/d_S allow gathering information on the mode of evolution for entire species. Both applications hinge on the assumption that d_S and mean d_S (~branch length) are neutral and adequately control for variation in substitution rates across genes and across organisms, respectively. We here explore the validity of this assumption using empirical data based on whole-genome protein sequence alignments between human and 15 other vertebrate species and several simulation approaches. We find that d_N/d_S does not appropriately reflect the action of selection as it is strongly influenced by its denominator (d_S). Particularly for closely related taxa, such as human and chimpanzee, d_N/d_S can be misleading and is not an unadulterated indicator of selection. Instead, we suggest that inconsistencies in the behavior of d_N/d_S are to be expected and highlight the idea that this behavior may be inherent to taking the ratio of two randomly distributed variables that are nonlinearly correlated. New null hypotheses will be needed to adequately handle these nonlinear dynamics.

Introduction

The extent to which selection affects genes and genomes is a key question in genetics and molecular evolution. Selection may modulate gene sequence evolution in different ways, for example, by constraining potential changes of amino acid sequences (purifying or negative selection) or by favoring new and adaptive genetic variants (positive selection). To quantify selection in the simplest case, the number of nonsynonymous differences in a pair of protein-coding sequences can be estimated. However, substitution rates vary across the genome and between species that makes direct comparisons solely based on nonsynonymous substitutions difficult. To control for variation in the underlying mutation rate, a standard way is to take the ratio of the number of nonsynonymous differences per total number of possible nonsynonymous changes (d_N) to the number of synonymous differences per total number of synonymous changes (d_S). This ratio (d_N/d_S) is then used as a measure of "functional divergence" that accounts for the underlying local or regional variation in the substitution rate for which d_S is taken as a proxy.

The application of d_N/d_S has a strong tradition in evolutionary research, notably for the identification of genes with a history of positive selection (e.g., Nielsen 2005). With the recent advances in sequencing technology, we are now at the wake of an era that will allow comparative genomic analysis across large evolutionary timescales where summary statistics like mean d_N/d_S potentially make it possible to gather information on the mode of evolution for any entity from gene families to chromosomes to entire species. This can address questions about the relative importance of negative and positive selection and about the influence of parameters such as life-history traits or effective population sizes that covary with patterns of molecular evolution (Wright and Andolfatto 2008; Ellegren 2009).

Despite the extensive use of d_N/d_S, there are substantial uncertainties associated with its basic properties. Estimates of mean d_N/d_S in sets of human–chimpanzee orthologous genes for instance have varied from 0.64 (Eyre-Walker and Keightley 1999) and 0.34 (Fay et al. 2001) to about 0.20–0.25 (CSAC 2005; Arbiza et al. 2006; Bakewell et al. 2007; RMGSC 2007). Moreover, based on alignments of sequences from several mammalian genomes, mean d_N/d_S has recently been found to vary among different branches of the mammalian tree (Kosiol et al. 2008). Although some of the variation may be attributed to technical problems like sequence quality and alignment inaccuracies (Schneider et al. 2009), the interpretation and validity of d_N/d_S as a tool for locating genes affected by selection have also been questioned on theoretical grounds. Recent studies convincingly suggest that d_N/d_S shows time dependency (Rocha et al. 2006), that within-population variation can cause a nonmonotonic relationship of the selection strength and d_N/d_S (Kryazhimskiy and Plotkin 2008), and that gene conversion may potentially mimic the effects of selection in the genome (Berglund et al. 2009). There is further a growing literature on the effects of negative selection on d_S that can erroneously mimic signatures of positive selection (Chamary et al. 2006). A detailed understanding of the factors influencing d_N/d_S is of crucial importance as it strongly bears on our ability to make inferences about the role of selection in evolution.

In this study, we focus on the idea that d_N/d_S will be an adequate estimator of functional divergence only if local variation in substitution rates equally affects both synonymous and nonsynonymous sites. Hence, it is of crucial

Key words: positive selection, negative selection, protein evolution, selection models, d_N/d_S ratio, neutral theory, adaptive evolution, melanocortin-1-receptor.

E-mail: wolf@evolbio.mpg.de.

Genome. Biol. Evol. Vol. 2009:308–319.
doi:10.1093/gbe/evp030
Advance Access publication August 13, 2009

importance to understand how d_N scales with d_S. We use simulations in combination with gene sequences available from the genomes of a wide range of vertebrate species to investigate the relationship between d_N and d_S and how this relationship affects their ratio (d_N/d_S and mean d_N/d_S).

Materials and Methods

Terminology

Throughout the manuscript, we adhere to the following terminology: the ratio of d_N and d_S for a single gene is denoted ω, the arithmetic mean of ω across genes is denoted $\bar{\omega}$, the ratio of the sum of d_N (across genes), and the sum of d_S (across genes) is denoted by ψ.

We expand on this in little more detail below as it recurrently emerges as an issue. Mean d_N/d_S can be computed in two ways. One can either calculate ω for each gene and take the average across all genes or calculate the sum of d_N and the sum of d_S across all genes and take the ratio of these two sums. Although the two approaches look similar at a first glance, they are not equal. With a few exceptions, the expectation of a ratio of two random variables is generally not equal to the ratio of the expectation of the two random variables (Hejmans 1999). We can denote

$$\bar{\omega} = \sum_{i \in C} [\frac{d_{N,i}}{d_{S,i}}]/n, \quad (1)$$

and

$$\psi = \frac{\sum_{i \in C} d_{N,i}}{\sum_{i \in C} d_{S,i}}, \quad (2)$$

where the set C contains all genes with $d_S > 0$, n is the number of genes in C, and the summation is over the genes in the set C (note that we could not include $d_S = 0$ when computing ψ, but we use the same set C for both calculations to be able to compare the values directly). To assess the level of difference between $\bar{\omega}$ and ψ, we performed simulations under a simple sequence evolution model (see Results on simulations). A third option would be to concatenate all coding sequences in a genome and estimate mean d_N/d_S directly. Although we expect this to be very similar to ψ, in-depth analysis of the relative performance of these measures may be warranted in future studies.

Data Extraction and Parameter Estimates

Pairwise and Multiple Comparisons with Human and Several Other Vertebrate Species

Full-coding sequences for human and 15 additional species (see table 1) were downloaded from the BioMart database (ENSEMBL 50), and information about pairwise 1:1 orthologues was extracted (http://www.biomart.org). Pairwise alignments with human were generated for all species on protein sequences using MAFFT Version 6.606b (Katoh and Toh 2008) and back translated to DNA sequences for subsequent analysis. Alignments are available upon request. Estimates for d_N, d_S, and ω were computed for each gene using a maximum likelihood (ML) method (Goldman and Yang 1994) and several counting methods (Nei and Gojobori 1986; Li 1993; Yang and Nielsen 2000) implemented in the CODEML program of the PAML package Version 4.1 (Yang 2007). ML analysis was performed with runmode-2. We used a method that takes nucleotide frequencies at each codon position into account and thereby controls for an artificial signature of ω that may be due to differences in the effective number of codons (Albu et al. 2008). Coding sequence alignments where d_N, d_S, or ω exceeded 5 were excluded from all downstream analyses (excluding all values >3 qualitatively yields the same results). We report the results from the ML method. Note that the maximum estimator is asymptotically unbiased. The distributional properties of d_N/d_S we expand on below are thus unlikely to be produced by an estimation bias but will most likely be inherent in the parameter as such. This is partially supported by the fact that counting methods yielded similar results.

In a first step, estimates were only based on pairwise alignments between human and all other species (fig. 1*A* and *B*) instead of branch-specific estimates based on multiple alignments (fig. 1*C*). This allows evaluating the effect of different gene sets across evolutionary distance and avoids potential bias from ancestral reconstruction. The drawback of this approach is that the same starting point (human) is repeatedly used what essentially results in pseudoreplication and may lead to properties specific to the primate lineage being overrated in the result. Explicit comparative contrasts cannot be used to control for it because evolutionary distance (branch length) is the parameter of interest here. We therefore replicated the analyses with mouse as a starting point (supplementary fig. S1, Supplementary Material online).

To further ensure that a single influential branch in the primate lineage does not introduce a systematic bias in the repeated pairwise comparisons with human, we also constructed multiple alignments for 4,181 genes common to all 11 species from human until opossum (11-way core set, see above). As for pairwise alignments, multiple alignments were generated using MAFFT Version 6.606b (Katoh and Toh 2008) and back translated to DNA sequences for subsequent analysis. A total of 3,866 alignments could be used for subsequent analyses. d_N, d_S, and ω were estimated for each gene using the ML method from Yang (2007) implemented in CODEML (model = 1; user tree specified according to Miller et al. [2007]). A threshold of <5 on d_N, d_S, and ω reduced the final data set to 826 estimates.

Pairwise Comparisons between Zebra Finch and Chicken

Consideration of d_N, d_S, and ω involving several species can be influenced by differences in N_e or lineage-specific substitution rates. To exclude the effects of N_e or substitution rate priori, we constructed pairwise alignments between chicken and zebra finch orthologues. We made use of the fact that in birds, there is a large variation in substitution rates across chromosomes and investigated the relationship of mean d_N, d_S, and ψ across chromosomes. We downloaded the zebra finch protein sequences (ZEBRA_FINCH_1, 2009; ENSEMBL 53) from the BioMart

homepage (http://www.biomart.org) and the chicken protein sequences from the inparanoid database that yielded a total of 17,148 sequences from zebra finch and 16,715 sequences from chicken, respectively. 1:1 orthology for these two proteomes was established by reciprocal blasting using inparanoid 3.0 (O'Brien et al. 2005). The program identified 11,413 groups of orthologs, of which 11,309 groups could be shown to have 1:1 orthologue relationships. From this set of genes, we constructed codon-based alignments using MUSCLE (Edgar 2004) followed by the calculation of d_N and d_S using the CODEML program of the PAML 4.1 package (see above). d_N, d_S, or $\omega > 3$ were discarded for subsequent analyses that reduced the data to a remaining 11,107 pairwise d_N and d_S values.

Pairwise and Multiple Alignments of Passerine MC1R Sequences

We also assessed the relationship between d_N, d_S, and ω on a single-gene basis. We chose a gene (*MC1R*) with a prominent role in evolutionary research. Full passerine *MC1R* sequences were obtained from the National Center for Biotechnology Information database (for GenBank accession numbers, see fig. 2). Codon-based pairwise alignments were constructed with the chicken *MC1R* sequence (GenBank accession number: AB201628) and each of the passerine sequences. d_N and d_S were estimated from each alignments using CODEML program. d_N, d_S, or $d_N/d_S > 3$ were not discarded to present the relationship across the full range of observed d_S values. Qualitatively, the results do not change if discarded. In a second step, multiple alignments between all 22 passerine sequences were obtained by MUSCLE (codon based). From this alignment, an ML phylogenetic tree was constructed using PhyML (Guindon and Gascuel 2003). d_N and d_S were estimated with CODEML, applying the free-ratio model to calculate the estimates from individual branches.

Statistical Analyses

Statistical analyses were performed in R 2.8.0 (R Development Core Team 2006). Model selection based on Akaike's information criterion (AIC), Bayesian information criterion (BIC), and backward selection was used to find the best description of the relationship between ψ (or $\bar{\omega}$) with evolutionary distance and the relationship of single gene ω with d_S. A log-log fit described the relationship better than a linear fit (cf. table 2) and is reported throughout the results.

Splines, or piecewise polynomials, were used to fit smoothing curves through the scatterplot data of all genes in pairwise comparison (fig. 3; supplementary figs. S2–S5, Supplementary Material online). We used B-splines as implemented in the "splines package." To decide on the number of knots for the final graphical representation of the splines, we used the BIC, which penalizes the number of parameters more strongly than AIC. As splines can be unduly influenced by values at the extreme of the ranges, we also fitted local regression algorithms (lowess in the "base package"). The shape of the curves was very robust to changes in the number of knots in the regression splines or the smoother span in the lowess algorithm. Bivariate histograms for the heatmaps in figure 3 and supplementary figures S2–S5 (Supplementary Material online) were generated by an in-house script making use of the "fields package."

An ML approach implemented in the "MASS package" was used to fit the best univariate density function from a range of distributions (gamma, Gaussian, uniform, and Poisson) to empirical d_N and d_S distributions. The gamma distribution was found to give the best fit (supplementary fig. S6, Supplementary Material online).

A Model for Pairs of Genes with Synonymous and Nonsynonymous Sites

This section contains a summary of the model used to simulate data from a simple population divergence model. A more detailed description can be found in the Supplementary Material online.

Let us consider a particular gene for which orthologous genes exist in a pair of species and that these two species diverged T_D units of time ago (time is measured in units of N generations and N denotes the population size). For this particular gene, the total substitution rate for synonymous sites is denoted $r_S/2$, and the total substitution rate for nonsynonymous sites is denoted $r_N/2$. We can view these two sets of sites as evolving independent of each other. We will let the sites evolve under rates that are similar to empirically observed rates (a lower rate for the nonsynonymous sites compared with synonymous sites—a difference likely to be caused by purifying selection acting on nonsynonymous sites).

Let's assume that we have sampled one lineage from each species and that substitutions are added to a lineage proportional to the length of the branch. In other words, the number of substitutions M of a branch of length t is Poisson distributed with parameter $r/2t$, $M \sim \mathrm{Po}(r/2t)$. The time till coalescence for two lineages (after they have entered the ancestral population) is denoted T_2. This waiting time is exponentially distributed $T_2 \sim \mathrm{Exp}(1)$, with parameter 1. The total coalescence time for the two lineages is $T_D + T_2 = T$. Assuming no recombination within a gene, all sites in a particular gene (both synonymous and nonsynonymous) evolve according to the same genealogy, that is, all sites within a gene have the exact same coalescent times. We show in the Supplementary Material online that allowing T_2 to vary has negligible impact on the variables that we are interested in here; therefore, we assume that all genes have the same divergence time T.

Results and Discussion

We produced pairwise coding sequence alignments between the complete set of human protein-coding sequences and the orthologous sequences of 15 species, chosen such that they cover a large part of the vertebrate evolutionary history. The number of genes obtained with a stringent 1:1 orthologue relationship ranges between 17,226 for human–chimpanzee and 936 for human–zebra fish (table 1). A total of 105 orthologous genes appear in all 15 pairwise

Table 1
Compilation of Parameters Derived from Pairwise Comparisons between the Human Genome and the Genomes of 15 Other Species

Common Species Name	Binomial Nomenclature	Number of Orthologues with Human	Number of Genes with $\omega > 1$	Number of Genes with $\omega \leq 1$	Probability of Genes with $\omega > 1$ (%)	Branch Length in PAML	Branch Length after Miller et al. (2007)	Mean d_S	Mean d_N	ψ	Spearman's r $d_N \sim d_S$	Spearman's r $\omega \sim d_N$	Spearman's r $\omega \sim d_S$
Chimp	*Pan troglodytes*	17,226	1,422	15,804	0.083	0.028	0.0136	0.020	0.006	0.306	0.287	0.847	−0.178
Macaque	*Macaca mulatta*	16,196	334	15,862	0.021	0.136	0.0640	0.106	0.028	0.260	0.461	0.875	0.034
Mouse lemur	*Microcebus murinus*	13,921	132	13,789	0.009	0.363	0.2237[a]	0.327	0.059	0.181	0.358	0.881	−0.071
Bush baby	*Otolemur garnettii*	12,936	98	12,838	0.008	0.421	0.2565	0.361	0.068	0.188	0.379	0.900	−0.008
Dog	*Canis familiaris*	13,145	11	13,134	0.001	0.490	0.3350	0.468	0.072	0.154	0.449	0.873	0.015
Elephant	*Loxodonta africana*	11,946	74	11,872	0.006	0.479	0.3381	0.427	0.075	0.176	0.352	0.891	−0.054
Rabbit	*Oryctolagus cuniculus*	11,592	43	11,549	0.004	0.506	0.3504	0.487	0.072	0.148	0.353	0.883	−0.073
Cow	*Bos taurus*	14,148	12	14,136	0.001	0.523	0.3423	0.506	0.075	0.149	0.391	0.880	−0.036
Mouse	*Mus musculus*	15,093	5	15,088	0.000	0.705	0.4532	0.670	0.091	0.137	0.397	0.923	0.060
Rat	*Rattus norvegicus*	13,904	3	13,901	0.000	0.734	0.4613	0.690	0.097	0.141	0.412	0.918	0.066
Opossum	*Monodelphis domestica*	12,283	2	12,281	0.000	1.224	0.7114	1.256	0.134	0.107	0.446	0.842	−0.048
Platypus	*Ornithorhynchus anatinus*	8,527	0	8,527	0.000	1.465	0.9674	1.615	0.149	0.092	0.419	0.828	−0.107
Chicken	*Gallus gallus*	8,485	0	8,485	0.000	1.637	1.0869	1.772	0.157	0.089	0.481	0.873	0.043
Xenopus	*Xenopus tropicalis*	3,575	2	3,573	0.001	2.208	1.5278	2.485	0.178	0.072	0.440	0.870	−0.004
Zebra fish	*Danio rerio*	936	0	936	0.000	2.623	1.8287	3.041	0.201	0.066	0.256	0.914	−0.104

[a] Branch length for mouse lemur could not directly be obtained from the study by Miller et al. (2007). We could, however, make use of the strong correlation between branch length values obtained by the CODEML package and those derived from Miller et al. (2007; $R^2 = 0.96$, $P < 0.001$; $d_S = -0.03 + 1.46 \times$ Miller branch length) to predict the branch length of mouse lemur.

comparisons, representing a common core set of genes shared between all the studied vertebrate species. For each gene, we estimated the number of nonsynonymous changes per nonsynonymous site (d_N), the number of synonymous changes per synonymous site (d_S), and their ratio $\omega = d_N/d_S$.

As there is theoretical motivation that the degree of divergence between two lineages can affect mean d_N/d_S (Rocha et al. 2006; Kryazhimskiy and Plotkin 2008), we initially chose to use external estimates of divergence (branch length estimates from Miller et al. [2007]) to explore its relationship with mean d_N/d_S. However, it turns out that this measure can basically be equated with mean d_S as estimated from our data ($R^2 = 0.96$, $P < 0.001$; $d_S = -0.03 + 1.46 \times$ Miller branch length). Therefore, d_S serves as a good proxy for branch length and all conclusions based on branch length will be true for d_S as well. This is of importance as we later propose that it is really d_S that influences d_N/d_S and not the concept of divergence per se.

Mean d_N/d_S Depends on Branch Length and the Set of Genes Used

Mean d_N/d_S, measured by the unbiased estimator ψ, strongly decreases with branch length (fig. 1*A*; log-log regression: $P < 0.001$, $R^2_{adj} = 0.89$). For example, ψ is 0.31 for human–chimpanzee, 0.14 for human–mouse, and 0.07 for human–zebra fish comparisons. An intuitive explanation for this relationship is that the set of orthologues of increasingly distant species comparisons contain an increasing fraction of conserved genes that are involved in basic biological processes and molecular functions shared among many vertebrate species. Low ω values in distant comparisons could thus be seen to represent genes evolving under strong purifying selection. This effect of the selected genes becomes clear if we use different sets of 1:1 orthologues that are present in all species under consideration. For example, figure 1*B* shows that the relationship between ψ and branch length is shifted toward higher ψ values when based on alignments of genes found in all comparisons from human–chimpanzee until human–opossum (11-way core set: 4,181 genes), compared with when based on genes found in all comparisons from human–chimpanzee until human–zebra fish (15-way core set: 105 genes).

Irrespective of which core set of common genes is used, ψ still decreases with branch length (fig. 1*B*; log-log regression: 11-way core set: $P < 0.001$, $R^2_{adj} = 0.91$; 15-way core set: $P < 0.001$, $R^2_{adj} = 0.89$; similar relationships are obtained with all other possible core sets; data not shown). This suggests that the decrease in ψ over time in pairwise comparisons is not only a consequence of using gene sets that are increasingly enriched for genes with conserved functions but rather that there is an additional factor influencing ψ. It can be argued that alignment length can influence estimates of ω potentially explaining the behavior of ψ. The argument goes that less constrained parts of a gene could be increasingly difficult to align for increasingly distant lineages, leading to gaps in the alignment, whereas more conserved parts of the gene are still easily aligned in distant species comparisons. However, we find no

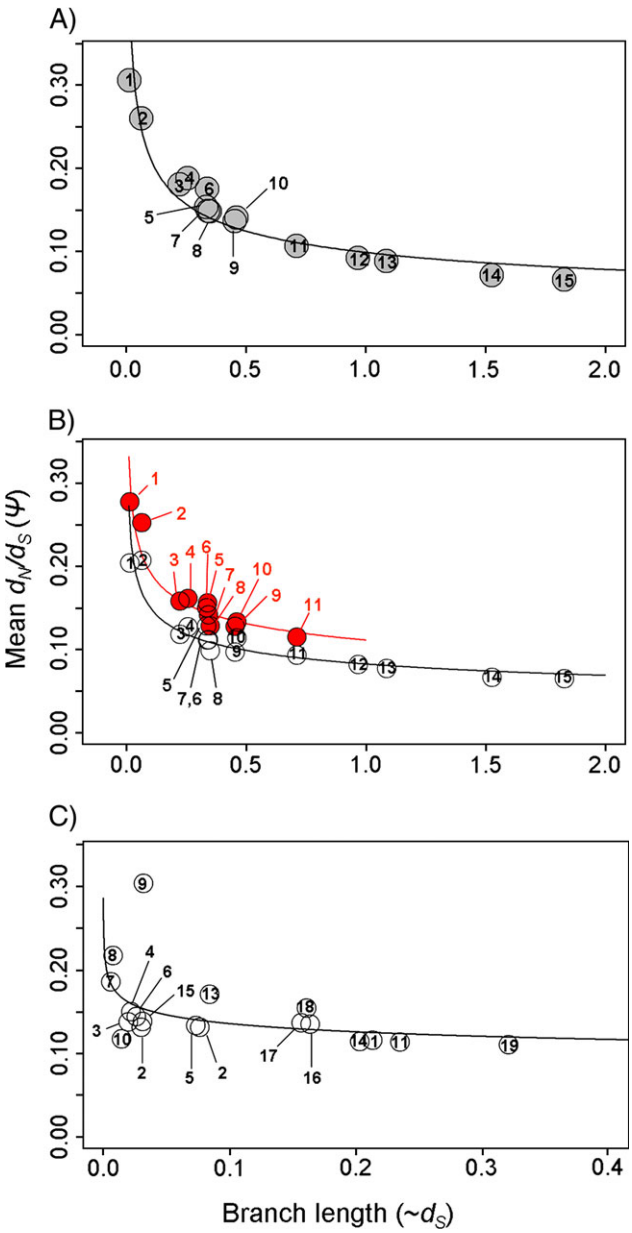

FIG. 1.—Relationship of ψ and branch length based on estimates from Miller et al. (2007). (*A*) Pairwise alignments of human and 15 other species where all possible orthologues between two species are included (compare table 1). (*B*) Pairwise alignments of human and 15 other species restricted to core sets of genes that are common to all species pairs under consideration. "Red": 11-way core set of 4,181 orthologues genes retrieved from all possible pairwise comparisons from human–chimpanzee to human–opossum. "Black": 15-way core set of 105 genes common to all possible pairwise comparisons from human–chimpanzee to human–zebra finch. The fitted lines are based on log-log regression models. "Number code": 1: chimp; 2: macaque; 3: mouse lemur; 4: bush baby; 5: dog; 6: elephant; 7: cow; 8: rabbit; 9: mouse; 10: rat; 11: opossum; 12: platypus; 13: chicken; 14: xenopus; and 15: zebra fish. (*C*) Relationship of ψ and branch length based on multiple alignment of the 11-way core set including a total of 3,866 genes. Individual data points represent estimated values of ψ for both terminal and internal branches after ancestral reconstruction. Numbers encode branch identity (see tree supplementary fig. S7, Supplementary Material online). Branches with the highest ψ 7, 8, 9 are the terminal branches of human, chimpanzee, and rhesus macaque, respectively.

relationship between alignment length and evolutionary distance in either of the core sets (11-way core set: $R^2_{adj}=0.05$, $P = 0.25$; 15-way core set: $R^2_{adj}=0.03$, $P = 0.26$).

The observed pattern could potentially be produced by some primate lineage-specific properties. If d_N/d_S was exceedingly high for an internal branch in the primate lineage (which is repeatedly included in all more distant pairwise comparisons), the observed negative correlation could in fact simply reflect the dilution of this high value with increasing branch length. To rule this out, we replicated the analyses with mouse as a starting point. We obtain the same results suggesting that the pattern is not an artifact of using human as a starting point (supplementary fig. S1, Supplementary Material online). To further exclude the influence of pseudoreplicated branches, we constructed multiple alignments for all species and genes included in the 11-way core set, from which we obtained branch-specific estimates of ψ for a total of 826 genes (see supplementary fig. S7, Supplementary Material online). Compared with pairwise estimates, a similar, but less pronounced decay of mean d_N/d_S with evolutionary distance is observed (fig. 1*C*; $R^2_{adj}=0.23$, $P < 0.05$). The by far shortest branches are the terminal branches of human, chimpanzee, and rhesus macaque (7, 8, and 9 in fig. 1*C*). It is apparent that the upward shift in ψ is most strongly pronounced for these. Still, a negative linear relationship between ψ and branch length persists for the remaining branches ($R^2_{adj}=0.23$, $P < 0.05$).

The dependency of d_N/d_S on its denominator can be observed even in pairwise comparisons within the same species where additional effects such as differences in N_e or substitution rate can be excluded a priori. We made use of the fact that in birds, there is a large variation in substitution rate across chromosomes and constructed pairwise alignments between chicken and zebra finch orthologues. The same significant negative correlation between ψ and mean d_S per chromosome is observed when ψ and mean d_S are calculated for each chromosome separately (data will be presented elsewhere).

We also note by passing that the way mean d_N/d_S is estimated strongly influences its relationship with evolutionary distance; the correlation between $\bar{\omega}$ and branch length is slightly stronger (log-log regression: pairwise 11-way core set: $P < 0.001$, $R^2_{adj}=0.98$, 15-way core set: $P < 0.001$, $R^2_{adj}=0.97$; branch specific: $P < 0.001$, $R^2_{adj}=0.59$) than the correlation between ψ and branch length (see above). However, $\bar{\omega}$ can often be a misleading and upwardly biased statistic for evaluating mean d_N/d_S. Simulations show that the level of bias of $\bar{\omega}$ varies considerably depending on substitution rate assumptions (see supplementary figs. S12 and S15, Supplementary Material online). In summary, mean d_N/d_S depends on several factors including the way it is estimated, the analyzed set of genes, and evolutionary distance between two lineages.

Interpretation and Implications for Comparative Studies

Recently, Rocha et al. (2006) presented a model predicting that mean d_N/d_S depends on time since divergence

of two lineages. The expected negative relationship between divergence time and mean d_N/d_S was both analytically derived for an island model with infinite population sizes and illustrated by simulation in a model incorporating genetic drift. Rocha et al. (2006) find that data from bacterial genomes follow their theoretical predictions. Here, we find a qualitatively similar decrease of mean d_N/d_S for increasing evolutionary distance (fig. 1). However, the effect described by Rocha et al. (2006) is only expected to be influential for very closely related lineages with divergence at least one order of magnitude lower than what we observe here. The relative slowdown of this process due to small effective population sizes of vertebrates compared with bacteria is unlikely to entirely make up for this difference. Likewise, Kryazhimskiy and Plotkin (2008) suggest that for very closely related species ω may be upward biased if slightly deleterious mutations prevail. In a population genetic framework, where most of the observed nucleotide differences are polymorphisms rather than substitutions, they show that the effect of selection is not appropriately captured by ω. For closely related lineages, the proportion of within-species variation to between-species variation can be substantial. For the human–chimpanzee comparison roughly, 10–15% of all observed nucleotide changes will be polymorphic in one of the species (CSAC 2005). Hence, this effect may contribute to the observed increase in ψ. Although the results from Rocha et al. (2006) and Kryazhimskiy and Plotkin (2008) possibly explain parts our observation of an initial strong decrease in ψ, between the human–chimpanzee and potentially also human–rhesus macaque, their models unlikely explain the continuing decrease over longer timescales. A tentative biological explanation may be sought in the effects of epistasis that could effectively shelter deleterious alleles from selection for very long times. According to this way of reasoning, selection coefficients of individual mutations may be low with purifying selection not acting until the cumulative effects of several slightly deleterious alleles reach a certain threshold. However, neither this explanation nor any of the discussed models can explain that the same pattern is observed across chromosomes in the same pairwise comparison of the same two species (chicken and zebra finch) where differences in N_e and evolutionary trajectory can be excluded a priori. This seems to be a general pattern at least in birds. A recent genome-wide study in 11 bird species reveals the same strong relationship between mean d_N/d_S and mean d_S per chromosome (Künstner A, Wolf JBW, Backstrom N, Wilson RK, Jarvis E, Warren WC, Ellegren H, unpublished data).

How does the relationship between mean d_N/d_S and evolutionary distance affect studies using mean d_N/d_S in a comparative framework? Taken to the extreme, it may invalidate intertaxa comparisons that simply use point estimates of mean d_N/d_S instead of time trajectories (cf. Rocha et al. 2006). Point estimates of mean d_N/d_S as a proxy for average selection pressure in specific species have recently been used to demonstrate a negative correlation between mean d_N/d_S and N_e with the interpretation that small populations face difficulty in removing slightly deleterious nonsynonymous mutations thereby leading to elevated ψ (Popadin et al. 2007; Wright and Andolfatto 2008; Ellegren 2009). These papers argue that the findings comply with Ohta's model of nearly neutral molecular evolution (e.g., Ohta and Ina 1995). It will be important for future studies that aim to relate the role of natural selection in molecular evolution to various features of life history to control for the effects of the dependency of mean d_N/d_S on evolutionary distance.

In Pairwise Comparisons of Closely Related Species, High d_N/d_S Is Not Only Driven by Positive Selection on d_N

The individual gene-centered estimates of ω, d_N, and d_S in a pairwise comparison are the raw material for the estimation of mean d_N/d_S. The behavior of these parameters is therefore connected to the behavior of mean d_N/d_S. As an example, we chose the gene *MC1R* that has been in focus in numerous evolutionary studies, being a determinant of pigmentation phenotypes (e.g., Nadeau et al. 2007). We obtained both pairwise d_N and d_S estimates between chicken and 22 passerine bird species and branch-specific estimates based on a phylogenetic tree reconstruction of the same species (supplementary fig. S8, Supplementary Material online). In accordance with what we observed for mean d_N/d_S, ω is negatively correlated with d_S for both pairwise ($P < 0.001$, $R^2_{adj}=0.86$; fig. 2*A*) and branch-specific estimates ($P < 0.05$, $R^2_{adj}=0.33$; fig. 2*B*). Moreover, note that analogous to the above observations on mean d_N/d_S, the inclination is stronger for low values of d_S. This observation will be discussed in-depth below.

Such gene-specific estimates are often used in genome scans for positively selected genes, which is probably the most common application of ω. It is generally assumed that high ω values are driven by a comparatively high number of nonsynonymous changes. However, low d_S can obviously also give rise to high ω values. In the following, we will explore this notion in more detail and we see that for closely related taxa such as human and chimpanzee, high ω values are often not the result of unusually high d_N values but unusually low d_S values.

We investigate the relationship of d_N and d_S from two different perspectives: a longitudinal approach following specific orthologous genes across a broad evolutionary time frame and a cross-sectional approach exploring the relationship of d_N and d_S for all genes in every pairwise alignment with the human sequence. For the longitudinal approach, we used the two core sets of genes described above, that is, genes found in all alignments of increasingly distant common ancestors of species pairs, up till human–opossum (11-way core set: 4,181 genes) and up till human–zebra finch (15-way core set: 105). For every gene in the data sets, we fitted several candidate functions through the 15 (15-way core set) and 11 (11-way core set) data points of d_N and d_S. This procedure was repeated for each pairwise alignments (table 2). A model selection approach based on AIC was used to determine the best fit (model selection based on the more conservative BIC, where the number of parameters is more penalized than for AIC, yielded the same results). Under mutation–selection–drift equilibrium, the neutral theory predicts a positive correlation between d_N and d_S (Ohta and Ina 1995), which can indeed be observed in all the 15 pairwise comparisons (mean $r_{Spearman} = 0.39$,

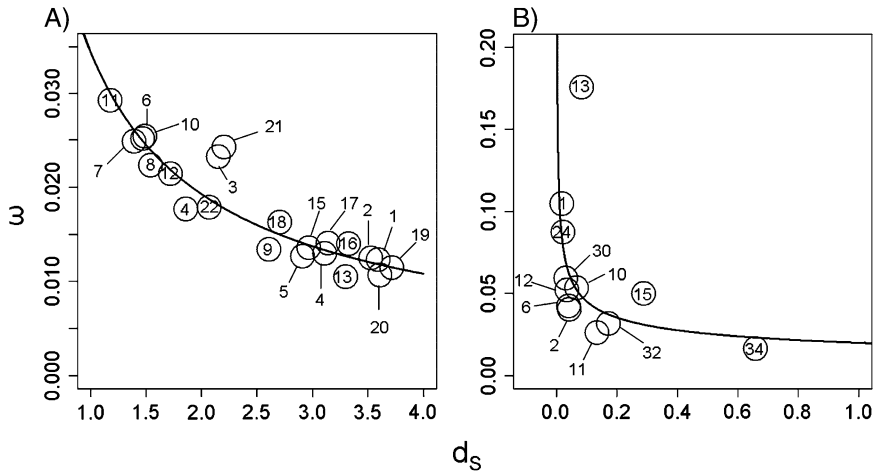

FIG. 2.—Relationship between ω and d_S estimated for the gene *MC1R*. (*A*) Estimates based on pairwise comparisons between chicken and 22 passerine bird species. Number code 1: *Lepidothrix serena* (DQ388331); 2: *Lepidothrix coronata* (DQ388330); 3: *Malurus leucopterus* (AY614610); 4: *Phylloscopus chloronotus* (AY308751); 5: *Phylloscopus humei* (AY308750); 6: *Phylloscopus tytleri* (AY308753); 7: *Phylloscopus fuscatus* (AY308754); 8: *Phylloscopus pulcher* (AY308752); 9: *Phylloscopus collybita* (AY308747); 10: *Seicercus burkii* (AY308757); 11: *Seicercus xanthoschistus* (AY308756); 12: *Phylloscopus trochiloides* (AY308749); 13: *Coereba flaveola* (AF362601); 14: *Tangara cucullata* (AF362606); 15: *Vermivora peregrina* (AY308755); 16: *Passerina cyanea* (EU191783); 17: *Passerina caerulea* (EU191787); 18: *Passerina amoena* (EU191785); 19: *Cyanocompsa cyanoides* (EU191789); 20: *Passerina rositae* (EU191788); 21: *Corvus corone* (EU348729); and 22: *Perisoreus infaustus* (DQ643387). (*B*) Branch-specific estimates from a phylogenetic reconstruction of the bird species in (*A*). Numbers encode branch identity (see tree supplementary fig. S8, Supplementary Material online).

see table 1). However, this relationship is nonlinear for basically all genes that have been explored in both core sets. Instead continuously decreasing functions or slightly *U*-shaped functions (for the parameter space of the data) for the ω–d_S relationships showed closer fits to the data than linear fits (table 2). This observation indicates that the relationship between d_N and d_S is better described by more parameter-rich models leading to ω being a nonlinear function of d_S. Note that d_S can effectively be seen as a proxy for evolutionary distance. The relationship of ω and d_S thus mirrors the decrease of ψ with evolutionary time (fig. 1).

Why would ω for the same gene be lower for more distantly related species? A closer look on the distributions of d_N and d_S in pairwise comparisons is insightful (fig. 3*A*–*C*; supplementary figs. S2–S4, Supplementary Material online). The first observation is that the proportion of genes that show $\omega > 1$, a traditional threshold for interpreting positive selection, strongly declines with evolutionary distance (logistic regression, $P < 0.001$, null deviance: 5284.15, residual deviance: 294.2). For example, in the human–chimpanzee comparison, ~8.3% of all genes have $\omega > 1$; this proportion quickly drops to ~2% for human–rhesus macaque, falls below 1% for comparisons with bush baby, and basically equals zero for more distant lineage comparisons (table 1, fig. 3*A*–*C*). Closer inspection of the distributions shows that the relationship between d_N and d_S is nonlinear and that the relationship changes with evolutionary distance (fig. 3 left; supplementary figs. S2–S4, Supplementary Material online). This nonlinear relationship leads to ω depending on d_N (fig. 3 center) and d_S (fig. 3 right) in ways that are hard to predict (cf. Wyckoff et al. 2005). Overall, ω is correlated with d_N (in each of the 15 pairwise alignments with human, there is a strong positive correlation between ω and d_N; mean $r_{Spearman} = 0.88$, table 1), whereas no overall correlation between ω and d_S is found, except a negative correlation for closely related species (table 1, mean $r_{Spearman} = -0.031$). However, there is an intricate relationship between ω and d_N and between ω and d_S that is exposed by nonparametric smoothing (fig. 3 center, right). Model selection approaches, based on AIC and BIC, corroborate that parameter local regressions provide a much better fit than linear regressions (fig. 3; supplementary figs. S2–S4, Supplementary Material online).

It has been argued that the observed initial positive correlation between ω and d_S (for $d_S < 1$ Wyckoff et al. 2005) may point toward a higher potential for adaptive evolution (indicated by ω) in loci with higher mutation rates (indicated by d_S). The inverse correlation between ω and d_S for closely related lineages has been ascribed to sampling variance (Wyckoff et al. 2005; Vallender and Lahn 2007). Indeed, if we assume a Poisson process generating the differences giving rise to d_S, it intuitively makes sense that the high variance at low d_S is associated with an increase in ω. However, if reduction in variance with increasing d_S accounted for the decline of ω, this effect should even be stronger for d_N. Yet d_N shows the opposite pattern of a positive correlation with ω across the whole range of species comparisons (table 1, fig. 3*A*–*C*). Thus, sampling variance is insufficient for explaining the observed inverse correlation between ω and d_S for closely related species. Combined with the observation of a nonlinear fit between ω and d_S (fig. 3 right) with an initial positive correlation that turns to be negative at higher d_S makes a biological explanation of the relationship less straightforward.

Stochastic Properties of d_N, d_S and ω

To further explore the properties of ω, we assume that d_N and d_S are random variables with some distribution.

Table 2
Candidate Functions That Describe Possible Relationships between d_N and d_S and the Resulting Relationship between ω and d_S

Relationship $d_N \sim d_S$	None (null model) $d_N = a$	Linear $d_N = a + b * d_S; a = 0$	Allometric (=linear log-log) $d_N = a * d_S^b$	Exponential $d_N = a + b * e^{d_S}$	Quadratic $d_N = a + b * d_S^2 + c * d_S$	Third-Order Polynomial $d_N = a + b * d_S^3 + c * d_S^2 + d * d_S$
Relationship $\omega \sim d_S$	Hyperbolic	None	Continuously decreasing with lower asymptote	Slightly *U* shaped	Slightly *U* shaped	Slightly *U* shaped depending on parameters
	$\omega = \frac{a}{d_S}$	$\omega = b$	$\omega = \frac{a*d_S^b}{d_S}$	$\omega = \frac{a+b*e^{d_S}}{d_S}$	$\omega = \frac{a+b*d_S^2+c*d_S}{d_S}$	$\omega = \frac{a+b*d_S^3+c*d_S^2+d*d_S}{d_S}$
Number of genes (core sets 1/2)	0/192	0/20	52/2,302	3/152	16/557	11/357
Proportion of genes (core sets 1/2)	0.00/0.05	0.00/0.01	0.63/0.64	0.04/0.04	0.20/0.16	0.13/0.10

NOTE.—The numbers of genes that are best described by a given function are listed for two core sets containing 105 and 4,181 genes, respectively. Note that the number of genes will not sum to the number of genes in the core sets because genes were only counted when one model was clearly preferred ($AIC_c > 1$).

We fitted gamma distributions to the observed d_N and d_S data as they provide a reasonable fit over a broad range of pairwise comparisons (supplementary fig. S6, Supplementary Material online). For a particular species comparison, drawing a pair of values from these distributions will represent a pair of d_N and d_S values for a hypothetical gene. In a first case, we will assume that there is no correlation between d_N and d_S. For the human–chimpanzee comparison, the fitted gamma distribution is $\Gamma(0.923, 123.8)$ for d_N and $\Gamma(1.416, 70.0)$ for d_S. Drawing a number of d_N and d_S values from these distributions and plotting ω versus d_N and d_S shows that the relationship between the simulated ω and d_N and the simulated ω and d_S are remarkably similar to the observed data (fig. 3*D*; see also supplementary fig. S5A, Supplementary Material online). It is worth mentioning that this pattern is inherent in random sampling of two distributions because similar patterns can be produced across a wide range of distributions that showed a poor fit to the observed distributions of d_N and d_S (we tested uniform, Poisson, and Gaussian; data not shown). The fact that we can mirror the empirical dependency of d_N, d_S, and ω and that we can produce high ω values by randomly drawing from two distributions suggests that at least part of the genes that would be ranked as potential candidates for positive selection in an empirical study could be stochastic artifacts. Still, the proportion of simulated genes with $\omega > 1$ is more than 18% as opposed to observed ~8% from the empirical human–chimpanzee data. From the empirical data, we know that d_N and d_S are positively correlated, which will affect the behavior of ω. We can introduce a covariance structure between the two gamma distributions leading to multivariate gamma distributions (Minhajuddin et al. 2004). Unfortunately, at present, multivariate gamma distributions are limited to two distributions with the same parameters. We therefore explored multivariate normal distributions again fitted on the two differing underlying empirical distributions of d_N and d_S and despite the bad fit of these distributions to the data, they mimic the empirical pattern for closely related species reasonably well (supplementary fig. S5B, Supplementary Material online). None of the approaches, however, yields an initial positive correlation between ω and d_S.

It is clear that this line of reasoning merely constitutes a stochastically informed verbal argument and requires in-depth modeling in the future. Nonetheless, the relationship between d_N and d_S will be a crucial predictor for how ω differs with evolutionary distance. Many parameters that shape the distributions of d_N and d_S themselves differ in their behavior with evolutionary distance. Mean (median) of d_N is on average 7.9 (8.11) times lower than the mean and median of d_S and the difference increases with evolutionary time (log-log regression: $R^2_{adj}=0.89$, $P < 0.001$). Both d_N and d_S show a strong degree of right skew that is alleviated with increasing evolutionary distance (log-log regression: d_N $R^2_{adj}=0.67$, d_S $R^2_{adj}=0.81$, $P_{both} < 0.001$). On average, the coefficient of variation of d_N exceeds that of d_S by 0.35, both increasing by the same relative amount for more closely related species.

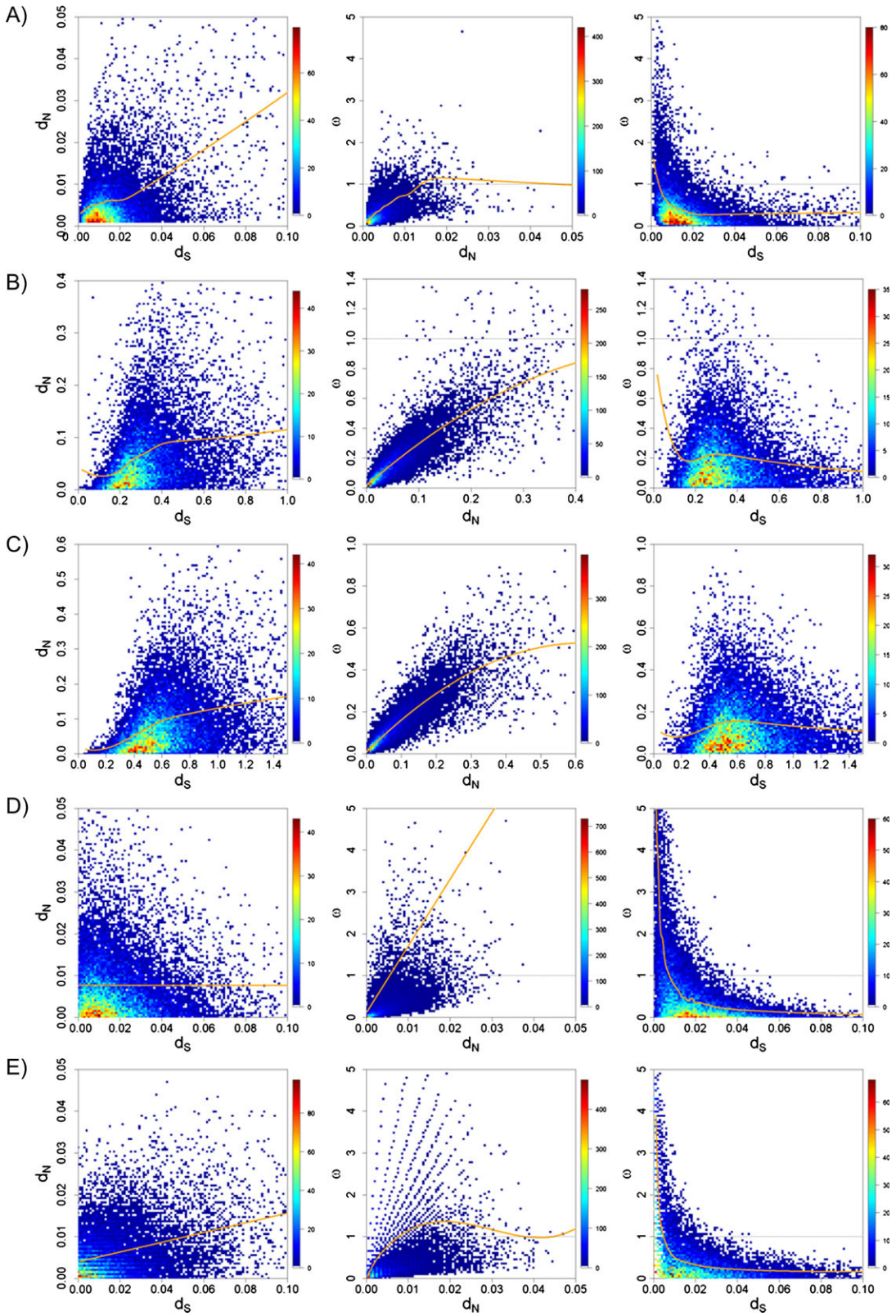

FIG. 3.—Relationship between measures of protein evolution. Left: d_N versus d_S, Middle: ω versus d_N, and Right: ω versus d_S. The relationships are depicted as heatmaps and summarized by regression splines selected by BIC model selection (orange line). The number of genes found in each pixel is symbolized by the different colors. The first three panel sets (*A–C*) show actual genome data, the last two panels (*D–E*) are based on simulations mimicking the human–chimpanzee comparison and should be evaluated in comparison with (*A*). (*A*) Human–chimpanzee comparison, (*B*) human–bush baby comparison, (*C*) human–mouse comparison, (*D*) uncorrelated draws from two multivariate gamma distributions with shape and rate parameters estimated from human–chimpanzee d_N and d_S values, and (*E*) simulated d_N and d_S values based on a Poisson process of accumulating mutations with varying substitution rates (gamma distributed) and a similar degree of correlation between d_N and d_S as in the empirical data ($\rho = 0.4$; see supplementary material, Supplementary Material online). Note that the axis scales differ owing to the large data ranges.

Simulations

To get an additional perspective on the relationship between d_N, d_S, and ω, we simulated data representing orthologous genes from a pair of species. These simulated genes contain 1,000 synonymous sites and 3,000 nonsynonymous sites that could be hit by a substitution. Substitutions are added to the two gene copies by random draws from a Poisson distribution with mean equal to the particular substitution rate (one for nonsynonymous sites and one for synonymous sites) times the time to divergence. The process of adding substitutions to the two sets of sites is independent of each other (except in one case, when the substitution rates were set to be correlated—see below). We also investigated a model that included recombination between genes and found that recombination had a very small effect on the parameters of interest (see supplementary material 2.4, Supplementary Material online; note also that our simulation model differs from the assumption in PAML of free recombination between sites). We assume that there is a weak purifying selection acting on the nonsynonymous sites resulting in a substitution rate that is 0.3 of the rate for synonymous sites. For several different assumptions about the relationship of the synonymous substitution rate and the nonsynonymous substitution rate (fixed rates, variable and uncorrelated rates, and variable and correlated rates), we computed d_N, d_S, ω, and mean d_N/d_S as $\bar{\omega}$ and ψ. A detailed description of the simulations can be found in the Supplementary Material online.

Using a range of assumptions about the relationship of the substitution rates, our simulations are able to capture a number of features of the empirical data, such as the positive correlation of ω and d_N (see e.g., supplementary figs. S11B and S14B and table S1, Supplementary Material online) and the distributions of d_S, d_N, and ω (see e.g., supplementary fig. S13, Supplementary Material online). Some differences between the simulations and the empirical data are found. For example, in the simulation when both the substitution rates are fixed, we find a greater negative correlation between d_S and ω than in the empirical data (supplementary table S1, Supplementary Material online), and in the simulation when the substitution rates are variable, the correlation of d_N and ω is lower than in the empirical data (supplementary tables S2 and S3, Supplementary Material online).

It is clear from our simulations that the level of bias of using $\bar{\omega}$ to measure mean d_N/d_S varies depending on substitution rate assumptions, for example, in the case with fixed substitution rates, the bias decreases with divergence time (supplementary fig. S12 and table S1, Supplementary Material online), and for the case with variable substitution rates, the bias is >40% for the examined interval of divergence times and the bias increases with divergence time (supplementary fig. S15 and tables S2 and S3, Supplementary Material online).

Because high values of ω are taken as evidence of positive selection, it is important to know the distribution of ω. In our simulations, the expectation of ω is set to 0.3, and we assume that a gene with $\omega > 1$ (this cutoff value is arbitrary) would potentially be flagged as a region of interest. In the simulations with fixed substitution rates, we find 0.86% of the genes having $\omega > 1$ when the divergence time is 6 My and the fraction of genes with $\omega > 1$ decreases with increasing divergence time just as observed for the empirical data (supplementary table S1, Supplementary Material online). When the substitution rate is allowed to vary, we find that 8–19% of the genes have $\omega > 1$ (supplementary tables S2 and S3, Supplementary Material online). In other words, assuming a model where nonsynonymous sites are affected by week purifying selection, a substantial fraction of the genes has high ω values, potentially being marked as genes under positive selection. Qualitatively, this resembles the empirical data and supports the result that high ω values can be produced by draws from two randomly distributed variables (fig. 3*E*).

Implications for Inferring Positive Selection

Positive selection is generally evaluated by comparing the likelihood of ω being larger than in a neutral or nearly neutral scenario (Nielsen and Yang 1998). However, likelihood ratio tests do not allow the intricate relationships between ω and d_N or d_S as described above for both empirical data and for simulations. For closely related species, such as human and chimpanzee, current methods may therefore partly identify genes having unusually low d_S rather than genes being molded by true positive selection (comparatively high d_N). We reanalyzed genome scan data from two well-known studies on human–chimpanzee evolution to explore this possibility further.

Nielsen et al. (2005) provided a list with the top 50 candidates showing the strongest evidence for positive selection based on pairwise estimates of ω with subsequent likelihood ratio tests. Mean d_S of this set of candidate genes is 10 times lower than d_S of all other remaining 13,617 genes under consideration (Wilcoxon rank sum test, $W = 146727.5$, $P < 0.001$). The majority of candidate genes do not show a single synonymous substitution. Having a closer look at the residuals of contingency tables suggests that almost half of the candidate genes have an unexpectedly low number of synonymous substitutions compared with the genomic background (supplementary table S5, Supplementary Material online; Fisher's exact test $P < 0.001$). This finding supports the idea that a nonnegligible proportion of genes that have been characterized as being positively selected may be biased toward genes with low d_S which is line with the distributional artifact described above. In biological terms, it could suggest that positive selection preferably acts on slowly evolving genes. It could also point to a strong role in purifying selection on d_S that seems to be essential in several ways, for example, to maintain splicing site accuracy (Parmley et al. 2006). Because most purifying selection on d_S is usually limited to localized windows within a gene (Parmley and Hurst 2007), we would, however, expect that it does not fully account for the observed pattern.

Although Nielsen et al. (2005) chose pairwise alignments between human and chimpanzee for the initial evaluation of candidate genes, Arbiza et al. (2006) pursued a different strategy. They used branch-specific models on

the human, chimpanzee, and their ancestral lineages derived from a common ancestor with mouse and rat. Their inferences are therefore based on d_N and d_S values that are two orders of magnitude higher than those of Nielsen et al. (2005). According to our prediction, artificial inflation of ω by low d_S is much less of a problem here. Indeed, the set of 108 and 577 positively selected genes flagged by Arbiza et al. (2006) for the human and chimpanzee lineage do not have lower d_S than the total set of genes. Accordingly, local purifying selection on d_S seems thus not to show at the level of the gene and does probably not play a major role in the misidentification of positively selected genes. On the contrary, it strengthens the view that most of the genes with unusually low d_S found in the study by Nielsen et al. (2005) are rather a product of the distributional artifact than of purifying selection on d_S.

Conclusion

Using empirical data and simulations, we show that d_N/d_S is not an unadulterated measure of selection but instead depends on d_S or its correlates such as branch length. Under certain conditions, this dependency bears on the outcome of genome scans for positive selection because commonly applied likelihood ratio tests do not explicitly control for this dependency. Inferences drawn from comparative studies using mean "species" d_N/d_S as an indication for the mode of protein evolution across evolutionary timescale (Popadin et al. 2007; Wright and Andolfatto 2008; Ellegren 2009) will be different when branch length is included as a covariate. Furthermore, it is questionable if estimates of the fixation rate of adaptive substitutions based on comparisons between fixed interspecies differences (d_N/d_S) and intraspecific polymorphism (p_N/p_S; Fay et al. 2001; Smith and Eyre-Walker 2002; CSAC 2005) will suffer from a comparable inherent problem. The systematic bias is not limited to genome-wide approaches. Comparative studies of single genes relying on inferences based on d_N/d_S are likely to also be affected.

The ratio of nonsynonymous to synonymous substitutions d_N/d_S has proven to be an important measure in evolutionary studies and will undoubtedly remain to be so. Still, to make best use of it, we will need to understand its properties and the factors that influence it in more detail. Ideally, we can develop new null hypotheses that take into account the influence of various factors including the proportion of polymorphisms to fixed differences (Kryazhimskiy and Plotkin 2008), time trajectories (Rocha et al. 2006), gene conversion (Berglund et al. 2009), and the intricate relationship of d_N and d_S examined here.

Funding

Swedish Research Council (to H.E.); VolkswagenStiftung grant I/83 496 (to J.W.); and FORMAS (to M.J.).

Supplementary Material

Supplementary materials, tables S1–S5 and figures S1–S15 are available at *Genome Biology and Evolution* online (http://www.oxfordjournals.org/our_journals/gbe/).

Acknowledgments

We thank Carina Mugal and Benoit Nabholz for helpful comments. J.W., A.K., M.J., and H.E. conceived of the study. J.W., H.E., and M.J. wrote the manuscript. A.K. was largely responsible for empirical data retrieval, alignments, and data analysis with help from K.N. J.W., A.K., and M.J. conducted statistical analyses and stochastic simulations.

Literature Cited

Albu M, Min XJ, Hickey D, Golding B. 2008. Uncorrected nucleotide bias in mtDNA can mimic the effects of positive Darwinian selection. Mol Biol Evol. 25:2521–2524.

Arbiza L, Dopazo J, Dopazo H. 2006. Positive selection, relaxation, and acceleration in the evolution of the human and chimp genome. PLoS Comput Biol. 2:288–300.

Bakewell MA, Shi P, Zhang JZ. 2007. More genes underwent positive selection in chimpanzee evolution than in human evolution. Proc Natl Acad Sci USA. 104:7489–7494.

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

Chamary JV, Parmley JL, Hurst LD. 2006. Hearing silence: nonneutral evolution at synonymous sites in mammals. Nat Rev Genet. 7:98–108.

CSAC. 2005. Initial sequence of the chimpanzee genome and comparison with the human genome. Nature. 437:69–87.

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

Ellegren H. 2009. A selection model of molecular evolution incorporating the effective population size. Evolution. 63:301–305.

Eyre-Walker A, Keightley PD. 1999. High genomic deleterious mutation rates in hominids. Nature. 397:344–347.

Fay JC, Wyckoff GJ, Wu CI. 2001. Positive and negative selection on the human genome. Genetics. 158:1227–1234.

Goldman N, Yang ZH. 1994. A codon-based model of nucleotide substitution for protein-coding DNA sequences. Mol Biol Evol. 11:725–736.

Guindon S, Gascuel O. 2003. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. Syst Biol. 52:696–704.

Hejmans R. 1999. When does the expectation of a ratio equal the ratio of the expectations? Stat Pap. 40:107–115.

Katoh K, Toh H. 2008. Recent developments in the MAFFT multiple sequence alignment program. Brief Bioinform. 9:286–298.

Kosiol C, et al. 2008. Patterns of positive selection in six mammalian genomes. PLoS Genet. 4:e1000144.

Kryazhimskiy S, Plotkin JB. 2008. The population genetics of dN/dS. PLoS Genet. 4:e1000304.

Miller W, et al. 2007. 28-way vertebrate alignment and conservation track in the UCSC Genome Browser. Genome Res. 17:1797–1808.

Minhajuddin ATM, Harris IR, Schucany WR. 2004. Simulating multivariate distributions with specific correlations. J Stat Comput Simul. 74:599–607.

Nadeau NJ, Burke T, Mundy NI. 2007. Evolution of an avian pigmentation gene correlates with a measure of sexual selection. Proc R Soc Lond B Biol Sci. 274:1807–1813.

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

Nielsen R. 2005. Molecular signatures of natural selection. Annu Rev Genet. 39:197–218.
Nielsen R, et al. 2005. A scan for positively selected genes in the genomes of humans and chimpanzees. PLoS Biol. 3: 976–985.
Nielsen R, Yang ZH. 1998. Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene. Genetics. 148:929–936.
O'Brien KP, Remm M, Sonnhammer ELL. 2005. Inparanoid: a comprehensive database of eukaryotic orthologs. Nucleic Acids Res. 33:D476–D480.
Ohta T, Ina Y. 1995. Variation in synonymous substitution rates among mammalian genes and the correlation between synonymous and nonsynonymous divergences. J Mol Evol. 41:717–720.
Parmley JL, Chamary JV, Hurst LD. 2006. Evidence for purifying selection against synonymous mutations in mammalian exonic splicing enhancers. Mol Biol Evol. 23: 301–309.
Parmley JL, Hurst LD. 2007. How common are intragene windows with K-A > K-S owing to purifying selection on synonymous mutations? J Mol Evol. 64:646–655.
Popadin K, Polishchuk LV, Mamirova L, Knorre D, Gunbin K. 2007. Accumulation of slightly deleterious mutations in mitochondrial protein-coding genes of large versus small mammals. Proc Natl Acad Sci USA. 104:13390–13395.
R Development Core Team. 2006. R: A language and environment for statistical computing. Vienna (Austria): R Foundation for Statistical Computing.
RMGSC. 2007. Evolutionary and biomedical insights from the rhesus macaque genome. Science. 316:222–234.
Rocha EPC, et al. 2006. Comparisons of dN/dS are time dependent for closely related bacterial genomes. J Theor Biol. 239:226–235.
Schneider A, et al. 2009. Estimates of positive Darwinian selection are inflated by errors in sequencing, annotation, and alignment. Genome Biol Evol. 2009:114–118.
Smith NG, Eyre-Walker A. 2002. Adaptive protein evolution in *Drosophila*. Nature. 415:1022–1025.
Vallender EJ, Lahn BT. 2007. Uncovering the mutation-fixation correlation in short lineages. BMC Evol Biol. 7.
Wright SI, Andolfatto P. 2008. The impact of natural selection on the genome: emerging patterns in *Drosophila* and *Arabidopsis*. Annu Rev Ecol Evol Syst. 39:193–213.
Wyckoff GJ, Malcom CM, Vallender EJ, Lahn BT. 2005. A highly unexpected strong correlation between fixation probability of nonsynonymous mutations and mutation rate. Trends Genet. 21:381–385.
Yang ZH. 2007. PAML 4: phylogenetic analysis by maximum likelihood. Mol Biol Evol. 24:1586–1591.
Yang ZH, Nielsen R. 2000. Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models. Mol Biol Evol. 17:32–43.

Laurence Hurst, Associate Editor

Accepted August 10, 2009