Estimating the Rate of Irreversibility in Protein Evolution

Onuralp Soylemez[1,2] and Fyodor A. Kondrashov[1,2,3,*]

[1]Bioinformatics and Genomics Programme, Centre for Genomic Regulation (CRG), Barcelona, Spain

[2]Universitat Pompeu Fabra (UPF), Barcelona, Spain

[3]Institució Catalana de Recerca i Estudis Avançats (ICREA), Barcelona, Spain

*Corresponding author: E-mail: Fyodor.Kondrashov@crg.eu.

Accepted: October 30, 2012

Abstract

Whether or not evolutionary change is inherently irreversible remains a controversial topic. Some examples of evolutionary irreversibility are known; however, this question has not been comprehensively addressed at the molecular level. Here, we use data from 221 human genes with known pathogenic mutations to estimate the rate of irreversibility in protein evolution. For these genes, we reconstruct ancestral amino acid sequences along the mammalian phylogeny and identify ancestral amino acid states that match known pathogenic mutations. Such cases represent inherent evolutionary irreversibility because, at the present moment, reversals to these ancestral amino acid states are impossible for the human lineage. We estimate that approximately 10% of all amino acid substitutions along the mammalian phylogeny are irreversible, such that a return to the ancestral amino acid state would lead to a pathogenic phenotype. For a subset of 51 genes with high rates of irreversibility, as much as 40% of all amino acid evolution was estimated to be irreversible. Because pathogenic phenotypes do not resemble ancestral phenotypes, the molecular nature of the high rate of irreversibility in proteins is best explained by evolution with a high prevalence of compensatory, epistatic interactions between amino acid sites. Under such mode of protein evolution, once an amino acid substitution is fixed, the probability of its reversal declines as the protein sequence accumulates changes that affect the phenotypic manifestation of the ancestral state. The prevalence of epistasis in evolution indicates that the observed high rate of irreversibility in protein evolution is an inherent property of protein structure and function.

Key words: irreversibility, epistasis, genetic diseases, protein evolution, ancestral state reconstruction.

Introduction

In the course of evolution, novel phenotypes and genotypes are produced. However, to what degree the acquired novelties in an evolving lineage can revert back to the states found in its direct ancestors remains an open and debated question (Dollo 1893; Muller 1939; Bull and Charnov 1985; Teotónio and Rose 2001; Chippindale et al. 2004; Collin and Miglietta 2008; Bridgham et al. 2009; Tan et al. 2011). Dollo formulated what is now known as the Dollo's law stating that "an organism is unable to return, even partially, to a previous stage already realized in the ranks of its ancestors" (Dollo 1893). Dollo's statement refers to irreversibility on the level of phenotypes. On the genotype level, Dollo's law can be adapted to describe a situation when a substantial fraction of substitutions that revert the genotype to the ancestral state are under strong negative selection. Many nucleotide changes, both synonymous and nonsynonymous, are expected to be unequivocally neutral because they never lead to any form of phenotypic change in the organism. However, there must also be genotypic changes that have profound effects on a phenotypic level, and whether or not such changes may be reversed in the course of evolution is the issue at hand when considering Dollo's law on the genotypic level. The irreversibility of amino acid changes has been tested empirically in only a handful of specific cases, and the overall rate of irreversibility of evolutionary change on the molecular level has not been comprehensively addressed (Bull and Charnov 1985; Kondrashov et al. 2002; Zufall and Rausher 2004; Poon and Chao 2005; Bridgham et al. 2009; Lunzer et al. 2010; de Visser et al. 2011; Tan et al. 2011). Here, we test the overall rate of irreversibility of amino acid substitutions by comparing data on pathogenic mutations in humans with sequence divergence data between human genes and orthologs from mammalian species.

Materials and Methods

Mutation Data and Orthologs

We obtained data on missense and nonsense pathogenic mutations from the Human Genome Mutation Database 2011.1 (HGMD) and SwissVAR databases (Stenson et al. 2009; Mottaz et al. 2010, respectively). For our data set, we selected all the 221 protein coding genes with at least 50 missense mutations in HGMD. For these genes, we retrieved sequences of orthologous proteins from all available vertebrate species available in GenBank (Benson et al. 2006) using a two-directional best blast hit approach (Tatusov et al. 1997). Multiple alignments of orthologous proteins were created using MUSCLE (Edgar 2004), and these alignments are refined by removing incompletely sequenced orthologs or sequences with more than 20% alignment gaps with respect to the human sequence. A subset of genes in which pathogenic mutations lead to early onset and severe pathologies was created using Online Mendelian Inheritance in Man descriptions.

Phylogeny and Ancestral Reconstruction

We constructed a phylogeny of all vertebrate species following previously published information on the phylogeny of separate clades that included all species with an ortholog to at least one gene out of the entire 221 gene data set. Reports on the phylogeny of Primata (Chatterjee et al. 2009; Fabre et al. 2009), Cetartiodactyla (Arnason et al. 2004; Hernández Fernández and Vrba 2005; Price et al. 2005; Steiner et al. 2005; Tanaka et al. 2011), Carnivora (Higdon et al. 2007), Chiroptera (Jones et al. 2002, 2005), Metatheria (Nilsson et al. 2010), Aves (Dimcheff et al. 2002; Livezey and Zusi 2007), Crocodylidae (Meganathan et al. 2010), Testudines (Krenz et al. 2005; Naro-Maciel et al. 2008; Barley et al. 2010), Amphibia (Zardoya and Meyer 2001), Sarcopterygii (Zardoya and Meyer 1996), Actinopterygii (Meyer et al. 1994; Westneat and Alfaro 2005; Mabuchi et al. 2007; Espiñeira et al. 2008; Chiba et al. 2009; Gaubert et al. 2009; Little et al. 2010; Tang et al. 2010), Chondrichthyes (Rocco et al. 2007; Inoue et al. 2010; Vélez-Zuazo and Agnarsson 2011), and Vertebrata (Meyer and Zardoya 2003) were used to create the global phylogeny that included all species in our sample (supplementary information, Supplementary Material online). Using this vertebrate super tree, we obtained trees for each of the 221 human genes and their orthologs. Ancestral states on these trees were reconstructed using the codeml package in PAML (Yang 2007). For the analysis of the rate of irreversibility, we used only the mammalian tree because of a higher fraction of the most probable ancestral states in that tree with a high posterior probability in the lineage leading up to humans (98.7% of all states had >0.95 posterior probability). When we estimated the raw number of irreversible states on the phylogeny, we only considered those irreversible states that had a posterior probability on the placental tree of >0.95.

We calculated the fraction of irreversible states among all ancestral states at each ancestral node as average of $\sum P_i/(L-\sum P_a)$ across all genes where P_i was the posterior probability of all irreversible states in one gene, P_a was the posterior probability of all ancestral states that were identical to the human sequence, and L was the length of the gap-free multiple alignment of each gene. Table 2 reports data for the fraction of irreversible states across the mammalian phylogeny.

Correcting for the Sparseness of the Mutational Data

To estimate the true fraction of pathogenic mutations in each gene, we employed a method described in Kondrashov et al. (2002). We calculated S_m and S_n, the total number of observed missense and nonsense mutations in HGMD, respectively. Assume that all nonsense mutations in each gene in our data set (T_n) is pathogenic, which was calculated directly for each human gene from the nucleotide sequence. Assuming further that missense mutations and nonsense mutations are ascertained with an equal probability, the ascertainment depth of missense mutations is S_n/T_n. Thus, the total fraction of irreversible states was estimated as the average across all genes of the observed fraction of irreversible states divided by S_n/T_n, the ascertainment depth at each gene. For 10 genes with no loss-of-function mutations, we assumed that the ascertainment bias is equal to the average measure across the other 211 genes.

Correspondence between Nonsynonymous Polymorphisms and Mutation Data

Data on single-nucleotide polymorphisms (SNPs) were obtained from dbSNP (Sherry et al. 2001) with handles of TSC-CSHL, SC_SNP, SC, WUGSC_SSAHASNP, WI_SSAHASNP, SSAHASNP, CSHL-HAPMAP, BCM_SSAHASNP, and SC_JCM and from the May 2011 ALL.2of4intersection SNPs calls from the 1000 Genomes Project (The 1000 Genomes Project Consortium 2010), the latter data set included data on SNP frequencies.

McDonald–Kreitman Test

To estimate the fraction of substitutions fixed by positive selection, we used the McDonald–Kreitman test (McDonald and Kreitman 1991) as implemented by Smith and Eyre-Walker (2002) by sequence divergence data between available 218 human and macaque orthologs and the SNP data from 1000 Genomes Project. For each of the 218 pairwise orthologous alignments, constructed by MUSCLE (Edgar 2004), we used codeml from the PAML package (Yang 2007) to estimate the number of nonsynonymous and synonymous substitutions, D_n and D_s, respectively, and calculated the number of nonsynonymous and synonymous polymorphisms, P_n and P_s, respectively. We then estimated the fraction of amino acid substitutions under positive selection in the human–macaque

divergence as $\alpha = 1 - (\overline{Ds}/\overline{Dn})(\overline{Pn/(Ps+1)})$ (supplementary table S4, Supplementary Material online). We estimated α using polymorphisms with >5% and >1% frequency in the human population. For our final analysis, we used α estimated using the data for SNPs >5% frequency. We estimated the fraction of irreversible states among all phenotypically relevant substitutions as the estimated fraction of all irreversible states divided by α. For the subset of 51 genes with at least one irreversible state in the placental phylogeny, we estimated α using D_n, D_s, P_n, and P_s values obtained for their gene subset of 41 genes with a human–macaque ortholog.

Results

We obtained data on pathogenic mutations in humans from the HGMD (2011.1) (Stenson et al. 2009) and SwissVar (Mottaz et al. 2010) for 221 genes that had at least 50 missense mutations in HGMD (fig. 1) and obtained orthologous sequences from vertebrates available in GenBank (see Materials and Methods). We then aligned the orthologous amino acid sequences (Edgar 2004) and determined the posterior probabilities of ancestral states using a maximum likelihood approach (Yang 2007) for all genes along the mammalian phylogeny (see Materials and Methods). We found a total of 98 cases in 51 genes for HGMD data and 49 out of 23 genes for SwissVar data where a known pathogenic state corresponded to a reconstructed ancestral state in the phylogeny with >0.95 posterior probability (table 1, see Materials and Methods), representing ~1% of the total number of the ancestral states different from the human sequence across the entire mammalian phylogeny (fig. 2 and table 2). Similar numbers were obtained when considering only those ancestral states that were supported by more than one outgroup species, confirming the accuracy of the ancestral state reconstruction (supplementary table S1, Supplementary Material online).

In principle, ancestral amino acid states that are identical to known pathogenic human mutations indicate cases of irreversibility of genotype as presently such mutations cannot be fixed in the human population due to the drastically deleterious nature of the disease phenotypes. Thus, all cases of an ancestral amino acid states matching a known human pathogenic mutations may be considered irreversible states, meaning that a return to those ancestral states is impossible at present. To make sure that our data truly identify irreversible states and are not grossly affected by biases in the mutational data, we performed two additional tests. First, we took a subset of 57 genes with early onset pathology and high probability of fatal outcome in prereproductive age (supplementary table S2, Supplementary Material online) to test the possibility that the high fraction of ancestral states matching pathogenic mutations is a factor of pathogenic mutations with weak effects or those linked with late onset diseases. The fraction of ancestral states identical to the human pathogenic mutations for this subset of genes was similar (table 1), broadly rejecting the hypothesis that late onset pathogenic mutations, or those with weak effects, disproportionally contribute to our data.

Second, we considered the possibility that pathogenic mutations matching ancestral states may represent benign polymorphisms erroneously identified as pathogenic. We found that pathogenic mutations that match an ancestral state have a higher probability of corresponding to a known SNP than an average pathogenic mutation (table 3). However, only a small fraction of the identified putative irreversible states matched a state described as polymorphic. Moreover, genes linked with recessive disorders were more likely to show an irreversible state (23 out of 47 genes) relative to genes linked to dominant disorders or X-linked genes (14 out of 78, Fisher's *t*-test, $P < 0.006$; 59 genes were linked to diseases with an uncertain mode of inheritance) indicating that irreversible evolution is more likely to proceed when the deleterious state segregate at a higher frequency, perhaps through a compensatory mechanism described by Kimura (1985). We conclude that neither the weak phenotypic manifestation of pathogenic mutations or erroneous description of benign polymorphisms as pathogenic mutations substantially contributes to our data and that instances when an ancestral state matches a human pathogenic mutation can generally be considered to be irreversible states.

The number of known pathogenic mutations represents a small fraction of all possible pathogenic mutations, and, therefore, the estimate that ~ 1% of all ancestral states as irreversible is a proportionate underestimate. To obtain an estimate of the total rate of irreversibility among the genes in our data set, we used an existing approach for estimating the fraction of known pathogenic mutations (Kondrashov et al. 2002). This method works on the assumption that in genes with known pathogenic loss-of-function mutations, all nonsense mutations are pathogenic. Out of the 221 genes, only 10 contribute to disease only through change-of-function mutations (COMP, FGFR2, GFAP, NOTCH3, PSEN1, RYR2, SCN4A, STAT3, UMOD, and TTR) implying that we can apply this assumption to a majority of genes in our data set. For the 211 genes, we calculated the fraction of all possible nonsense mutations that have been described (on average 0.10, table 4, supplementary table S3, Supplementary Material online). This ratio indicates the fraction of pathogenic missense with the same degree of pathogenicity as nonsense mutations that have been described in HGMD (see Materials and Methods). Thus, only 10% of all possible pathogenic mutations have been described in HGMD for the genes in our data set, although this measure differed gene by gene (supplementary table S2, Supplementary Material online).

The estimate of the fraction of described pathogenic mutations can be used to obtain the total rate of irreversibility for genes in our data set. However, in a large fraction of genes, we found zero irreversible states. Because of the sparseness of the available data on pathogenic mutations, it is not clear

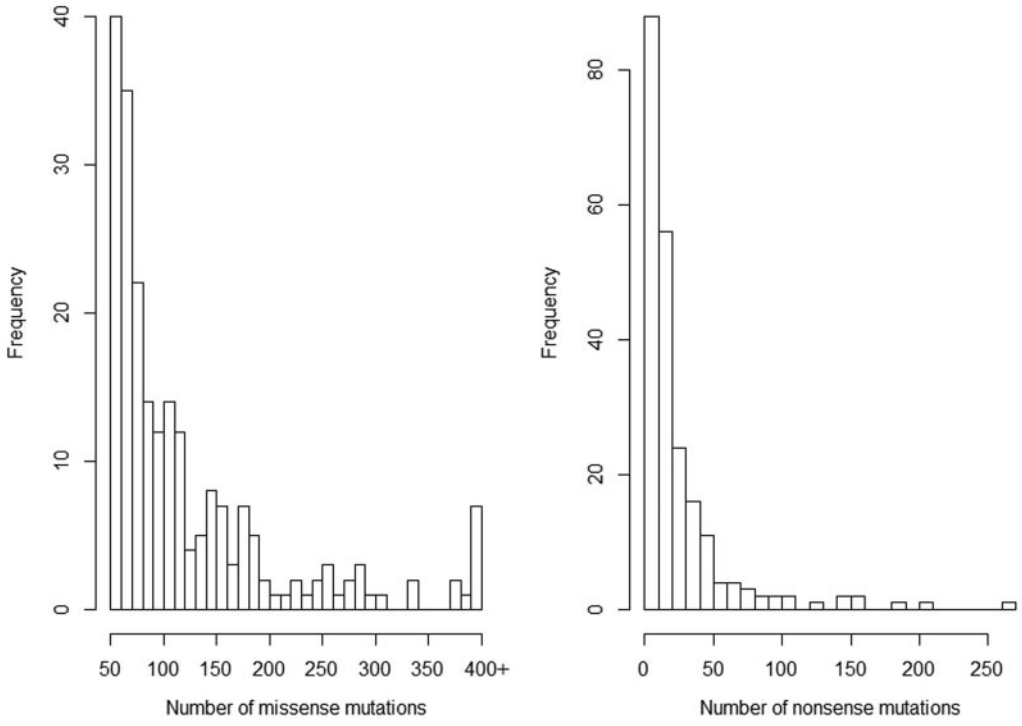

Fig. 1.—Distribution of pathogenic mutations in our data set. Two histograms show the distribution of missense and nonsense mutations listed in HGMD for 221 genes included in our study.

Table 1

Number and Fraction of Irreversible States that Represent Cases of a Pathogenic Mutation Matching an Ancestral State that Has a >0.95 Posterior Probability in the Mammalian Phylogeny

	HGMD	SwissVar	HGMD with Severe Phenotypes	SwissVar with Severe Phenotypes
Total number of pathogenic mutations (genes)	29,143 (221)	14,481 (221)	6,892 (57)	4,490 (57)
Number of irreversible states (genes)	98 (51)	49 (23)	22 (11)	16 (5)

Table 2

Estimating the Total Fraction of Irreversible States Out of Phenotypically Relevant Substitutions on the Placental Phylogeny

	Fraction of Irreversible States Out of All Substitutions	Estimated Total Fraction of Irreversible States Out of All Substitutions	Corrected Total Fraction of Irreversible States among Beneficial Substitutions
HGMD data for all genes in the data set	1.2% (221)	10.5% (221)	18.4% (221)
HGMD data for genes with at least one irreversible state in the placental phylogeny	4.8% (51)	43% (51)	81.1% (51)
SwissVar data for all genes in the data set	0.8% (221)	5.3% (221)	9.3% (221)
SwissVar data for genes with at least one irreversible state in the placental phylogeny	7.2% (23)	45.4% (23)	85.6% (23)

Note.—The estimated total fraction of irreversible states was obtained by correcting for the sparseness of the mutational data as described in section Materials and Methods. The fraction of irreversible states among beneficial substitution is estimated by dividing the value in the second column by α (0.57 for 221 genes and 0.53 for 51 genes), the fraction of beneficial substitutions in evolution from the McDonald–Kreitman test (supplementary table S4, Supplementary Material online).

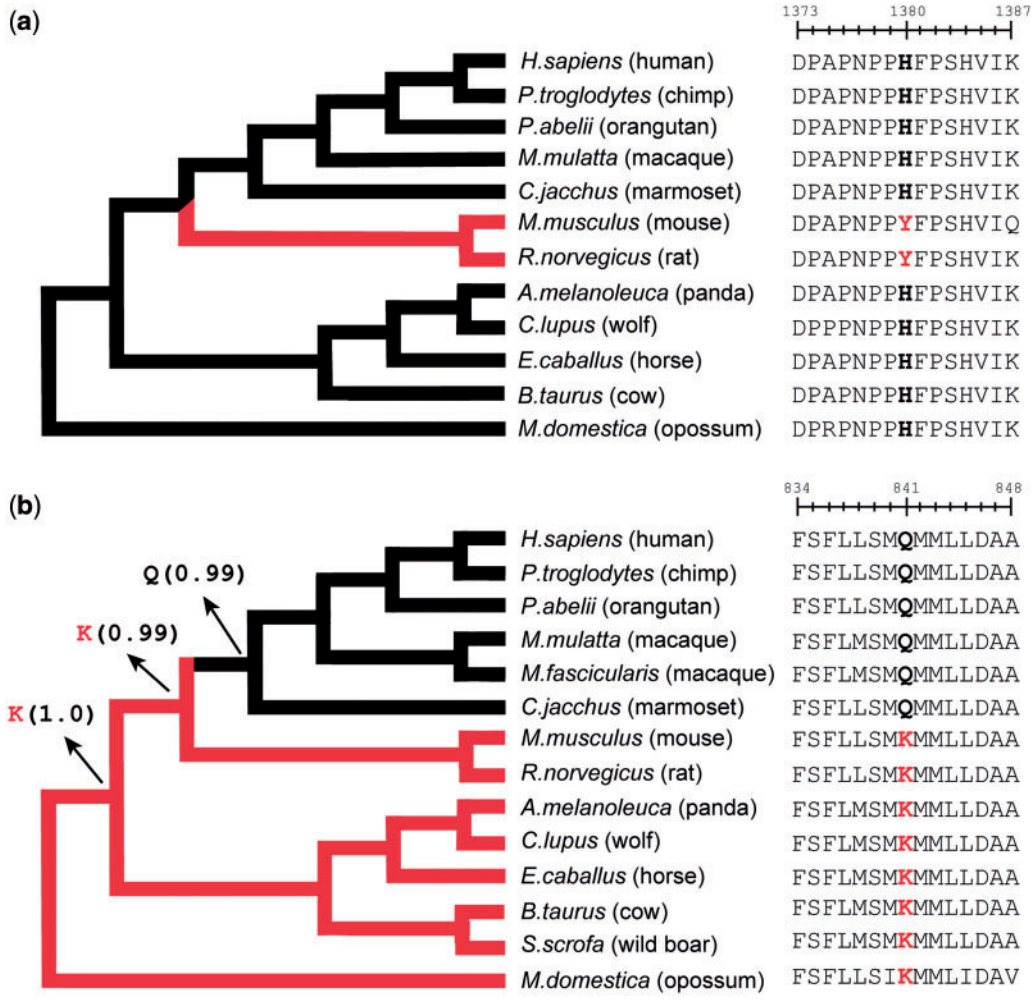

Fig. 2.—Amino acid states that match known human pathogenic mutations. States that match known human pathogenic mutations were found in closely related species without the drastic phenotypic manifestations observed in humans. Such amino acid states may be confined to a sister clade (*a*) and, therefore, while being indicative of compensatory evolution (Kondrashov et al. 2002) do not represent cases of evolutionary irreversibility. Alternatively, some pathogenic mutations may match the ancestral state at some point along the phylogenetic lineage leading to the human branch (*b*), and such cases necessarily represent cases of evolutionary irreversibility on the genotype level as such mutations cannot be currently fixed in the human population. For both cases, H1380Y in ATM gene leading to breast cancer susceptibility (*a*) and Q841K in ABCA4 gene leading to Stargardt disease (*b*) the multiple alignment with seven amino acids on each side of the site in question is shown. Human wild-type and pathogenic states are represented by black and red colors, respectively. Posterior probabilities associated with irreversible states are shown in parentheses.

Table 3
Pathogenic Mutations Matching Known Human SNPs from 1000 Genomes Data and dbSNP

	HGMD Mutations Matching a Known SNP (%)	Irreversible States from HGMD Matching a Known SNP (%)	SwissVar Mutations Matching a Known SNP (%)	Irreversible States from SwissVar Matching a Known SNP (%)
1000 genomes data	368/29,143 (1.26)	15/98 (15.3)	181/14,481 (1.25)	4/49 (8.1)
dbSNP	85/29,143 (0.29)*	6/98 (6.6)*	50/14,481 (0.33)*	4/49 (8.1)*

*All comparisons were statistically significant (Fisher's exact test, $P < 0.0001$).

whether such genes never evolve in an irreversible fashion, or zero irreversible states in these genes were observed due to lack of data on pathogenic mutations. In the latter case, taking into account genes with zero irreversible states would substantially underestimate the estimated total rate of irreversibility when correcting for the sparseness of the mutational data. We thus tested whether or not 51 genes with at least one identified irreversible state are representative of the rate

Table 4
Pathogenic Mutation Data and Estimated Fraction of Missense Mutations that Are Pathogenic

	Missense Mutations	Nonsense Mutations
Average number of mutations	131.9 (123.4)	26.5 (38.6)
Average number of total possible mutations	5,498.8 (5,050.1)	332.3 (330.5)
Estimated average number of pathogenic mutations	2,133 (2,704.1)	332.3 (330.5)
Average fraction out of all possible mutations	0.037 (0.042)	0.099 (0.102)
Average fraction of described mutations out of estimated pathogenic mutations	0.107 (0.99)	0.099 (0.102)
Estimated average fraction of pathogenic mutations among all possible mutations	0.431 (0.253)	1.00 (N.A.)

NOTE.—The table reports averages across 221 genes in our data sets and standard deviations in parentheses.

of irreversibility of the entire set. For each gene, we obtained a phylogeny spanning the vertebrate clade and identified irreversible states on nodes before the evolution of placental mammals. We found that genes with irreversible states in placental mammals were more likely to have another irreversible state deeper in the phylogeny compared with genes with zero irreversible states in the placental lineage (35 vs. 55 irreversible states in nonplacental nodes for 170 and 51 genes, respectively, P value < 0.001, Fisher's exact test) indicating that the rate of irreversibility of genes with at least one irreversible state in the placental phylogeny is higher overall but genes with zero irreversible states in the placental phylogeny can also evolve in an irreversible fashion. Thus, we estimated the expected total rate of irreversible evolution using two separate sets of genes, all genes in our data set (221 genes), and genes that showed at least one ancestral state in the placental phylogeny (51 genes). The latter estimate is representative only for ~20% of genes with especially high rates of irreversible evolution; however, the total rate of irreversibility of an average gene is likely to be in between these two estimates.

Assuming that known and undescribed pathogenic mutations have an equal probability of being found as an ancestral state, we estimated the overall rate of irreversible evolution. To obtain this estimate, we divided the fraction of irreversible states by the observed fraction of known nonsense pathogenic mutations for each gene (see Materials and Methods). The resulting fraction of irreversible ancestral states was ~10 times larger, such that for all genes in our data set, a total of ~10% of all ancestral states in the placental phylogeny were irreversible, whereas for a subset of genes with particularly high rates of irreversible evolution, this fraction was 43% (table 2).

Dollo's law is concerned with the irreversibility of evolutionary innovations on a phenotypic level. Keeping with the spirit of Dollo's law, we estimated the fraction of irreversible evolution among phenotypically relevant amino acid substitutions. We used the McDonald–Kreitman test (Smith and Eyre-Walker 2002) to estimate the fraction of amino acids in evolution that were selectively neutral and likely to be phenotypically silent (see Materials and Methods). Excluding rare SNPs (Smith and Eyre-Walker 2002) from the polymorphism data (The 1000 Genomes Project Consortium 2010) for each gene, we estimated the fraction of neutral amino acid substitutions since the human–macaque divergence (supplementary table S4, Supplementary Material online, Materials and Methods). We found that approximately half of all amino acid substitutions since the human–macaque divergence evolved under positive selection (0.57 and 0.53 on average for 221 and 51 gene data sets, respectively). Thus, we estimate that for an average gene, 18.4% of all phenotypically relevant amino acid substitutions are irreversible, whereas for our subset of 51 genes with a higher rate of evolution up to 80% of all such substitutions are irreversible (table 2).

Discussion

The presence of allele states pathogenic to humans in genomes of other species without any apparent deleterious consequences has been described previously as compensated pathogenic deviations (CPDs, Schaner 2001; Kondrashov et al. 2002; Gao and Zhang 2003; Ferrer-Costa et al. 2007; Baresić et al. 2010). The analysis presented here uses a similar approach to estimate the fraction of irreversible evolution by relating data on CPDs to their phylogenetic distribution, which led to an estimate of the fraction of irreversible substitutions in protein evolution. Additionally, the present analysis is based on seven times more genes with at least 50 missense mutations than was available for the original analysis of Kondrashov et al. (2002) a decade ago.

The fraction of irreversible states between two genomes is expected to increase with divergence as long as they represent cases of Dobzhansky–Muller incompatibilities (Dobzhansky 1936; Muller 1939; Orr 1995; Kondrashov et al. 2002; Povolotskaya and Kondrashov 2010). Alternatively, if the accumulation of irreversible states is driven by positive selection (Kondrashov et al. 2002), or they represent cases where a combination of two deleterious alleles make a neutral state (Kimura 1985), their fraction may be constant with sequence divergence (Kondrashov et al. 2002; Kulathinal et al. 2004; Meer et al. 2010). We estimated the fraction of irreversible states relative to genetic distance between humans and the common ancestor of six clades of placental mammals: Hominini, Hominidae, Catarrhini, Simiiformes, Euarchontoglires,

and Boreoeutheria (fig. 3). We found a lower fraction of irreversible states in the common ancestors within the great apes but a constant fraction of irreversible states across nodes deeper in the phylogeny making it unclear whether these data favor one type of mechanism over the others (Kimura 1985; Orr 1995; Kondrashov et al. 2002). However, our data show that the rate of irreversibility was a constant factor in evolution, and its estimate does not depend on genetic distance beyond a minimal level of sequence divergence.

The reason why a large fraction of amino acid states are irreversible may be explained by two nonmutually exclusive mechanisms. An adaptive explanation holds that irreversible states represent cases of adaptation to different environmental conditions between the extant and ancestral species. Thus, a mutation is pathogenic because it represents a reversal to a state that was adapted to a different environmental condition that was realized among ancestral species (Bull and Charnov 1985; Chippindale et al. 2004; Collin and Miglietta 2008; Bridgham et al. 2009). At the heart of the second mechanism lies the molecular nature of the genotype–phenotype relationship that leads to epistatic interactions among amino acid states within the same protein or between interacting proteins. If protein evolution is a walk in sequence space (Maynard Smith 1970) along a rugged fitness landscape (Wright 1932) then as amino acid substitutions accumulate in the course of evolution they affect each other's fitness impact (Kondrashov et al. 2002; DePristo et al. 2005; Povolotskaya and Kondrashov 2010). Thus, one substitution may be pathogenic in one genetic context, whereas being benign in another, leading to cases where an amino acid state was benign in an ancestor and became pathogenic further along the evolutionary lineage due to other, compensatory, changes in the same protein (Kondrashov et al. 2002; Choi et al. 2005; DePristo et al. 2005; Poon and Chao 2005; Bridgham et al. 2009; Baresic et al. 2010; Lunzer et al. 2010; de Visser et al. 2011; Tan et al. 2011) or its interaction partners (Kondrashov et al. 2002; Tokuriki and Tawfik 2009; Burga et al. 2011; de Visser et al. 2011; Poelwijk et al. 2011).

The drastic phenotypic manifestation of many pathogenic mutations indicates that the adaptive mechanism is unlikely to explain a substantial fraction of the irreversible states in our data. Pathogenic phenotypes are unlikely to resemble ancestral phenotypes, and, therefore, pathogenic phenotypes do not represent reversals to the ancestral phenotypes. Thus, the respective amino acid substitutions on the placental phylogeny are unlikely to represent cases of adaptive evolution.

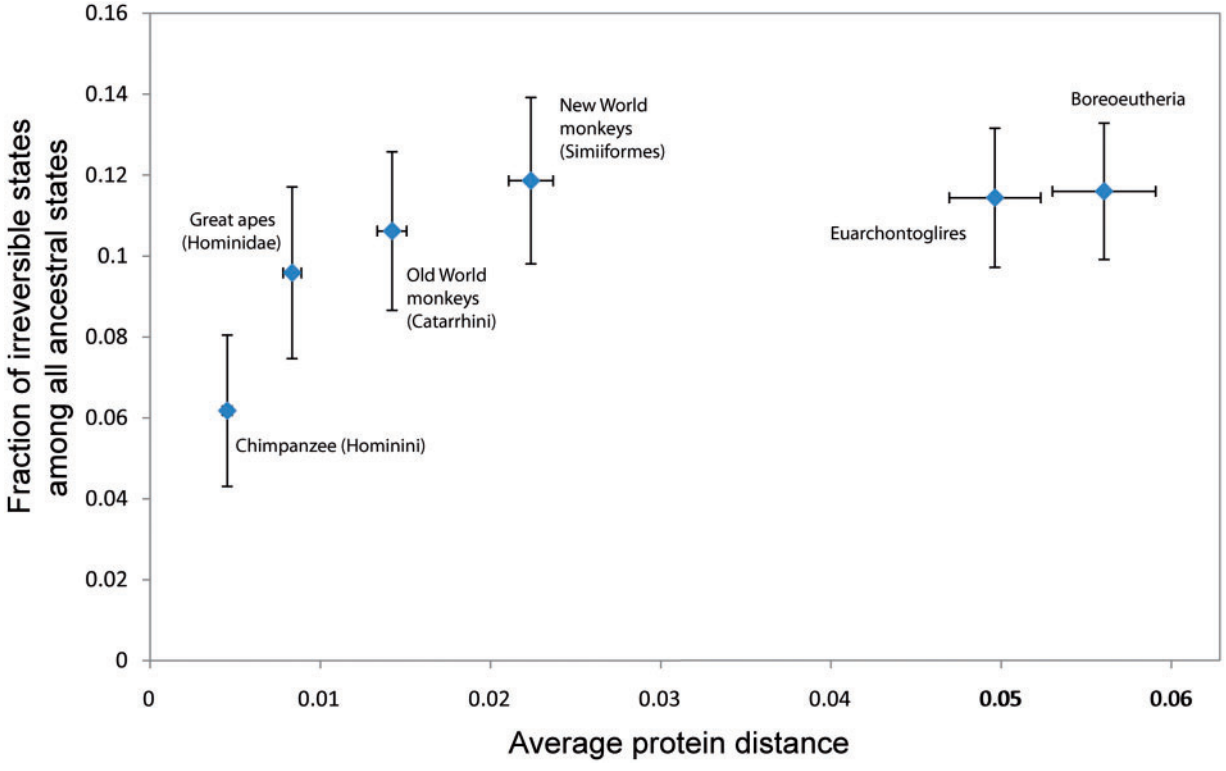

Fig. 3.—Estimating the overall rate of amino acid irreversibility along the placental mammal phylogeny. We estimated the fraction of irreversible states between the amino acid sequence of humans and their common ancestor in six clades of placental mammals. Protein distance was calculated as the fraction of different amino acid states between the human sequences and that in the ancestral node. Vertical (horizontal) error bars represent standard error for the fraction of irreversible states (protein distance).

In contrast, the proposed compensatory nature of protein evolution has a strong basis in our structural (Kimura 1985; Kondrashov et al. 2002; Choi et al. 2005; DePristo et al. 2005; Poon and Chao 2005; Bridgham et al. 2009; Baresic et al. 2010; Lunzer et al. 2010; Povolotskaya and Kondrashov 2010; Tan et al. 2011) and functional (Poon and Chao 2005; Bridgham et al. 2009; Tokuriki and Tawfik 2009; Lunzer et al. 2010; Burga et al. 2011; de Visser et al. 2011, Poelwijk et al. 2011) understanding of proteins and their network interactions. It is, therefore, probable that the irreversibility of amino acid states is a consequence of evolution among a rugged fitness landscape, such that the irreversibility of genotypes is not a predetermined consequence of adaptive evolution but rather a probabilistic walk through sequence space among a rugged fitness landscape (Maynard Smith 1970; Kimura 1985; Orr 1995; Kondrashov et al. 2002; DePristo et al. 2005, 2007; Meer et al. 2010; Povolotskaya and Kondrashov 2010; de Visser et al. 2011; Tan et al. 2011). In that case, once an amino acid substitution occurs, the probability that this event becomes irreversible is a linearly increasing function of the number of other substitutions in the same protein or the entire genotype (Maynard Smith 1970; Orr 1995; Kondrashov et al. 2002; Povolotskaya and Kondrashov 2010; Tan et al. 2011). Similarly, if a particular substitution is observed to be irreversible at a given moment, the maintenance of its irreversibility in the future is a probabilistic function of potential compensatory substitutions. Estimating the probability that a currently irreversible amino acid state becomes available to evolution in the future may be the next question in the study of irreversibility on the genotype level.

Our estimate of the propensity of irreversible evolution on the molecular level of protein sequence takes into consideration and corrects for the incomplete mutational data and the estimate of proportion of phenotypically relevant amino acid substitutions. Potential problems with these corrections may bias our estimate. First, genes included in our data set are selected on the basis of number of pathogenic mutations they harbor and may not be representative of the rates of irreversibility found in an average gene in the genome. Second, the results of the McDonald–Kreitman test are surprising as the numbers are higher than previously reported (Gojobori et al. 2007; Boyko et al. 2008) although consistent with observations for other species (Andolfatto 2005; Eyre-Walker 2006; Eyre-Walker and Keightley 2009; Halligan et al. 2010). One possibility is that pervasive epistasis biases the McDonald–Kreitman test toward overestimating the true fraction of positive selection in evolution, and, therefore, the test may not be appropriate if epistasis is a common feature of protein evolution. To our knowledge, the issue of the impact of epistasis on the McDonald–Kreitman test has not been considered either empirically or theoretically. Finally, substantial biases in the relative rates of reporting of nonsense to missense mutations may bias our correction for the sparseness of missense mutational data. However, we are not aware of estimates for the propensity of under- or over-reporting of either types of pathogenic mutations. At present, we cannot make an estimate of how such potential bias may affect our results.

Pathogenic mutations in our data set represent cases of intermediate impact on phenotype. Many mutations are slightly deleterious in that they are highly unlikely to achieve fixation but have only a mild effect on phenotype, such that they are unlikely to be registered as pathogenic mutations. Similarly, an unknown fraction of mutations is lethal and not ascertained as pathogenic. If mechanisms of irreversible evolution apply to lethal or slightly deleterious alleles then the high fraction of irreversible states reported here is a lower bound estimate.

Supplementary Material

Supplementary tables S1–S4 are available at *Genome Biology and Evolution* online (http://www.gbe.oxfordjournals.org/).

Acknowledgments

This work was supported by Plan Nacional grant BFU2009-09271 from the Spanish Ministry of Science and Innovation and by FPU (Formación del Profesorado Universitario) program grant AP2008-01888 from the Spanish Ministry of Education to O.S. F.A.K. is a European Molecular Biology Organization Young Investigator and Howard Hughes Medical Institute International Early Career Scientist.

Literature Cited

Andolfatto P. 2005. Adaptive evolution of non-coding DNA in Drosophila. Nature 437(7062):1149–1153.

Arnason U, Gullberg A, Janke A. 2004. Mitogenomic analyses provide new insights into cetacean origin and evolution. Gene 333:27–34.

Baresić A, Hopcroft LE, Rogers HH, Hurst JM, Martin AC. 2010. Compensated pathogenic deviations: analysis of structural effects. J Mol Biol. 396:19–30.

Barley AJ, Spinks PQ, Thomson RC, Shaffer HB. 2010. Fourteen nuclear genes provide phylogenetic resolution for difficult nodes in the turtle tree of life. Mol Phylogenet Evol. 55:1189–1194.

Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Wheeler DL. 2006. GenBank. Nucleic Acids Res. 34:D16–D20.

Boyko AR, et al. 2008. Assessing the evolutionary impact of amino acid mutations in the human genome. PLoS Genet. 4:1–13.

Bridgham JT, Ortlund EA, Thornton JW. 2009. An epistatic ratchet constrains the direction of glucocorticoid receptor evolution. Nature 461: 515–519.

Bull JJ, Charnov EL. 1985. On irreversible evolution. Evolution 39: 1149–1155.

Burga A, Casanueva MO, Lehner B. 2011. Predicting mutation outcome from early stochastic variation in genetic interaction partners. Nature 480:250–253.

Chatterjee HJ, Ho SY, Barnes I, Groves C. 2009. Estimating the phylogeny and divergence times of primates using a supermatrix approach. BMC Evol Biol. 9:259.

Chiba SN, Iwatski Y, Yoshino T, Hanzawa N. 2009. Comprehensive phylogeny of the family Sparidae (Perciformes: Teleostei) inferred from mitochondrial gene analyses. Genes Genet Syst. 84:153–170.

Chippindale PT, Bonett RM, Baldwin AS, Wiens JJ. 2004. Phylogenetic evidence for a major reversal of life-history evolution in plethodontid salamanders. Evolution 58:2809–2822.
Choi SS, Li W, Lahn BT. 2005. Robust signals of coevolution of interacting residues in mammalian proteomes identified by phylogeny-aided structural analysis. Nat Genet. 37:1367–1371.
Collin R, Miglietta MP. 2008. Reversing opinions on Dollo's law. Trends Ecol Evol. 23:602–609.
DePristo MA, Hartl DL, Weinreich DM. 2007. Mutational reversions during adaptive protein evolution. Mol Biol Evol. 24:1608–1610.
DePristo MA, Weinreich DM, Hartl DL. 2005. Missense meanderings in sequence space: a biophysical view of protein evolution. Nat Rev Genet. 6:678–687.
Dobzhansky T. 1936. Studies on hybrid sterility. II. Localization of sterility factors in Drosophila pseudoobscura hybrids. Genetics 21:113–135.
de Visser JA, Cooper TF, Elena SF. 2011. The causes of epistasis. Proc Biol Sci. 278:3617–3624.
Dimcheff DE, Drovetski SV, Mindell DP. 2002. Phylogeny of Tetraoninae and other galliform birds using mitochondrial 12S and ND2 genes. Mol Phylogenet Evol. 24:203–215.
Dollo L. 1893. The laws of evolution. Bull Soc Bel Geol Paleontol. 7: 164–166.
Edgar RC. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 32:1792–1797.
Espiñeira M, González-Lavín N, Vieites JM, Santaclara FJ. 2008. Development of a method for the genetic identification of flatfish species on the basis of mitochondrial DNA sequences. J Agric Food Chem. 56:8954–8961.
Eyre-Walker A. 2006. The genomic rate of adaptive evolution. Trends Ecol Evol. 21:569–575.
Eyre-Walker A, Keightley PD. 2009. Estimating the rate of adaptive molecular evolution in the presence of slightly deleterious mutations and population size change. Mol Biol Evol. 26:2097–2108.
Fabre PH, Rodrigues A, Douzery EJ. 2009. Patterns of macroevolution among Primates inferred from a supermatrix of mitochondrial and nuclear DNA. Mol Phylogenet Evol. 53:808–825.
Ferrer-Costa C, Orozco M, de la Cruz X. 2007. Characterization of compensated mutations in terms of structural and physico-chemical properties. J Mol Biol. 365:249–256.
Gao L, Zhang J. 2003. Why are some human disease-associated mutations fixed in mice? Trends Genet. 19:678–681.
Gaubert P, Denys G, Oberdorff T. 2009. Genus-level supertree of Cyprinidae (Actinopterygii: Cypriniformes), partitioned qualitative clade support and test of macro-evolutionary scenarios. Biol Rev Camb Philos Soc. 84:653–689.
Gojobori J, Tang H, Akey JM, Wu CI. 2007. Adaptive evolution in humans revealed by the negative correlation between the polymorphism and fixation phases of evolution. Proc Natl Acad Sci U S A. 104: 3907–3912.
Halligan DL, Oliver F, Eyre-Walker A, Harr B, Keightley PD. 2010. Evidence for pervasive adaptive protein evolution in wild mice. PLoS Genet. 6(1): e1000825.
Hernández Fernández M, Vrba ES. 2005. A complete estimate of the phylogenetic relationships in Ruminantia: a dated species-level supertree of the extant ruminants. Biol Rev Camb Philos Soc. 80: 269–302.
Higdon JW, Bininda-Emonds OR, Beck RM, Ferguson SH. 2007. Phylogeny and divergence of the pinnipeds (Carnivora: Mammalia) assessed using a multigene dataset. BMC Evol Biol. 7:216.
Inoue JG, et al. 2010. Evolutionary origin and phylogeny of the modern holocephalans (Chondrichthyes: Chimaeriformes): a mitogenomic perspective. Mol Biol Evol. 27:2576–2586.
Jones KE, Bininda-Emonds OR, Gittleman JL. 2005. Bats, clocks, and rocks: diversification patterns in Chiroptera. Evolution 59:2243–2255.
Jones KE, Purvis A, MacLarnon A, Bininda-Emonds OR, Simmons NB. 2002. A phylogenetic supertree of the bats (Mammalia: Chiroptera). Biol Rev Camb Philos Soc. 77:223–259.
Kimura M. 1985. The role of compensatory neutral mutations in molecular evolution. J Genet. 64:7–19.
Krenz JG, Naylor GJ, Shaffer HB, Janzen FJ. 2005. Molecular phylogenetics and evolution of turtles. Mol Phylogenet Evol. 37:178–191.
Kondrashov AS, Sunyaev S, Kondrashov FA. 2002. Dobzhansky-Muller incompatibilities in protein evolution. Proc Natl Acad Sci U S A. 99: 14878–14883.
Kulathinal RJ, Bettencourt BR, Hartl DL. 2004. Compensated deleterious mutations in insect genomes. Science 306:1553–1554.
Little AG, Lougheed SC, Moyes CD. 2010. Evolutionary affinity of billfishes (Xiphiidae and Istiophoridae) and flatfishes (Plueronectiformes): independent and trans-subordinal origins of endothermy in teleost fishes. Mol Phylogenet Evol. 56:897–904.
Livezey BC, Zusi RL. 2007. Higher-order phylogeny of modern birds (Theropoda, Aves: Neornithes) based on comparative anatomy. II. Analysis and discussion. Zool J Linn Soc. 149:1–95.
Lunzer M, Golding GB, Dean AM. 2010. Pervasive cryptic epistasis in molecular evolution. PLoS Genet. 6:e1001162.
Mabuchi K, Miya M, Azuma Y, Nishida M. 2007. Independent evolution of the specialized pharyngeal jaw apparatus in cichlid and labrid fishes. BMC Evol Biol. 7:10.
Maynard Smith J. 1970. Natural selection and the concept of a protein space. Nature 225:563–564.
McDonald JH, Kreitman M. 1991. Adaptive protein evolution at the Adh locus in Drosophila. Nature 351:652–654.
Meganathan PR, Dubey B, Batzer MA, Ray DA, Haque I. 2010. Molecular phylogenetic analyses of genus Crocodylus (Eusuchia, Crocodylia, Crocodylidae) and the taxonomic position of Crocodylus porosus. Mol Phylogenet Evol. 57:393–402.
Meer MV, Kondrashov AS, Artzy-Randrup Y, Kondrashov FA. 2010. Compensatory evolution in mitochondrial tRNAs navigates valleys of low fitness. Nature 464:279–282.
Meyer A, Morrissey JM, Schartl M. 1994. Recurrent origin of a sexually selected trait in Xiphophorus fishes inferred from a molecular phylogeny. Nature 368:539–542.
Meyer A, Zardoya R. 2003. Recent advances in the (molecular) phylogeny of vertebrates. Annu Rev Ecol Evol Syst. 34:311–338.
Mottaz A, David FP, Veuthey AL, Yip YL. 2010. Easy retrieval of single amino-acid polymorphisms and phenotype information using SwissVar. Bioinformatics 26:851–852.
Muller HJ. 1939. Reversibility in evolution considered from the standpoint of genetics. Biol Rev Camb Philos Soc. 14:261–280.
Naro-Maciel E, Le M, FitzSimmons NN, Amato G. 2008. Evolutionary relationships of marine turtles: a molecular phylogeny based on nuclear and mitochondrial genes. Mol Phylogenet Evol. 49:659–662.
Nilsson MA, et al. 2010. Tracking marsupial evolution using archaic genomic retroposon insertions. PLoS Biol. 8:e1000436.
Online Mendelian Inheritance in Man, OMIM. Baltimore (MD): McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University. Available from: http://omim.org/ (last accessed September 2011).
Orr HA. 1995. The population genetics of speciation: the evolution of hybrid incompatibilities. Genetics 139:1805–1813.
Poelwijk FJ, de Vos MG, Tans SJ. 2011. Tradeoffs and optimality in the evolution of gene regulation. Cell 146:462–470.
Poon A, Chao L. 2005. The rate of compensatory mutation in the DNA bacteriophage phiX174. Genetics 170:989–999.
Povolotskaya IS, Kondrashov FA. 2010. Sequence space and the ongoing expansion of the protein universe. Nature 465:922–926.
Price SA, Bininda-Emonds OR, Gittleman JL. 2005. A complete phylogeny of the whales, dolphins, and even-toed hoofed mammals (Cetartiodactyla). Biol Rev Camb Philos Soc. 80:445–473.

Rocco L, et al. 2007. Molecular and karyological aspects of Batoidea (Chondrichthyes, Elasmobranchi) phylogeny. Gene 389:80–86.
Schaner P, et al. 2001. Episodic evolution of pyrin in primates: human mutations recapitulate ancestral amino acid states. Nat Genet. 27: 318–321.
Sherry ST, et al. 2001. dbSNP: the NCBI database of genetic variation. Nucleic Acids Res. 29:308–311.
Smith NG, Eyre-Walker A. 2002. Adaptive protein evolution in Drosophila. Nature 415:1022–1024.
Steiner C, Tilak MK, Douzery EJ, Catzeflis FM. 2005. New DNA data from a transthyretin nuclear intron suggest an oligocene to miocene diversification of living South America opossums (Marsupialia: Didelphidae). Mol Phylogenet Evol. 35:363–379.
Stenson PD, et al. 2009. The Human Gene Mutation Database: 2008 update. Genome Med. 1:13.
Tan L, Serene S, Chao HX, Gore J. 2011. Hidden randomness between fitness landscapes limits reverse evolution. Phys Rev Lett. 106: 198102.
Tanaka K, et al. 2011. Polymorphisms in the bovine hemoglobin-beta gene provide evidence for gene-flow between wild species of Bos (Bibos) and domestic cattle in Southeast Asia. Anim Sci J. 82:36–45.
Tang KL, et al. 2010. Systematics of the subfamily Danioninae (Teleostei: Cypriniformes: Cyprinidae). Mol Phylogenet Evol. 57:189–214.
Tatusov RL, Koonin EV, Lipman DJ. 1997. A genomic perspective on protein families. Science 278:631–637.
Teotónio H, Rose MR. 2001. Perspective: reverse evolution. Evolution 55: 653–660.
1000 Genomes Project Consortium. 2010. A map of human genome variation from population-scale sequencing. Nature 467:1061–1073.
Tokuriki N, Tawfik DS. 2009. Chaperonin overexpression promotes genetic variation and enzyme evolution. Nature 459:668–673.
Vélez-Zuazo X, Agnarsson I. 2011. Shark tales: a molecular species-level phylogeny of sharks (Selachimorpha, Chondrichthyes). Mol Phylogenet Evol. 58:207–217.
Westneat MW, Alfaro ME. 2005. Phylogenetic relationships and evolutionary history of the reef fish family Labridae. Mol Phylogenet Evol. 36: 370–390.
Wright S. 1932. The roles of mutation, inbreeding, crossbreeding, and selection in evolution. In: Jones DF, editor. Proceedings of the Sixth International Congress on Genetics. Vol. 1. Bethesda (MD): Genetics Society of America. p. 356–366.
Yang Z. 2007. PAML 4: phylogenetic analysis by maximum likelihood. Mol Biol Evol. 24:1586–1591.
Zardoya R, Meyer A. 1996. Evolutionary relationships of the coelacanth, lungfishes, and tetrapods based on the 28S ribosomal RNA gene. Proc Natl Acad Sci U S A. 93:5449–5454.
Zardoya R, Meyer A. 2001. On the origin of and phylogenetic relationships among living amphibians. Proc Natl Acad Sci U S A. 98: 7380–7383.
Zufall RA, Rausher MD. 2004. Genetic changes associated with floral adaptation restrict future evolutionary potential. Nature 428: 847–850.

Associate editor: George Zhang