Relaxed Purifying Selection and Possibly High Rate of Adaptation in Primate Lineage-Specific Genes

James J. Cai* and Dmitri A. Petrov
Department of Biology, Stanford University
*Corresponding author: E-mail: jamescai@stanford.edu.

Accepted: 2 May 2010

Abstract
Genes in the same organism vary in the time since their evolutionary origin. Without horizontal gene transfer, young genes are necessarily restricted to a few closely related species, whereas old genes can be broadly distributed across the phylogeny. It has been shown that young genes evolve faster than old genes; however, the evolutionary forces responsible for this pattern remain obscure. Here, we classify human–chimp protein-coding genes into different age classes, according to the breadth of their phylogenetic distribution. We estimate the strength of purifying selection and the rate of adaptive selection for genes in different age classes. We find that older genes carry fewer and less frequent nonsynonymous single-nucleotide polymorphisms than younger genes suggesting that older genes experience a stronger purifying selection at the protein-coding level. We infer the distribution of fitness effects of new deleterious mutations and find that older genes have proportionally more slightly deleterious mutations and fewer nearly neutral mutations than younger genes. To investigate the role of adaptive selection of genes in different age classes, we determine the selection coefficient ($c = 2N_e s$) of genes using the MKPRF approach and estimate the ratio of the rate of adaptive nonsynonymous substitution to synonymous substitution ($x_A$) using the DoFE method. Although the proportion of positively selected genes ($\gamma > 0$) is significantly higher in younger genes, we find no correlation between $x_A$ and gene age. Collectively, these results provide strong evidence that younger genes are subject to weaker purifying selection and more tenuous evidence that they also undergo adaptive evolution more frequently.

Key words: lineage-specific genes, nonsynonymous polymorphism, evolutionary age, adaptive selection.

Introduction
Fully sequenced genomes from a wide range of species allow us to determine the phylogenetic distribution of protein-coding genes in the genomes of these species. The phylogenetic distribution of a gene contains information about the gene’s evolutionary age (i.e., the time when the gene first appeared in some ancestral genomes) and the gene’s propensity to persist in genomes. Without horizontal gene transfer, genes with broader and deeper phylogenetic distributions are necessarily older and more persistent than genes that are originated recently or do not tend to persist for long periods. Young genes, also termed lineage-specific genes, tend to have either restricted or patchy phylogenetic distributions.


Despite the same pattern repeatedly found in various organisms, the underlying evolutionary forces responsible for such a phenomenon remain obscure. Specifically, it is not clear whether the anticorrelation between evolutionary age and protein-divergence rate are due to the variation in the strength of purifying selection or due to the variation in the rate of adaptive evolution. Distinguishing these two causes is of fundamental importance and provides clues...
about the role of newly created genes. Weaker purifying selection in young or lineage-specific genes would imply that these genes are less “important” in the sense that defects in these genes have less effect on fitness. The alternative, although not a mutually exclusive, possibility is that genes recently added to the genome participate more in the lineage-specific adaptive evolution.

Here, we study the molecular evolution of genes in humans and chimps through the window of phylogenetic profile of these genes. We investigate all human–chimp genes over the same evolutionary distance in way that is both consistent and avoids problems with saturation. Particularly, comparing sequences of human and chimp orthologs in conjunction with interrogating sequence polymorphisms in humans, we estimate the rates of evolution at synonymous and nonsynonymous sites and the levels of selective constraint for all human–chimp protein-coding genes. To estimate relative prevalence of positive selection of different age classes, we calculate the ratio of adaptive nonsynonymous-to-synonymous substitution rates (Eyre-Walker and Keightley 2009) and the scaled selection coefficient (Bustamante et al. 2002, 2005).

To classify genes, we use three phylogenetic measures: lineage specificity (LS), phylostratum level (PL), and the number of gene losses (GLs). Each measure represents a different age-classifying system, capturing a unique feature of phylogenetic profiles of genes. LS measures the breadth and depth of the phylogenetic profiles but focuses only on genes that have nonpatchy phylogenetic distributions (Cai, Woo, et al. 2006). PL focuses on homologs and determines the age of the gene family by strict parsimony that assumes that a gene family can be lost but cannot reevolve independently in different lineages or be horizontally transferred (Domazet-Loso et al. 2007). GL captures the patchiness of phylogenetic profiles for genes that have the same age measured using strict parsimony. We obtain qualitatively identical results using all three measures of gene age.

We confirm that younger and less broadly distributed genes evolve faster at the protein level. We determine that these genes are subject to weaker purifying selection in humans and provide some evidence that positive selection does play a role in the faster evolution of younger genes. We discuss implications of these results for the understanding of human evolution and human health. We also put these results in the context of classical models of molecular evolution.

Materials and Methods

**LS of Genes**

LS describes how specifically a gene and orthologs of the gene are distributed on a given phylogeny (Cai, Woo, et al. 2006). If a gene and its orthologs are present in the species all belong to a single lineage, the gene is considered specific to this particular lineage. On the other hand, if a gene and its orthologs are present in all the species of all lineages on the phylogeny, the gene is a “common” gene not specific to any lineages. Most of genes, however, have a certain level of LS laying between those of two “extremes” scenario—they are present in some but not all species.

To calculate LS for human genes in regard to the primate lineage, we used the phylogeny of 11 eukaryotic species, including *Homo sapiens*, *Pan troglodytes*, *Mus musculus*, *Bos Taurus*, *Gallus gallus*, *Xenopus tropicalis*, *Danio rerio*, *Ciona intestinalis*, *Drosophila melanogaster*, *Caenorhabditis elegans*, and *Saccharomyces cerevisiae* (fig. 1A). The topology of the phylogeny, supported by a range of molecular and morphological data (Blair et al. 2002; Bourlat et al. 2006; Nikolaev et al. 2007; Dunn et al. 2008), was retrieved from the National Center for Biotechnology Information (NCBI) Taxonomy database (Wheeler et al. 2000). The distribution of each human gene (i.e., the presence–absence pattern of its orthologs) on the tree forms the phylogenetic profile of the gene. We obtained phylogenetic profiles of genes from PhyloPat v41 (Hulsen et al. 2006) (http://www.cmbi.ru.nl/pw/phylopats). Phylopats uses information of orthologs predicted in Ensembl compara database ( Birney et al. 2006) to construct phylogenetic profile for each gene according to the presence or absence of orthologs of the gene in other species. The software pipeline of Ensembl compara database collected gene pairs of the best reciprocal hits and best score ratio values from a WU-Blastp or Smith–Waterman whole-genome comparisons and then created a graph of gene relations, followed by a clustering step. The clusters were then applied to build a multiple alignment and a phylogenetic tree, which is reconciled with the species tree. From each reconciled gene tree, the orthologous relationships were inferred (for details, see [Vilella et al. 2009] and http://www.ensembl.org/info/docs/compara/homology_method.html). In figure 1A, we represented the phylogenetic profiles of human genes with a string of 11 symbols: ● and ○ indicate the presence and absence of ortholog in the corresponding species, respectively. In our data analysis, we only included genes whose string representation of phylogenetic profile belongs to one of ten “regular” given patterns, in which ● and ○ are constitutively assigned to these regular patterns and hence to genes whose phylogenetic profiles match these patterns. We discarded 10,032 (out of 20,150) genes that showed “irregular” phylogenetic profiles.

**PL of Genes**

Phylostratum is a set of genes from an organism that coalesce to founder genes having common phylogenetic origin (Domazet-Loso et al. 2007). Using a phylostratigraphic approach, Domazet-Loso and Tautz (2008) assigned all human protein-coding genes into 19 phylostrata. Here, we describe the procedure they used to determine gene’s PL.

---

Briefly, BlastP algorithm (E value cutoff 0.001) was used to search human proteins against the NCBI nonredundant (NR) database to determine the presence/absence of homologs. The 0.001 cutoff value presents a good compromise between specificity and sensitivity of sequence similarity search (Domazet-Loso and Tautz 2003). Before the sequence search, the NR database was cleaned up with respect to sequences of uncertain or missing taxonomic status, unreliable phylogenetic positions, and filled up with complete genomes that were absent in the database but otherwise were publicly available. In addition, TBLASTN search was conducted against substantial trace and expressed sequence tags archives of Porifera, Cyclostomata, and Chondrichtyes as for these internodes complete annotated genomes were lacking. We obtained PL of all human genes from the supplementary data of Domazet-Loso and Tautz (2008). The numbers of genes in different PL groups vary (supplementary fig. S1, Supplementary Material online). To facilitate data analysis, we merged 19 PL groups into nine bins (fig. 2A). Empirically, neighboring phylostrata (e.g., phylostrata 9–11) with few genes were merged into one bin. Phylostrata with more genes (e.g., phylostrata 8 and 6) were assigned as two separate bins. The “empirically populated binning” procedure ensured the difference of gene number between bins was less substantial (supplementary fig. S1, Supplementary Material online). We also binned genes using an alternative procedure—“equally populated binning.” We added a random variable, e ~ norm(0, 0.001), to PL value of each gene, making PL a continuous variable. The value of e was small such as to not change the original rank of PL of genes substantially, but, by adding an e, each gene obtained a distinct rank. To generate equally populated bins, we adjusted the width of bins so that the same number of genes would fall into each bin. Two binning procedures produce similar results in subsequent data analyses. We only report results obtained using the first procedure. Note that this analysis was performed with genes in Ensembl database v45, for which the PL values were originally estimated (Domazet-Loso and Tautz 2008).

**Gene Sets and Divergence Rate**

To estimate the divergence rate of human protein-coding genes, we retrieved coding sequences of 20,150 human genes and their corresponding orthologs in chimpanzee genome from Ensembl database (Birney et al. 2006; Hubbard et al. 2007). Nonsynonymous substitution rate (K_a) and synonymous substitution rate (K_s) for human–chimpanzee orthologous pairs were calculated using the counting method of Nei and Gojobori (1986) implemented in MBE-Toolbox (Cai et al. 2005; Cai, Smith, et al. 2006). Before comparing median values of K_a and K_s between different groups of genes, we removed data points of 70 ribosomal genes (that are extremely slow-evolving genes that lack divergence information), 228 sex-chromosomal genes (that are under different strength of selection compared with autosomal genes), and 1,997 pseudogenes listed in http://Pseudogene.org (Karro et al. 2007). We further removed data points with K_a ≥ 0.05 and/or K_s ≥ 0.05 from 523 human–chimpanzee gene pairs to reduce the problem introduced by misalignment of coding sequences. Similar overall results were obtained when K_a and K_s (or denoted as dN and dS) were calculated using the maximum-likelihood method (Goldman and Yang 1994) implemented in PAML (Yang 1997). Given the fact that chimpanzees are the most closely related species to humans and divergence between human–chimp ortholog pairs may be extremely low, it is possible that the low values of K_a and K_s for some genes may just be an artifact of the choice of species to do the comparison. To eliminate the concern, we recomputed K_a and K_s with human–macaque ortholog pairs. The sequences of the corresponding macaque orthologs were again obtained from Ensembl database (Binney et al. 2006; Hubbard et al. 2007). We obtained similar result as above (supplementary fig. S2, Supplementary Material online). In addition, instead of computing K_a and K_s for individual genes, we summed divergence sites of genes in the same age classes and calculated pooled K_a and K_s for each age class (same procedure as described below). We obtained qualitatively unchanged results (data not shown).

We also obtained the numbers of nonsynonymous and synonymous sites (N and S) and the numbers of nonsynonymous (D_n) and synonymous (D_s) differences in coding sequences of human–chimpanzee genes resequenced in the study of Bustamante et al. (2005), who applied exon-specific polymerase chain reaction amplification to 20,362 loci in 39 humans and one chimpanzee to obtain sequence variants in these regions. We summed up D_n (and D_s) and N (and S) for genes in the same group to calculate pooled K_a and K_s (i.e., K_a and K_s for groups, in which sequences of genes are essentially concatenated). To obtain the 95% confidence interval (CI) of pooled K_a and K_s, we used the bootstrapping approach. For each group of genes, we constructed 10,000 resamples of the observed gene sets (and of equal size to the observed gene sets), each of which was obtained by randomly sampling with replacement from the original genes in the group. The CIs of pooled K_a and K_s were then obtained using percentile method from values of pooled K_a and K_s for resamples.

**SNPs and Polymorphic Consequences**

We computed A* (the number of nonsynonymous single-nucleotide polymorphism [nSNP] per nonsynonymous site) as the ratio between total nSNPs and total number of nonsynonymous sites of genes in the same class, and S* (the number of synonymous SNPs [sSNPs] per synonymous site) as the ratio between total sSNPs and total number of synonymous sites of genes in the same class. To determine the nonsynonymous and synonymous status of SNPs, we
mapped SNPs (from dbSNP or Perlegen) onto the coding regions of the longest transcripts of human genes using functions in PGEToolbox (Cai 2008). For Applera SNPs, we used the nonsynonymous and synonymous classification from the original study (Bustamante et al. 2005). We systematically repeated SNP-related analyses using three SNP data sets: 1) Applera SNPs in the 11,624 resequenced genes from the study of Bustamante et al. (2005), 2) SNPs in dbSNP build 126, and 3) Perlegen SNPs (Hinds et al. 2005). Applera SNPs were identified by resequencing and thus suffer from less severe ascertainment biases. For dbSNP data, we removed SNPs that are not validated (dbSNP category 0) and retained NCBI designated “double hit” or “submitter validated” SNPs, which are more likely to be real. For Perlegen data, we used all SNPs ascertained in all populations. We obtained allele frequencies of Applera SNPs and Perlegen SNPs to perform allele frequency spectrum analysis.

Robustness Tests for SNP Analyses
To conduct robustness tests against several confounding factors, we identified singleton genes, masked CpG-related SNPs, grouped genes according to functional categories, and classified genes according to their local genomic guanine-cytosine (GC) content. To identify singleton genes, we retrieved the data set of predicted human paralogous genes from Ensembl database (Birney et al. 2006; Hubbard et al. 2007). Singleton genes are those that do not have any paralogs.

Single-nucleotide mutations at CpG sites are much more frequent than at other sites (Cooper and Krawczak 1989; Sved and Bird 1990). To control for the effect of CpG-related SNPs, we generated the subset of SNPs excluding all CpG-related SNPs. To do so, we downloaded neighboring nucleotides of each SNP from human genome sequence from Ensembl v41 (Birney et al. 2006). SNPs were considered potentially CpG related in cases when: 1) A/G SNPs proceeded by a C or 2) C/T SNPs followed by G (Webster and Smith 2004). CpG related in cases when: 1) A/G SNPs proceeded by a C or 2) C/T SNPs followed by G (Webster and Smith 2004). CpG related in cases when: 1) A/G SNPs proceeded by a C or 2) C/T SNPs followed by G (Webster and Smith 2004).

The association between Gene Ontology (GO) terms and individual genes was obtained from FatiGO (Al-Shahrour et al. 2004). We used genes whose functional annotation contains the same GO term and tested whether the pattern between \(K_r/K_s\) (or \(A^S/S^S\)) and LS remains the same as that obtained from all genes. We obtained the result for GO terms: “cellular physiological process,” “metabolism,” “regulation of cellular process,” “regulation of physiological process,” and “cell communication,” which were associated with at least 25 genes in all of LS groups.

To test the effect of regional nucleotide composition of genes, we obtained the isochore map of human genome from the supplementary table of Costantini et al. (2006). Genes within isochore family L1 and L2 were considered in the GC-poor regions; genes within isochore families H1, H2, and H3 were considered in the GC-rich regions.

We obtained global mRNA expression data from Gene Expression Atlas (http://wombat.gnf.org) (Su et al. 2004). We included normal adult samples from 54 tissue types in the analysis. The expression level of each probe set in a given tissue was calculated as the mean of log (base 2) signal intensities of all samples after GC-robust multi-array average normalization (Wu et al. 2004). When multiple probes were mapped onto the same gene, the probe with the highest expression level was used as the report probe for this gene. We calculated the mean expression level of a gene (aveExp) as the mean of log signal intensities of probe sets across all tissues. We also calculated the maximum expression level among all tissues (maxExp) and the heterogeneity of expression level across all tissues (hetExp) for all genes with expression data available according to our previous study (Cai et al. 2009).

Estimation of the Ratio of the Rate of Adaptive Nonsynonymous Substitution to Synonymous Substitution (\(\omega_A\))

The distribution of fitness effects (DFEs) of new deleterious mutations and \(\omega_A\) were estimated by using the method of Eyre-Walker and Keightley (2009) implemented in the program DoFE v2.1 (http://www.lifesci.susx.ac.uk/home/Adam_Eyre-Walker/Website/Software.html). To make the input file, we compiled Applera data obtained from Bustamante et al. (2005) and Lohmueller et al. (2008), which contain the numbers of nucleotides divergent between human and chimp and the site frequency spectra (SFS) for sSNPs and nSNPs, respectively. Genes of different LS groups were separated into different sets. Each set comprises the numbers of selected and neutral divergence polymorphism sites, as well as SFS vectors. DoFE provides an option for excluding singletons in SFS. Our result was unaffected qualitatively using the analysis with this option, thus only the result obtained without excluding singletons is reported below.

Estimation of Selection Coefficient (\(\gamma\)) using MKPRF
We used the program MKPRF (http://cbsuapps.tc.cornell.edu/mkprf.aspx) to estimate the selection coefficient \(\gamma (=2N_{As})\) of genes. The program MKPRF samples from the posterior distribution of parameters in the MKPRF models of (Bustamante et al. 2002) and (Barrier et al. 2003) using a Markov Chain Monte Carlo algorithm based on Poisson random field (PRF) theory (Sawyer and Hartl 1992). We used a subset of genes with at least two variable nonsynonymous sites in the alignment (i.e., \(P_r + D_r \geq 2\)). Using exactly the same approach as implemented in Bustamante et al. (2005), we applied the nonhierarchical model by specifying the flag FIXED_VARIANCE = 1, such that a single Gaussian prior of \(\gamma\) with a mean of 0 and the standard deviation (SD) (\(\sigma\)) of 8 is set for all loci (Bustamante CD, personal communication). Slightly deleterious SNPs can lead to an underestimate of the rate of adaptive evolution because they contribute to
polymorphism but rarely become fixed. The effects of these slightly deleterious mutations can be partially controlled by removing mutations segregating at low frequencies (Fay et al. 2001). To circumvent the problem of unequal levels of slightly deleterious polymorphisms present in genes of different LS groups, we used two related procedures: 1) removing all SNPs whose derived alleles are at low frequencies (derived allele frequency [DAF] < 0.15) and 2) subsampling nSNPs at low frequencies (DAF < 0.15), such that SFS distributions across all LS classes match each other. Before subsampling, we calculated the fractions of these low-frequency nSNPs (lfnSNPs) in all nSNPs for genes in ten LS groups. The nSNPs in LS 10 group genes have the lowest fraction of lfnSNPs. We thus kept lfnSNPs in LS 10 group unchanged and subsampled lfnSNPs in the rest of the LS groups. For each LS group 1–9, we computed how many lfnSNPs (φ%) should be removed in order to make the final fraction of lfnSNPs equal to that of LS 10. Then we randomly purged φ% of lfnSNPs from each of LS groups 1–9. As a result, allele frequency spectra of the subsampled nSNPs in all ten LS groups became similar to each other (supplementary fig. S3, Supplementary Material online). We also used two different DAF cutoffs (<0.05 and <0.20) to define lfnSNPs. The effect of subsampling on final MKPRF results does not change depending on the exact values of the DAF cutoffs.

Li et al. (2008) found that results of MKPRF are sensitive to the model, and the value of σ used to estimate γ values. To get a sense of the robustness of γ estimation, we reran our analyses using hierarchical model (FIXED_VARIANCE = 0) with all default parameters and also using nonhierarchical model with the values of σ set at 1, 4, and 16.

**Results**

**Classifications of Human–Chimp Genes**

We first classified human genes into ten groups based on their LS (Cai, Woo, et al. 2006). Here, we considered the
primate lineage containing human and chimpanzee in the phylogeny of total 11 eukaryotic species (fig. 1A). We used LS to quantify the extent to which orthologs of a human gene are distributed in the species close to the primate lineage. Human genes whose orthologs are present in few species closely related to human and chimp are more primate lineage-specific than those genes whose orthologs are present in more distantly distributed species. We only assessed genes of ten LS categories that have nonpatchy distributions (fig. 1A). The numbers of genes in the LS categories 1–10 are 737, 848, 355, 458, 1,377, 294, 361, 834, 219, and 1,817, respectively.

We then classified human genes based on their PL, which quantifies evolutionary age of genes in terms of the most diverged lineage in which the homologs of those genes can be detected using Blast (see Materials and Methods and [Domazet-Loso et al. 2007]). As noted in Domazet-Loso and Tautz (2008), genes that share a particular protein domain will have the same PL based on when this domain emerged first, even though a particular gene may have evolved later (e.g., due to gene duplication or exon shuffling). Because a protein domain is usually linked to a certain function, PL is used to trace the origin of this function, irrespective of the further origin of paralogs. We classified human genes into nine PL groups (Materials and Methods) (fig. 2A). LS and PL measures two proprieties of phylogenetic distribution of genes—breadth and depth, which are related to each other. Lineage-specific genes with high LS levels necessarily have high PL levels and vice versa. Indeed, the Spearman’s correlation coefficient between LS and PL is 0.65, and it is highly significant ($P < 0.001$).

Finally, we identified genes that were present in human, chimp, and yeast but that varied in their presence and absence in the “intermediate” lineages between human/chimp and yeast (fig. 3A). We classified these genes into four groups according to the number of GLs: (0) genes that have
never been lost, (1) genes that have been lost once, (2) twice, and (3) three times or more (fig. 3A). These genes belong to the same PL group but vary in the propensity for loss. In such a setting, all human/chimp genes are successfully detected in the distantly related species (such as, mouse, cow, and chicken). Same notation is used as in figure 1. The number of GLs is counted in species between human and yeast. (B) Median values of divergence rates (pooled \(K_a\), \(K_s\), and \(K_a/K_s\) derived from the Applera divergence data [Bustamante et al. 2005]) for groups of genes whose loss counts are 0, 1, 2, and \(\geq 3\). Error bars indicate 95% CIs calculated from the 10,000 bootstrap replications. For individual genes, the \(K_a\) and \(K_a/K_s\) values vary marginally significantly among different GL groups (\(\chi^2 = 16.04\) and 15.85, respectively, degrees of freedom \(df = 3, P = 0.001\) in both cases, KW test). The difference in \(K_s\) is not significant among GL groups (\(\chi^2 = 0.55, df = 3, P = 0.908, KW\) test). Both \(K_a\) and \(K_a/K_s\) are positively correlated with GL (Spearman’s \(r = 0.068\) and 0.067, respectively, \(P < 0.001\) in both cases), whereas \(K_s\) shows no correlation (Spearman’s \(r = 0.001, P = 0.939\)).

Higher Rates of Protein Evolution in Lineage-Specific Genes

We use the nonsynonymous substitution rate (\(K_a\)) and the ratio of nonsynonymous to synonymous substitution rate (\(K_a/K_s\)) between the human and chimpanzee orthologs to measure the rate of protein evolution. \(K_a\) and \(K_s\) values for the genes in each group (LS, PL, or GL) were calculated in two ways. First, we calculated \(K_a\) and \(K_s\) values for the individual genes and estimated the median value within each group. Second, we concatenated the sequences of all genes in the same group and calculated pooled \(K_a\) and \(K_s\) values for each group in this manner. Two approaches gave essentially identical values. For brevity, we report the results of the pooled \(K_a\) and \(K_s\) values.

The \(K_a\) and \(K_a/K_s\) values vary significantly among different LS, PL, and GL groups (Kruskal–Wallis [KW] test, \(P \leq 0.001\) in all cases) (figs. 18, 2B, and 3B). \(K_a\) and \(K_a/K_s\) are positively correlated with the LS and PL and negatively with GL values (Spearman’s \(r = 0.503\) and 0.507 [LS] and \(r = 0.215\) and 0.206 [PL] for \(K_a\) and \(K_s\) respectively, \(P < 0.001\)), whereas \(K_s\) shows either no correlation (LS and GL, Spearman’s \(r = 0.068\) and 0.067, respectively, \(P < 0.001\) in both cases), whereas \(K_s\) shows no correlation (Spearman’s \(r = 0.001, P = 0.939\)).

Note that in every case, we retained only those human genes that have clear orthologs present in the chimpanzee genome. In this way, for all genes in the study, the rates of molecular evolution are estimated through human–chimp comparison in a consistent and reliable manner.

**Fig. 3.**—Protein divergence rates (\(K_a\) and \(K_a/K_s\)) as a function of number of GL. (A) Phylogenetic profiles of human genes that are present in human, chimpanzee, and yeast but vary in their presence and absence intermediate species (such as, mouse, cow, and chicken). Same notation is used as in figure 1. The number of GLs is counted in species between human and yeast. (B) Median values of divergence rates (pooled \(K_a\), \(K_s\), and \(K_a/K_s\) derived from the Applera divergence data [Bustamante et al. 2005]) for groups of genes whose loss counts are 0, 1, 2, and \(\geq 3\). Error bars indicate 95% CIs calculated from the 10,000 bootstrap replications. For individual genes, the \(K_a\) and \(K_a/K_s\) values vary marginally significantly among different GL groups (\(\chi^2 = 16.04\) and 15.85, respectively, degrees of freedom \(df = 3, P = 0.001\) in both cases, KW test). The difference in \(K_s\) is not significant among GL groups (\(\chi^2 = 0.55, df = 3, P = 0.908, KW\) test). Both \(K_a\) and \(K_a/K_s\) are positively correlated with GL (Spearman’s \(r = 0.068\) and 0.067, respectively, \(P < 0.001\) in both cases), whereas \(K_s\) shows no correlation (Spearman’s \(r = 0.001, P = 0.939\)).
orthologs were again obtained from Ensembl database (Birney et al. 2006; Hubbard et al. 2007). We obtained similar result as above (Supplementary fig. S2, Supplementary Material online). These results are consistent with previous reports in humans or species in other domains (Krylov et al. 2003; Alba and Castresana 2005; Cai, Woo et al. 2006).

Robustness to Potential Inability to Detect Fast-Evolving Genes in Distant Lineages

Inability of Blast to detect orthologs of fast-evolving genes in distant lineages can in principle contribute to the inverse relationship between the protein-divergence rate and LS (or PL) (Elhaik et al. 2006). This problem is unlikely to be severe, as simulations of the evolution of protein sequences with the same rates and among-site heterogeneity as those estimated from real mammalian protein-coding genes demonstrated that most functional genes could be detected by Blast in comparisons of even very distantly related genomes (e.g., fungi or plants vs. mammals) (Alba and Castresana 2007). The simulated procedure is exactly the same as the one used for the determination of PL values and should have as much sensitivity as the one employed in the LS statistic. Furthermore, the GL statistic, which measures the rate of GL, should be the least susceptible to this problem. This is because in the case of GL, proteins are first detected between the most distant lineages (humans and yeast) and thus are unlikely to be missed in the comparisons of more closely related species. Finally, in our analyses, the strongest signal comes from the comparisons of fairly young genes (e.g., high LS groups, fig. 1) and thus should be more robust to this problem.

We also conducted an additional test. We divided all genes into two groups: the slower evolving group (\(K_a \leq 0.007\)) and the faster evolving half (\(K_a > 0.007\)). The cutoff \(K_a = 0.007\) is the 72th percentile of ordered \(K_a\) values, which was chosen to split genes into two groups in such a way that each group contains enough (>30) genes in each of 10 LS or 9 PL classes. The positive correlation between \(K_a/K_s\) and LS and PL values is evident in both data sets \((P < 0.001\) in both cases, fig. 4). This demonstrates that the detected pattern is not due to the unusual behavior of the fastest evolving genes, which are the likeliest genes to go undetected in distant comparisons.

Overall, we believe it is highly unlikely that the problem of detecting proteins in distantly related lineages is sufficiently severe to invalidate the described analyses. Note that the slow rate of protein evolution is evident in the older and more broadly distributed genes independently of whether we classify genes using the LS, PL, or GL statistics. All the analyses described below can be reproduced using any of the statistics. In the rest of the manuscript, for brevity, we only report the results derived using the LS classification.

Robustness to the Variation in the Levels of Gene Expression

The level of gene expression is strongly anticorrelated with the protein divergence rate (Drummond et al. 2006; Pal et al. 2006). It is therefore important to determine whether the correlation between \(K_a\) or \(K_a/K_s\) and LS categories is due to the variation in expression levels of genes of different LS categories.

We consider three expression-related measures across all tissues: average expression (aveExp), the maximum expression (maxExp), and the heterogeneity of expression (hetExp) (see Materials and Methods). As expected, aveExp is anticorrelated with \(K_a/K_s\) (Spearman’s \(\rho = -0.256, P < 0.001\)) and \(K_a\) (Spearman’s \(\rho = -0.241, P < 0.001\)). The genes in higher LS categories have lower levels of expression (Spearman’s \(\rho = -0.230, P < 0.001\)). After controlling for expression level, the correlations between LS and \(K_a/K_s\) or \(K_a\) remain positive and highly significant [corr\((K_a/K_s,LS)\)aveExp] = 0.238 and corr\((K_a,LS)\)aveExp] = 0.222, both \(P < 0.001\), Spearman partial correlation test]. Thus, correlation between LS and the rate of protein evolution is not entirely driven by lower expression levels of high LS genes. The other two variables, maxExp and hetExp, have similar relationships with LS and \(K_a/K_s\) or \(K_a\) as aveExp does, which is not unexpected given the strong correlations between maxExp and hetExp with aveExp (Spearman’s \(\rho = 0.772\) and 0.294, respectively [both \(P < 0.001\)].

We conducted a linear multiple regression in the forward stepwise manner to examine the contributions of LS, aveExp, maxExp, and hetExp on the variation in log\((K_a)\). The regression model defines log\((K_a)\) as a function of all corresponding variables \(X_{LS}, X_{aveExp}, X_{maxExp},\) and \(X_{hetExp}\):

\[
\log(K_a) = \beta_0 + \beta_{LS}X_{LS} + \beta_{aveExp}X_{aveExp} + \beta_{maxExp}X_{maxExp} + \beta_{hetExp}X_{hetExp}
\]

Table 1 gives the result of the modeling procedure. The final model gives a global \(R^2 = 0.131\) \((P < 0.001)\), that is, more than 13% of the variation in log\((K_a)\) is explained by this model. During the construction of the final model, two predictors most highly correlated with log\((K_a)\)—LS and aveExp—were kept. The remaining variables, which have minor roles in overall regression, were excluded. The standardized coefficients were examined to determine the relative importance of the significant predictors. LS contributes more to the model than does aveExp, as shown by its larger absolute standardized coefficient 0.350 and t statistic of 20.712, compared with values of 0.041 and 2.414 for aveExp. This analysis suggests that LS is the most relevant predictor of the rate of protein divergence.

Weaker Purifying Selection in Lineage-Specific Genes

The slower protein evolution of older and more broadly distributed genes is most likely due to stronger purifying
selection acting on these genes. Indications for this should include a lower density of nSNPs and lower frequencies of derived alleles of these nSNPs. We use human SNP data and human/chimp divergence to investigate this prediction.

We analyze three SNP data sets: 1) SNPs in genes resequenced by Applera (Bustamante et al. 2005), 2) SNPs in NCBI dbSNP build 126, and 3) Perlegen SNPs (Hinds et al. 2005). Coding SNPs were split into sSNPs and nSNPs. For each LS category, we computed A* (the number of nSNP per nonsynonymous site) and S* (the number of sSNP per synonymous site) (see Materials and Methods). Figure 5 shows the results derived from all three data sets. None of the conclusions related to the assessment of the strength of purifying selection (measured by using the ratio A*/S*) change qualitatively depending on the investigated SNP data sets. Below we only describe Applera results.

First, we test whether older genes are subject to stronger purifying selection in humans. Values of A* and A*/S* correlate strongly and positively with the value of LS (Spearman’s $r = 0.881$ and 0.874, $P < 0.005$ and 0.001, respectively). The values of A*/S* increase almost 2-fold (0.31–0.57) from the lowest to the highest values of LS (fig. 5A). In addition, DAFs of SNPs in the genes from the lower LS groups are skewed toward rare alleles relative to FIG. 4.—Median $K_a/K_s$ as a function of LS and PL for slowly and fast-evolving genes. (A) Median $K_a/K_s$ for ten LS groups; (B) Median $K_a/K_s$ for nine PL groups. Genes are grouped into the slowly ($K_a < 0.007$) and fast ($K_a > 0.007$)-evolving ones.

Table 1

<table>
<thead>
<tr>
<th>Included variables</th>
<th>Overall Contribution of Variable ($R^2$)*</th>
<th>Incremental Contribution of Variable ($\Delta R^2$)</th>
<th>Order of Entryb</th>
<th>Unstandardized coefficient ($\beta$) ± standard error</th>
<th>Standardized Coefficient ($\beta$)</th>
<th>t*</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td></td>
<td></td>
<td></td>
<td>−5.929 ± 0.097</td>
<td>−1.293</td>
<td>&gt;0.0001</td>
<td></td>
</tr>
<tr>
<td>LS</td>
<td>0.129</td>
<td>0.129</td>
<td>1</td>
<td>0.132 ± 0.006</td>
<td>0.350</td>
<td>&lt;0.0001</td>
<td></td>
</tr>
<tr>
<td>avExp</td>
<td>0.006</td>
<td>0.002</td>
<td>2</td>
<td>−0.301 ± 0.125</td>
<td>−0.041</td>
<td>&lt;0.016</td>
<td></td>
</tr>
<tr>
<td>Excluded variables</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>hetExp</td>
<td>0.005</td>
<td></td>
<td>3</td>
<td>−0.015</td>
<td>−0.858</td>
<td>&gt;1</td>
<td></td>
</tr>
<tr>
<td>maxExp</td>
<td>0.006</td>
<td></td>
<td>4</td>
<td>−0.008</td>
<td>−0.306</td>
<td>&gt;0.1</td>
<td></td>
</tr>
</tbody>
</table>

* $R^2$ is the proportion of variation in the dependent variable (log($K_{a}$)) explained by the regression model constructed from the individual variable, indicating the independent contribution of each variable to explain the global variance of log($K_{a}$).

b Order of variables entered into the model at each step.

c The $t$ statistic indicates the relative importance of each variable in the model.
that of higher LS genes (fig. 6). Specifically, the proportion of rare nSNPs (DAF < 0.15) is negatively correlated with LS values (Spearman’s ρ = −0.794, P < 0.01), whereas, in contrast, the proportion of rare sSNPs are not correlated with the LS values (P = 0.45, fig. 6). The proportion of rare nSNPs is significantly greater than the proportion of rare sSNPs for the LS groups 1 through 5 (P < 0.05 for all tests, G-test with Bonferroni correction) but not for the LS groups 6 through 10 (P > 0.05 for all tests, G-test). We also tried different DAF cutoffs (0.05, 0.1, and 0.2) as well as investigated the pattern derived from minor allele frequencies (MAFs). Regardless of the different cutoffs of DAF and the use of derived or MAF, results are similar (supplementary figs. S4 and S5, Supplementary Material online). Thus, compared with younger genes, older genes contain fewer nSNPs per site, and the frequencies of derived or minor alleles of these SNPs are lower. These results indicate that younger genes are subject to weaker purifying selection at the protein-coding level.

Robustness Tests for Polymorphism Patterns
We conducted several tests to demonstrate that the relationship between the proportion of A*/S* and LS cannot be explained entirely by a number of confounding factors. Specifically, we demonstrated that stronger purifying selection acting on older genes can be detected within subsets of the data defined by 1) whether a gene belongs to a particular functional (GO) gene group (supplementary table S1, Supplementary Material online), 2) whether a gene has
a duplicate elsewhere in the human genome (supplementary table S2, Supplementary Material online), 3) whether SNPs are due to C to T transitions at CpG sites (supplementary table S3, Supplementary Material online), and 4) whether a gene resides in a genomic region of high or low GC content (supplementary table S4, Supplementary Material online) (for details, see Materials and Methods).

The correlation between A*/S* and LS is unlikely due to different LS groups containing peculiar assemblages of genes defined by function, presence of a duplicate, number of CpG sites, or GC content.

We also verified that variation in levels of purifying selection in genes of different LS categories is not due to the variation in the level of gene expression, although A*/S* is strongly influenced by the expression pattern and breadth (Osada 2007). The partial correlation between A*/S* and LS is positive and significant after controlling for the different average expression levels of genes in different LS categories (i.e., corr(A*/S*,LS|aveExp) = 0.830, P < 0.001, partial Spearman correlation test with Applera data).

Rate of Adaptive Evolution in Lineage-Specific Genes

The restricted phylogenetic distribution of young genes implies that many of them are probably involved in lineage-specific adaptive processes. So, in addition to experiencing weaker purifying selection, young genes might be expected to experience higher rates of positive selection. To test this conjecture, we first used the method of Eyre-Walker and Keightley (2009) implemented in DoFE v2.1 to estimate \( \omega_A \) for each LS categories. We compared the estimated fractions of mutations in different \( N_S \) ranges among LS categories (fig. 7). There is a wide range of fractions of amino acid-changing mutations that behave as effectively neutral (0 \(< N_S <= 1\) among LS categories, ranging from 15.7% for LS 2 to 52.3% for LS 9. The fraction of effectively neutral mutations is correlated with LS significantly (Kendall’s \( \tau = 0.73, P < 0.005 \), that is, the younger the genes the more effectively neutral mutations they have proportionally. The combined fractions of mutations with \( N_S \) ranges 1–10 and 10–100 are for slightly deleterious mutations. The fraction is negatively correlated with LS （\( r = -0.76, P = 0.017 \), Pearson correlation test after excluding outlier data point at LS 5） (fig. 7 and supplementary fig. S6, Supplementary Material online). This is consistent with previous results that old genes have proportionally more slightly deleterious mutations.

We also use the MKPRF analysis (Sawyer and Hartl 1992; Bustamante et al. 2002, 2005) to assess the proportion of genes showing evidence of positive selection in each LS categories. First, we ran MKPRF with a nonhierarchical model.
and a SD (σ = 8) of Gaussian prior, using exactly the same settings as in Bustamante et al. (2005). For each gene, we estimated the value of population-effective selection coefficient, γ (= 2Neεs, where Ne is the effective population size and s is the selection coefficient in a Wright–Fisher genic selection model). The values of γ vary significantly among genes of different LS groups (P < 0.001, KW test) and are positively correlated with LS values (Spearman’s ρ = 0.706, P < 0.001). We also obtained the 95% CIs of γ. If a gene has its 95% CIs of γ completely above 0, the gene appears to have been evolving under positive selection. On the other hand, if the 95% CIs of γ are completely below 0, the gene appears to be under negative selection and have a high proportion of weakly deleterious nonsynonymous polymorphisms. We found that the proportion of positively selected genes (\( f_{γ>0} \)) increases with the increment of LS values (Spearman’s ρ = 0.927, P < 0.001; Kendall’s τ = 0.778, P < 0.001) (fig. 8A, left panel), and the proportion of negatively selected genes (\( f_{γ<0} \)) decreases with the increment of LS values (Spearman’s ρ = −0.924, P < 0.001; Kendall’s τ = −0.778, P < 0.001) (fig. 8A, right panel). We also calculated the ratios of the numbers of positively selected and negatively selected genes to the numbers of neutrally evolving genes (\( f_{γ>0} \) and \( f_{γ<0} \), respectively). Similar to \( f_{γ>0} \) and \( f_{γ<0} \), \( f_{γ<0} \) correlates positively with the LS values (Spearman’s ρ = 0.927; P < 0.001; Kendall’s τ = −0.778, P < 0.001) (fig. 8A, left panel), and \( f_{γ<0} \) correlates negatively with the LS values (Spearman’s ρ = −0.915; Kendall’s τ = −0.778, P < 0.001) (fig. 8A, right panel). Note that, as in Bustamante et al. (2005), we focused our analysis only on the potentially informative loci that contain enough polymorphism and/or divergence events to have a chance of showing signals of selection. Specifically, we included informative loci with \( P_n + D_n ≥ 2 \) (Bustamante et al. 2005). We also carried out MKPRF analysis in two ways: either analyzing all genes together or analyzing genes from each LS group separately. The results remain virtually unchanged. Therefore, MKPRF analysis with full Applera SNPs suggested that younger genes experience a higher rate of adaptive evolution.

Slightly deleterious SNPs lead to an underestimation of the rate of adaptive evolution because they make a substantial contribution to polymorphism but fix at a much lower rate compared with neutral polymorphisms (Crow and Kimura 1970; Eyre-Walker and Keightley 1999; Eyre-Walker 2002; Eyre-Walker et al. 2002). From allele frequency analysis of nSNPs and the MKPRF analysis, we know that genes from lower LS groups have more slightly deleterious polymorphisms on average. This is indicated both by the higher proportion of rare nSNPs (fig. 6) and by the higher proportion of genes for which the MKPRF estimate of γ is negative for the low LS genes. Unlike the DoFE method of Eyre-Walker and Keightley (2009), in which the effect of slightly deleterious mutations are attempted to be controlled, MKPRF analysis per se does not control for this effect. Thus, the finding of a smaller proportion of genes experiencing positive selection in low LS group genes might be an artifact of the larger proportion of slightly deleterious polymorphisms in these genes.

We try to rule out this possibility using two approaches: 1) removing SNPs of low frequencies (Fay et al. 2001) prior to the analysis and 2) subsampling nSNPs to ensure that nSNPs in different LS groups have same frequency distribution and thus have the same bias (see Materials and Methods). The first procedure attempts to limit the effect of slightly deleterious
deleterious SNPs by focusing on more frequent SNPs under the assumption that slightly deleterious SNPs are unlikely to reach intermediate frequencies (Fay et al. 2001). After removing the SNPs at less than 15% frequency and rerunning the MKPRF procedure, the positive correlation between LS and portion of genes under positive selection remains virtually unchanged (fig. 8B, Spearman corr(LS, $f_{c>0}$) = 0.93, P < 0.0001, and corr(LS, $f_{c<0}$) = 0.94, P < 0.0001). The second,
subsample procedure attempts to subsample the nSNPs such that the DFE of all of them is similar and is affected by the slightly deleterious nSNPs both as little as possible, but more importantly, to the same extent across all the LS groups. We subsampled the nSNPs in all groups using the distribution in the LS 10 group. After subsampling, all LS groups have indistinguishably similar proportions of slightly deleterious nSNPs no longer decreases with the increment of LS when MKPRF is carried out with the subsampled nSNPs (fig. 8C). Importantly, the proportion of genes under positive selection remains higher in the higher LS groups (fig. 8C, Spearman corr(LS, f_{c}>0) = 0.94, P < 0.0001, and corr(LS, f_{c}>0) = 0.95, P < 0.0001).

Results of the MKPRF analysis might change depending on the prior and the model used in the analysis (Li et al. 2008). To explore these effects, we ran the MKPRF analysis using the nonhierarchical model with three additional σ values (1, 4, and 16). We found that although the absolute values of proportion of genes in each age groups change, the positive correlation between LS and f_{c}>0 (or f_{c}>0) remains, especially in the cases of high σ values (e.g., σ = 16, Spearman corr(LS, f_{c}>0) = 0.93, P = 0.0001, and corr(LS, f_{c}>0) = 0.88, P = 0.008) (supplementary fig. S7, Supplementary Material online). We also ran MKPRF analysis using the hierarchichal model (see Materials and Methods). The positive correlation between LS and f_{c}>0 (or f_{c}>0) is weaker in this case but remains significant after removing rare frequent nSNPs (P = 0.004, Spearman correlation test) and after the subsampling (P = 0.003, Spearman correlation test) (supplementary fig. S8, Supplementary Material online). Overall, the MKPRF results do suggest that younger genes tend to experience positive selection more frequently, although to a modest degree. The faster evolution of younger genes appears to be attributable almost entirely to the weaker purifying selection acting on these genes.

Discussion

Genes in the human genome vary in their evolutionary age. A considerable proportion of human genes (e.g., ~10%, even only considering “strict orthologs” with unambiguous one-to-one relationships [Berglund et al. 2008]) can be detected in the yeast genome, implying that they originated before the common ancestor of human and yeast diverged more than 1.5 billion years ago. On the other hand, human genome contains a small fraction of genes found in only one or a few closely related species, such as, mammals- or primate-specific genes (e.g., morphicus [Johnson et al. 2001] and spanx [Kouprina et al. 2004]). Recent bioinformatics analysis revealed 270 primate-specific and 364 mammal-specific genes; some of them may have originated de novo (Toll-Riera et al. 2009; Toll-Riera, Castelo, et al. 2009). Indeed, there is increasing experimental evidence for emergence of new genes from noncoding mammalian genomic regions (Heinen et al. 2009; Knowles and McLyshagh 2009).

We have classified human/chimp genes based on the breadth and the depth of their phylogenetic distributions in 11 eukaryotic genomes using three related but distinct metrics that quantify the breadth (LS), the depth (PL), and the rate of GL (Krylov et al. 2003; Alba and Castresana 2005; Cai, Woo, et al. 2006). We confirmed that younger and less broadly distributed proteins evolved at distinctly higher divergence rates than older and broadly distributed genes (Domazet-Loso and Tautz 2003; Daubin and Ochman 2004; Alba and Castresana 2005; Wang et al. 2005; Cai, Woo, et al. 2006; Kuo and Kissinger 2008). This pattern is very pronounced: for instance, the correlation coefficient between one of the measures of the phylogenetic breadth and depth (LS) and the rate of protein evolution between humans and chimpanzees (K_0, K_1) is higher than 0.5. Another illustration of the strength of this signal is that human/chimp genes that cannot be detected in the mouse genome and beyond have been evolving approximately 4 times faster between humans and chimps than the human/chimp genes whose presence can be detected all the way to yeast. In addition, this effect is robust to the variation in levels of gene expression, existence of paralogs, relative abundance of CpG sites, GC content of genomic regions, and classes of gene functions (i.e., GO annotations). The age of a gene or the breadth of its phylogenetic distribution is thus one of the best predictors of its rate of evolution (Alba and Castresana 2005; Cai, Woo, et al. 2006).

The fast evolution of genes that have a restricted phylogenetic distribution raises a possibility that even old and broadly distributed but fast-evolving genes might be misclassified as young and lineage specific due to our inability to detect their orthologs in distant species (Elhaik et al. 2006). Fortunately, this entirely reasonable concern appears not to generate severe ascertainment problems in practice. Alba and Castresana (2007) simulated the evolution of protein genes using the same overall evolutionary rates and the same among-site rate heterogeneity as observed in mammalian genes. They found that Blast could detect practically all genes in this analysis all the way to the level of divergence between yeast and mammals. This is probably because even fast-evolving proteins tend to contain some conserved segments. These conserved segments, even if they are fairly short, can still be detected by the local alignment algorithm.
of Blast. One of our phylogenetic measures, PL, exclusively depends on Blast to determine gene age and should be reliable based on the simulations of Alba and Castresana (2007). One of the other measures, LS, should be at least as sensitive as PL and thus should not be affected severely either. We provided two additional lines of evidence that our results are not artifactual. First, we split the genes into two groups based on their rate of evolution between humans and chimps. We were able to detect faster evolution of younger and more narrowly distributed genes within each group and most importantly within the group of slowly evolving genes. The second line of evidence is based on the use of the number of GL measure. This measure classifies genes based on the detected number of losses in the phylogeny for genes that can be detected in the most distant taxa, in our case human/chimp and yeast. In the case of GL, all human/chimp genes can be detected in yeast making it very unlikely that the apparent absence of these genes in much closer related lineages was due to the failure of detection and not due to their true absence.

The faster protein evolution of younger or more narrowly distributed genes must be due to changes in the way natural selection operates on mutations in these genes. It is not due to the difference of mutation rates because the patterns of evolution at synonymous sites in younger genes are indistinguishable from those in older genes. In addition, these patterns are robust to the variation in GC content across the human/chimp genomes, which in principle could generate spurious signals. But what are these changes in the natural selection? There are two nonmutually exclusive possibilities: 1) younger genes can be subject to weaker selective constraint (weaker purifying selection) and/or 2) younger genes are subject to positive selection more frequently.

We used two approaches (DoFE [Eyre-Walker and Keightley 2009] and MKPRF [Sawyer and Hartl 1992; Bustamante et al. 2002, 2005]) to test the second possibility, namely that younger genes experience a higher rate of positive selection. Using DoFE, we estimated $\omega_A$ for each LS class of genes. We detected no correlation between LS and corresponding $\omega_A$ for genes in LS classes, providing no evidence of higher prevalence positive selection in younger genes. However, using MKPRF, we did find some evidence that there were proportionally more genes showing signs of positive selection ($\gamma > 0$) in younger age classes. The proportion of genes with a positive $\gamma$ goes from $\sim 1-2\%$ in the oldest genes to $\sim 6-12\%$ in the more lineage-specific genes (LS groups 7 through 10). Because this result can be biased by the higher prevalence of slightly deleterious nSNPs in the older genes, we reran the analysis either after eliminating rare (<15%) SNPs (Fay et al. 2001) or after subsampling nSNPs in different LS categories to match that in the youngest and the least biased gene category. Furthermore, MKPRF results might be affected by the choice of a different prior and the use of different models (hierarchical vs. non-hierarchical) (Li et al. 2008). In all these additional analyses, MKPRF results continue to suggest that a higher proportion of younger genes exhibit signs of positive selection. The inconsistent results produced by two methods emphasize that
the evidence for higher rate of adaptive evolution among younger genes is tentative. Among other things, the difficulty of detecting the difference could be due to a weak genome-wide signal of positive selection associated with human protein-coding genes in general.

Nevertheless the higher rate of adaptation in the young genes might be consistent with the ideas that lineage-specific genes may drive morphological specification, enabling organisms to adapt to changing conditions (Khalturin et al. 2009) and also with the observation that young genes tend to be less functionally important. Fisher’s geometric models of adaptation predicts that small phenotypic changes should have a higher probability of being advantageous (Fisher 1939) (but see [Kimura 1983; Orr 2002]). If mutations in younger genes tend to have more subtle phenotypic effects, then such effects would be both less likely to be deleterious and more likely to be adaptive. In this way, older, indispensable proteins would form the conserved, ancient, unchanging core of functionality of the cell and the organism, whereas the newly added and patchily distributed genes would not only contribute to generic and functional diversity among lineages directly but also disproportionately underlie their continuous adaptation to environmental changes. Furthermore, if adaptation preferentially takes place in young and lineage-specific genes while deleterious mutations preferentially land in ancient and shared genes, then the ways organisms fail would bear more resemblance with each other than the ways in which they adapt. The case in point is that most human genes with known disease-causing mutations do tend to be old (Domazet-Loso and Tautz 2008; Cai et al. 2009). This is good news for the investigation of human disease through the investigation of even distantly related animal models.

**Supplementary Material**

Supplementary figures S1–S8 and tables S1–S4 are available at Genome Biology and Evolution online (http://www.oxfordjournals.org/our_journals/gbe/).

**Acknowledgments**

We thank Philipp Messer, Adam Eyre-Walker, Marta Wayne, and anonymous reviewers for valuable comments on the manuscript; Carlos Bustamante, Allen Orr, Aaron Hirsh, David Lawrie, Ruth Hershberg, Josefina Gonzalez, and M. Mar Albà for helpful discussions; and Amit Indap and Adam Boyko for helping with data analysis. Part of this work was carried out by using the resources of the Computational Biology Service Unit from Cornell University, which is partially funded by Microsoft Corporation. We also thank the Stanford Beckman Service Center for Bioinformatics Resources for technical support and Stanford Bio-X2 cluster (funded by the National Science Foundation award CNS-0619926) for computer resources.

**Literature Cited**


Elhaik E, Sabath N, Graur D. 2006. The “inverse relationship between evolutionary rate and age of mammalian genes” is an artifact of
increased genetic distance with rate of evolution and time of divergence. Mol Biol Evol. 23:1–3.


Associate editor: Marta Wayne