Young proteins experience more variable selection pressures than old proteins

Anchal Vishnoi, Sergey Kryazhimskiy, Georgii A. Bazykin, et al.

Genome Res. 2010 20: 1574-1581 originally published online October 4, 2010
Access the most recent version at doi:10.1101/gr.109595.110

Young proteins experience more variable selection pressures than old proteins

Anchal Vishnoi,[1] Sergey Kryazhimskiy,[1] Georgii A. Bazykin,[2] Sridhar Hannenhalli,[3,4] and Joshua B. Plotkin[1,4]

[1]*Department of Biology, University of Pennsylvania, Philadelphia, Pennsylvania 19104, USA;* [2]*Institute for Information Transmission Problems of the Russian Academy of Sciences, Moscow 127994, Russia;* [3]*Department of Genetics, University of Pennsylvania, Philadelphia, Pennsylvania 19104, USA*

It is well known that young proteins tend to experience weaker purifying selection and evolve more quickly than old proteins. Here, we show that, in addition, young proteins tend to experience more variable selection pressures over time than old proteins. We demonstrate this pattern in three independent taxonomic groups: yeast, *Drosophila*, and mammals. The increased variability of selection pressures on young proteins is highly significant even after controlling for the fact that young proteins are typically shorter and experience weaker purifying selection than old proteins. The majority of our results are consistent with the hypothesis that the function of a young gene tends to change over time more readily than that of an old gene. At the same time, our results may be caused in part by young genes that serve constant functions over time, but nevertheless appear to evolve under changing selection pressures due to depletion of adaptive mutations. In either case, our results imply that the evolution of a protein-coding sequence is partly determined by its age and origin, and not only by the phenotypic properties of the encoded protein. We discuss, via specific examples, the consequences of these findings for understanding of the sources of evolutionary novelty.

[Supplemental material is available online at http://www.genome.org.]

Protein sequences vary by three orders of magnitude in their characteristic rates of evolution. What determines a protein's rate of evolution has been debated for several decades. Early hypotheses suggested that a protein's evolution is governed by at least two factors: the protein's level of functional constraint (i.e., the density of functionally important residues) (Ingram 1961) and the protein's overall importance to the organism (Wilson et al. 1977). Most research on the determinants of protein evolution has interrogated these structural/functional hypotheses by comparing inferred substitution rates with high-throughput phenotypic measurements. Researchers have considered many phenotypic covariates of evolutionary rates, including the measured number of physical and genetic interactions of a protein, its codon usage, the fitness consequences of its knockout, its sequence length (Lipman et al. 2002), its mRNA level, and its protein level (see Pal et al. 2006 and references therein). Of these covariates, expression levels show the strongest correlations with evolutionary rates (Pal et al. 2001; Drummond et al. 2006), leading to the view that selection against mistranslation-induced misfolding plays a dominant role in shaping the evolutionary rates of protein sequences (Drummond and Wilke 2008; Lobkovsky et al. 2010). However, the interdependencies among these phenotypic features, and the variable amounts of noise with which they have been measured, complicate the causal interpretation of their observed correlations with evolutionary rate (Plotkin and Fraser 2007; Wolf et al. 2009).

Most comparative analyses, including the studies referenced above, assume that the characteristic rate at which a protein sequence evolves depends on structural, functional, and expression constraints that are intrinsic to the encoded protein and that can be measured by phenotypic assays within a given organism. However, some studies have raised an intriguing possibility that the rate of a protein's evolution may partly be determined by its evolutionary origin, independent of its intrinsic characteristics (Domazet-Loso and Tautz 2003; Daubin and Ochman 2004; Alba and Castresana 2005, 2007; Garcia-Vallve et al. 2005; Wolf et al. 2009). After all, genomes are comprised of heterogeneous sets of genes that differ not only in their functions, but also in their evolutionary histories. Some genes in a genome are "old," in the sense that they have identifiable orthologs across a diverse range of species spanning vast evolutionary distance. Other genes are "young" in the sense that orthologs are identifiable only in closely related species. Young genes can arise in a genome by duplication, followed by accelerated substitutions, so that similarity to the ancestral copy becomes undetectable (Ohno 1970; Long 2001; Lynch and Katju 2004), by conversion of noncoding sequences (Long et al. 2003; Levine et al. 2006; Cai et al. 2008; Heinen et al. 2009) or by other mechanisms (Toll-Riera et al. 2009).

It is well established that a gene's origin can influence its molecular evolution: In bacteria, the so-called "ORFan" genes (genes with no homologs outside of a closely related group of strains or species) are less likely to be lost in evolution than other genes (Van Passel et al. 2008); synonymous substitution rates of newly acquired genes in bacteria and viruses depend on the nucleotide composition of their source genomes (Lawrence and Ochman 1997; Kryazhimskiy et al. 2008); in both prokaryotes and eukaryotes, genes that have duplicated recently experience elevated substitution rates compared with those that duplicated in the distant past (Lynch and Conery 2000; Jordan et al. 2004).

More generally, old genes are known to evolve more slowly and experience stronger purifying selection than young genes (Domazet-Loso and Tautz 2003; Daubin and Ochman 2004; Alba and Castresana 2005, 2007; Garcia-Vallve et al. 2005; Wolf et al.

[4]Corresponding authors.
E-mail jplotkin@sas.upenn.edu.
E-mail sridharh@pcbi.upenn.edu.
Article published online before print. Article and publication date are at http://www.genome.org/cgi/doi/10.1101/gr.109595.110.

20:1574–1581 © 2010 by Cold Spring Harbor Laboratory Press; ISSN 1088-9051/10; www.genome.org

2009). Since gene age is defined by the phylogenetic breadth of significant BLAST hits, some have argued that such a result is tautological and circular (Elhaik et al. 2006). However, others have convincingly demonstrated that this pattern is nontrivial: The characteristic evolutionary rate of a protein in recent times, i.e., within one taxonomic group such as *Aspergillus*, depends on whether the protein has a homolog in distant species outside of the taxonomic group (Alba and Castresana 2007; Wolf et al. 2009). Moreover, the differences in evolutionary rates between young and old genes are significant even when controlling for expression level and functional characteristics (Wolf et al. 2009). This observation is intriguing in light of the ongoing debate about the determinants of evolutionary rates, because it suggests that a protein retains some "memory" of its age and that its molecular evolution is influenced by its time of origin, not only by its function or expression level.

Here, we demonstrate a further nontrivial relationship between gene age and gene evolution: During relatively recent evolutionary timescales, young proteins exhibit more variable selection pressures (d_N/d_S) than old proteins. We show that this pattern is highly significant even after controlling for the fact that (1) old proteins tend to experience stronger purifying selection overall (Alba and Castresana 2005), and (2) old proteins tend to be longer (Wolf et al. 2009). We interpret these results in light of simple models of protein evolution under constant and variable selective regimes. On the one hand, our findings suggest that young genes tend to experience variable selective regimes because they lose their function along some lineages and gain new functions along others. At the same time, some of our results are consistent with the idea that newly arisen genes serve the same function across divergent lineages, but nevertheless appear to evolve under temporally variable selection pressures, as measured by d_N/d_S, due to depletion of sites that confer adaptive benefits (see below).

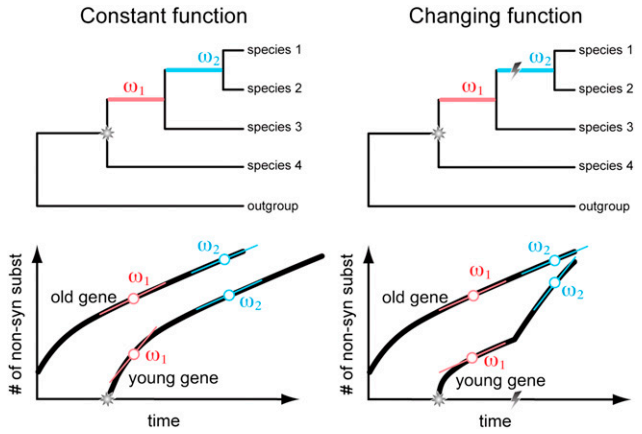

Figure 1. According to conceptual models of protein evolution, young genes experience more variable selection pressures over time than old genes. The two examples shown here illustrate a young gene that arose at the common ancestor of species 1–4 (stars), compared with an old gene that arose in the distant past. (*Bottom*) The cumulative number of nonsynonymous substitutions over time, the slope of which is proportional to the d_N/d_S value. If the young gene retains its original function (*left*), it rapidly depletes adaptive sites, resulting in a dramatic slowdown in the nonsynonymous substitution rate: $\omega_2 < \omega_1$. If the young gene experiences a functional change in a lineage (either relaxed constraint or neofunctionalization; *right*), as indicated by a lighting bolt, it rapidly accumulates nonsynonymous substitutions, resulting in $\omega_2 > \omega_1$. In both cases the young gene experiences variable selection pressures (i.e., large $|\omega_1 - \omega_2|$), whereas the old gene experiences roughly the same selection pressure in recent times, because it retains its function and most of its adaptive sites have already been depleted: $\omega_2 \approx \omega_1$.

Results

Conceptual background

We first discuss some conceptual models of protein evolution that motivate our study. We outline two mechanisms (illustrated in Fig. 1) by which young genes might be expected to experience more variables selection pressure than old genes.

First, since a newly arisen gene is likely to perform a redundant (e.g., if created by duplication) or highly specialized (if created de novo or by horizontal transfer) function (Domazet-Loso and Tautz 2003; Daubin and Ochman 2004), it is more prone than an old gene either to lose its function or to acquire novel functions in subsequent lineages (Wolf et al. 2009). Under these scenarios, the selective constraint on the newly arisen protein will either be relaxed or increased along subsequent lineages in the phylogeny. Consequently, the d_N/d_S ratio of a young gene is expected to increase or decrease along temporally subsequent branches in a phylogeny. In contrast, an old gene is more likely to be highly optimized for its (typically essential) function (Wolf et al. 2009), resulting in a relatively constant d_N/d_S ratio over recent times.

Aside from the mechanism above, there is a second, more subtle reason why a newly arisen gene might experience more variable d_N/d_S ratios—a reason that applies even if the newly arisen gene serves the same, constant function in all species descendant from the species in which it originated. In the case of a constant function, a gene will gradually accumulate adaptive substitutions in all lineages, which increase organismal fitness by improving, say, the protein's enzymatic activity. Neutral (typically synonymous) mutations will accrue at a constant rate, but, under most fitness landscapes, adaptive (typically nonsynonymous) mutations will eventually be depleted, resulting in a gradual slowdown in the rate of nonsynonymous substitutions (Hartl et al. 1985). The quantitative properties of such a substitution pattern have recently been studied theoretically (Kryazhimskiy et al. 2009), and such patterns have been directly observed in specific bacterial enzymes (Hartl et al. 1985). As these theories and experiments show, a gene under a constant selective regime (i.e., selection for a constant function) will initially accrue beneficial substitutions rapidly, but eventually such substitutions will accrue more slowly as the supply of adaptive sites is depleted. As a result of this process, the d_N/d_S ratio will tend to decrease over evolutionary time—a pattern that we might interpret as "changing selection pressures," even though the protein is adapting on a static fitness landscape.

Under this scenario of constant function, a young gene initially has a large supply of available adaptive mutations. By substituting some of these mutations, the gene's function is slowly optimized and the supply of further adaptive mutations is diminished; thus, we expect that the d_N/d_S value of a young gene will decrease over time. In contrast, an old gene, being highly optimized to its function, is likely to have already exhausted all beneficial mutations by recent times; we expect it to evolve under negative selection and fix only neutral and/or nearly neutral mutations. Thus, old genes tend to fall on the asymptote of the substitution trajectory, and so, their d_N/d_S values will remain roughly constant over recent times. Thus, under this simple model of an adaptive landscape (Hartl et al. 1985), a newly arisen gene that serves a constant function over all descendant species should be expected to exhibit descending d_N/d_S values along temporally

subsequent branches in a recent phylogeny, whereas an old gene of constant function is expected to have relatively constant d_N/d_S values within a recent phylogeny.

In summary, according to the conceptual models discussed above, young genes, whether they change their function or retain the same function subsequent to their origin, are expected to experience more variable d_N/d_S ratios over time than old genes. In subsequent sections we compare these theoretical expectations with sequence data.

Branch-specific d_N/d_S values in yeast, *Drosophila*, and mammals

We studied temporal variation in the selective constraints on proteins in three independent taxonomic groups: yeast, *Drosophila*, and mammals. Within each of these groups we identified five species or groups of species (see Methods). The species were chosen so as to ensure appropriate divergence along all branches in the five-species phylogenetic tree in order to obtain reliable estimates of branch-specific d_N/d_S ratios (Kryazhimskiy and Plotkin 2008). Within each taxonomic group, for each orthologous gene family with exactly one ortholog per species, we performed a multiple alignment or obtained alignments from public databases, as described in the Methods section. For each alignment we used PAML to estimate four or five d_N/d_S values (Fig. 2): ω_0 for the branch connecting the outgroup; ω_1 and ω_2 for two subsequent internal branches, ω_3 for most terminal branches, and ω_4 for additional internal branches (for *Drosophila*). Genes with d_S estimates smaller than 0.02 in any branch were excluded from further analysis (other cutoffs produced similar results). This criterion yielded 2807 (of originally 3747) ortholog groups in yeast; 9118 (of 9657) ortholog groups in *Drosophila*; and 10,637 (of 13,771) ortholog groups in mammals.

Old proteins are longer and exhibit stronger purifying selection than young proteins

We categorized each orthologous group as either "old" or "young" depending on whether or not a homolog was identifiable in bacteria (see Methods). Alternative definitions of old and young genes based on a more closely related outgroup species produced similar results (see below).

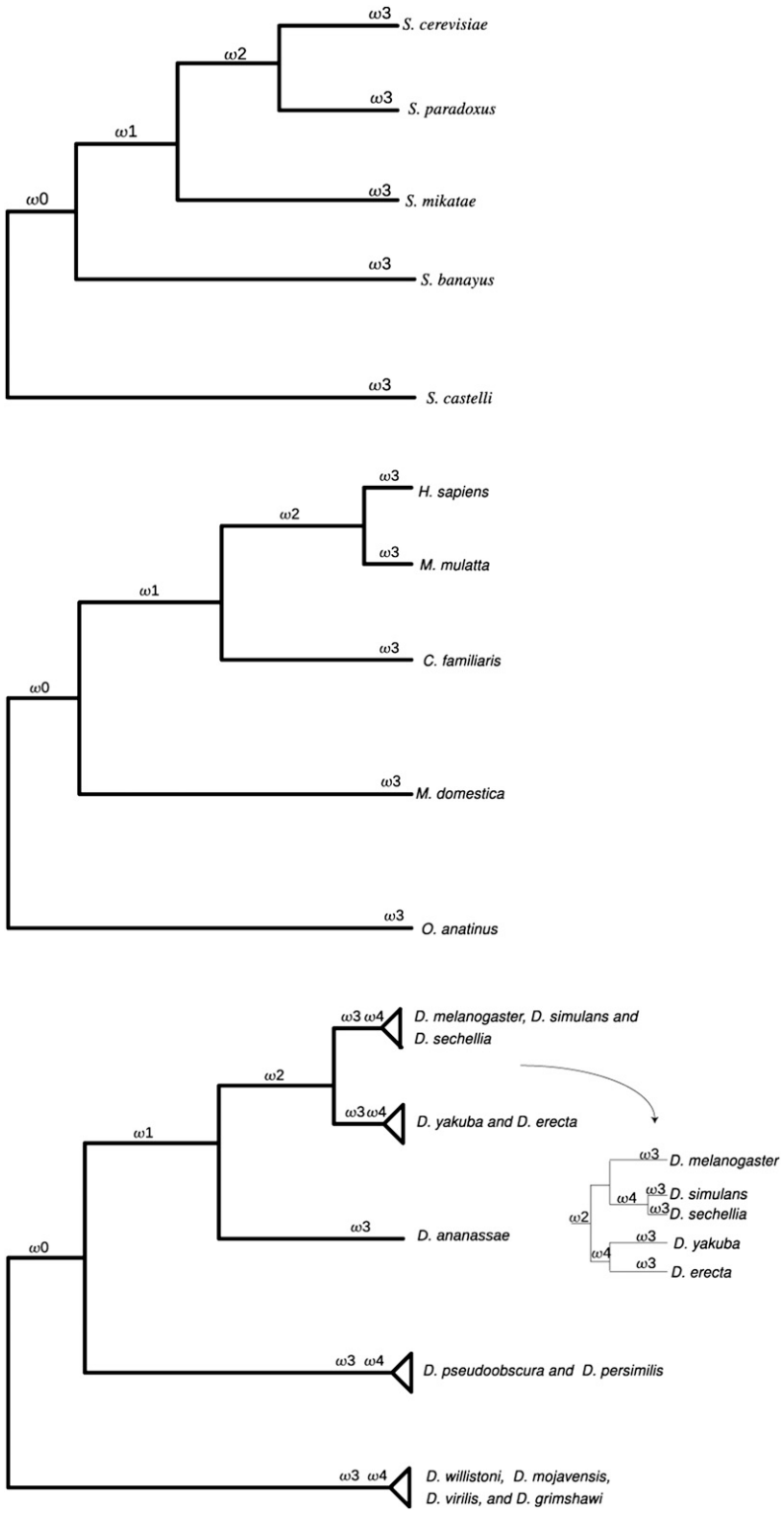

Figure 2. The five-species phylogeny for each of the taxa analyzed in our study (yeast, mammals, *Drosophila*). For each ortholog we estimated branch-specific d_N/d_S values along two subsequent branches (ω_1, ω_2), as well as separate values for most terminal branches (ω_3), the outgroup branch (ω_0), and additional internal branches in *Drosophila* (ω_4). We quantified the temporal variability in d_N/d_S as the absolute difference $\nu = |\omega_1 - \omega_2|$.

As was reported previously (Wolf et al. 2009), we observed that old genes exhibit stronger purifying selection in recent times than young genes do, on average. In other words, in each of the three taxonomic groups the distribution of ω_1 is skewed toward smaller values among old genes than among young genes (and similarly for ω_2). Despite this trend, there is a large overlap in the d_N/d_S values between old and young genes, as reported previously (Fig. 3; Wolf et al. 2009).

In addition, we observed that old genes are longer, on average, than young genes, as reported previously (Wolf et al. 2009). Nonetheless, there is a substantial overlap in the lengths of old and young genes as well (data not shown).

Young proteins experience more variable selection pressure than old proteins

For each of the three taxonomic groups we studied, the five-species phylogeny contains two internal, temporally subsequent branches (Fig. 2). For each orthologous protein alignment, we calculated the corresponding d_N/d_S values along these subsequent branches, denoted ω_1 and ω_2. In order to inspect the temporal variation in selection pressures we studied the absolute difference in d_N/d_S between these two branches: $v = |\omega_1 - \omega_2|$. (A related measure of variability in d_N/d_S, namely $\max\{\omega_1/\omega_2, \omega_2/\omega_1\}$, yields similar results.)

A meaningful comparison of variability in d_N/d_S between old and young genes requires that we first control for the fact that old genes tend to have smaller ω_1 values. After all, d_N/d_S values are constrained to be nonnegative, and so a gene's ω_1 value will naturally influence the value of $v = |\omega_1 - \omega_2|$. Similarly, we must also control for the fact that old genes are typically longer than new genes, as length could also influence the value of v. In fact, we control for the joint distribution of ω_1 and length in order to control for the aforementioned effects as well as any possible interaction between length and ω_1 that might influence variability of d_N/d_S.

In order to control for these potential biases, we sampled an equal number of old and young genes in such a way so as to ensure that the joint distribution of ω_1 values and lengths among the old genes in the sample was the same as the joint distribution among the young genes in the sample (see Methods). This conservative sampling approach discards lots of data, resulting in 710, 5346, and 8088 sampled genes in yeast, *Drosophila*, and mammals, respectively (in each case, half of the sampled genes were old and half young). Using Peacock's two-dimensional version of the Komolgorov-Smirnov test (Peacock 1983), we verified that the joint distributions of ω_1 and length did not differ between the two classes of sampled genes (Peacock test: $P = 0.9$ for yeast, $P = 0.2$ for mammals, $P = 0.5$ for *Drosophila*). (The distributions of ω_1 and v values among the 8088 sampled mammalian genes are shown as Supplemental Figs. S1 and S2.) We then compared the values of $v = |\omega_1 - \omega_2|$ between these old and young genes, using the Wilcoxon test (one-sided alternative). This procedure provides a test of whether young genes exhibit significantly greater temporal variation in selection pressures (i.e., larger values of v) than old genes, controlling for both length and ω_1.

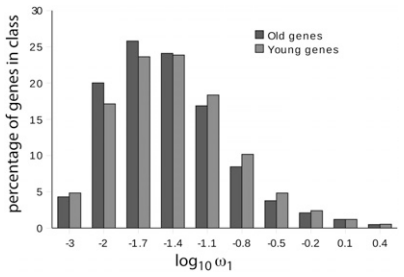

Figure 3. The distribution of d_N/d_S values (ω_1) among old and young mammalian genes. The old genes are skewed toward lower d_N/d_S values, but there is a substantial overlap. When comparing old and young genes we control for this bias by analyzing a subset of old and young genes with the same distributions of ω_1 values.

In mammals, we found that young genes exhibited significantly more variable selection pressures, i.e., larger values of v, than old genes ($P = 8.8 \times 10^{-5}$, Wilcoxon test). In particular, the average variability in selection pressures was 25% higher among young genes compared with old genes. Similarly, in *Drosophila* young genes also exhibited significantly more variable selection pressures than old genes (average v was 26% greater among young genes, $P = 7.5 \times 10^{-7}$) and also in yeast (average v was 56% greater among young genes, $P = 5.0 \times 10^{-3}$). Thus, in all three taxa we find the same phenomenon: Young genes experience significantly more variable selection pressures, on average, than old genes (Table 1).

Table 2 shows the proportion of sampled genes, in either the young or old categories, whose d_N/d_S values increased, decreased, or remained roughly constant over time. As the analysis above would suggest, young genes were more likely to exhibit changing d_N/d_S values, compared with old genes. When d_N/d_S values did change over time, there was a bias toward increasing values, especially among the young genes.

Table 2 shows that there is a greater tendency toward $\omega_2 > \omega_1$ in young genes than there is in old genes. This implies that "depletion of adaptive sites" is not the primary mechanism responsible for the greater variability in d_N/d_S that we have observed among young genes. Nevertheless, the depletion mechanism is likely to contribute to the increased rate variability in some of the young genes.

Alternative definitions of gene age produce the same qualitative pattern

In the analysis above we used the existence of a homolog in bacteria to define "old" versus "young" genes. For *Drosophila* and mammalian genomes, the choice of bacterial species as the outgroup set is quite conservative, producing a stringent definition of an old gene. To ensure that our results are robust with respect to outgroup choice, we repeated the above analyses using a diverse set of fungi species (*Saccharomyces cerevisiae*, *Debaryomyces hansenii*, *Shizosaccharomyces pombe*, and *Cryptococcus neoformans*) as the outgroup set for defining old *Drosophila* genes, and the *Anopheles gambiae*, *Drosophila melanogaster* and *Caenorhabditis elegans* genomes as the outgroup set for defining old mammalian genes. These new criteria resulted in a different number of sampled old–young gene pairs that control for the joint distributions of length and ω_1 (3008 *Drosophila* pairs and 2754 mammalian pairs), but nonetheless highly significant differences in the variability of selection pressures in old versus young genes (Wilcoxon test, 5×10^{-6} for *Drosophila*, and 4×10^{-21} for mammals). Thus, even with a more inclusive definition of "old," the data support our contention that old genes tend to experience less variable selection pressures than young genes.

In addition, we randomly permuted the "old" and "young" labels of genes, as a negative control, and repeated the analysis above. In this case, none of the Wilcoxon tests was significant ($P = 0.54$ for yeast, $P = 0.78$ for *Drosophila*, $P = 0.55$ for mammals), as expected.

Table 1. The number of genes used in our statistical analysis, the mean variability in d_N/d_S ($\nu = |\omega_1 - \omega_2|$) among old and young genes, the proportional elevation of variability among young genes, and the corresponding Wilcoxon *P*-value

	Mammals	*Drosophila*	Fungi
No. of old-young pairs	4044	2682	355
Mean of young ν	0.1454	0.0559	0.1184
Mean of old ν	0.1165	0.0444	0.0760
Difference in mean ν	25%	26%	56%
Wilcoxon *P*-value	8.8×10^{-5}	7.5×10^{-7}	5.0×10^{-3}

In each of the three taxonomic groups young genes are significantly more variable in their d_N/d_S values, on average, than old genes.

Gene age, but not housekeeping functionality, determines variation in selection pressures

Old genes are known to be enriched for housekeeping functions (Wolf et al. 2009). Therefore, the possibility remains that gene function, as opposed to gene age itself, underlies the observed differential variability in selection pressures between old and young genes. To test for this possibility we asked whether categorization of genes by function (housekeeping or not) would lead to the same qualitative results as those for categorization by age (old vs. young).

To perform this analysis, we considered the previously identified set of 1364 human housekeeping genes (Podder and Ghosh 2010). Such genes are enriched for critical biological functions, including translation, cellular biosynthesis, etc. All other human genes were categorized as "nonhousekeeping." We repeated the analysis above on the resulting 1281 pairs of housekeeping/nonhousekeeping genes that had been chosen to match their distributions of ω_1, as above. We found no significant difference in the variability of selection pressures (ν) between the two functional classes of genes (Wilcoxon test, $P = 0.46$) despite the fact that a comparison between old and young genes of the same power (1281 old/young pairs) yields a highly significant difference (Wilcoxon test, $P = 2.1 \times 10^{-3}$). Thus, the observed difference in selection pressures between old and young genes is unlikely to be caused by differences in housekeeping function, but rather should be attributed to differences in gene age itself.

Gene age, but not expression level, determines variation in selection pressures

Because old genes are more often essential (Wolf et al. 2009), they may be expressed at higher levels on average than young genes. Some models for the relationship between expression level and evolutionary rate suggest that genes expressed at high levels should exhibit both lower d_N/d_S values and less variation in d_N/d_S values compared with genes expressed at low levels (Gout et al. 2010). Therefore, much like the concern about housekeeping functions above, the possibility remains that expression level, as opposed to gene age itself, underlies the observed differential variability in selection pressures between old and young genes. To test for this possibility we asked whether categorization of genes by expression level (high vs. low) would lead to the same qualitative results as categorization by age (old vs. young).

We performed this analysis for the yeast and *Drosophila* because the putative relationships between expression level and rate variability do not apply to mammals (Gout et al. 2010). Ideally, we would like to categorize all genes into those expressed at high and those expressed at low level in natural conditions. As is common practice, we used a gene's codon adaptation index (CAI) as a surrogate for its typical expression level in nature (Sharp and Li 1987). We divided *S. cerevisiae* genes into those expressed at high levels (top 5% of CAI values) and the remaining genes. We repeated the analysis above on the resulting 110 pairs of high/low CAI genes that had been matched to control for their ω_1 and length distributions, as above. We found no significant difference in the variability of selection pressures (ν) between these two classes of genes (Wilcoxon test, $P = 0.09$), despite the fact that a comparison between old and young yeast genes of the same power (110 old/young pairs) yields a significant difference (Wilcoxon test, $P = 0.01$). The same results hold in *Drosophila* as well: There is no significant difference in the variability of selection pressures (ν) between 425 pairs of high/low CAI genes matched to control for their ω_1 and length distributions (Wilcoxon test, $P = 0.38$), despite the fact that a comparison between old and young *Drosophila* genes of the same power (425 old/young pairs) yields a significant difference (Wilcoxon test, $P = 0.008$). Thus, the observed difference in selection pressures between old and young genes is unlikely to be caused by differences in expression levels.

Illustrative examples of variable selection pressures on young genes

The statistical analyses above reveal a significant tendency for young proteins to experience more variable selection pressures over time, compared with old proteins. Below, we discuss three examples of young genes drawn from the mammalian group, which illustrate the various evolutionary scenarios that may be responsible for the observed overall pattern. We emphasize that these three examples are included below only to illustrate the idea behind possible underlying mechanisms, and they do not constitute proof of each mechanism. Although the statistical trends we have observed are highly significant, it is easy to find examples that do or do not follow the general trend. Nonetheless, the anecdotal examples are useful for illustrative purposes.

Example 1: $\omega_2 > \omega_1$ due to relaxed negative selection

OR5A1 is an olfactory receptor belonging to the largest multigene family, expressed in a wide variety of metazoan species (Buck and Axel 1991; Freitag et al. 1995; Barth et al. 1997). The d_N/d_S value of

Table 2. The proportion of old and young sampled genes in each of three taxa that exhibited increasing, decreasing, or roughly constant d_N/d_S values

	Old genes	Young genes
Mammals (4044 old/young pairs)		
$\omega_2 > \omega_1$	32%	34%
$\omega_2 < \omega_1$	14%	16%
$\omega_2 \approx \omega_1$	54%	50%
Drosophila (2682 old/young pairs)		
$\omega_2 > \omega_1$	11%	17%
$\omega_2 < \omega_1$	9%	8%
$\omega_2 \approx \omega_1$	80%	75%
Yeast (355 old/young pairs)		
$\omega_2 > \omega_1$	26%	44%
$\omega_2 < \omega_1$	16%	7%
$\omega_2 \approx \omega_1$	58%	49%

$\omega_1 \approx \omega_2$, if the two values fall within 5% of each other. In each of the taxa, a greater proportion of young genes exhibit changing d_N/d_S values than the corresponding proportion of old genes.

OR5A1 increases along the primate lineage (ω_1 = 0.03 and ω_2 = 0.91), consistent with the known relaxation of negative selection on olfactory receptors in primates compared with other mammals (Glusman et al 2001). Thus, OR5A1 exemplifies the case of a young protein whose temporal variation in selection pressures is probably caused by loss of functionality in a very recent lineage.

Example 2: $\omega_2 > \omega_1$ due to increased positive selection

Lysozyme (LYZ) cleaves the peptidoglycan of the bacterial cell wall. Several sites in the gene encoding this enzyme are known have experienced positive along the primate lineage (Messier and Stewart 1997; Yang 1998), resulting in a new function that allows primates to digest bacteria (Leonard 2002). As expected, we observed an increase in d_N/d_S along the primate lineage within mammals (ω_2 = 1.01 vs. ω_1 = 0.11). Thus, *LYZ* exemplifies that the case of a young gene with temporal variation in selection pressures may be due to changing functionality across lineages.

Example 3: $\omega_2 < \omega_1$ due to possible depletion of adaptive sites

Oligodendrocyte myelin glycoprotein (OMG) is an anchor protein that originated along the mammalian lineage after mammals and fish diverged from their common ancestor (Kehrer-Sawatzki et al. 1998). The protein generally experiences strong purifying selection in mammals, presumably because of its important role in brain development (Kuma et al. 1995; Duret and Mouchiroud 2000; Wang et al. 2007). OMG is required for the production of the myelin sheath that allows rapid conduction in axons, and it serves this same function in all mammals (Wang et al. 2002). Since this young gene arose recently (i.e., near the base of the mammalian clade) and serves the same function across all mammalian species, we might expect that d_N/d_S will decrease along subsequent mammalian branches due to depletion of adaptive sites. Indeed, this pattern is observed (ω_1 = 0.44 and ω_2 = 0.17), which is consistent with the plausible mechanism of depleting adaptive sites.

Discussion

We have demonstrated that the age of a gene influences the tempo of its molecular evolution: Young genes experience more variable d_N/d_S values than old genes. We have observed this pattern in three independent taxonomic groups—mammals, *Drosophila*, and yeast—suggesting that the phenomenon is quite general. Although the gene age has a significant effect on the mean variability in rates, controlling for length and ω_1, age alone does not explain a large proportion of all variation in rate variability (<2% in each taxa).

Our results complement previous findings that young genes evolve faster, on average, than old genes (Domazet-Loso and Tautz 2003; Daubin and Ochman 2004; Alba and Castresana 2005, 2007; Garcia-Vallve et al. 2005; Wolf et al. 2009). Alba and Castresana (2005) proposed two alternative explanations for this trend (see also Domazet-Loso and Tautz 2003). Their "constant constraint" model posits that each gene has an intrinsic level of constraint (d_N/d_S) that it maintains throughout its life, but because faster evolving genes are generally more dispensable, they are more often lost in evolution. This model would predict no systematic differences in rate variability between old and young genes. Our analysis reveals such systematic differences and is thus evidence against the "constant constraint" model.

The "increasing constraint" model of Alba and Castresana (2005) posits that the strength of selection on a gene changes through its lifetime: Immediately after birth, a gene evolves under weak negative or positive selection, while later in life it evolves primarily under strong negative selection. The depletion of adaptive sites (Hartl et al. 1985), discussed above, provides a simple mechanistic basis for the model of "increasing constraint." Under this model, the substitution rate of a protein should decrease over time. Since the substitution rate cannot decrease indefinitely, we expect that the decrease will be faster at initial stages of a gene's life and slower at later stages. Indeed, many young genes experience substantial decreases in their d_N/d_S ratios over time (Table 2), and many old genes show no substantial changes in their d_N/d_S ratio over time, but the difference in these trends between old and young genes is not statistically significant in all taxa (data not shown)—suggesting that "increasing constraint" alone cannot explain the elevated variability of d_N/d_S ratios among young genes.

We have also observed many young genes whose d_N/d_S ratios increase over time (see Table 2), which is inconsistent with either of the models suggested by Alba and Castresana (2005). We therefore propose to extend their "increased constraint" model to incorporate a period of high variability in a gene's evolutionary rate immediately after it is born. We have conceived of two sources of such initial variability. First, a newborn gene is likely to perform a redundant or highly specialized function (Domazet-Loso and Tautz 2003), which can easily become unnecessary in a changing environment. This would lead to a relaxed constraint on the young gene and result in an increase of its d_N/d_S ratio. Later, the same gene may be co-opted to another function and evolve under positive selection, which would further elevate its d_N/d_S ratio. Alternatively, even if a newborn gene maintains its original function, it gradually exhausts available adaptive mutations, and so its d_N/d_S ratio decreases over time. According to both of these mechanisms, young genes will experience more variable d_N/d_S values over recent times than old genes, as we have observed.

Methods

Estimating d_N/d_S

In each of the broad taxonomic groups we studied—yeast, *Drosophila*, and mammals—we analyzed five species, or groups of species, including an unambiguous outgroup (see Fig. 2). The species were chosen so that the average d_S along all branches would exceed 0.02, as is required to obtain reliable estimates of d_N/d_S. In addition, species were chosen so that no branch length would exceed d_S = 1.0, on average, in order to avoid the effects of saturation. For the yeast analysis we used S*accharomyces cerevisiae, S. paradoxus*, *S. mikatae*, and *S. bayanus*, with *S. castelli* as an outgroup. For mammals, we used *Homo sapiens*, *Macaca mulatta*, *Canis familiaris*, and *Monodelphis domestica*, with *Ornithorhynchus anatinus* as an outgroup. For *Drosophila*, we grouped the 12 species with sequenced genomes into five groups: *Drosophila melanogaster*, *D. simulans*, and *D. sechellia* comprised the first group; *D. yakuba* and *D. erecta* comprised the second group; *D. ananassae* comprised the third group; *D. pseudoobscura* and *D. persimilis* comprised the fourth group; *D. willistoni*, *D. virilis*, *D. mojavensis*, and *D. grimshawi* comprise the fifth group, and was used as the outgroup.

In yeast, we identified orthologous open reading frames from ftp://ftp.ncbi.nih.gov/genomes/Fungi/. We then aligned the orthologous yeast ORFs using MUSCLE, and superimposed this alignment onto the amino acid sequence using pal2nal (Suyama et al. 2006). For mammals, we used the orthologs and alignments from the UCSC Genome Browser. For *Drosophila*, we used the orthologs and alignments from FlyBase (ftp://ftp.flybase.net/genomes/12_species_analysis/clark_eisen/).

The orthologous gene alignments for yeast, *Drosophila*, and mammals were each analyzed using PAML to estimate the d_N/d_S

values in the five-species tree: ω_0 for the branch connecting the outgroup; ω_1 and ω_2 for two subsequent internal branches, ω_3 for most terminal branches, and ω_4 for additional internal branches (if any). All d_N/d_S values were estimated using maximum likelihood (PAML model 2). Genes with $\omega_1 > 5$ or $\omega_2 > 5$ were discarded, because these genes were typically associated with imprecise d_S estimates. Note that we did not use the d_N/d_S values associated with terminal branches in our analysis because they are known to be biased due to weakly selected mutations that segregate in the population of each species. Treating such segregating mutations as fixed differences, as PAML does, leads to an upward bias in the d_N/d_S value along a terminal branch (Rocha et al. 2006; Kryazhimskiy and Plotkin 2008).

Definitions of old and young genes

Genes were categorized as either old or young, depending on the presence of a homolog in a distant reference species or set of species. Using a BLAST *E*-value threshold of 10^{-6}, as in Wolf et al. (2009), any gene with an ortholog in bacteria (ftp://ftp.ncbi.nih.gov/genomes/Bacteria/) was considered to be old, while all other genes were considered to be young.

For *Drosophila* and mammals we explored additional alternative definitions of old and young genes. In particular, for *Drosophila* we alternatively used the presence/absence of a BLAST hit in *S. cerevisiae*, *D. hansenii*, *S. pombe*, and *C. neoformans* or *C. albicans* as the definition of old/young genes. And for mammals, we alternatively used the presence/absence of a BLAST hit in *A. gambiae*, *D. melanogaster*, or *C. elegans* as the definition of old/young genes (see main text).

Sampling old and young genes to control for ω_1 and length

In order to control for differences in the joint distributions of ω_1 and the length between old and young genes, we sampled a subset of genes for which the joint distributions were virtually identical using the following sampling procedure. We first binned gene lengths and ω_1 values into 200 arithmetic bins each, resulting in a two-dimensional array of 40,000 cells. Each one of these cells represented a narrow range of allowable lengths and ω_1 values and contained a certain number of corresponding old genes, n, and a certain number of corresponding young genes, m. Considering each cell in turn, we sampled without replacement from each class the number of genes equal to the lesser of n and m (e.g., in the case $n < m$ we sampled all n of the old genes in the cell and n of the young genes, discarding the remaining young genes). After performing this procedure for each cell, we obtained an equal number of sampled old and young whose joint distributions of lengths and ω_1 values were guaranteed to be nearly identical. We verified that the resulting sampled genes did not differ in their joint distributions using Peacock's two-dimensional Komolgorov-Smirnov test (Peacock 1983).

Applying this procedure to the mammalian genes produced 4044 sampled genes (half of these young and half old). As desired, we found no significant difference in the joint distribution of ω_1 values and lengths between the sampled old and young mammalian genes (Peacock test, $P = 0.2$). Similarly, for *Drosophila* we obtained 2682 sampled genes without significant differences in their joint distributions of ω_1 and length (Peacock test, $P = 0.5$). For yeast this procedure discarded many genes because of significant differences between the original joint distributions, and resulted in 355 sampled genes. Again, there was no significant difference in their joint distributions of ω_1 and length between the old and young sampled genes (Peacock test, $P = 0.9$). In each taxonomic group, we used the sampled genes when comparing the variability in selection pressures, $v = |\omega_1 - \omega_2|$, between old and young genes.

Our sampling procedure is partly random, and so it produces slightly different sets of sampled genes for each pseudorandom seed. Nonetheless, repeating the procedure above with different pseudorandom seeds does not change our qualitative results.

Acknowledgments

We thank the Plotkin and Hannenhalli labs for productive conversations and input. J.B.P. acknowledges support from Burroughs Wellcome Fund, the David and Lucile Packard Foundation, the James S. McDonnell Foundation, the Alfred P. Sloan Foundation, and the Defense Advanced Research Projects Agency (HR0011-05-1-0057). J.B.P., S.H., and A.V. acknowledge support from NIGMS R01GM085226.

References

Alba MM, Castresana J. 2005. Inverse relationship between evolutionary rate and age of mammalian genes. *Mol Biol Evol* **22:** 598–606.

Alba MM, Castresana J. 2007. On homology searches by protein Blast and the characterization of the age of genes. *BMC Evol Biol* **7:** 53. doi: 10.1186/1471-2148-7-53.

Barth AL, Dugas JC, Ngai J. 1997. Noncoordinate expression of odorant receptor genes tightly linked in the zebrafish genome. *Neuron* **19:** 359–369.

Buck L, Axel R. 1991. A novel multigene family may encode odorant receptors: A molecular basis for odor recognition. *Cell* **65:** 175–187.

Cai J, Zhao R, Jiang H, Wang W. 2008. De novo origination of a new protein-coding gene in *Saccharomyces cerevisiae*. *Genetics* **179:** 487–496.

Daubin V, Ochman H. 2004. Bacterial genomes as new genes homes: The genealogy of ORFans in *E. coli*. *Genome Res* **14:** 1036–1042.

Domazet-Loso T, Tautz D. 2003. An evolutionary analysis of orphan genes in *Drosophila*. *Genome Res* **13:** 2213–2219.

Drummond DA, Wilke CO. 2008. Mistranslation-induced protein misfolding as a dominant constraint on coding-sequence evolution. *Cell* **134:** 341–352.

Drummond DA, Raval A, Wilke CO. 2006. A single determinant dominates the rate of yeast protein evolution. *Mol Biol Evol* **23:** 327–337.

Duret L, Mouchiroud D. 2000. Determinants of substitution rates in mammalian genes: Expression pattern affects selection intensity but not mutation rate. *Mol Biol Evol* **17:** 68–74.

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.

Freitag J, Krieger J, Strotmann J, Breer H. 1995. Two classes of olfactory receptors in *Xenopus laevis*. *Neuron* **15:** 1383–1392.

Garcia-Vallve S, Alonso A, Bravo IG. 2005. Papillomaviruses: Different genes have different histories. *Trends Microbiol* **13:** 514–521.

Glusman G, Yanai I, Rubin I, Lancet D. 2001. The complete human olfactory subgenome. *Genome Res* **11:** 685–702.

Gout J-F, Kahn D, Duret L, *Paramecium* Post-Genomics Consortium. 2010. The relationship among gene expression, the evolution of gene dosage, and the rate of protein evolution. *PLoS Genet* **6:** e1000944. doi: 10.1371/journal.pgen.1000944.

Hartl DL, Dykhuizen DE, Dean AM. 1985. Limits of adaptation: The evolution of selective neutrality. *Genetics* **111:** 655–674.

Heinen TJAJ, Staubach F, Haming D, Tautz D. 2009. Emergence of a new gene from an intergenic region. *Curr Biol* **18:** 1527–1531.

Ingram VM. 1961. Gene evolution and the haemoglobins. *Nature* **189:** 704–708.

Jordan IK, Wolf YI, Koonin EV. 2004. Duplicated genes evolve slower than singletons despite the initial rate increase. *BMC Evol Biol*. **4:** 22. doi: 10.1186/1471-2148-4-22.

Kehrer-Sawatzki H, Maier C, Moschgath E, Elgar G, Krone W. 1998. Genomic characterization of the neurofibromatosis type 1 gene of Fugu rubripes. *Gene* **222:** 145–153.

Kryazhimskiy S, Plotkin JB. 2008. The population genetics of d_N/d_S. *PLoS Genet* **4:** e1000304. doi 10.1371/journal.pgen.1000304.

Kryazhimskiy S, Bazykin GA, Dusshoff J. 2008. Natural selection for nucleotide usage at synonymous and nonsynonymous sites in influenza A virus genes. *J Virol* **82:** 4938–4945.

Kryazhimskiy S, Tkacik G, Plotkin JB. 2009. The dynamics of adaptation on correlated fitness landscapes. *Proc Natl Acad Sci* **106:** 18638–18643.

Kuma K, Iwabe N, Miyata T. 1995. Functional constraints against variations on molecules from the tissue level: Slowly evolving brain-specific genes demonstrated by protein kinase and immunoglobulin supergene families. *Mol Biol Evol* **12:** 123–130.
Lawrence JG, Ochman H. 1997. Amelioration of bacterial genomes: Rates of change and exchange. *J Mol Evol* **44:** 383–397.
Leonard WR. 2002. Food for thought. Dietary change was a driving force in human evolution. *Sci Am* **287:** 106–115.
Levine MT, Jones CD, Kern AD, Lindfors HA, Begun DJ. 2006. Novel genes derived from noncoding DNA in *Drosophila melanogaster* are frequently X-linked and exhibit testis-biased expression. *Proc Natl Acad Sci* **103:** 9935–9939.
Lipman DJ, Souvorov A, Koonin EV, Panchenko AR, Tatusova TA. 2002. The relationship of protein conservation and sequence length. *BMC Evol Biol* **2:** 20. doi: 10.1186/1471-2148-2-20.
Lobkovsky AE, Wolf YI, Koonin EV. 2010. Universal distribution of protein evolution rates as a consequence of protein folding physics. *Proc Natl Acad Sci* **107:** 2983–2988.
Long M. 2001. Evolution of novel genes. *Curr Opin Genet Dev* **11:** 673–680.
Long M, Betran E, Thornton K, Wang W. 2003. The origin of new genes: Glimpses from the young and old. *Nat Rev Genet* **4:** 865–875.
Lynch M, Conery JS. 2000. The evolutionary fate and consequences of duplicate genes. *Science* **290:** 1151–1155.
Lynch M, Katju V. 2004. The altered evolutionary trajectories of gene duplicates. *Trends Genet* **20:** 544–549.
Messier W, Stewart C-B. 1997. Episodic adaptive evolution of primate lysozyme. *Nature* **385:** 151–154.
Ohno S. 1970. *Evolution by gene duplication*. Springer-Verlag, Berlin/ Heidelberg/New York.
Pal C, Papp B, Hurst LD. 2001. Highly expressed genes in yeast evolve slowly. *Genetics* **158:** 927–931.
Pal C, Papp B, Lercher MJ. 2006. An integrated view of protein evolution. *Nat Rev Genet* **7:** 337–348.
Peacock JA. 1983. Two-dimensional goodness-of-fit testing in astronomy. *Mon Not R Astron Soc* **202:** 615–627.
Plotkin JB, Fraser HB. 2007. Assessing the determinants of evolutionary rates in the presence of noise. *Mol Biol Evol* **24:** 1113–1121.
Podder S, Ghosh TC. 2010. Exploring the differences in evolutionary rates between monogenic and polygenic disease genes in human. *Mol Biol Evol* **27:** 934–941.
Rocha EP, Smith JM, Hurst LD, Holden MT, Cooper JE, Smith NH, Feil EJ. 2006. Comparisons of dN/dS are time dependent for closely related bacterial genomes. *J Theor Biol* **239:** 226–235.
Sharp PM, Li WH. 1987. The codon adaptation index—a measure of directional synonymous codon usage bias, and its potential applications. *Nucleic Acids Res* **11:** 1281–1295.
Suyama M, Torrents D, Bork P. 2006. PAL2NAL: Robust conversion of protein sequence alignments into the corresponding coding alignments. *Nucleic Acids Res* **34:** W609–W612.
Toll-Riera M, Bosch N, Bellora N, Castelo R, Armengol L, Estivill X, Alba MM. 2009. Origin of primate orphan genes: A comparative genomics approach. *Mol Biol Evol* **26:** 603–612.
Van Passel MWJ, Marri PR, Ochman H. 2008. The emergence and fate of horizontally acquired genes in *Escherichia coli*. *Plos Comput Biol*. **4:** e1000059. doi: 10.1371/journal.pcbi.1000059.
Wang KC, Koprivica V, Kiam JA, Sivasankaran R, Guo Y, Neve RL, He Z. 2002. Oligodendrocyte-myelin glycoprotein is a Nogo receptor ligand that inhibits neurite outgrowth. *Nature* **417:** 941–944.
Wang HY, Chien HC, Osada N, Hashimoto K, Sugano S, Gojobori T, Chou CK, Tsai SF, Wu CI, Shen CKJ. 2007. Rate of evolution in brain-expressed genes in humans and other primates. *PLoS Biol* **5:** e13. doi: 10.1371/ journal.pbio.0050013.
Wilson AC, Carlson SS, White TJ. 1977. Biochemical evolution. *Annu Rev Biochem* **46:** 573–639.
Wolf YI, Novichkov PS, Karev GP, Koonin EV, Lipman DJ. 2009. The universal distribution of evolutionary rates of genes and distinct characteristics of eukaryotic genes of different apparent ages. *Proc Natl Acad Sci* **106:** 7273–7280.
Yang Z. 1998. Likelihood ratio tests for detecting positive selection and application to primate lysozyme evolution. *Mol Biol Evol* **15:** 568–573.

Received April 23, 2010; accepted in revised form September 1, 2010.