Evolutionary switches between two serine codon sets are driven by selection

Igor B. Rogozin[a], Frida Belinky[a], Vladimir Pavlenko[a], Svetlana A. Shabalina[a], David M. Kristensen[b], and Eugene V. Koonin[a,1]

[a]National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD 20894; and [b]Biomedical Engineering Department, University of Iowa, Iowa City, IA 52242

Contributed by Eugene V. Koonin, September 22, 2016 (sent for review July 7, 2016; reviewed by Kenneth H. Wolfe and Jianzhi Zhang)

Serine is the only amino acid that is encoded by two disjoint codon sets so that a tandem substitution of two nucleotides is required to switch between the two sets. Previously published evidence suggests that, for the most evolutionarily conserved serines, the codon set switch occurs by simultaneous substitution of two nucleotides. Here we report a genome-wide reconstruction of the evolution of serine codons in triplets of closely related species from diverse prokaryotes and eukaryotes. The results indicate that the great majority of codon set switches proceed by two consecutive nucleotide substitutions, via a threonine or cysteine intermediate, and are driven by selection. These findings imply a strong pressure of purifying selection in protein evolution, which in the case of serine codon set switches occurs via an initial deleterious substitution quickly followed by a second, compensatory substitution. The result is frequent reversal of amino acid replacements and, at short evolutionary distances, pervasive homoplasy.

tandem nucleotide substitutions | serine codon sets | purifying selection | positive selection | homoplasy

Point mutations are generally assumed to involve substitution of single nucleotides, an assumption that underlies many phylogenetic approaches (1–5). However, the presently characterized mechanisms of mutation do not exclude the possibility that two or more adjacent nucleotides change simultaneously in a single mutational event (6–13). There is experimental evidence that some replication and repair enzymes, for example, mammalian DNA polymerases η (14) and ζ (15, 16), produce excessive amounts of double (tandem) substitutions. Runs of consecutive mismatches between closely related genomes rapidly decrease in frequency with increasing length, following an exponential distribution and thus indicating that nucleotides are mostly substituted one at a time (17). Nevertheless, even for spontaneous mutations, there are indications that several adjacent nucleotides can change simultaneously in a single mutational event (2, 6–8, 18). In particular, Averof and colleagues analyzed nucleotide substitutions in primate noncoding sequences and found a high frequency of double (tandem) substitutions in an approximately twofold excess over the expected frequencies (2).

Averof and colleagues further analyzed evolutionary switches between TCN and AGY (where N is any nucleotide and Y is a pyrimidine) codons that encode strictly conserved serine residues in multiple protein families from diverse organisms (2). Serine is unique among amino acids in that it is encoded by two disjoint codon sets, TCN and AGY, which cannot be interconverted by a single-nucleotide mutation. Accordingly, switches between serine codons from the two sets can occur either directly, by simultaneous double (tandem) mutation (TC > AG), or indirectly, via two consecutive single-nucleotide substitutions (TC > AC > AG or TC > TG > AG). In the latter case, the switch would involve as an intermediate either threonine ACN or cysteine TGY, amino acid residues with properties substantially different from those of serine (19), so that such changes are unlikely to be tolerated at critical functional or structural sites of a protein. Nevertheless, TCN > AGY switches have been observed at sites encoding extremely conserved serine residues, for example, in ubiquitins (20) and in the active sites of serine proteases (21). Given the likely low fitness of threonine or cysteine intermediates, switches at such sites could be attributed to simultaneous double (tandem) mutations, which in this context are synonymous and most likely selectively neutral. Alternatively, mechanisms have been hypothesized to explain switches in serine codon sets through nondeleterious intermediates (21–23). For example, a transient substitution of serine by another amino acid could be complemented by the presence of a neighboring serine residue (23), or the presence of serine codons from different sets in the same essential site might reflect independent origins from different ancestral amino acids (21).

We sought to investigate the evolutionary factors that affect serine codon set switches on the genome scale in a broad range of organisms. Contributions from both selection and mutational processes were identified, but the impact of selection seems to be much more pronounced, especially in prokaryotes.

Results

Double and Single Mutations in Serine Codons in Prokaryotes. To accurately reconstruct the nucleotide substitutions that lead to serine codon switches, we analyzed triplets of species with unambiguous phylogeny for which the substitutions can be polarized (Fig. 1). Specifically, the analysis involved 37 triplets of closely related prokaryotic genomes from the Alignable Tight Genome Clusters (ATGC) database and two groups of eukaryotes (see *Materials and Methods* and *SI Appendix* for details). We first determined the frequencies of double substitutions between the two serine codon sets (Fig. 2). Altogether, there were 856 TCY > AGY substitutions and 809 AGY > TCY substitutions

Significance

When a rare evolutionary event is observed, such as substitution of two adjacent nucleotides, the question emerges whether such rare changes are caused by mutational bias or by selection. Here we address this question through genome-wide analysis of double substitutions that lead to switch of the codon sets for the amino acid serine, the only one that is encoded by two disjoint sets of codons. We show that selection is the primary factor behind these changes. These findings suggest that short-term evolution of proteins is subject to stronger purifying selection than previously thought and has significant implications for methods of phylogenetic analysis.

Author contributions: I.B.R. and E.V.K. designed research; I.B.R., F.B., and V.P. performed research; D.M.K. contributed new reagents/analytic tools; I.B.R., S.A.S., and E.V.K. analyzed data; and I.B.R. and E.V.K. wrote the paper.

Reviewers: K.H.W., University College Dublin; and J.Z., University of Michigan.

The authors declare no conflict of interest.

Freely available online through the PNAS open access option.

[1]To whom correspondence should be addressed. Email: koonin@ncbi.nlm.nih.gov.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1615832113/-/DCSupplemental.

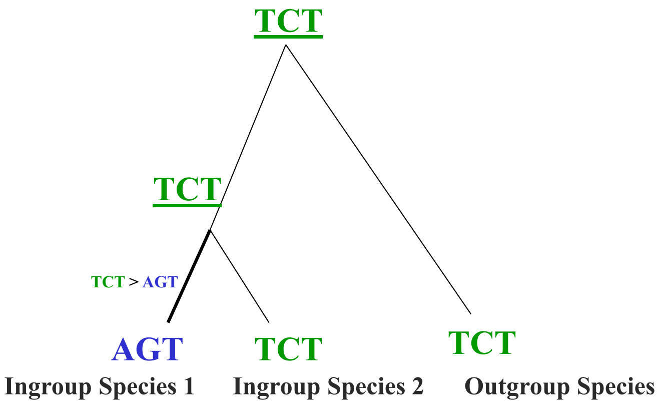

Fig. 1. Reconstruction of the history of nucleotide substitutions in three closely related genomes. The reconstructed ancestral codons are underlined. Under the parsimony principle, one or no mutation per site per branch was allowed.

(Fig. 2), numbers that are high enough to allow reliable statistical inferences. The frequencies of double substitutions in the first two serine codon positions were 0.002 for the TCY > AGY replacements and 0.0015 for the AGY > TCY replacements (Fig. 2). These frequencies are, respectively, 43 and 52 times greater than expected from the estimation of the double substitution frequencies as the products of the respective single-mutation frequencies (e.g., 0.002/(0.000043 + 0.000004) = 42.6) (*SI Appendix*, Fig. S1). Analysis of five individual ATGCs with the highest numbers of serine codon switches revealed considerable variability of the substitution frequencies (*SI Appendix*, Fig. S2), but the lowest excess of TCY ↔ AGY double mutations over the expected frequencies was more than 10-fold (*SI Appendix*, Fig. S2, ATGC199).

AGY ↔ TCY Mutations in Prokaryotic Noncoding DNA. In the noncoding ATGC sequences, we identified 53 TCY > AGY replacements and 49 AGY > TCY replacements (Fig. 3). The frequencies of these double substitutions (0.0002 for TCY > AGY and 0.0003 for AGY > TCY) were substantially and significantly [P(Fc_obs/exp > Fn_obs/exp) < 0.0001] lower than the respective frequencies in the protein-coding regions but still approximately threefold greater than expected from the frequencies of single substitutions (Fig. 3).

The small number of AGY ↔ TCY double substitutions in noncoding DNA (due to the generally low fraction of noncoding DNA in prokaryotic genomes) within the analyzed genome set calls for caution in interpretation. Therefore, we performed additional control measurements. Analysis of the AGR > TCR (R stands for a purine) substitutions detected no significant effect of the third position on the frequency of the double substitutions (*SI Appendix*, Fig. S3). Although the number of double substitutions was slightly greater than it was with pyrimidines in the third position [87 TCR > AGR substitutions and 63 AGR > TCR substitutions], these numbers still translate into frequencies that are 3–4 times greater than those estimated as the product of single-mutation frequencies (*SI Appendix*, Fig. S3). We also analyzed substitutions in the shuffled triplets GAY and CTY (*SI Appendix*, Fig. S4) and identified 62 CTY > GAY substitutions and 90 GAY > CTY substitutions, again corresponding to double-substitution frequencies that were 3–5 times greater than expected from the single substitution frequencies (*SI Appendix*, Fig. S4). Thus, the frequencies of double substitutions in different dinucleotides in noncoding DNA are closely similar (Fig. 3 and *SI Appendix*, Figs. S3 and S4) and reflect a three- to fivefold excess over the random expectation.

Double Substitutions in Serine Codons in Yeast and Mammals. We additionally analyzed a triplet of yeast species and two mammalian triplets. Altogether, there were 25 TCY > AGY substitutions and 16 AGY > TCY substitutions in yeast (Fig. 4). The frequencies of double substitutions in the first two positions were 0.0004 for both TCY > AGY and AGY > TCY replacements (Fig. 4). These frequencies are approximately 2–3 times greater than expected from the estimation of the double-substitution frequencies as the product of the respective single-substitution frequencies (Fig. 4). In the yeast noncoding DNA sequences, we identified 59 TCY > AGY mutations and 54 AGY > TCY mutations (Fig. 5). In this case, the frequencies of double substitutions (0.0009 for both TCY > AGY and AGY > TCY substitutions) were not significantly different [P(Fc_obs/exp > Fn_obs/exp) = 0.079] from those expected from the frequencies of single mutations (Fig. 5).

Despite the small numbers of double substitutions in the coding regions, a qualitatively similar result was obtained for primate and rodent species triplets, with the frequencies of double substitutions being about 2–8 times greater than expected from the respective single-mutation frequencies (*SI Appendix*, Fig. S5). The difference between primate coding and noncoding DNA was not statistically significant [P(Fc_obs/exp > Fn_obs/exp) = 0.827], but a marginally significant difference was found for rodent species [P(Fc_obs/exp > Fn_obs/exp) = 0.029].

Serine Codon Set Switches in Slow- and Fast-Evolving Genes. Under the hypothesis that the greater-than-expected rate of serine codon set switches results from selection pressure to maintain the serines, it can be predicted that this effect should be more pronounced in slow-evolving (subject to strong selection) than in fast-evolving (weakly selected) genes. We split genes in each ATGC (*SI Appendix*, Table S1) into two bins that included, respectively, “conserved,” slow-evolving genes with *dN*/*dS* (ratio of nonsynonymous-to-synonymous substitution rates) equal or smaller than the median and the remaining “nonconserved,” fast-evolving genes (Fig. 6). We found that the frequency of double (tandem) mutations was more than twofold higher in conserved genes than in nonconserved genes in full agreement with the selection hypothesis (Fig. 6).

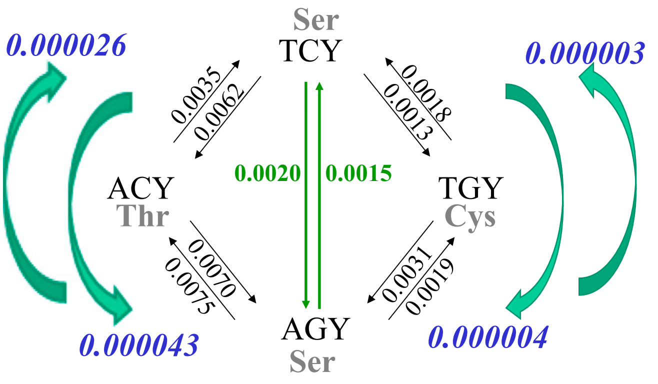

Fig. 2. Frequencies of double (tandem) nucleotide substitutions leading to serine codon set switches and the contributing single substitutions in protein-coding sequences in prokaryotic species triplets. The frequency of substitutions was estimated as the number of substitutions in a given codon divided by the number of the respective codons in the reconstructed ancestral sequence (Fig. 1). The expected number of two consecutive mutations A > B > C was calculated as a product of the frequencies of the mutations A > B and B > C (see *SI Appendix*, Fig. S1, for details). The frequencies of single substitutions are shown in black, the frequencies of double substitutions are shown in green, and the expected frequencies of two consecutive substitutions are shown in blue.

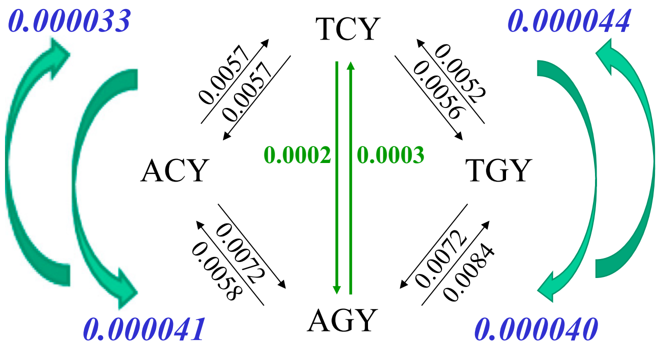

Fig. 3. Frequencies of double (tandem) and single substitutions in noncoding DNA from prokaryotic species triplets. The estimates and designations are as in Fig. 2.

Discussion

Analysis of the evolutionary switches between serines encoded by the codons of the two disconnected sets offers a rare opportunity to disambiguate the evolutionary effects of biased mutation and selection. The results of comparative analysis presented here indicate that switches from one serine codon set to another, which require two nucleotide substitutions in adjacent codon positions, occur in bacterial and archaeal protein-coding sequences up to 50 times more frequently than expected by chance and an order of magnitude more frequently than the corresponding tandem nucleotide substitutions in noncoding regions. These findings imply that, in microbial evolution, serine codon sets switch mostly by rapid succession of single-nucleotide substitutions driven by selection via a threonine and/or cysteine intermediates. In other words, once a serine is replaced by a threonine or cysteine, typically, there is a strong selective pressure for the reversal to serine. Such reversal involves either restoration of the original sequence, which is undetectable in genome comparisons, or a switch to a different codon set, which was detected in the analyses described here. The character of the selective processes that govern the serine codon set switch merits closer attention. The overall outcome of these processes can be naturally interpreted as purifying (stabilizing) selection whereby the selection pressure maintains the ancestral state (serine) because substitutions are often deleterious. Microscopically, however, the switches involve the initial, deleterious substitution resulting in the replacement of serine by threonine or cysteine that is rapidly followed by a second, beneficial substitution through which the site reverts to serine. This second step reflects positive, "compensatory" selection.

In the analyzed eukaryotic genomes (yeast and mammals), the excess of the frequencies of the observed double (tandem) mutations over expected frequencies and nucleotide substitutions in coding regions was much smaller than in prokaryotes whereas, in the noncoding regions, there was virtually no excess. These observations are compatible with the fundamental population-genetics theory (24–27) whereby eukaryotes have substantially smaller effective population sizes than prokaryotes, and the consequent decrease in the power of selection most likely accounts for weaker pressure for serine restoration. This hypothesis predicts that genes subject to stronger purifying selection should have a higher frequency of synonymous double (tandem) mutations in serine codons. Indeed, a comparison of the slow-evolving vs. fast-evolving prokaryotic genes revealed a twofold excess of serine codon set switches in the former compared with the latter (Fig. 6). These findings are fully consistent with the hypothesis that compensatory selection largely determines the frequency of serine codon set switches.

There are several potential methodological issues with the genome-wide analysis of nucleotide substitutions including effects of selection on noncoding DNA, violations of parsimony (more than one mutation per site per branch), and possible biases in the alignment procedures. We found that, in noncoding regions, the frequencies of tandem substitutions in dinucleotides of the same composition as the first two bases of serine codons are consistently several fold higher than predicted from the frequencies of the respective single mutations in prokaryotic genomes (this effect was much less pronounced in eukaryotes). This excess of double substitutions apparently reflects the aggregate effect of various potential biases along with mutational processes that result in simultaneous substitution of two bases (11). The existence of any factors that would specifically suppress tandem mutations of AG, GA, TC, and CT dinucleotides in noncoding regions and instead favor single mutations in these dinucleotides appears extremely unlikely (28). Thus, the much greater excess of tandem substitutions in the coding regions most likely reflects purifying selection that drives reversal to serine. The approach used here for the reconstruction of mutational events in triplets of species, based on the maximum parsimony principle, allows one to control for uncertainty in the branching order of large species trees and for saturation of substitutions at synonymous sites and noncoding DNA (29).

As discussed above, there are two paths for the serine codon set switch by two consecutive single-nucleotide substitutions, via cysteine and via threonine (Fig. 2). The serine-to-cysteine replacements are about threefold less frequent than the serine-to-threonine replacements (Fig. 2), arguably because the former are more likely to be deleterious than the latter. Thus, the path through threonine can be expected to be the principal route of serine codon set switching. The selective pressure for a cysteine to revert to serine is likely to be substantially stronger than that for threonine, potentially increasing the contribution of the cysteine route. Nevertheless, given that the frequency of double substitutions leading to switches is slightly greater than the frequency of serine-to-cysteine replacements (Fig. 2), the "smooth" serine–threonine–serine path appears to be dominant.

The results described here are not at odds with the previous findings and conclusions on serine codon set switches (2). Indeed, Averof and colleagues analyzed tandem nucleotide substitutions in codons for extremely conserved, most likely essential, serine residues (such as those in active sites of enzymes) for which the evolutionary trajectories through threonine or cysteine are likely to be inaccessible. The codon set switches in such positions indeed could be attributed to the moderate excess of double mutations that was observed both in the study of

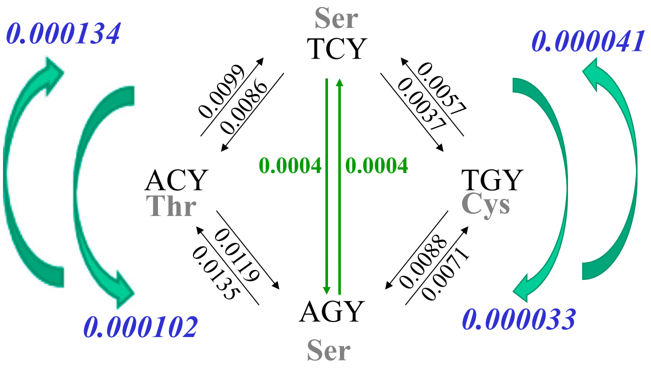

Fig. 4. Frequencies of double (tandem) nucleotide substitutions leading to serine codon set switches and contributing single substitutions in protein-coding sequences from three yeast species. The estimates and designations are as in Fig. 2.

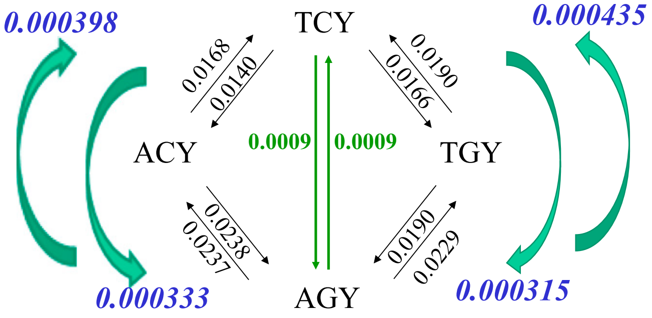

Fig. 5. Frequencies of double (tandem) AGY ↔ TCY substitutions and the contributing single substitutions in noncoding DNA from three yeast species. The estimates and designations are as in Fig. 2.

Averof et al. (2) and in this work. In contrast, here we analyzed all serines that are conserved in multiple triplets of closely related species for many (if not most) of which the threonine and cysteine-containing states are likely to be only mildly deleterious, and the evolutionary routes through these intermediates are accordingly passable. The results of this work indicate that, when all serines are analyzed, the vast majority of double (tandem) mutations leading to codon set switches is explained by the pressure of purifying selection.

In the phylogenetic context, the high frequency of serine codon set switches (Fig. 2) corresponds to an extremely high frequency of reversals that are an important source of homoplasy at the sequence level (30–34). Furthermore, it has been shown that extensive parallel changes in the evolution of *Drosophila* species are largely explainable by effects of purifying selection driven by similar functional constraints in closely related organisms (35) in agreement with the conclusions of the present study. From the data in Fig. 2 and *SI Appendix*, Fig. S1, it can be estimated that in prokaryotic genomes, a substantial fraction of serine-coding sites that mutate to either threonine or cysteine [0.0015/(0.0075 + 0.0019) = 16% for the AGY to TCY switch and 0.002/(0.0062 + 0.0013) = 27% for the TCY to AGY switch (*SI Appendix*, Fig. S1)] have reverted to serine. In the setting of this analysis, we detect about half of the serine-to-serine reversals, namely those that lead to codon set switch, whereas the other half, those that involve return to the original codon, remain invisible. Furthermore, there is no reason to believe that serines are in any way special with respect to the frequency of reversals. Reversals are expected to be as common for other amino acids, only not as easily detectable because, apart from serine, all amino acids are encoded by sets of codons that are connected through single substitutions. Homoplasy, arguably, is the principal impediment for all (broadly defined) cladistic approaches in phylogenetics although it creates numerous problems for phylogenetic methods in general (31, 34–39). The results of this work provide a strong indication that homoplasy is extremely common in the evolution of closely related species. A related but more general conclusion is that a large fraction, possibly the majority of amino acid residues in proteins, is subject to purifying selection that is strong enough to drive frequent reversal on the scale of divergence of closely related species.

The results of our genome-wide analysis of serine codon set switches indicate that, at least in prokaryotes, double-nucleotide substitutions required for such switches occur primarily via rapid succession of two single-nucleotide substitutions. The first of these substitutions results in the replacement of serine by threonine or cysteine and is often deleterious, whereas the second one restores serine and is apparently driven by positive, compensatory selection. We estimated that the fraction of serine-to-threonine and serine-to-cysteine replacements that revert to serine is quite substantial, over 15% under conservative estimates. These findings imply unexpectedly high levels of selection that result in extensive homoplasy in short-term evolution, potentially leading to artifacts in phylogenetic analysis.

Materials and Methods

Genomic data for bacteria and archaea were obtained from an updated version of the ATGC. The ATGCs consist of bacterial and archaeal genomes with a high level of conserved gene synteny and a synonymous substitution rate (*dS*) <1.5 (40, 41). To reconstruct mutations in protein-coding and noncoding DNA, we used triplets of closely related species (Fig. 1) (29). The triplets of species were extracted from the ATGC database as follows. We aimed for a pair of species with *dS* in the range of 0.25–1.0, preferentially choosing those as close to 0.25 as possible, to balance the requirements for a sufficient number of substitutions for reliable analysis and for the lack of mutational saturation. Then, the third species was chosen such that the distance from each member of the initially selected pair of species was at least 1.2 (preferably 1.5) times greater than the distance within the pair, so that it would represent an unambiguous outgroup. The *dS* values were maximum-likelihood estimates calculated using the CODML program of the PAML4.8 package (42). The *dN*/*dS* value (the measure of protein-level selection) was estimated using the Pamilo-Bianchi-Li method (43).

Alignments of protein-coding and noncoding DNA sequences from 37 triplets of closely related species were analyzed (*SI Appendix*, Table S1). Sequences were aligned using the MUSCLE program (44). The datasets are available from ftp.ncbi.nlm.nih.gov/pub/shabalin/SERINE/. Evolutionary distances for the intergenic regions were generally smaller compared with fourfold degenerate sites, suggesting that on average purifying selection is stronger in the noncoding regions than in the synonymous sites of the coding sequences (*SI Appendix*, Table S1). The smallest distances for fourfold degenerate sites and intergenic spacers between the two in-groups were 0.065 and 0.031, respectively, and the largest distances were 0.295 and 0.152, respectively (*SI Appendix*, Table S1). The smallest numbers of protein-coding sites and intergenic positions among the 37 triplets of species were 639,171 and 19,962, respectively, and the largest numbers were 4,542,510 and 279,940, respectively (*SI Appendix*, Table S1. Given that the triplets of species included closely related sequences, we used the maximum-parsimony approach to infer mutational events (45) (Fig. 1). Specifically, the frequencies of the codon substitutions that involve exchanges between serine and threonine as well as serine and cysteine and the serine codon set switches were calculated by dividing the number of the corresponding nucleotide

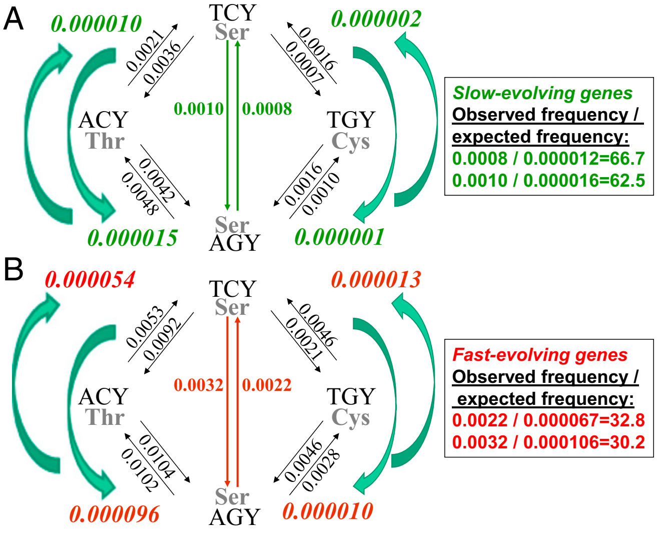

Fig. 6. Frequencies of double (tandem) AGY↔TCY substitutions and the contributing single substitutions in protein-coding sequences of slow-evolving (*A*) and fast-evolving (*B*) genes. The combined protein-coding genes from all ATGCs were split into a slow-evolving (*dN*/*dS* equal to or less than the ATGC-specific median, green) and fast-evolving (*dN*/*dS* greater than the median, red) gene bins. The other designations and estimates are as in Fig. 2.

substitutions in the ungapped alignment columns by the number of the respective codons in the ancestral sequence that was reconstructed using maximum parsimony (43). Alignments of three yeast species (*Saccharomyces cerevisiae*, *S. paradoxus*, and *S. mikatae*) were extracted from the Saccharomyces Genome Database (46); protein-coding regions and 3′ regions were used for analysis. Mammalian alignments (human–green monkey–bush baby and mouse–rat–rabbit triplets of species) of protein-coding genes and 3′ regions (5,000 bases downstream from stop codons) were downloaded from the University of California at Santa Cruz Table Browser (47).

The significance of the differences between the ratios of observed and expected nucleotide substitution frequencies in coding and noncoding DNA was estimated using random sampling of positions from multiple alignments of the respective sequences. For each triplet of species, 10,000 sampled alignments were produced. Alignment columns containing AGY or TCY triplets were sampled from coding or noncoding DNA sequences depending on the sizes of the datasets. If the number of AGY↔TCY changes was greater in the coding sequences, we sampled columns from coding alignments until the number of AGY↔TCY changes was the same in coding and noncoding sequences. Conversely, if the number of AGY↔TCY changes was greater in the noncoding sequences, the respective columns from the noncoding sequence alignments were sampled until the number of AGY↔TCY changes was the same in the coding and noncoding sequences. For each sampled alignment, the observed and expected frequencies (Fc and Fn for coding and noncoding sequences, respectively) of the AGY↔TCY substitutions were calculated, and the probability P(Fc_obs/exp > Fn_obs/exp) was calculated based on 10,000 sampled alignments.

ACKNOWLEDGMENTS. We thank Michael Galperin, Kira Makarova, David Landsman, Joshua Cherry, and Yuri Wolf for helpful discussions and Olga Batyreva for statistical advice. This research was supported in part by the Intramural Research Program of the National Library of Medicine at the National Institutes of Health.

1. Li WH (1997) *Molecular Evolution* (Sinauer, Sunderland, MA).
2. Averof M, Rokas A, Wolfe KH, Sharp PM (2000) Evidence for a high frequency of simultaneous double-nucleotide substitutions. *Science* 287(5456):1283–1286.
3. Nei M, Kumar S (2001) *Molecular Evolution and Phylogenetics* (Oxford Univ. Press, Oxford), p 333.
4. Blair C, Murphy RW (2011) Recent trends in molecular phylogenetic analysis: Where to next? *J Hered* 102(1):130–138.
5. Yang Z, Rannala B (2012) Molecular phylogenetics: Principles and practice. *Nat Rev Genet* 13(5):303–314.
6. Seidman MM, Bredberg A, Seetharam S, Kraemer KH (1987) Multiple point mutations in a shuttle vector propagated in human cells: Evidence for an error-prone DNA polymerase activity. *Proc Natl Acad Sci USA* 84(14):4944–4948.
7. Pavlov YI, Shcherbakova PV, Rogozin IB (2006) Roles of DNA polymerases in replication, repair, and recombination in eukaryotes. *Int Rev Cytol* 255:41–132.
8. Chen Z, Feng J, Buzin CH, Sommer SS (2008) Epidemiology of doublet/multiplet mutations in lung cancers: Evidence that a subset arises by chronocoordinate events. *PLoS One* 3(11):e3714.
9. Drake JW, Bebenek A, Kissling GE, Peddada S (2005) Clusters of mutations from transient hypermutability. *Proc Natl Acad Sci USA* 102(36):12849–12854.
10. Drake JW (2007) Too many mutants with multiple mutations. *Crit Rev Biochem Mol Biol* 42(4):247–258.
11. Chan K, Gordenin DA (2015) Clusters of multiple mutations: Incidence and molecular mechanisms. *Annu Rev Genet* 49:243–267.
12. Burch LH, et al. (2011) Damage-induced localized hypermutability. *Cell Cycle* 10(7): 1073–1085.
13. Chen JM, Férec C, Cooper DN (2015) Complex multiple-nucleotide substitution mutations causing human inherited disease reveal novel insights into the action of translesion synthesis DNA polymerases. *Hum Mutat* 36(11):1034–1038.
14. Matsuda T, et al. (2001) Error rate and specificity of human and murine DNA polymerase eta. *J Mol Biol* 312(2):335–346.
15. Harfe BD, Jinks-Robertson S (2000) DNA polymerase zeta introduces multiple mutations when bypassing spontaneous DNA damage in Saccharomyces cerevisiae. *Mol Cell* 6(6):1491–1499.
16. Stone JE, et al. (2009) Low-fidelity DNA synthesis by the L979F mutator derivative of Saccharomyces cerevisiae DNA polymerase zeta. *Nucleic Acids Res* 37(11):3774–3787.
17. Silva JC, Kondrashov AS (2002) Patterns in spontaneous mutation revealed by human-baboon sequence comparison. *Trends Genet* 18(11):544–547.
18. Buettner VL, Hill KA, Scaringe WA, Sommer SS (2000) Evidence that proximal multiple mutations in Big Blue transgenic mice are dependent events. *Mutat Res* 452(2): 219–229.
19. Grantham R (1974) Amino acid difference formula to help explain protein evolution. *Science* 185(4154):862–864.
20. Sharp PM, Li WH (1987) Ubiquitin genes as a paradigm of concerted evolution of tandem repeats. *J Mol Evol* 25(1):58–64.
21. Brenner S (1988) The molecular evolution of genes and proteins: A tale of two serines. *Nature* 334(6182):528–530.
22. Irwin DM (1988) Evolution of an active-site codon in serine proteases. *Nature* 336(6198):429–430.
23. Koonin EV, Gorbalenya AE (1989) Tale of two serines. *Nature* 338(6215):467–468.
24. Lynch M, Conery JS (2003) The origins of genome complexity. *Science* 302(5649): 1401–1404.
25. Lynch M (2007) The frailty of adaptive hypotheses for the origins of organismal complexity. *Proc Natl Acad Sci USA* 104(Suppl 1):8597–8604.
26. Charlesworth B (2009) Fundamental concepts in genetics: Effective population size and patterns of molecular evolution and variation. *Nat Rev Genet* 10(3):195–205.
27. Loewe L, Hill WG (2010) The population genetics of mutations: Good, bad and indifferent. *Philos Trans R Soc Lond B Biol Sci* 365(1544):1153–1167.
28. Rogozin IB, Pavlov YI (2003) Theoretical analysis of mutation hotspots and their DNA sequence context specificity. *Mutat Res* 544(1):65–85.
29. Jordan IK, et al. (2005) A universal trend of amino acid gain and loss in protein evolution. *Nature* 433(7026):633–638.
30. Wake DB, Wake MH, Specht CD (2011) Homoplasy: From detecting pattern to determining process and mechanism of evolution. *Science* 331(6020):1032–1035.
31. Rogozin IB, Thomson K, Csürös M, Carmel L, Koonin EV (2008) Homoplasy in genome-wide analysis of rare amino acid replacements: The molecular-evolutionary basis for Vavilov's law of homologous series. *Biol Direct* 3:7.
32. Pelosi L, et al. (2006) Parallel changes in global protein profiles during long-term experimental evolution in Escherichia coli. *Genetics* 173(4):1851–1869.
33. Bazykin GA, et al. (2007) Extensive parallelism in protein evolution. *Biol Direct* 2:20.
34. Rokas A, Carroll SB (2008) Frequent and widespread parallel evolution of protein sequences. *Mol Biol Evol* 25(9):1943–1953.
35. Zou Z, Zhang J (2015) Are convergent and parallel amino acid substitutions in protein evolution more prevalent than neutral expectations? *Mol Biol Evol* 32(8):2085–2096.
36. Reyes A, Pesole G, Saccone C (2000) Long-branch attraction phenomenon and the impact of among-site rate variation on rodent phylogeny. *Gene* 259(1-2):177–187.
37. Philippe H, Zhou Y, Brinkmann H, Rodrigue N, Delsuc F (2005) Heterotachy and long-branch attraction in phylogenetics. *BMC Evol Biol* 5:50.
38. Philippe H, Lartillot N, Brinkmann H (2005) Multigene analyses of bilaterian animals corroborate the monophyly of Ecdysozoa, Lophotrochozoa, and Protostomia. *Mol Biol Evol* 22(5):1246–1253.
39. Delsuc F, Brinkmann H, Philippe H (2005) Phylogenomics and the reconstruction of the tree of life. *Nat Rev Genet* 6(5):361–375.
40. Novichkov PS, Ratnere I, Wolf YI, Koonin EV, Dubchak I (2009) ATGC: A database of orthologous genes from closely related prokaryotic genomes and a research platform for microevolution of prokaryotes. *Nucleic Acids Res* 37(Database issue):D448–D454.
41. Puigbò P, Lobkovsky AE, Kristensen DM, Wolf YI, Koonin EV (2014) Genomes in turmoil: Quantification of genome dynamics in prokaryote supergenomes. *BMC Biol* 12:66.
42. Yang Z (2007) PAML 4: Phylogenetic analysis by maximum likelihood. *Mol Biol Evol* 24(8):1586–1591.
43. Pamilo P, Bianchi NO (1993) Evolution of the Zfx and Zfy genes: Rates and interdependence between the genes. *Mol Biol Evol* 10(2):271–281.
44. Edgar RC (2004) MUSCLE: Multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res* 32(5):1792–1797.
45. Rogozin IB, Babenko VN, Wolf YI, Koonin EV (2005) *Dollo Parsimony and Reconstruction of Genome Evolution. Parsimony, Phylogeny, and Genomics*, ed Albert VA (Oxford Univ. Press, Oxford), pp 190–200.
46. Kellis M, Patterson N, Endrizzi M, Birren B, Lander ES (2003) Sequencing and comparison of yeast species to identify genes and regulatory elements. *Nature* 423(6937): 241–254.
47. Karolchik D, et al. (2004) The UCSC Table Browser data retrieval tool. *Nucleic Acids Res* 32(Database issue):D493–D496.