MBE Advance Access published July 30, 2008

Synonymous and Nonsynonymous Polymorphism vs. Divergence in Bacterial Genomes

Austin L. Hughes*, Robert Friedman, Pierre Rivailler, and Jeffrey O. French

Department of Biological Sciences, University of South Carolina, Columbia SC 29208 USA

*Author for correspondence at Department of Biological Sciences, Coker Life Sciences Building, 715 Sumter St., University of South Carolina, Columbia SC 29208 USA. Email: austin@biol.sc.edu. Tel.: 1-803-777-9186. Fax: 1-803-777-4002.

Keywords: bacterial evolution, homologous recombination, McDonald-Kreitman test, neutrality index, nucleotide substitution

Abstract

Comparison of the ratio of nonsynonymous to synonymous polymorphisms within species with the ratio of nonsynonymous to synonymous substitutions between species has been widely used as a supposed indicator of positive Darwinian selection, with the ratio of these two ratios being designated as a neutrality index (NI). Comparison of genome-wide polymorphism within 12 species of bacteria with divergence from an outgroup species showed substantial differences in NI among taxa. A low level of nonsynonymous polymorphism at a locus was the best predictor of NI < 1, rather than a high level of nonsynonymous substitution between species. Moreover, genes with NI < 1 showed a strong tendency toward the occurrence of rare nonsynonymous polymorphisms, as expected under the action of ongoing purifying selection. Thus our results are more consistent with the hypothesis that a high relative rate of between-species nonsynonymous substitution reflects mainly the action of purifying selection within species to eliminate slightly deleterious mutations rather than positive selection between species. This conclusion is consistent with previous results highlighting an important role of slightly deleterious variants in bacterial evolution and suggests caution in the use of the McDonald-Kreitman test and related statistics as tests of positive selection.

Introduction

Numerous studies have compared the patterns of synonymous and nonsynonymous polymorphism within species with the patterns of synonymous and nonsynonymous divergence between species in order to obtain information regarding the action of natural selection on protein-coding genes; but the interpretation of such data is complicated (Hughes 2007). McDonald and Kreitman (1991) suggested that a higher ratio of nonsynonymous to synonymous divergence than of nonsynonymous to synonymous polymorphism might provide evidence of positive Darwinian selection that favored adaptive amino acid changes between species; and they proposed a statistical test (the McDonald-Kreitman or MK test) of this hypothesis based on a contingency table. Several other authors have used similar reasoning to define as a "neutrality index" (NI) the ratio of nonsynonymous to synonymous polymorphism divided by the ratio of nonsynonymous to synonymous divergence, which may or may not be used in conjunction with the MK test (Rand and Kann 1996; Bazin et al. 2006). Alternatively, the inverse of NI may be defined as a "fixation index" (Shapiro et al. 2007).

It has been argued that the MK test and related methods may be subject to distortions caused by the presence of slightly deleterious nonsynonymous mutations in natural populations. As has frequently been suggested, these methods may not be able distinguish clearly between fixation of nonsynonymous substitutions due to positive selection and the fixation of slightly deleterious nonsynonymous mutations during a bottleneck accompanying speciation (Eyre-Walker 2002; Hughes et al. 2006; Hughes 2007; Ohta 1993). Furthermore, slightly deleterious nonsynonymous polymorphisms can confound the MK test and related methods in other ways. An abundance of slightly deleterious polymorphisms, subject to ongoing purifying selection, within a species might mask the effect of between-species divergence (Hughes 2007). Conversely, effective removal of slightly deleterious variants by purifying selection might lead to a false inference of positive selection between species because within-species nonsynonymous polymorphism is reduced.

There are two aspects of purifying selection that leave signatures detectable by nucleotide sequence analysis: (1) certain deleterious mutations have previously been eliminated from the population, thereby contributing the fact that d_N is less than d_S in most pairwise comparisons of coding sequences; and (2) certain slightly deleterious mutations, still present in populations, are subject to ongoing purifying selection that acts to lower their frequencies in comparison to neutral variants (Hughes et al. 2003; Hughes and Hughes 2007a,b; Hughes and Piontkivska 2008; Irausquin and Hughes 2008). Since nonsynonymous mutations are more likely to be slightly deleterious than synonymous mutations, a signature of ongoing purifying selection is that gene diversity ("heterozygosity") is reduced at nonsynonymous polymorphic sites in comparison to that at synonymous polymorphic sites (Hughes et al. 2003). Likewise, ongoing purifying selection will influence statistics that compare the pairwise nucleotide difference with the number of segregating sites and thereby identify the presence of rare variants (Tajima 1989).

Several lines of evidence suggest that slightly deleterious variants, subject to ongoing purifying selection, are widespread in protein-coding genes of many bacterial species. A survey of 149 datasets of bacterial sequence polymorphism revealed an excess of rare nonsynonymous variants but not of synonymous variants; the former are likely to represent slightly deleterious alleles whose frequency in the population has decreased as a result of ongoing purifying selection (Hughes 2005). Similar results were reported in a study *of Escherichia coli* and *Salmonella enterica*, even when singletons (possibly due to sequencing errors) were removed from the data (Charlesworth and Eyre-Walker 2006). Moreover, Rocha et al. (2006) presented evidence that the ratio of the number of nonsynonymous substitutions per nonsynonymous site (d_N) to the number of synonymous substitutions per synonymous site (d_S) tends to decrease as the evolutionary time between two related bacterial genomes increases.

The efficiency with which slightly deleterious variants are removed is correlated with effective population size. Although there is evidence that effective population sizes of bacteria are in general much greater than those of multicellular eukaryotes (Lynch and Conery 2003), little is known about variations among bacterial species with respect to effective population size. Another factor that will affect these analyses is the rate of homologous recombination, since recombination is necessary to purge deleterious variants from a population ("Mueller's ratchet"; Lynch et al. 1993; Lynch 2007); and recombination rates are known to vary among bacterial species (Feil et al. 2001). If nonsynonymous polymorphism involves mainly slightly deleterious variants subject to ongoing purifying selection, species with large effective population sizes and/or efficient recombination may be expected to have reduced within-species nonsynonymous polymorphism relative to within-species synonymous polymorphism. As a result, the MK test and related methods may tend to identify positive selection in such species. Moreover, within a given species, the MK test and related methods may tend to infer positive selection on individual genes subject to strong ongoing purifying selection and thus having low ratios of nonsynonymous to synonymous polymorphism.

Here we address this question by examining within-species polymorphism in protein-coding genes in the completely sequenced genomes of 12 species of bacteria, including important pathogens of humans, domestic animals, and plants. Using the closest available outgroup species, we estimate between-species synonymous and nonsynonymous substitutions and compare the pattern of net between-species divergence with that of within-species polymorphism. In particular, we examine the neutrality index (NI), defined as the ratio of nonsynonymous to synonymous polymorphism divided by the ratio of nonsynonymous to synonymous divergence (Rand and Kann 1996). In addition, in the case of three species for which polymorphism data were available on a substantial number of genes in four or more genomes (*Burkholderia pseudomallei*, *Staphylococcus aureus*, and *Streptococcus pyogenes*), we test for the presence of rare nonsynonymous variants likely to be subject to ongoing purifying selection. By comparing the patterns of within-species purifying selection within species with those of between-species divergence, we test for the role of slightly deleterious nonsynonymous variants on MK and related methods.

Methods

Sequences Analyzed

We analyzed polymorphism within 12 species of Bacteria for which at least three complete genome sequences were available and for which the complete genome sequence of a congeneric outgroup species was available (for accession numbers see

Supplementary Table S1). The species (with numbers of sequences) and outgroups were as follows: *Burkholderia pseudomallei* (4) and *B. thailandensis*; *Campylobacter jejuni* (3) and *C. fetus*; *Chlamydophila pneumoniae* (4) and *C. felis*; *Ehrlichia ruminantium* (3) and *E. chaffeensis*; *Helicobacter pylori* (3) and *H. acinonychis*; *Pseudomonas syringae* (3) and *P. fluorescens*; *Shigella flexneri* (3) and *S. sonnei*; *Staphylococcus aureus* (9) and *S. epidermidis*; *Streptococcus pneumoniae* (3) and *S. sanguinis*; *Streptococcus pyogenes* (12) and *S. agalactiae*; *Yersinia pestis* (5) and *Y. pseudotuberculosis*. We designate the species for which multiple genomes were analyzed as the ingroup species. Gene families in a species and its outgroup were identified by applying the Blastclust software (Altschul et al. 1997) to predicted protein translations**.** Putative orthologs were identified as families with exactly one representative per genome. The parameters L (minimum length coverage) and S (similarity threshold measured as BLAST similarity score divided by the alignment length) were chosen for each pair of species so as to yield as many orthologs as possible, given the level of sequence divergence between ingroup and outgroup species (Supplementary Table S1). Orthologs were aligned at the amino acid level using Clustal W (Thompson et al. 1994), and the alignment was imposed on the DNA sequences.

The species concept in Bacteria has been controversial, because genetic exchange can occur beyond the boundaries of named species (Vulić et al. 1997; Majewski 2001; Cohan 2002; Fraser et al. 2007). Furthermore, named bacterial species do not always correspond to monophyletic groups. We used preliminary phylogenetic analyses to ensure that each of the twelve ingroup species used here constituted a monophyletic group, at least as regards the majority of its protein-coding genes. Likewise, we used phylogenetic analyses to test the hypothesis that each of the outgroup species constituted an outgroup to the ingroup species with which it was compared, at least as regards the majority of its protein-coding genes. These phylogenetic analyses were based on the aligned sets of orthologous amino acid sequences using the following methods: (1) maximum parsimony (MP) using branch-and-bound search (Swofford 2002); (2) neighbor-joining (NJ; Saitou and Nei 1987) based on the JTT model (Jones et al. 1992) with the gamma correction for rate variation among sites, using the MEGA 3 program (Kumar et al. 2004); and (3) the quartet maximum likelihood method (QML) using the TREE-PUZZLE 5.2 program (Schmidt et al. 2002) based on the JTT model with gamma correction for rate variation among sites. We only included in our analyses species for which these analyses supported monophyly and the validity of the outgroup.

Nucleotide Diversity and Divergence

The number of synonymous nucleotide substitutions per synonymous site (d_S) and the number of nonsynonymous nucleotide substitutions per nonsynonymous site (d_N) were estimated for all pairwise comparisons of orthologs by Yang and Nielsen's (2000) method. This method takes into account nucleotide content and mutational biases (Yang and Nielsen 2000). For each ortholog, we computed the mean of d_S values in all pairwise comparisons within the ingroup species (i.e., the synonymous nucleotide diversity, symbolized π_S). Likewise, for each ortholog, we computed the mean of d_N values in all pairwise comparisons within the ingroup species (i.e., the nonsynonymous nucleotide diversity, symbolized π_A). For each ortholog, we computed the net synonymous (k_S) and nonsynonymous (k_A) nucleotide divergences between the ingroup and outgroup species following Nei and Jin (1989; see also Nei and Kumar 2000). To compute k_S, we first computed the mean of d_S for all comparisons between ingroup and outgroup (d_{Sb}); then $k_S = d_{Sb} - \pi_S$. Similarly, to compute k_A, we first computed the mean of d_N for all comparisons between ingroup and outgroup (d_{Nb}); then $k_A = d_{Nb} - \pi_A$.

Different authors have used different methods based on the original idea of McDonald and Kreitman (1991) to compare synonymous and nonsynonymous polymorphism and divergence (e.g., Bazin et al. 2006; Bustamante et al. 2002; Rand and Kann 1996; Smith and Eyre-Walker 2002). Some of these methods have been developed because counting of synonymous and nonsynonymous differences between species may underestimate the amount of substitution if the evolutionary time has been long and because, when there are multiple substitutions per codon, the numbers of synonymous and nonsynonymous substitutions depend on the pathway taken by evolution (Whittam and Nei 1991). In preliminary analyses, we found that the results of all these methods are highly correlated. For example, we reconstructed the ancestral sequences of the set of orthologs for selected ingroup species by maximum parsimony and estimated d_S and d_N between the outgroup species and the reconstructed ancestor of the ingroup species. In addition, we used the maximum likelihood method to reconstruct internal branches within the trees. The resulting values of these preliminary analyses were highly correlated with k_S and k_A computed as described above. Therefore we used the latter as measures of between-species divergence because they involve fewer assumptions than methods dependent on ancestral reconstruction, while taking into account both multiple hits and the probability of different evolutionary pathways (Yang and Nielsen 2000).

Because of homologous recombination, certain loci may have evolutionary histories that contrast markedly with those of other genes in the genomes. We used a twopronged approach to exclude these genes from our data set. First, we used a k-means clustering algorithm to identify genes with unusual patterns of d_S in pairwise comparisons among ingroup members (Hughes and French 2007; Hughes and Friedman 2004, 2005; Hughes and Langley 2007). Of 16,776 orthologs identified by homology search, 73 (0.4%) showed highly unusual patterns and were excluded from further analyses. Second, we excluded all genes with $\pi_S > d_{Sb}$ (162 such genes or 1.0% of the total) and all genes with $\pi_A > d_{Nb}$ (232 such genes or 1.4% of the total). Note that the fact that such a small proportion of genes showed greater within-species polymorphism than between-species divergence provides further support for the conclusion that the outgroup species chosen constituted genuine outgroups to the ingroup species.

We computed the following quantity:

$$\mathrm{NI} = (\pi_A/\pi_S)/(k_A/k_S) \qquad (1)$$

NI (the "neutrality index"; Rand and Kann 1996) is greater than 1 when the ratio of nonsynonymous to synonymous polymorphism exceeds the ratio of nonsynonymous to synonymous divergence; this is the condition of the absence of positive selection increasing between-species divergence at nonsynonymous sites according to the assumptions of McDonald and Kreitman (1991). By contrast NI < 1 is taken to be indicative of such selection (McDonald and Kreitman 1991), although other interpretations are possible (Eyre-Walker 2002; Hughes et al. 2006; Hughes 2007; Ohta 1993). In the present analyses, we excluded genes for which NI was undefined; 5956 (35.5%) of orthologs identified by homology search were excluded for this reason. NI was defined for relatively few genes when the available genomes for a given ingroup species were very closely related because in those species $\pi_S = 0$ in the case of many genes. *Chlamydophila pneumoniae* and *Yersinia pestis* were the genomes that had the highest numbers of genes excluded for this reason.

Codon Usage

Within each genome, we computed for each predicted protein-coding gene five quantities summarizing codon usage in that gene. For a given gene, let $nTC2f$ = the number of two-fold degenerate sites using T or C; $nAG2f$ = the number of two-fold degenerate sites using A or G; and $n4f$ = the number of four-fold degenerate sites. Let $nC2f$ = the number of occurrences of C at two-fold degenerate sites; $nG2f$ = the number of occurrences of G at two-fold degenerate sites; and $nC4f$, $nG4f$, and $nA4f$ designate, respectively, the numbers of occurrences of C, G, and A at four-fold degenerate sites. Then define $T2 = nT2f/nTC2f$; $A2 = nA2f/nAG2f$; $T4 = nT4f/n4f$; $C4 = nC4f/n4f$; and $A4 = nA4f/n4f$. Taken together, these five codon usage variables ($pC2f$, $pG2f$, $pC4f$, $pG4f$, and $pA4f$) provide a measure of nucleotide usage, including both nucleotide content and content skewness, at almost all synonymous sites in the gene (excluding only the three-fold degenerate sites in isoleucine codons and the very rare C/A synonymous mutations in the first positions of certain arginine codons). The advantage of this approach is that it describes codon usage in only five linearly independent variables that are not dependent on amino acid composition (Hughes and Langley 2007). Moreover, these five variables are not subject to the stochastic errors due to small sample size seen in variables measured on each amino acid or each codon.

In order to compare these five variables across species, we computed standard normal deviates for each of these quantities within each species. We computed the standard normal deviate for a given value by subtracting from that value the mean for the species and dividing that difference by the standard error for the species. In order to identify genes with unusual codon usage for the species, we first computed the absolute value of the standard normal deviates. Then we used principal components, extracted from the correlation matrix, to reduce dimensionality. The first principal component (PC1) accounted for 35.1% of the variance in the correlation matrix, and PC1 was strongly positively correlated with the absolute values of all five standard normal deviates. We therefore used PC1 as an overall measure of atypical codon usage patterns. In preliminary analyses, separate analyses of all five standard normal deviates yielded similar results to those based on PC1 (data not shown); for simplicity, we report below only the results based on PC1.

Within-Species Polymorphism

. For selected genomes, we compared the pattern of synonymous and nonsynonymous polymorphism within individual genes. There were three species for which we had sequences from four or more genomes and on a substantial number of genes including both synonymous and nonsynonymous polymorphic sites: *Burkholderia pseudomallei* (4 genomes, 1710 genes); *Staphylococcus aureus* (9 genomes; 1317 genes) and *Streptococcus pyogenes* (11 genomes, 875 genes). For these genes, we computed the single-locus gene diversity (“heterozygosity”) independently for each polymorphic site by the following formula:

$$1-\sum_{i=1}^{n} x_i^{\,2} \qquad (2)$$

where n is the number of alleles and x_i is the population frequency of the i^{th} allele (Nei 1987, p. 177). Single nucleotide polymorphisms were classified either as synonymous or nonsynonymous depending on their effect of the encoded nucleotide sequence; we excluded ambiguous sites at which both synonymous and nonsynonymous variants occurred or at which the polymorphism could be considered synonymous or nonsynonymous depending on the pathway taken by evolution. There were 4316 such ambiguous polymorphic sites out of 177,038 total polymorphic sites (2.4%).

In order to examine the relative frequency of rare alleles at synonymous and nonsynonymous sites, we compared the average number of nucleotide differences and the number of segregating sites (Tajima 1989) separately for synonymous and nonsynonymous sites (Rand and Kann 1996; Hughes 2005; Hughes and Hughes 2007a,b). For each gene, we computed the difference *Ks* – *S*s*. *Ks* is the mean number of synonymous nucleotide differences for all pairwise comparisons among the *n* allelic sequences in the data set. If *Ss* is the number of synonymous segregating sites, then

$$S^*s = Ss/a_1 \qquad (3)$$

The divisor in equation (3) is a factor providing an adjustment for sample size (n) and is given by the following (Tajima 1989):

$$a_1 = \sum_{i=1}^{n-1} 1/i \qquad (4)$$

Similarly, for each gene, we computed the difference $Kn - S^*n$. Kn is the mean number of nonsynonymous nucleotide differences for all pairwise comparisons among the n allelic sequences in the data set; and, if Sn is the number of segregating nonsynonymous sites

$$S^*n = Sn/a_1 \qquad (5)$$

The differences $Ks - S^*s$ and $Kn - S^*n$ constitute the numerator of Tajima's (1989) D statistic computed separately for synonymous and nonsynonymous polymorphisms, respectively. We then computed the ratio of this difference to the absolute value of the minimum possible value of the difference, which would occur if all polymorphisms were singletons (Schaeffer 2002). We designate this ratio Q_{syn} in the case of synonymous polymorphisms and Q_{non} in the case of nonsynonymous polymorphisms. Comparing Q_{syn} and Q_{non} provides an index of the relative abundance of rare alleles at synonymous and nonsynonymous sites, with a strongly negative value indicating an abundance of rare alleles (Hughes and Hughes 2007a,b). Note that, unlike Tajima's D, these ratios are independent of sample size and thus can be compared between datasets of different size.

Tajima's D was designed to provide a test of neutrality, but that test depends on the assumption of mutation-drift equilibrium (Tajima 1989; Nei and Kumar 2000). For example, a recent population bottleneck can yield a negative value of D in the absence of purifying selection. However, in the present analyses, we did not test for neutrality but merely used Q_{syn} and Q_{non} as indices of the relative abundance of rare variants at synonymous and nonsynonymous sites, respectively. Note also that, since a bottleneck should affect both synonymous and nonsynonymous polymorphisms equally (Tajima

1989), a marked and consistent difference between Q_{syn} and Q_{non} is likely to be due to selection on nonsynonymous sites.

In order to examine the relationships among NI, the action of purifying selection, and the results of MK tests (McDonald and Kreitman 1991), we constructed for each gene in our data sets for *Burkholderia pseudomallei*, *Staphylococcus aureus*, and *Streptococcus pyogenes* a 2 X 2 contingency table including the following quantities: *Ss*; *Sn*; the estimated number of synonymous differences between the ingroup and outgroup species; and the estimated number of nonsynonymous differences between the ingroup and outgroup species. The latter two quantities were estimated by Yang and Nielsen's (2000) method and rounded to the nearest integer. Note that this approximation is probably reasonably valid in the case of these three species because in each of these species the outgroup was relatively close to the ingroup (see Results). For each contingency table, we conducted G-tests of independence using Williams' correction (Sokal and Rohlf 1981).

Other Statistical Methods

In order to compare π_S, π_A, k_S, and k_A across species, we computed standard normal deviates for each of these quantities within each species. We computed the standard normal deviate for a given value by subtracting from that value the mean for the species and dividing that difference by the standard error for the species; we designate the standard normal deviates as $\pi_S dev$, $\pi_A dev$, $k_S dev$, and $k_A dev$, respectively.

Since the variables analyzed were not normally distributed, we used for hypothesis testing nonparametric methods of statistical analysis, which make no assumptions regarding the form of the underlying distribution (Hollander and Wolfe 1973). However, in preliminary analyses, methods assuming a normal distribution generally yielded essentially identical results.

Results

Comparisons among Taxa

Table 1 summarizes means of synonymous (π_S) and nonsynonymous (π_A) nucleotide diversity for sets of orthologous genes from 12 bacterial species, along with mean net synonymous (k_S) and nonsynonymous (k_A) divergence from an outgroup species. The synonymous nucleotide diversity within species ranged from less than 1% to over 50% (in *Pseudomonas syringae*; Table 1). Likewise, available outgroups covered a wide range of divergences. For example, *Campylobacter fetus* was very distant from the three *C. jejuni* genomes analyzed, with mean k_S estimated at nearly three substitutions per site (Table 1). By contrast, *Yersinia pseudotuberculosis* was very close to the five *Y. pestis* genomes analyzed (mean $k_S = 0.0186$; Table 1).

As predicted by Rocha et al. (2006), mean π_A / π_S for the 12 taxa was negatively correlated with mean π_S ($r_S = -0.699$; $P = 0.011$; Figure 1A). *Pseudomonas syringae* provided an influential point in this relationship, with the lowest mean π_A / π_S (0.0232) and by far the highest mean π_S (0.5481; Figure 1A). Nonetheless, even when *Pseudomonas* was excluded from the analysis, the negative relationship between mean π_A / π_S and mean π_S remained ($r_S = -0.609$; $P = 0.047$). Besides *Pseudomonas*, the next lowest value of π_A / π_S was that seen in *Campylobacter* (0.0800; Figure 1A).

There was a highly significant difference among taxa with respect to the portion of genes with NI ≥ 1 ($\chi^2 = 1158.5$; 11 d.f.; $P < 0.001$; Figure 1B). The proportion of orthologs with NI ≥ 1 exceeded the proportion with NI < 1 in four of the 12 taxa, while in other eight taxa the proportion of orthologs with NI < 1 exceeded the proportion with NI ≥ 1 (Figure 1B). There were two taxa in which the proportion of genes with NI ≥ 1 was strikingly low. In *Campylobacter*, there were only 150 of 934 orthologs (16.1%) with NI ≥ 1; and in *Pseudomonas*, there were only 94 of 1005 orthologs (9.4%) with NI ≥ 1 (Figure 1B). In the remaining 10 taxa, the lowest proportion of orthologs with NI ≥ 1 was 38.5% (363 of 942) in *Neisseria* (Figure 1A). Combining data for all taxa, there were 4551 (44.0%) of 10353 orthologs with NI ≥ 1 (Figure 1A). But when *Campylobacter* and *Pseudomonas* were excluded, 4307 of the remaining 8414 genes (51.2%) had NI ≥ 1 (Figure 1A).

If values of NI < 1 are due to positive selection leading to amino acid sequence divergence between species, it would be expected that species with NI < 1 would be species with unusually high k_A values. As an initial test of this prediction, we computed for each species the number of genes falling in the top quartile (25%) of values of π_A / π_S across all species; then we compared this value with the number of genes expected to be in the top quartile if the distribution of π_A / π_S were uniform across species (i.e., 25% of the number of genes for the species; Figure 2A). Likewise, we computed for each species the number of genes falling in the top quartile (25%) of values of k_A / k_S across all species; then we compared this value with the number of genes expected to be in the top quartile if the distribution of k_A / k_S were uniform across species (i.e., 25% of the number of genes for the species; Figure 2B).

Pseudomonas and *Campylobacter* had by far the lowest ratios of observed to expected gene numbers in the top quartile of π_A / π_S values (Figure 2A). In the case of *Pseudomonas*, only 4 of 1005 genes (0.4%) were in the top quartile of π_A / π_S values; and in *Campylobacter*, only 80 of 934 (8.6%) were in the top quartile of π_A / π_S values (Figure 2A). *Burkholderia* (52 of 1711 or 3.1%) had the lowest proportion of genes in the top quartile of k_A / k_S values (Figure 2B). *Pseudomonas* also had a very low proportion of genes (42 of 1005 or 4.2%) in the top quartile of k_A / k_S values (Figure 2B). By contrast, the two highest proportions of genes in the top quartile of k_A / k_S values were seen in *Yersinia* (31 of 65 or 47.7%) and *Campylobacter* (405 of 934 or 43.4%; Figure 2B).

Thus, contrary to the usual interpretation of the NI statistic, *Pseudomonas,* the taxon with the highest proportions of genes with NI < 1 (Figure 1B), showed exceptionally few genes with high k_A / k_S values (Figure 2B). On the other hand, *Pseudomonas* also showed exceptionally few genes with high π_A / π_S values (Figure 2A). Since NI is a ratio, these results suggest that low NI values in the case of *Pseudomonas* were due to low numerators rather than to high denominators.

Campylobacter, the taxon with the second highest proportions of genes with NI < 1 (Figure 1B), did show a high proportion of genes with high k_A / k_S values (Figure 2B). But, like *Pseudomonas, Campylobacter* also showed exceptionally few genes with high of π_A / π_S values (Figure 2A). Thus, the factors contributing to values of NI < 1 were not consistent between the two taxa with the highest proportions of such genes.

Codon Usage

Synonymous codon usage is one factor that might be expected to influence the pattern of nucleotide substitution (Sharp 1991); but the method we used to estimate synonymous and nonsynonymous substitutions is designed to take into account differing patterns of synonymous codon usage (Yang and Nielsen 2000). Moreover, there was no obvious relationship between codon usage and the frequency of genes with NI < 1. *Campylobacter jejuni* has an AT-rich genome, while *Pseudomonas syringae* is GC-rich (Supplementary Table S2). However, the species with lower frequencies of genes with NI < 1 included both very AT-rich species (for example, *Ehrlichia ruminantium*) and GC-rich species (for example, *Burkholderia pseudomallei*; Supplementary Table S2).

Rank Partial Correlations

In order to examine the factors influencing the value of NI for individual genes, we used standard normal deviates as a measure of the deviation of π_A, π_S, k_A, and k_S for a given gene from the mean values for the species. We designated the standard normal deviates as $\pi_A dev$, $\pi_S dev$, $k_A dev$, and $k_S dev$, respectively. We used partial rank correlation to analyze the relationship between these variables and NI. We also tested for an independent effect of atypical patterns of codon usage, using the first principal component (PC1) extracted from the correlation matrix of absolute values of the standard normal deviates of five variables describing nucleotide content at synonymous sites (see Methods). We computed partial rank correlation coefficients between NI and each of five predictor variables ($\pi_A dev$, $\pi_S dev$, $k_A dev$, $k_S dev$, and PC1) simultaneously controlling in each case for the other four predictor variables (Table 2). Because certain ingroup species showed much higher k_S values in comparison with the outgroup than did others, we analyzed the taxa with close outgroups (*Burkholderia*, *Helicobacter*, *Neisseria*, *Shigella*, and *Yersinia*; Table 1) separately from the other taxa (Table 2). However, the results were similar when all species were considered together (data not shown).

Given that NI is a ratio of two ratios, it was unsurprising that there were significant positive partial rank correlation coefficients both between $\pi_A dev$ and NI and between $k_S dev$ and NI, a pattern seen both in genes from species with close outgroups and in genes from species with distant outgroups (Table 2). Likewise, it was unsurprising that $\pi_S dev$ and $k_A dev$ showed significant negative partial rank correlations with NI both for genes from species with close outgroups and for genes from species with distant outgroups (Table 2). However, in the case of both close and distant outgroups, the strongest partial rank correlation between NI and any of the predictor variables was that with $\pi_A dev$ (Table 2). In fact, whether positive or negative, all the other partial rank correlations differed significantly in absolute value from that between NI and any of the predictor variables was that with $\pi_A dev$ (Table 2). Thus $\pi_A dev$, rather than $k_A dev$, was the strongest predictor of NI, implying that low NI was primarily a consequence of unusually low within-species nonsynonymous nucleotide diversity rather than unusually high nonsynonymous between-species divergence. This result is contrary to the expectation that a low value of NI is evidence of positive selection favoring between-species divergence at the amino acid level.

In the case of species with close outgroups, there was not a significant partial rank correlation between NI and PC1 (Table 2). By contrast, when the outgroup was distant, there was a highly significant negative partial rank correlation between NI and PC1. Thus, in these species NI tended to be decreased when codon usage was unusual, independently of the estimates of nucleotide sequence polymorphism and divergence. It seems plausible that this result may have occurred because of the difficulty of estimating k_S in the case of distant outgroups, where synonymous sites approached saturation (Table 1), which may be especially problematic when codon usage is highly unusual.

We tested this interpretation by computing rank partial correlations between PC1 and $k_S dev$, simultaneously controlling for NI, $\pi_A dev$, $\pi_S dev$, and $k_A dev$. In the case of species with close outgroups, the partial rank correlation between PC1 and $k_S dev$ was –0.034 ($P = 0.025$), whereas in the case of species with distant outgroups the partial rank correlation between PC1 and $k_S dev$ was –0.143 ($P < 0.001$). The difference between these two correlation coefficients was highly significant ($P < 0.001$). Thus, unusual nucleotide content tended to cause a reduction in the estimate of k_S, and this effect was much more pronounced when the outgroup was distant.

Within-Species Polymorphism

For genes of *Burkholderia pseudomallei*, *Staphylococcus aureus*, and *Streptococcus pyogenes*, we computed mean gene diversity separately for synonymous and nonsynonymous polymorphic sites (Table 3). In each species, the median of the gene diversity values at synonymous sites was significantly greater than that at nonsynonymous sites (Sign test; $P < 0.001$ in each case; Table 3). The fact that gene diversities at nonsynonymous polymorphic sites tended to be lower than those at synonymous polymorphic sites in the same genes was evidence that nonsynonymous variants in these species include many that are subject to ongoing purifying selection. Additional evidence in support of this interpretation was provided by comparisons of Q_{syn}

and Q_{non} (Table 3). In all three species, median Q_{non} was significantly lower than median Q_{syn} (Sign test; $P < 0.001$ in each case; Table 3). In all three species, median Q_{non} was strongly negative (Table 3). Median Q_{syn} was negative in *Burkholderia* and in *Staphylococcus*, although less negative than Q_{non}, whereas in *S. pyogenes* median Q_{syn} was actually slightly positive (Table 3). Thus, in all three species, nonsynonymous polymorphic sites showed a greater bias toward rare polymorphisms than did synonymous sites, implying the presence of stronger ongoing purifying selection on the former than on the latter.

In order to examine the relationship between the value of NI and purifying selection on nonsynonymous polymorphic variants, we compared Q_{syn} and Q_{non} between genes with NI < 1 and genes with NI ≥ 1 separately for the three species. In each species, median Q_{non} values differed significantly between these two categories of genes (Figure 3). For all three species, median Q_{non} was lower in genes with NI < 1 than in genes with NI ≥ 1. Median Q_{syn} also differed significantly between the two categories of genes in all three species (Figure 3). However, in the case of Q_{syn}, median values for genes with NI < 1 were consistently higher than those for genes with NI ≥ 1 (Figure 3). Thus, genes with NI < 1 were characterized by an excess of rare nonsynonymous variants but not of synonymous variants.

MK Tests

In order to examine the relationship between within-species polymorphism and the results of the MK test, we compared numbers of polymorphic synonymous and nonsynonymous sites within *Burkholderia pseudomallei*, *Staphylococcus aureus*, and

Streptococcus pyogenes with the estimated numbers of synonymous and nonsynonymous substitutions between each of these species and the appropriate outgroup species. According to the usual interpretation of the MK test, genes with NI $\geq$ 1 and a significant G-test represent genes with a significant excess of nonsynonymous polymorphisms, whereas genes with NI < 1 and a significant G-test represent cases of positive selection between species. In the present data, the latter were much less frequent than the former. Of 1580 genes with NI < 1, the G-test was significant at the 5% level in only 50 cases (3.2%). By contrast, of 2322 genes with NI $\geq$ 1, the G-test was significant at the 5% level in 499 cases (21.5%).

Among genes with NI $\geq$ 1, Q_{non} was greater than zero in 103 of 499 genes with significant G-tests (20.6%). By contrast, of 1823 genes with NI $\geq$ 1 and non-significant G-tests, Q_{non} was greater than zero in only 276 (15.1%) genes. The difference between proportions was highly significant ($\chi^2 = 8.7$; 1 d.f.; $P < 0.001$). This difference is expected since a significant G-test with NI $\geq$ 1 implies an excess of within-species nonsynonymous polymorphisms. Among genes with NI < 1, Q_{non} was greater than zero in 15 of 50 (30.0%) genes with significant G-tests (30%), but in only 146 of 1530 (9.5%) of genes with non-significant G-tests. Again the difference in proportions was highly significant ($\chi^2 = 22.1$; 1 d.f.; $P < 0.001$). The latter result is surprising because a significant G-test for genes with NI < 1 is typically interpreted as indicating positive selection between species. Our results suggest on the contrary that the test may be influenced by the pattern of within-species polymorphism. In particular, because a high value of Q_{non} implies few nonsynonymous variants of low frequency, our results suggest that the absence of such variants is likely to lead to a significant result in the MK test.

We used discriminant analysis to examine further the relationship between Q_{non} and the outcome of the G-test. For genes with NI < 1, a linear discriminant function using Q_{non} as the sole predictor correctly predicted the outcome of the G-test (significant or non-significant) in 80.6% of cases. By contrast, in genes with NI ≥ 1, a linear discriminant function with Q_{non} as the sole predictor successfully predicted the outcome of the G-test in only 36.2% of cases. This analysis further supported the conclusion that the pattern of within-species nonsynonymous polymorphism can strongly influence the MK G-test when NI < 1.

Discussion

Comparison of the ratio of nonsynonymous to synonymous nucleotide diversity within species (π_A / π_S) with the ratio of nonsynonymous to synonymous substitutions between species (k_A / k_S) has been widely used as an indicator of positive Darwinian selection favoring amino acid changes between species. Comparison of genome-wide polymorphism within 12 species of bacteria with divergence from an outgroup species yielded results that cannot easily be explained on this supposition. Two species, *Campylobacter jejuni* and *Pseudomonas syringae*, showed extremely high frequencies (83.9% and 90.6%, respectively) of genes for which NI (the ratio of π_A / π_S to k_A / k_S) was less than 1 (Figure 1B). By contrast, in the 10 other species analyzed, NI was less than 1 in only about 48% of genes. According to the usual interpretation of NI, these results might be taken to indicate an extraordinarily high frequency of between-species positive selection in *Campylobacter* and *Pseudomonas*, but there is no obvious reason to expect that these two taxa are subject to unusually strong positive selection.

As an alternative to the hypothesis of positive selection, it might be proposed that the differences among taxa are artifacts of the estimation of synonymous and nonsynonymous substitutions. For example, it might be suggested that factors such as an unusual pattern of codon usage or a high rate of synonymous substitution have biased the estimates of synonymous and nonsynonymous substitution in certain taxa. However, the method used here to estimate numbers of synonymous and nonsynonymous substitutions per site is designed to take into account factors such as nucleotide usage and mutational bias and appears to be robust even when the rate of synonymous substitution is high (Yang and Nielsen 2000). In the present data, the level of synonymous substitution between ingroup and outgroup species was high in *Campylobacter* and *Pseudomonas*, but not unusually so in comparison to the other species analyzed (Table 1). Moreover, *Campylobacter* and *Pseudomonas* did not share any atypical patterns of codon usage that set them apart from the other species analyzed.

On the other hand, our results showed that *Campylobacter* and *Pseudomonas* shared a high proportion of genes with an unusually low π_A / π_S (Figure 2A). Moreover, *Pseudomonas* had by far the lowest overall mean π_A / π_S of the 12 taxa analyzed (Figure 1A). Thus, contrary to the usual interpretation of NI < 1 as evidence of positive selection favoring amino acid sequence divergence between species, our results suggested that in our data this statistic mainly reflected a low π_A / π_S within species rather than a high k_A / k_S.

This interpretation was strongly supported by partial correlation analyses showing that an unusually low value of π_A was a significantly better predictor of a low NI than was an unusually high k_A (Table 2). The same pattern was seen both in species with close outgroups and in species with distant outgroups. Thus it seemed unlikely that problems in estimating k_S when synonymous sites were near saturation had a major impact on NI. The fact that, in the case of distantly related outgroups, there was a significant negative correlation between NI and a measure of unusual codon usage suggested that there may have been some problems with estimating k_S between distantly related species, particularly in the case of genes with unusual patterns of codon usage. But this effect was quite minor in comparison with the strong positive relationship between π_A and NI.

Factors that might cause relatively low π_A in a given gene include the effects of purifying selection acting to eliminate or reduce in frequency slightly deleterious nonsynonymous variants. In order to examine purifying selection further, we analyzed synonymous and nonsynonymous polymorphisms in three species (*Burkholderia pseudomallei*; *Staphylococcus aureus*; and *Streptococcus pyogenes*) for which we had sequences of at least three genomes with numerous polymorphic sites. All of these species had a majority of genes with NI ≥ 1, but a substantial number with NI < 1 (Figure 1B). All three of these genomes showed an excess of rare nonsynonymous polymorphisms, indicative of ongoing purifying selection acting to eliminate slightly deleterious variants (Hughes et al. 2003; Hughes and Hughes 2007a,b; Hughes and Piontkivska 2008; Irausquin and Hughes 2008). Moreover, the tendency toward rare nonsynonymous variants was strongest in genes with NI < 1.

Across the 12 taxa in our data set, there was a negative correlation between mean π_A / π_S and mean π_S (Figure 1A). Taking π_S as a proxy for average time since the common ancestor of a set of sequences, these results are consistent with those of Rocha et al. (2006) and support those authors' hypothesis that this effect is due to a time lag in the removal of slightly deleterious nonsynonymous variants by purifying selection. In our data set, *Pseudomonas* showed by far the highest mean π_S, the lowest mean π_A / π_S, and the highest proportion of genes with NI < 1.

Our results thus imply that the major factor in causing NI < 1 was not positive selection between species but effective purifying selection within species, lowering π_A / π_S and thus lowering NI. Moreover, certain results suggested that the MK test (G-test) is sensitive to the effectiveness of purifying selection acting to eliminate slightly deleterious nonsynonymous variants within populations. In the usual interpretation of the MK test, genes with NI < 1 and a significant G- test are considered to be subject to positive selection causing the fixation of nonsynonymous differences between species. However we found an association in genes with NI < 1 between a significant G-test and a positive Q_{non}. Since a strongly negative Q_{non} characterizes a gene containing numerous nonsynonymous variants of low frequency, this result implies that the G-test is more likely to be significant when such low-frequency nonsynonymous variants are lacking. The absence of rare nonsynonymous polymorphisms evidently causes the level of nonsynonymous between-species divergence to be relatively large. Since rare nonsynonymous variants are often slightly deleterious (Hughes et al. 2003), their absence reflects the action of purifying selection. An ironic aspect of the MK test and related statistics thus appears to be that they mistakenly identify as positively selected between species the very genes that are most stringently negatively selected within species.

These considerations can explain, without resorting to the hypothesis of positive selection, the finding of lower NI in the mitochondrial genomes of invertebrates than of vertebrates (Bazin et al. 2006). On the assumption that invertebrates tend to have larger effective population sizes than vertebrates, this result is easily explained if low NI is generally due to purifying selection within a species reducing π_A / π_S, since a species with a larger effective population size will in general be more efficient in eliminating slightly deleterious variants. The same reasoning may also explain the detection of more “adaptive evolution” by the MK test in regions of the *Drosophila* genome with normal recombination than in those with low recombination (Shapiro et al. 2007), since purifying selection is expected to more effective in removing deleterious variants in the former regions than in the latter.

There are additional factors affecting NI that could not be addressed easily by our analyses. One such factor might be the fixation, during a population bottleneck, of slightly deleterious nonsynonymous variants, thereby increasing k_A (Eyre-Walker 2002; Hughes et al. 2006; Hughes 2007; Ohta 1993). On the other hand, it is possible that given the presumably very large effective population sizes of the geographically widespread bacterial species analyzed here that no extreme bottleneck occurred in speciation. And of course, it is possible that certain nonsynonymous substitutions between species in our data set were indeed fixed by positive selection (Charlesworth and Eyre-Walker 2006). Indeed, if fixation of slightly deleterious mutations is a widespread phenomenon, one might expect it to be accompanied by a certain degree of positive selection favoring “compensatory mutations” that ameliorate the effect of deleterious alleles (Charlesworth and Eyre-Walker 2007; Sawyer et al. 2007).

Nonetheless, our results suggest that fixation of nonsynonymous substitutions between species – whether by drift or by positive selection – has a minor impact on the MK test and related methods. Rather, the strongest effect is that of ongoing purifying selection on within-species nonsynonymous polymorphism. Thus our results suggest caution regarding the use of the MK test and related methods in testing for positive selection. Moreover, our results provide support for the nearly neutral theory of Ohta (1973) and its prediction that slightly deleterious variants play a significant role in evolution.

Acknowledgments

This research was supported by grant GM43940 from the National Institutes of Health to A.L.H.

Literature Cited

Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 25: 3389–3402.

Bazin E, Glémin S, Galtier N. 2006. Population size does not influence mitochondrial genetic diversity in animals. Science 312: 570-572.

Bustamante CD, Nielsen R, Sawyer SA, Olsen KM, Purugganan MD, Hartl DL. 2002. The cost of inbreeding in *Arabidopsis*. Nature 416: 531-534.

Charlesworth J, Eyre-Walker A. 2006. The rate of adaptive evolution in enteric bacteria. Mol. Biol. Evol. 23: 1348-1356.

Charlesworth J, Eyre-Walker A. 2007. The other side of the nearly neutral theory, evidence of slightly advantageous back-mutations. Proc. Natl. Acad. Sci. USA 104: 16992-16997.

Cohan FM. 2002. What are bacterial species? Annu. Rev. Microbiol. 56: 457-487.

Dunn OJ. 1964. Multiple comparisons using rank sums. Technometrics 6: 241-252.

Eyre-Walker A. 2002. Changing effective population size and the McDonald-Kreitman test. Genetics 162: 2017-2024.

Feil EJ, Holmes EC, Bessen DE, Chan M-S, Day NP, Enright MC, Goldstein R, Hood DW, Kalia A, Moore CE, Zhou J, Spratt BG. 2001. Recombination within natural populations of pathogenic bacteria: short-term empirical estimates and long-term phylogenetic consequences. Proc. Natl. Acad. Sci. USA 98: 182-187.

Fraser C, Hanage WP, Spratt BG. 2007. Recombination and the nature of bacterial speciation. Science 315: 476-480.

Hollander M, Wolfe DA. 1973. *Nonparametric statistical methods*. New York: Wiley.

Hughes AL. 2005. Evidence for abundant slightly deleterious polymorphisms in bacterial populations. Genetics 169: 533-538.

Hughes AL. 2007. Looking for Darwin in all the wrong places: the misguided quest for positive selection at the nucleotide sequence level. Heredity 99: 364-373.

Hughes AL, French JO. 2007. Homologous recombination and the pattern of nucleotide substitution in *Ehrlichia ruminantium*. Gene 387: 31-37.

Hughes AL, Friedman R. 2004. Patterns of sequence divergence in 5’ intergenic spacers and linked coding regions in 10 species of pathogenic Bacteria reveal distinct recombinational histories. Genetics 168: 1795-1803.

Hughes AL, Friedman R. 2005. Nucleotide substitution and recombination at orthologous loci in *Staphylococcus aureus*. J. Bacteriol. 187: 2698-2704.

Hughes AL, Friedman R, Glenn NL. 2006. The future of data analysis in evolutionary genomics. Curr. Genomics 7: 227-234.

Hughes AL, Hughes MA. 2007a. More effective purifying selection on RNA viruses than in DNA viruses. Gene 404: 117-125.

Hughes AL, Hughes MA. 2007b. Coding sequence polymorphism in avian mitochondrial genomes reflects population histories. Mol. Ecol. 16: 1369-1376.

Hughes AL, Langley KL. 2007. Nucleotide usage, synonymous substitution pattern, and past recombination in genomes of *Streptococcus pyogenes*. Infect. Genet. Evol. 7: 188-196.

Hughes AL, Piontkivska, H. 2008. Nucleotide sequence polymorphism in circoviruses. Infect. Genet. Evol. 8:130-138.

Hughes AL, Packer B, Welsch R, Bergen AW, Chanock SJ, Yeager M. 2003. Widespread purifying selection at polymorphic sites in human protein-coding loci. Proc. Natl. Acad. Sci. USA 100: 15754-15757.

Irausquin SJ, Hughes AL. 2008. Distinctive pattern of sequence polymorphism in the NS3 protein of hepatitis C virus type 1b reflects conflicting evolutionary pressures. J. Gen. Virol. (in press).

Jones DT, Taylor WR, Thornton JM. 1992. The rapid generation of mutation data matrices from protein sequences. Comput. Appl. Biosci. 8: 275-282.

Kumar S, Tamura K, Nei M.. 2004. MEGA3: integrated software for Molecular Evolutionary Genetics Analysis and sequence alignment. Brief. Bioinform. 5: 150-163.

Lynch M. 2007. *The origins of genome architecture*. Sunderland, MA: Sinauer.

Lynch M, Conery S. 2003. The origins of genome complexity. Science 302: 1401-1404.

Lynch M, Bürger R, Butcher D, Gabriel W. 1993. The mutational meltdown in asexual populations. J. Hered. 84: 339-344.

Majewski J. 2001. Sexual isolation in bacteria. FEMS Microbiol. Lett. 199: 161-169.

McDonald JH, Kreitman M. 1991. Adaptive protein evolution at the *Adh* locus in *Drosophila*. Nature 351: 114-116.

Nei M. 1987. *Molecular evolutionary genetics*. New York: Columbia University Press.

Nei M, Jin L. 1989. Variances of the average numbers of nucleotide substitutions within and between populations. Mol. Biol. Evol. 6: 290-300.

Nei M, Kumar S. 2000. *Molecular evolution and phylogenetics*. New York: Oxford University Press.

Ohta T. 1973. Slightly deleterious mutant substitutions in evolution. Nature 246: 96-98.

Ohta T. 1993. Amino acid substitution at the Adh locus of Drosophila is facilitated by small population size. Proc. Natl. Acad. Sci. USA 90: 4548-4551.

Rand DM, Kann LM. 1996. Excess amino acid polymorphism in mitochondrial DNA: contrasts among genes from *Drosophila*, mice, and humans. Mol. Biol. Evol. 13: 735-748.

Rocha EP, Maynard Smith J, Hurst LD, Holden MT, Cooper JE, Smith NH, Feil EJ. 2006. Comparisons of *dN*/*dS* are time dependent for closely related bacterial genomes. J. Theoret. Biol. 239: 226-235.

Saitou N, Nei M. 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol. Biol. Evol. 4: 406-425.

Sawyer SA, Parsch J, Zhang Z, Hart; DL. 2007. Prevalence of positive selection among nearly neutral amino acid replacements in *Drosophila*. Proc. Natl. Acad. Sci. USA 104: 6504-6510.

Schaeffer SW. 2002. Molecular population genetics of sequence length diversity in the Adh region of *Drosophila pseudoobscura.* Genet. Res. 80: 163-175.

Schmidt HA, Strimmer K, Vingron M, von Haeseler A. 2002. TREE-PUZZLE: maximum likelihood phylogenetic analysis using quartets and parallel computing. Bioinformatics 18: 502-504.

Shapiro J, Huang W, Zhang C, Hubisz MJ, Lu J, Turissini DA, Fang S, Wang H-Y, Hudson RR, Nielsen R, Chen Z, Wu C-I. 2007. Adaptive genic evolution in *Drosophila* genomes. Proc. Natl. Acad. Sci. USA 104: 2271-2276.

Sharp PM. 1991. Determinants of DNA sequence divergence between *Escherichia coli* and *Salmonella typhimurium*: codon usage, map position, and concerted evolution*.* J. Mol. Evol. 33: 23-33.

Smith NG, Eyre-Walker A. 2002. Adaptive evolution in *Drosophila*. Nature 415: 1022-1024.

Sokal RR, Rohlf FJ. 1981. *Biometry*. 2nd. Ed. New York: W.H. Freeman.

Swofford DL. 2002. PAUP*: phylogenetic analysis using parsimony (*and other methods). Sunderland, MA: Sinauer.

Tajima F. 1989. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics 123: 585-595.

Thompson JD, Higgins DG, Gibson TJ. 1994. CLUSTAL W: improvement of the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res. 22: 4673-4680.

Vulić M, Dionisio F, Taddei F, Radman M. 1997. Molecular keys to speciation: DNA polymorphism and the control of genetic exchange in enterobacteria. Proc. Natl. Acad. Sci. USA 94: 9763-9767.

Whittam TS, Nei M. 1991. Neutral mutation hypothesis test. Nature 354: 115-116.

Yang Z, Nielsen R. 2000. Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models. Mol. Biol. Evol.17: 32–43.

Table 1. Synonymous (π_S) and nonsynonymous (π_A) nucleotide diversity within 12 bacterial species; and mean net numbers of synonymous substitutions per synonymous site (k_S) and of nonsynonymous substitutions per nonsynonymous site (k_A) between each species and an outgroup species.

Species	Outgroup	No. genes	π_S ± S.E.	π_A ± S.E.	k_S ± S.E.	k_A ± S.E.
Burkholderia pseudomallei	*B. thailandensis*	1711	0.01591 ± 0.00023	0.00104 ± 0.00003	0.3436 ± 0.0030	0.0180 ± 0.0003
Campylobacter jejuni	*C. fetus*	934	0.06937 ± 0.00177	0.00477 ± 0.00022	2.8491 ± 0.0295	0.3818 ± 0.0053
Chlamydophila pneumoniae	*C. felis*	158	0.00232 ± 0.00011	0.00047 ± 0.00006	1.9790 ± 0.0527	0.2557 ± 0.0109
Ehrlichia ruminantium	*E. chaffeensis*	633	0.02913 ± 0.00078	0.00252 ± 0.00016	1.9871 ± 0.0314	0.1364 ± 0.0026
Helicobacter pylori	*H. acinonychis*	1130	0.11903 ± 0.00099	0.02118 ± 0.00049	0.1609 ± 0.0023	0.0336 ± 0.0009
Neisseria meningitidis	*N. gonorrhoeae*	942	0.07336 ± 0.00193	0.00806 ± 0.00026	0.0843 ± 0.0022	0.0117 ± 0.0003
Pseudomonas syringae	*P. fluorescens*	1005	0.54810 ± 0.00883	0.01169 ± 0.00030	0.9906 ± 0.0160	0.0477 ± 0.0007
Shigella flexneri	*S. sonnei*	562	0.00295 ± 0.00011	0.00081 ± 0.00006	0.0380 ± 0.0014	0.0040 ± 0.0002
Staphylococcus aureus	*S. epidermidis*	1463	0.03415 ± 0.00094	0.00273 ± 0.00010	2.6932 ± 0.0267	0.1646 ± 0.0026
Streptococcus pneumoniae	*S. sanguinis*	819	0.02362 ± 0.00149	0.00307 ± 0.00019	1.6791 ± 0.0242	0.1793 ± 0.0040
Streptococcus pyogenes	*S. agalactiae*	931	0.02924 ± 0.00114	0.00407 ± 0.00019	2.1018 ± 0.0335	0.2014 ± 0.0040
Yersinia pestis	*Y. pseudotuberculosis*	65	0.00119 ± 0.00011	0.00051 ± 0.00014	0.0186 ± 0.0027	0.0033 ± 0.0006

Table 2. Partial rank correlation coefficients between five predictor variables and NI, computed separately for genes from species with close outgroups and species with distant outgroups[1].

Predictor variable	Close outgroup (N = 4410)	Distant outgroup (N =5943)
$\pi_A dev$	0.747 [2]	0.684 [2]
$\pi_S dev$	-0.340 [2,3]	-0.523 [2,3]
$k_A dev$	-0.412 [2,3]	-0.422 [2,3]
$k_S dev$	0.305 [2,3]	0.400 [2,3]
PC1	-0.020 [3]	-0.102 [2,3]

[1] Each value is a fourth-order partial rank correlation between the predictor variable and NI, simultaneously controlling for the other four predictor variables.

[2] Tests of the hypothesis that a given partial rank correlation coefficient equals zero: $P < 0.001$.

[3] Tests of the hypothesis that a given partial rank correlation coefficient equals that between $\pi_A dev$ and NI: $P < 0.001$.

Table 3. Median gene diversity at polymorphic nonsynonymous and synonymous sites and median Q_{non} and Q_{syn} in three bacterial species.

Species (no. genes)	Gene Diversity		Q_{non}	Q_{syn}
	Nonsynonymous	Synonymous		
Burkholderia pseudomallei (1710)	0.375	0.377[1]	-0.501	-0.368[2]
Staphylococcus aureus (1317)	0.272	0.294[1]	-0.438	-0.145[2]
Streptococcus pyogenes (875)	0.254	0.299[1]	-0.289	0.039[2]

[1] Tests of the hypothesis that median gene diversity at nonsynonymous sites equals median gene diversity at synonymous sites: P < 0.001 (Sign test).

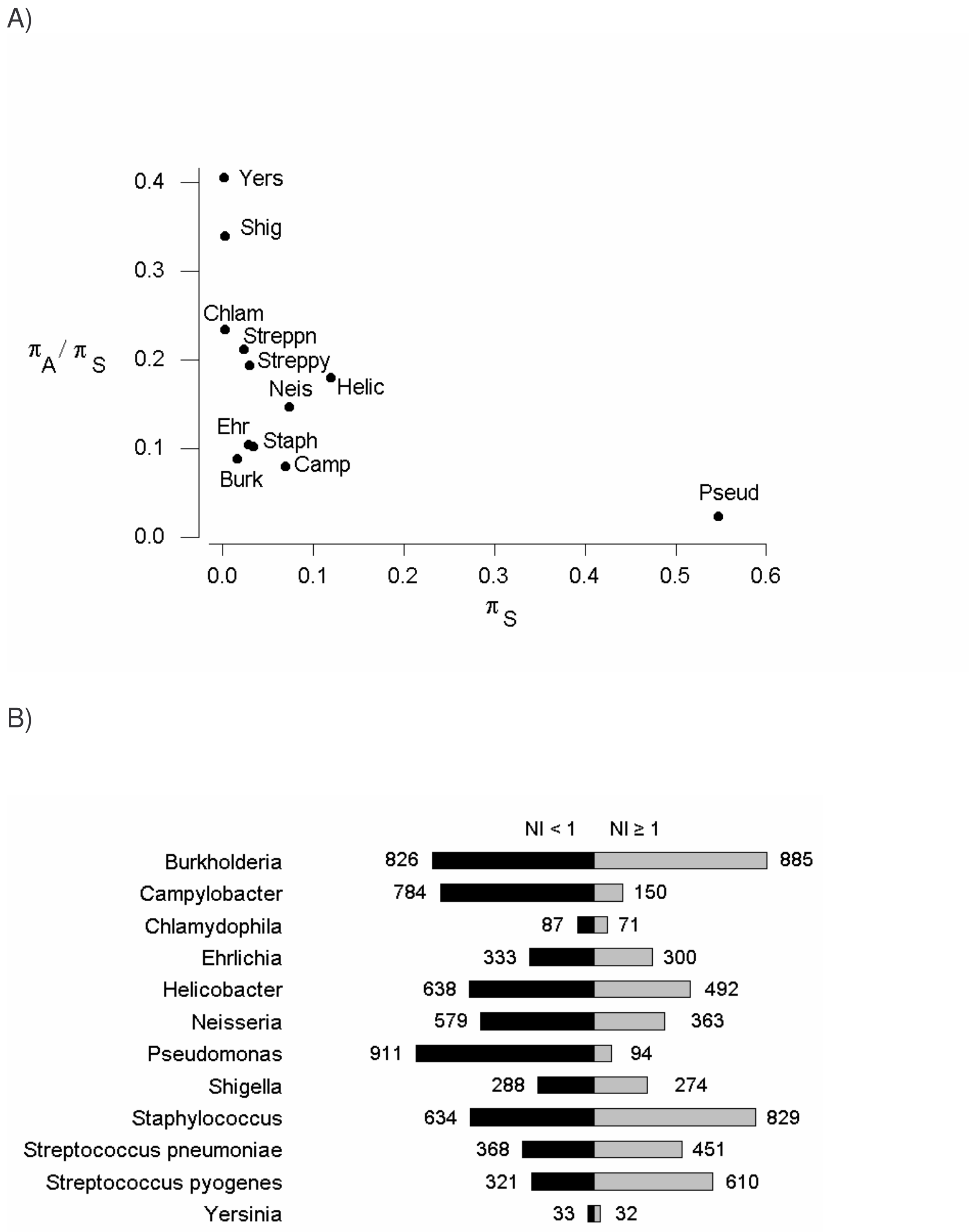
[2] Tests of the hypothesis that median Q_{non} equals median Q_{syn} : P < 0.001 (Sign test).

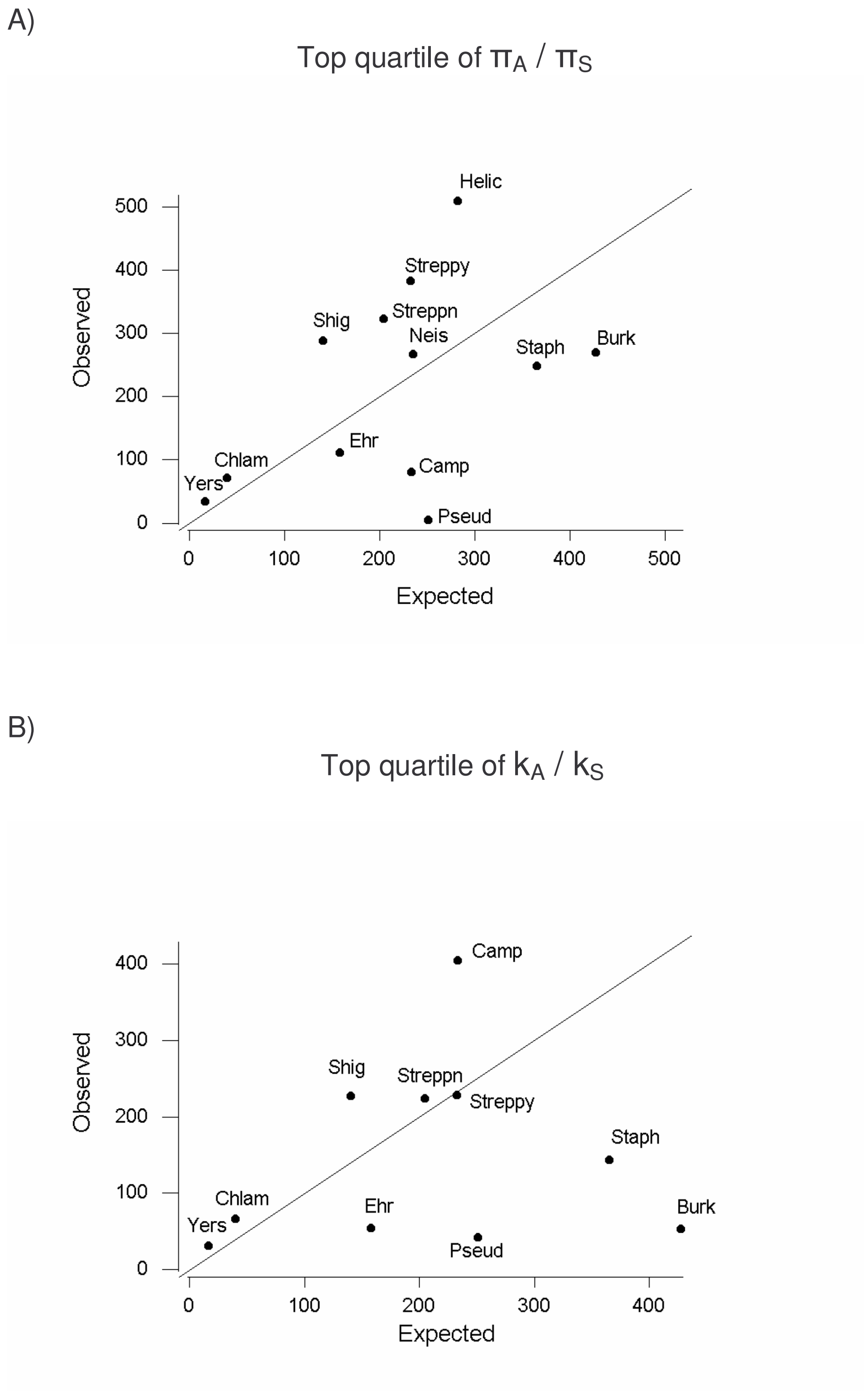
Figure Captions

Figure 1. (A) Plot of mean π_A / π_S vs. mean π_S for 12 bacterial taxa (r_S = -0.699; P = 0.011). Abbreviations for taxa: Burk (*Burkholderia*); Camp (*Campylobacter*): Chlam (*Chlamydophila*); Ehr (*Ehrlichia*); Helic (*Helicobacter*); Neis (*Neisseria*); Pseud (*Pseudomonas*); Shig (*Shigella*); Staph (*Staphylococcus*); Streppn (*Streptococcus pneumoniae*); Streppy (*Streptococcus pyogenes*); Yers (*Yersinia pestis*). (B) Numbers of genes with NI < 1 and NI ≥ 1. There was a significant difference among taxa with respect to the portion of genes with NI ≥ 1 (χ^2 = 1158.5; 11 d.f.; P < 0.001).

Figure 2. Observed vs. expected numbers of genes in the top quartile of values of (A) π_A / π_S and (B) k_A / k_S. Abbreviations for taxa are as in Figure 1A.

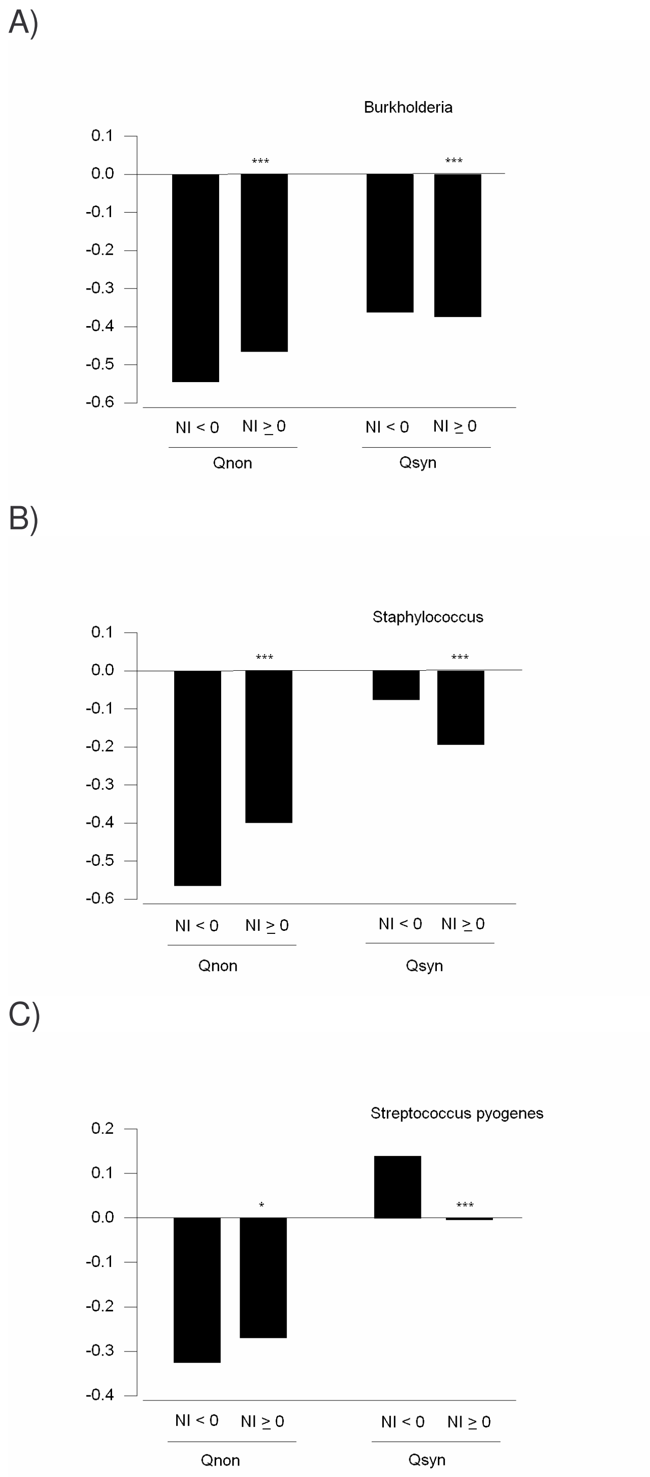
Figure 3. Median Q_{syn} and Q_{non} in genes with NI < 1 and NI ≥ 1 in (A) *Burkholderia*; (B) *Staphylococcus*; and (C) *Streptococcus pyogenes*. Mann-Whitney tests of the hypothesis that the median value for genes with NI < 1 equals that for genes with NI ≥ 1: * P < 0.05; *** P < 0.001.


A)
0.4
0.3
0.2
0.1
0.0
π_A / π_S
Yers
Shig
Chlam
Streppn
Streppy
Helic
Neis
Ehr
Staph
Camp
Burk
Pseud
0.0
0.1
0.2
0.3
0.4
0.5
0.6
π_S
B)
NI < 1
NI ≥ 1
Burkholderia 826 885
Campylobacter 784 150
Chlamydophila 87 71
Ehrlichia 333 300
Helicobacter 638 492
Neisseria 579 363
Pseudomonas 911 94
Shigella 288 274
Staphylococcus 634 829
Streptococcus pneumoniae 368 451
Streptococcus pyogenes 321 610
Yersinia 33 32



A) Top quartile of π_A / π_S

B) Top quartile of k_A / k_S


A)
Burkholderia
0.1
0.0
-0.1
-0.2
-0.3
-0.4
-0.5
-0.6
***
***
NI < 0
NI ≥ 0
NI < 0
NI ≥ 0
Qnon
Qsyn
B)
Staphylococcus
0.1
0.0
-0.1
-0.2
-0.3
-0.4
-0.5
-0.6
***
***
NI < 0
NI ≥ 0
NI < 0
NI ≥ 0
Qnon
Qsyn
C)
Streptococcus pyogenes
0.2
0.1
0.0
-0.1
-0.2
-0.3
-0.4
*
***
NI < 0
NI ≥ 0
NI < 0
NI ≥ 0
Qnon
Qsyn