Evidence for Abundant Slightly Deleterious Polymorphisms in Bacterial Populations

Austin L. Hughes¹

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

USA

¹ Corresponding author: Department of Biological Sciences, University of South Carolina, Coker Life Sciences Bldg., 700 Sumter St., Columbia SC 29208. Email: austin@biol.sc.edu.
Running Head: Polymorphism in Bacteria

Key Words: Bacteria, natural selection, nearly neutral theory, polymorphism, Tajima’s D

Address correspondence to:

Austin L. Hughes, Ph.D.
Department of Biological Sciences
University of South Carolina
Coker Life Sciences Bldg.
700 Sumter St.
Columbia SC 29208 USA
Tel: 1-803-777-9186
Fax: 1-803-777-4002
Email: austin@biol.sc.edu
Abstract

The nearly-neutral theory of molecular evolution predicts that slightly deleterious mutations subject to purifying selection are widespread in natural populations, particularly those of large effective population size. In order to test this hypothesis, the standardized difference between pairwise nucleotide difference and number of segregation sites (corrected for number of sequences) was estimated for 149 population data sets from 84 species of Bacteria. This quantity (Tajima’s D statistic) was estimated separately for synonymous (D_{syn}) and nonsynonymous (D_{non}) polymorphisms. D_{syn} was positive in 70% of data sets, and the overall median D_{syn} (0.873) was positive. By contrast D_{non} was negative in 68% of data sets, and the overall median D_{non} (-0.656) was negative. The preponderance of negative values of D_{non} is evidence that there are widespread rare nonsynonymous polymorphisms in the process of being eliminated by purifying selection, as predicted to occur in populations with large effective size by the nearly-neutral theory. The major exceptions to this trend were seen among surface proteins, particularly those of bacteria parasitic on vertebrates, which included a number of cases of polymorphisms apparently maintained by balancing selection.
The concept of purifying or negative natural selection – i.e., natural selection acting to
decrease the frequency of deleterious alleles – is one of the key ideas of modern
evolutionary biology (Kimura and Ohta 1974). As first pointed out by Kimura (1977),
strong evidence for the widespread occurrence of purifying selection is provided by the
fact that, in most comparisons between homologous genes, the number of synonymous
nucleotide substitutions per synonymous site (dS) generally exceeds the number of
nonsynonymous nucleotide substitutions per nonsynonymous site (dN) (Li et al. 1985).
However, this evidence provides no information regarding the overall strength of
purifying selection, which is an important factor in the extent to which nonsynonymous
polymorphisms are observed in populations.

If most nonsynonymous mutations are strongly deleterious, they will be
eliminated from populations very quickly. On the other hand, the fate of slightly
deleterious nonsynonymous mutations depends on effective population size (Ohta 1976).
Purifying selection against slightly deleterious mutations is not efficient in populations
with small effective size but is much more efficient as effective population size increases.
Ohta’s “nearly neutral” theory of molecular evolution emphasizes the importance of such
slightly deleterious mutations in the evolutionary process (Ohta 1973, 1976, 2002). Some
recent evidence in support of this theory is based on the gene diversity at single
nucleotide polymorphism (SNP) loci in the human population. Nonsynonymous SNPs –
particularly those having radical effects on protein structure – tend to have lower average
gene diversities than synonymous and non-coding SNPs in the same genes (Freudenberg-
Hua et al. 2003; Hughes et al. 2003; Sunyaev et al. 2003; Zhao et al. 2003). This pattern
suggests that slightly deleterious alleles, which had drifted to relatively high allelic
frequencies when the effective size of the human population was low (Harpending et al. 1998), have decreased in frequency as a result of purifying selection after population expansion (Hughes et al. 2003).

Tajima (1989) pointed out that important inferences regarding population processes can be obtained from a sample of allelic DNA sequences by comparing the average number of pairwise nucleotide differences ($\hat{k}$) with the number of segregating sites, corrected for the number of sequences compared. Specifically, if $S$ is the number of segregating (or polymorphic) sites, we define $S^* = S/a_1$, where

$$a_1 = \frac{n - 1}{\sum_{i = 1}^{n} \frac{1}{i}} \quad (1)$$

Both $\hat{k}$ and $S^*$ are estimators of the population parameter $M = 4Nu$, where $N$ is the effective population size and $u$ is the mutation rate per generation per sequence under investigation. However, when there are rare variants present in a population, $S^*$ will tend to be larger than $\hat{k}$. The difference between these two quantities, divided by its standard error, is known as Tajima’s D statistic; $D$ represents an index of the action of natural selection on a sample of allelic sequences (Tajima 1989). When $D$ is strongly positive, there are few rare variants, a situation found under balancing selection. When $D$ is strongly negative, rare variants are abundant, a situation indicating purifying selection (Tajima 1989).

Several authors have extended Tajima’s (1989) logic to examine synonymous and nonsynonymous polymorphisms separately (e.g., Rand and Kann 1996; Wise et al. 1998; Navarro-Sabaté et al. 2003). Here I applied this method to 149 nucleotide sequence data
sets from 84 species of Bacteria. In order to avoid ambiguity, the analysis was restricted
to codons at which only a single polymorphic site was observed. Computing the D
statistic separately for synonymous (D_{syn}) and nonsynonymous (D_{non}) polymorphisms
makes it possible to test the hypothesis that rare variants are particularly common at
nonsynonymous sites, indicating purifying selection against slightly deleterious alleles.

On the nearly-neutral theory, purifying selection against slightly deleterious
alleles is expected in Bacteria, because most bacterial species are believed to have very
large effective population sizes (Lynch and Conery 2003). Feil et al. (2003) recently
reported a high level of nonsynonymous nucleotide polymorphisms in *Staphylococcus
aureus* and suggested that many of these may be deleterious mutations in the process of
elimination by purifying selection. On the other hand, the bacterial species analyzed
include several that are parasitic on vertebrates; certain proteins, particularly surface
proteins exposed to the host immune system in these species are known to be subject to
positive selection (Reid et al. 1999). Thus, these data make it possible to compare the
effects on sequence polymorphism of both positive and purifying selection.

**Methods**

Datasets of protein-coding genes from Bacteria (eubacteria) were chosen from the
NCBI Popset database. No sets were included that represented laboratory isolates of a
single strain, but only sets that had been collected as patient or environmental isolates or
reference strains (e.g. ATCC). I included only sets datasets for which functional
information about the encoded protein was available; datasets containing at least four
allelic sequences; and data sets including both synonymous and nonsynonymous
polymorphisms at codns with only a single polymorphic site. Using these criteria, it was possible to find usable data sets for 7 out of 21 bacterial phyla. The mean number of sequences per data set was 13.11 (± 0.86 S.E.); the median was 11.00; the range 4-64.

The sequences were aligned at the amino acid level using the CLUSTALW program (Thompson et al. 1994), and the resulting alignment was imposed on the DNA. Portions of any alignment that appeared unreliable on account of large numbers of gaps or poor sequence conservation and codons including undetermined nucleotides were removed. The numbers of synonymous of substitutions per synonymous site (dS) and of nonsynonymous substitutions per nonsynonymous site (dN) were estimated by Nei and Gojobori’s (1986) method. In preliminary analyses, more complex methods of estimating dS and dN (Li 1993; Yang et al. 2000; Zhang et al. 1998) yielded essentially identical results, as is expected in a case like the present one where numbers of substitutions per site are low (Nei and Kumar 2000). dS and dN were estimated for each pairwise comparison between homologous sequences in two ways: (1) for the entire sequence; (2) for the sequence excluding codons with two or more polymorphic sites. From these values, \( \pi_S \) (the mean of all pairwise dS values) and \( \pi_N \) (the mean of all pairwise dN values) were estimated for each dataset.

After excluding codons with two or more polymorphic sites, the following quantities were computed: the average number of pairwise synonymous differences (\( ^\wedge k_S \)) and the average number of nonsynonymous pairwise differences (\( ^\wedge k_N \)); the number of synonymous segregating sites (\( S_S \)); and the number of nonsynonymous segregating sites (\( S_N \)). Then, I computed \( S^* = S_S/a_1 \) and \( S^*_N = S_N/a_1 \), where \( a_1 \) is as in equation (1) above. \( D_{syn} \) was defined as \( ^\wedge k_S - S^*_S \), divided by the standard error of that difference (computed...
as described by Tajima, 1989). Likewise, $D_{\text{non}}$ was defined as $k_N - S^*_N$, divided by the standard error of that difference (computed as described by Tajima, 1989).

Bacterial species were characterized as primarily parasitic on vertebrates if a vertebrate host constitutes a major portion of the species niche. Thus species which can be vertebrate pathogens but whose primary niche is not parasitism on vertebrates (e.g., *Bacillus anthracis* and *Vibrio cholerae*) were not included in the group of vertebrate parasites. Because surface proteins may be particularly important in interactions with the immune system of vertebrates, all proteins were categorized on the basis of surface or non-surface expression.

Because the distributions of $\pi_S$, $\pi_N$, $D_{\text{syn}}$, and $D_{\text{non}}$ across the 149 datasets deviated significantly from normality (Kolmogorov-Smirnov test), nonparametric methods were used for testing hypothesis regarding differences among datasets.

**Results**

The occurrence of polymorphic sites per codon was compared with that expected, assuming a Poisson distribution with $\mu$ equal to the observed frequency of polymorphic sites per codon (Figure 1). The hypothesis that the median difference between observed and expected proportions of codons having a given number of polymorphic sites equaled zero was tested by Wilcoxon signed-rank tests (Figure 1). There were significantly fewer than expected codons with zero or two substitutions, but significantly more codons than expected with one substitution (Figure 1). Thus, the distribution of polymorphic sites in these data showed overall a more dispersed pattern than a Poisson distribution.
For each dataset, similar results were obtained when $\pi_S$ and $\pi_N$ were estimated for the entire sequence and for the sequence excluding codons with two or three polymorphic sites. In both cases, median $\pi_S$ (0.0836 and 0.0835, respectively) was significantly greater than median $\pi_N$ (0.0070 and 0.0044, respectively). There were seven datasets in which $\pi_N$ exceeded $\pi_S$ by both methods of computation, but in none of these was there a significant difference between $\pi_N$ and $\pi_S$ (z-test).

$D_{syn}$ was positive in 104 (69.8%) of the 149 datasets, and median $D_{syn}$ (0.8726) was positive (Figure 2). On the other hand, $D_{non}$ was negative in 102 (69.4%) of data sets, and median $D_{non}$ (-0.6555) was negative (Figure 2). Median $D_{syn}$ exceeded $D_{non}$ in 123 (82.6%) of 149 data sets, and the two medians differed significantly from each other (Wilcoxon signed-rank test; $P < 0.001$) (Figure 2). Also both median $D_{syn}$ and median $D_{non}$ were significantly different from zero (Wilcoxon signed-rank test; $P < 0.001$ in each case).

Overall, $D_{syn}$ and $D_{non}$ were positively correlated (Spearman’s rank correlation coefficient, $r_S = 0.236; P = 0.004$) (Figure 2). An examination of the data showed six datasets with high positive values of both $D_{non}$ and $D_{syn}$. These unusual datasets were the following (with $D_{syn}$ and $D_{non}$ values, respectively, in parentheses): *Borrelia burgdoerferi* flagellin (8.8576, 2.4235); *Clostridium difficile* adhesin (4.2776, 2.9993) and S-layer protein (3.5102, 1.9122); *Streptococcus pneumoniae* cspB (4.9233, 1.6862); *Vibrio cholerae* toxin co-regulated pilus subunit (10.1966, 2.0389); and *Wolbachia pipientis* surface protein wsp (4.2608, 3.1798). When these six datasets were excluded, there was no longer a significant correlation between $D_{syn}$ and $D_{non}$ ($r_S = 0.147; n.s.$).
Bacteria were categorized in terms of parasitism on vertebrates and in terms of surface expression of the proteins, and median $\pi_S$ and $\pi_N$ were computed for each category (Figure 3). Figure 3 illustrates median $\pi_S$ and $\pi_N$ computed excluding codons with two or three polymorphic sites; similar results were seen when the latter codons were included (data not shown). There was not a significant difference with respect to median $\pi_S$ among categories (Figure 3). However, there was a highly significant difference among categories with respect to median $\pi_N$ (Kruskal-Wallis test; $P < 0.001$). Median $\pi_N$ was much higher for surface proteins from bacteria parasitic on vertebrates than for the other three categories (Figure 3B).

Similar patterns were seen when median $D_{\text{syn}}$ and $D_{\text{non}}$ were compared across categories (Figure 4). No significant differences among categories were observed with respect to median $D_{\text{syn}}$, but median $D_{\text{non}}$ differed significantly among categories (Kruskal-Wallis test; $P = 0.035$). Median $D_{\text{non}}$ was higher for surface proteins from bacteria parasitic on vertebrates than for the other three categories (Figure 4B). Four of the six datasets identified as having unusually high values of both $D_{\text{syn}}$ and $D_{\text{non}}$ were surface proteins of bacteria primarily parasitic on vertebrates; namely *Borrelia burgdoerferi* flagellin; *Clostridium difficile* adhesin and S-layer protein; and *Streptococcus pneumoniae* cspB. Yet even when these six datasets were excluded from the analysis, there was still a significant difference among categories with respect to median $D_{\text{non}}$ (Kruskal-Wallis test; $P = 0.04$), and the highest median $D_{\text{non}}$ (-0.1683) was still seen in surface proteins of bacteria primarily parasitic on vertebrates.

It is well known that bacterial species have characteristic patterns of codon usage, and it has been proposed that bacterial codon usage is selectively maintained (Grantham
et al. 1980). In order to test whether the trends in $\pi_N$ and $D_{non}$ (Figures 3B and 4B) were correlated with codon usage, the median percent G + C at third-codon positins (GC3) was compared across datasets (Figure 5). The data in Figure 5 were computed excluding codons with two or three polymorphic sites; similar results were seen when the latter codons were included (data not shown). Although there were significant differences among categories (Figure 5), there was no obvious relationship between the pattern of GC3 variation and that of $\pi_N$ and $D_{non}$.

Tajima (1989) provides a formula for estimating the total number of deleterious mutants per DNA sequence when the D statistic is negative. For codons with a single polymorphic site in the present data set, this method yielded the estimates that 25 of 5248 synonymous polymorphisms (0.48%) are deleterious and 58 of 1840 nonsynonymous polymorphisms (3.15%) are deleterious. Thus the ratio of the proportions of deleterious mutations and nonsynonymous sites to that at synonymous sites is about 6.6:1. This ratio is over twice the ratio of nonsynonymous to synonymous sites in the data set (79806:24621 or about 3.24:1).

**Discussion**

An examination of DNA sequence of polymorphism in 149 datasets from 84 bacterial species showed a significant preponderance of negative $D_{non}$ values but not of $D_{syn}$ values. This pattern indicates the widespread occurrence of relatively rare nonsynonymous polymorphisms but not of synonymous polymorphisms. The abundance of such rare polymorphisms at nonsynonymous but not synonymous sites in the same
genes is in turn evidence that bacterial populations harbor abundant slightly deleterious 
nonsynonymous allelic substitutions, which are subject to ongoing purifying selection.

It is believed that, in many bacterial species, synonymous codon usage is subject 
to selection relating to tRNA abundance or other factors (Grantham et al. 1980). Thus, 
 purifying selection at synonymous as well as nonsynonymous sites might be expected to 
occur in bacteria. The present results suggested that such selection on synonymous sites 
is a minor factor in comparison to that on nonsynonymous sites. Deleterious mutations 
were estimated to be present at only 0.48% of synonymous sites, in comparison to 3.15% 
of nonsynonymous sites. These results are consistent with earlier estimates that the 
selection coefficients at synonymous sites in Bacteria are quite low (Bulmer 1991).

Deleterious mutations present in the bacterial populations examined here 
presumably largely represent nearly neutral mutations; i.e., those whose selection 
coefficients are close to the reciprocal of the effective population size (Ohta and Gillespie 
1996) Deleterious synonymous mutations are expected to have very low selection 
coefficients (Bulmer 1991). By contrast, at least some nonsynonymous mutations are 
expected to be strongly deleterious because they damage protein structure and thereby 
render the cell nonviable. These mutations are likely to be eliminated very quickly and 
thus not appear as polymorphisms within populations. The low $d_N / d_S$ ratios reported for 
Bacteria largely reflect the elimination of such strongly deleterious mutations. For 
example, in a comparison of over 14,000 phylogenetically independent pairs of genes 
between closely related pairs of bacterial species, Friedman et al. (2004) found a mean 
$d_N / d_S$ ratio of about 0.14; and Jordan et al. (2002) reported similar values. On the other 
hand, the present results indicate that there are abundant nearly neutral nonsynonymous
mutations present in bacterial populations, where they are subject to ongoing purifying selection.

The eventual fate of nearly neutral mutations is predicted by the nearly-neutral theory of molecular evolution to be determined by a combination of selection and drift (Ohta 1973, 1976, 2002). The present results support the prediction of this theory that selection against slightly deleterious mutations is likely to be efficient in species with large effective population sizes, since most bacterial species have very large effective population sizes (Lynch and Conery 2003). Thus, the present results are consistent with Jordan et al.’s (2002) observation of a lower $d_N / d_S$ ratio in Escherichia coli than in three other bacterial species. Since E. coli was believed to have a larger effective population size than the other species examined, Jordan et al. (2002) interpreted their results as support for the prediction of the nearly neutral theory that purifying selection will be most effective in species with large effective population sizes.

In addition, the fact that evidence for purifying was selection was found in the case of nonsynonymous but not synonymous sites supports the hypothesis that slightly deleterious mutations are more likely to be nonsynonymous than synonymous because of the effect of the former on protein structure. In the present data, deleterious nonsynonymous mutations were estimated to exceed deleterious synonymous mutations by a ratio over twice the ratio of nonsynonymous to synonymous sites.

The fact that a majority of datasets analyzed had positive $D_{syn}$ is most likely to be due to a certain degree of population subdivision in the bacterial species studied (Kreitman 2000). In the case of bacteria, limits on the extent of recombination are expected to impose limits on gene flow among clonal lineages and thus to create
population subdivision. However, it is important to recognize that the present results are inconsistent with an entirely clonal population structure in the bacterial species examined. The evidence of purifying selection reducing the frequency of slightly nonsynonymous variants implies recombination, since without recombination such variants cannot be purged from a given genetic background. The present results are thus consistent with numerous other studies showing evidence of both small-scale and large-scale recombination events among bacterial genomes (e.g., Nelson et al. 1997; McGraw et al. 1999; Hughes and Friedman 2004).

A small number of the datasets analyzed here showed high positive values of both $D_{\text{non}}$ and $D_{\text{syn}}$, suggestive of balancing selection. Of the six datasets with the highest $D_{\text{non}}$ and $D_{\text{syn}}$, all encoded surface proteins, and four encoded surface proteins of bacterial species classified as parasitic on vertebrates. Because vertebrates possess an immune system capable of specific recognition of foreign proteins, it is expected that the parasitic organisms exposed to vertebrate immune mechanisms will be under selective pressure to evade this recognition (Frank 2002). As a result, organisms parasitic on vertebrates are expected to be subject to balancing selection favoring amino acid sequence diversity of immunogenic proteins, including cell-surface proteins (e.g., Hughes and Hughes 1995).

The data sets with high positive values of both $D_{\text{non}}$ and $D_{\text{syn}}$ that were not classified among surface proteins of vertebrate parasites were the toxin co-regulated pilus subunit of *Vibrio cholerae* and the *Wolbachia pipientis* surface protein wps. As mentioned previously (see Methods), *Vibrio cholerae* was not classified as a vertebrate parasite for purpose of analyses because the vast majority of known *V. cholerae* strains belong to the natural flora of the aquatic environment (Faruque and Mekalanos 2003).
However, strains pathogenic on humans have acquired by horizontal gene transfer plasmids, phages and chromosomal gene clusters ("pathogenicity islands.") The toxin co-regulated pilus subunit gene forms part of such a "pathogenicity island" and the protein it encodes is thus exposed to vertebrate immune recognition (Kotetishvili et al. 2003).

On the other hand *Wolbachia pipientis* is a maternally transmitted obligate parasite of insects that influences its hosts life history (Tusutsui et al. 2003). The present results provide evidence for balancing selection on a surface protein (wsp) of this species. This in turn suggests that polymorphism at the locus encoding this protein may play a role in the evolutionary dynamics of the host-parasite relationship.

**Acknowledgments**

This research was supported by grant GM43940 from the National Institutes of Health.

**References**


duplication in primate ribonuclease genes. Proc Natl Acad. Sci USA 98: 3708–
3713.

Zhao, Z., Y.-X. Fu, D. Hewett-Emmett, and E. Boerwinkle, 2003 Investigating single
nucleotide polymorphism (SNP) density in the human genome and its

**Figure Legends**

Figure 1. Median pairwise difference between observed and expected (assuming a
Poisson distribution) proportions of codons with 0-3 polymorphic nucleotides in the 149
datasets. Tests of the equality of the median proportion observed and the median
proportion expected (Wilcoxon signed-rank test): ** P < 0.01; *** P < 0.001.

Figure 2. Plot of $D_{\text{non}}$ vs. $D_{\text{syn}}$ for the 149 datasets. The line is a 45° line. Median $D_{\text{syn}}$ was
significantly greater than median $D_{\text{non}}$ (Wilcoxon signed-rank test; P < 0.001).

Figure 3. Median values of (A) $\pi_S$ and (B) $\pi_N$ for datasets categorized by parasitism on
vertebrates and surface expression of the protein. There was not a significant difference
among categories with respect to median $\pi_S$ (Kruskal-Wallis test), but median $\pi_N$ differed
significantly among categories (Kruskal-Wallis test; P < 0.001).
Figure 4. Median values of (A) $D_{\text{syn}}$ and (B) $D_{\text{non}}$ for datasets categorized by parasitism on vertebrates and surface expression of the protein. There was not a significant difference among categories with respect to median $D_{\text{syn}}$ (Kruskal-Wallis test), but median $D_{\text{non}}$ differed significantly among categories (Kruskal-Wallis test; $P < 0.001$).

Figure 5. Median percent G + C at third-codon positions (GC3) for datasets categorized by parasitism on vertebrates and surface expression of the protein. There was a significant difference among categories with respect to median GC3 (Kruskal-Wallis test; $P = 0.004$).
A

$D_{\text{Syn}}$

0.05

1.0

nonsurface (30)
not vert.
surface (25)
parasite
nonsurface (41)
vertebrate
surface (53)
parasite

B

$D_{\text{non}}$

-1.0

0

-0.5

nonsurface (30)
not vert.
surface (25)
parasite
nonsurface (41)
vertebrate
surface (53)
parasite