Strong Asymmetric Mutation Bias in Endosymbiont Genomes Coincide with Loss of Genes for Replication Restart Pathways

Lisa Klasson and Siv G. E. Andersson

Department of Molecular Evolution, Evolutionary Biology Center, Uppsala University, Uppsala, Sweden

A large majority of bacterial genomes show strand asymmetry, such that G and T preferentially accumulate on the leading strand. The mechanisms are unknown, but cytosine deaminations are thought to play an important role. Here, we have examined DNA strand asymmetry in three strains of the aphid endosymbiont *Buchnera aphidicola*. These are phylogenetically related, have similar genomic GC contents, and conserved gene order structures, yet *B. aphidicola* (Bp) shows a fourfold higher replication-induced strand bias than *B. aphidicola* (Sg) and (Ap). We rule out an increase in the overall substitution frequency as the major cause of the stronger strand bias in *B. aphidicola* (Bp). Instead, the results suggest that the higher GC skew in this species is caused by a different spectrum of mutations, including a relatively higher frequency of C to T mutations on the leading strand and/or of G to A mutations on the lagging strand. A comparative analysis of 20 γ-proteobacterial genomes revealed that endosymbiont genomes lacking *recA* and other genes involved in replication restart processes, such as *priA*, which codes for primosomal helicase PriA, displayed the strongest strand bias. We hypothesize that cytosine deaminations accumulate during single-strand exposure at arrested replication forks and that inefficient restart mechanisms may lead to high DNA strand asymmetry in bacterial genomes.

Introduction

Nucleotide composition biases are common in microbial genomes and reflect mutational as well as selective forces associated with fundamental cellular processes such as replication, transcription, and translation. Some biases in nucleotide frequency statistics are well studied, particularly the preferential usage of a subset of codons in free-living bacteria that arises through selection for efficient translation of highly expressed genes (Sharp et al. 2005). Codon usage patterns in other species are dominated by mutational biases, which may drive genomes toward extreme AT or GC richness, as often observed in intracellular bacteria (Andersson and Sharp 1996*a*, 1996*b*; Sharp et al. 2005).

Strand-specific mutational biases (also referred to as GC skew) are most often accredited to the asymmetry of the replication machinery (Lobry 1996; Frank and Lobry 1999; Lobry and Sueoka 2002; Rocha 2004). Because the leading strand is replicated continuously and the lagging strand is synthesized in discrete steps by the joining of Okazaki fragments, it is perhaps not surprising that nucleotides accumulate differently on the two strands. The strand bias is normally highest at synonymous third codon positions, implying that it reflects mutational processes rather than selective constraints (Lobry and Sueoka 2002; Rocha 2004).

Strand asymmetry was first examined in mitochondrial genomes (Andersson and Kurland 1991; Asakawa et al. 1991; Tanaka and Ozawa 1994; Perna and Kocher 1995) and later recognized in chloroplast genomes (Morton 1999). An exceptionally strong strand bias is observed in *Borrelia burgdorferi*, the causative agent of Lyme disease. Most, but not all, bacterial genomes exhibit a clear difference in G versus C content between the leading and lagging strands (Rocha, Danchin, and Viari 1999; Lobry and Sueoka 2002). With the exception of *Deinococcus*, *Thermotogales*, and *Aquifex*, nucleotide strand bias is observed in all bacterial subdivisions (Rocha 2004). Fewer than a dozen of the first 100 sequenced bacterial genomes show a complete lack of compositional strand bias (Rocha 2004).

In genomes that display nucleotide composition asymmetry, the leading strand is normally enriched in Gs over Cs and in Ts over As. The underlying causes for GC skew is not completely understood, but the chemical properties of the DNA combined with species-specific replication-repair systems are often implicated. Prime among potential mutagenesis mechanisms is the deamination of C to U and of 5-methyl-C to T, reactions that occur spontaneously under normal physiological conditions in all cell types. The rate of this hydrolytic deamination is enhanced by oxidative damage and ionizing radiation and occurs at up to a 100-fold higher frequency in single-stranded DNA (Frederico, Kunkel, and Shaw 1990). Because the leading strand spends longer time in single-stranded state during replication (Marians 1992), it is supposed to be particularly vulnerable to such attacks. If the deamination mutations are not corrected before the next round of replication, they will be fixed in the genome in the form of T:A base pairs in place of the original C:G base pairs.

Cytosine deamination of single-stranded DNA has been suggested as the major contributor to asymmetric strand bias in bacteria (Frank and Lobry 1999). However, TA skews are generally not as high as GC skews which suggests that C → T substitutions is not the only explanation (Rocha 2004). Another mutation that can also contribute to strand bias is the deamination of A, which yields hypoxanthine that binds to cytosine rather than to thymine. Assuming that also this deamination process occurs more frequently on single-stranded DNA, it will further contribute to the excess of G and T on the leading strand. However, the rate of deamination of A to hypoxanthine is much slower than the rate of C to U or T (Karran and Lindahl 1980), and this bias is therefore expected to contribute less to the overall GC skew than the cytosine deamination.

Many intracellular bacteria exhibit atypically strong GC skew (Rocha 2004) that has been attributed to the extreme stability of these bacterial genomes, which may allow the bias to gradually build up over time (Rocha 2004). Thus, the implicit assumption is that cytosine deaminations

Key words: *Buchnera*, molecular evolution, nucleotide composition, strand asymmetry.

E-mail: siv.andersson@ebc.uu.se.

Mol. Biol. Evol. 23(5):1031–1039. 2006
doi:10.1093/molbev/msj107
Advance Access publication February 13, 2006

Table 1
Contribution of Replication (B_I) and Transcription- and Translation- (B_{II}) Bias in γ-proteobacteria

Species	Accession	x_1	x_2	*t*-test	MW	y_1	y_2	*t*-test	MW	x_c	y_c	B_I	B_{II}
E. coli K12	NC_000913	0.503	0.454	***	***	0.397	0.410	***	***	0.478	0.403	0.051	0.099
X. axonopodis pv. *citri*	NC_003919	0.474	0.435	***	***	0.391	0.437	***	***	0.455	0.414	0.060	0.097
X. campestri ATCC	NC_003902	0.476	0.437	***	***	0.384	0.430	***	***	0.457	0.407	0.060	0.103
E. coli EDL933[a]	AE005174	0.512	0.454	***		0.398	0.415	***		0.483	0.407	0.060	0.095
E. coli O157:H7 RIMD	NC_002695	0.513	0.453	***	***	0.398	0.415	***	***	0.483	0.406	0.062	0.095
S. typhi CT18	NC_003198	0.518	0.457	***	***	0.385	0.404	***	***	0.487	0.395	0.064	0.106
Y. pestis KIM	NC_004088	0.514	0.449	***	***	0.398	0.421	***	***	0.481	0.409	0.070	0.092
S. typhimurium LT2	NC_003197	0.518	0.451	***	***	0.382	0.403	***	***	0.485	0.393	0.070	0.108
H. influenzae[a]	L42023	0.484	0.403	***		0.451	0.473	***		0.444	0.462	0.084	0.068
V. cholerae chrI[a]	AE003852	0.505	0.414	***		0.413	0.439	***		0.460	0.426	0.095	0.084
V. cholerae chrII[a]	AE003853	0.513	0.419	***		0.415	0.441	***		0.466	0.428	0.098	0.080
C. burnetti	NC_002971	0.499	0.400	***	***	0.441	0.450	***	0.01	0.450	0.446	0.100	0.074
B. aphidicola (Ap)	NC_002528	0.473	0.380	***	***	0.442	0.485	***	***	0.427	0.464	0.102	0.082
P. multocida PM70[a]	AE004439	0.515	0.395	***		0.456	0.470	***		0.455	0.463	0.121	0.058
B. aphidicola (Sg)	NC_004061	0.489	0.379	***	***	0.438	0.498	***	***	0.434	0.468	0.125	0.073
P. aeruginosa PAO1[a]	AE004091	0.433	0.376	***		0.408	0.539	***		0.405	0.474	0.143	0.099
P. luminicens	NC_005126	0.557	0.404	***	***	0.390	0.443	***	***	0.481	0.416	0.162	0.086
B. pennsylvanicus	NC_007292	0.536	0.370	***	***	0.433	0.465	***	***	0.453	0.449	0.170	0.069
F. tularensis	NC_006570	0.559	0.373	***	***	0.414	0.433	***	***	0.466	0.423	0.188	0.084
B. aphidicola (Bp)	NC_004545	0.658	0.262	***	***	0.421	0.517	***	***	0.460	0.469	0.408	0.050
B. floridanus	NC_005061	0.777	0.242	***	***	0.406	0.484	***	***	0.510	0.445	0.540	0.056

[a] Values were extracted from Lobry and Sueoka (2002).
*** P values < 0.001 in the unpaired *t*-test or the Mann-Whitney test (MW).

occur in all bacterial genomes, but that inversions, fusions, deletions, and insertions may confound the strength of the bias in free-living bacteria.

Buchnera are obligate intracellular mutualists of aphids belonging to the γ-proteobacterial group. Three complete genomes of different *Buchnera* subspecies have been published during the last few years of the gall-forming aphid *Baizongia pistaciae* (Bp) (van Ham et al. 2003), the greenbug aphid *Schizaphis graminum* (Sg) (Tamas et al. 2002), and the pea aphid *Acyrthosiphon pisum* (Ap) (Shigenobu et al. 2000). The estimated divergence date for *B. aphidicola* (Ap) and (Sg) is at least 50–70 Myr, whereas *B. aphidicola* (Bp) represents an earlier diverging lineage that separated from the other two ca. 150 Myr (Moran et al. 1993). These genomes are very small in size, less than 700 kb, and show almost perfect gene order synteny and gene content conservation (Tamas et al. 2002; van Ham et al. 2003).

Rispe and colleagues (2004) were the first to study comparatively codon usage patterns and show strand bias differences in the *B. aphidicola* genomes. However, possible reasons for differences in strand-composition bias among closely related bacterial species have so far not been much explored. To test the hypothesis that an increased frequency of cytosine deaminations is the main cause of the higher GC skew in *B. aphidicola* (Bp), we have here estimated the patterns and relative rates of sequence evolution in the three genomes. The results suggest that the higher GC skew of *B. aphidicola* (Bp), which coincides with the loss of genes in replication restart processes, may be explained by a higher frequency of cytosine deaminations.

Methods

Sequences

The genome and protein-coding gene sequences of *Buchnera aphidicola* subspecies (Ap) (Shigenobu et al. 2000), (Sg) (Tamas et al. 2002) and (Bp) (van Ham et al. 2003), *Wigglesworthia glossinidia* (Akman et al. 2002), and *Blochmannia floridanus* (Gil et al. 2003) were downloaded from NCBI RefSeq (Accession numbers NC_002528, NC_004061, NC004545, NC_004344, NC_005061). Orthologous genes where identified automatically using Blast searches (Altschul et al. 1990) between the genomes and in uncertain cases by manual inspection and by annotations. The two sets of orthologs in *Wigglesworthia-Buchnera* and *Blochmannia-Buchnera* were mostly derived from Klasson and Andersson (2004).

Classification of Genes on the Leading and Lagging Strands

The classification of genes on the leading and lagging strands was based on the GC skew in *B. aphidicola* (Bp). Genes located between position 291112 and 351842 on the chromosome (between *cls* and *htpX*) were excluded because of difficulties to discriminate between the leading and lagging strands at these positions. In total, 193 genes were inferred to be located on the lagging strand and 267 on the leading strand in *B. aphidicola* (Bp). Classification of genes on the leading and lagging strands in the other species (table 1) was done manually based on the cumulative GC skew using a sliding window with step size 1,000 bp and a window size of 10,000 bp.

Relative Rate Tests

As outgroups, we selected *W. glossinidia* (Akman et al. 2002) and *B. floridanus* (Gil et al. 2003) to represent species with weak versus strong DNA strand asymmetry, respectively. Both are close relatives to *B. aphidicola* and possess similar genomic GC contents. The orthologous protein sequences for the two sets, *Buchnera-Wigglesworthia* and *Buchnera-Blochmannia*, were aligned separately

using ClustalW (Thompson, Higgins, and Gibson 1994) and converted to nucleotide sequences. The relative rate test was preformed as implemented in RRTree (Robinson-Rechavi and Huchon 2000), including only nonsynonymous substitutions, calculated according to Li (1993). Synonymous sites were not considered because these are in most cases saturated between the species examined. The three *Buchnera* species were treated as different lineages with either *W. glossinidia* or *B. floridanus* as outgroup. The Bonferroni-Holm correction was used to adjust the *P* values for multiple tests (Wright 1992).

Analysis of GC skew

The positional variation of the GC skew and TA skew was calculated as $(G - C)/(G + C)$ and $(T - A)/(T + A)$, respectively, using a sliding window of 10,000 bp and a step size of 1,000 bp over the entire genomes and a sliding window of 5,000 bp and step size of 500 bp over the third codon positions. The strength of the GC skew was estimated by calculating two values, B_I and B_{II}, as described by Lobry and Sueoka (2002). The B_I value shows the replication-associated effects to GC skew at third codon position ($B_I = \sqrt{(x_2 - x_1)^2 + (y_2 - x_1)^2}$, where $x = G_3/(G_3 + C_3)$ and $y = A_3/(A_3 + T_3)$ and x_1, y_1 refers to the means of the leading strand genes and x_2, y_2 to the means of the lagging strand genes. The B_{II} values shows shifts that affect genes irrespectively of their strand location ($B_{II} = \sqrt{x_c - 0.5)^2 + (y_c - 0.5)^2}$, where $x_c = (x_1 + x_2)/2$ and $y_c = (y_1 + y_2)/2$). The latter bias in GC content at third codon position, which is not strand-specific, probably arises from differences at the transcriptional and/or translational levels. Following the routines of Lobry and Sueoka (2002), we used an unpaired *t*-test to analyze the significance between the values for x_1 and x_2 and for y_1 and y_2. However, because the data do not appear to be normally distributed we also used the Mann-Whitney test to examine whether the values for x_1 and x_2 and for y_1 and y_2 were significantly different.

Analysis of Substitution Patterns

The pattern of substitutions was examined using a parsimony approach. Nucleotide alignments with the three *B. aphidicola* strains and either of the two outgroup species, *W. glossinidia* or *B. floridanus*, were inspected. Positions at which one of the *B. aphidicola* sequences differed, whereas those from the two other *B. aphidicola* species and the outgroup species were identical, were counted as a change from the majority to the single different base. In addition, we estimated the substitution patterns in the ancestral lineage of *B. aphidicola* (Ap) and (Sg). In this analysis, we counted all positions where *B. aphidicola* (Ap) and (Sg) have a nucleotide that is different from the one in *B. aphidicola* (Bp) and the outgroup species. All orthologous genes, separated into the leading and lagging strands according to their position in *B. aphidicola* (Bp), were used and substitutions were calculated separately for each of the different codon positions in each species and for all codon positions jointly. From the absolute number of changes in all genes on each strand, individual frequencies were calculated and normalized with respect to base composition (Li, Wu, and Luo 1984). To infer the variation between genes, frequencies were calculated for each individual gene longer than 750 bp, from which the variance and standard deviation (SD) were estimated (supplementary table 1). To test if there is a significant difference in the pattern of substitutions, a multinomial model was assumed with the substitution frequencies as parameters. To test if two multinomial distributions differed significantly, we used the following formula:

$$\sum_{j=1}^{2}\sum_{i=1}^{k}\frac{[x_{ij} - n_j\{(x_{i1} + n_{i2})/(n_1 + n_2)\}]^2}{n_j\{(x_{i1} + n_{i2})/(n_1 + n_2)\}},$$

which under the null hypothesis (no difference) has a χ^2 distribution with $k-1$ degrees of freedom. x_{ij} is the number of substitutions in class i from distribution j and n_j is the total number of substitutions from distribution j. The values x_{ij} were calculated by multiplying the total number of substitutions, n_j, by the normalized frequencies, so that the observations were normalized by base composition.

A limitation is that the likelihood of multiple substitutions is high at synonymous codon positions that may be saturated for mutations in the comparisons made. For these reasons, we relied mostly on the patterns of substitutions at second codon positions. However, the method seemed robust, because the same patterns of substitutions were observed irrespectively of codon positions used and outgroups selected for the analysis. Furthermore, because the aim was to identify species- and strand-specific differences in substitution patterns rather than absolute substitution frequency values, slight under- or overestimates that are spread uniformly across either the two strands or three *B. aphidicola* lineages present less of a problem.

Results

Positional Variation of the Nucleotide Strand Bias

To quantify the strength of the asymmetric substitution pattern, we first estimated the contribution of the replication-induced bias at third codon positions (B_I) relative to the bias caused by transcription- and translation-associated effects (B_{II}) in 20 γ-proteobacterial genomes, including three *B. aphidicola* genomes (table 1). The starting point for our comparative analysis is that *B. aphidicola* (Bp) shows a fourfold higher replication-induced bias (B_I = 0.408) than *B. aphidicola* (Sg) and (Ap) (B_I = 0.125 and 0.102, respectively) (table 1) (Lobry and Sueoka 2002). Taken together, as much as 89% of the total GC bias at third codon positions in this species is explained by differences in the rates and/or patterns of nucleotide substitutions between the leading and the lagging strands.

A plot of the variation in GC skew along the chromosome in *B. aphidicola* (Bp) shows a rather uniform increase relative to the other two species (fig. 1A), although the GC skew is much stronger than the TA skew in each species (fig. 1B). Local variations within and among genomes were observed, with more than 10 segments showing an altered direction of the GC skew in *B. aphidicola* (Sg) and (Ap) that is evident irrespectively of whether all sites (fig. 1A) or only third codon positions are included in the analysis (fig. 1B).

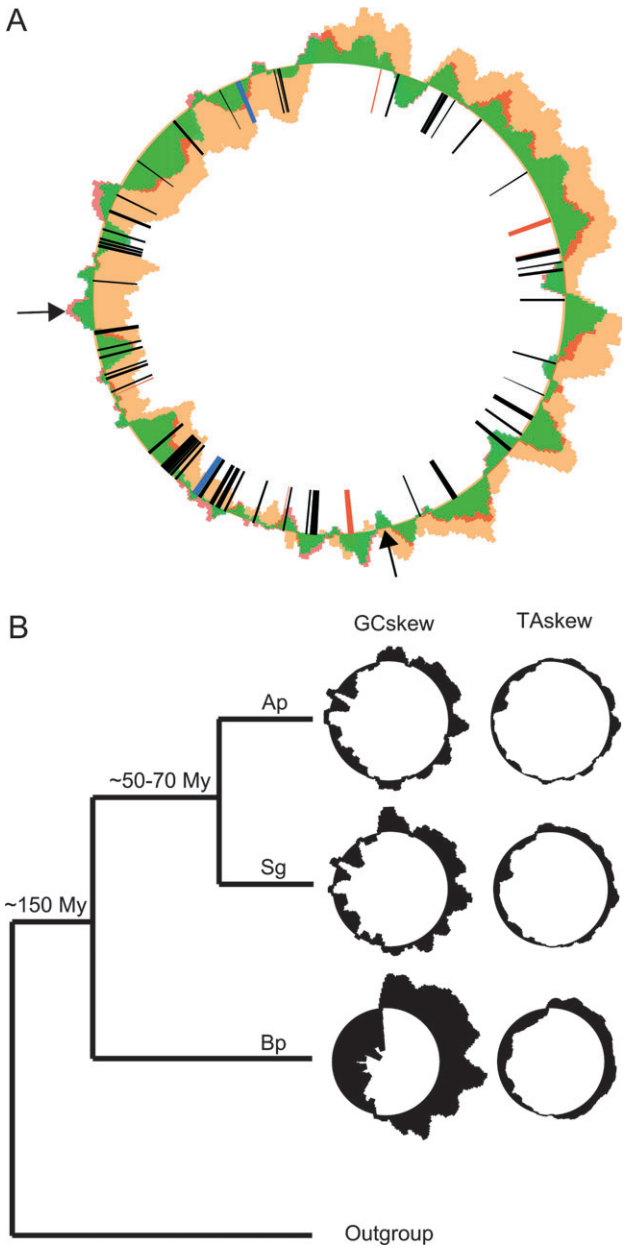

FIG. 1.—(*A*) Representation of the GC skew in *B. aphidicola* together with the positions of genes with significantly higher nonsynonymous substitution frequencies and the positions of genes lost in *B. aphidicola* (Bp) as compared with *B. aphidicola* (Sg). The colors in (*A*) show the chromosomal variation of the GC skew pattern in the different species: yellow (Bp), red (Sg), and green (Ap). Genes with significantly higher nonsynonymous substitution rate are all on the leading strand (blue lines). The black lines represent genes that are missing in *B. aphidicola* (Bp) compared to (Sg) and the red lines indicate the position of *priA*, *topA*, *dnaT*, *hinA*, *himD*, and *fis*. The arrows represent the location of the two inversions. (*B*) Representation of the GC skew and TA skew in *B. aphidicola* (Bp), (Ap), and (Sg) including third codon positions only.

Two of these segments coincide with chromosomal inversions, one of which covers six genes (*ygfZ*, *prfB*, *lysS*, *lysA*, *lgt*, and *thyA*) and the other a single gene (*pyrF*) (marked with arrows in fig. 1A). Because a reverse bias is observed at the longer segment compared to its flanking regions in *B. aphidicola* (Ap) and (Sg), we infer that the inversion probably occurred in an ancestor of these two species.

Relative Rates of Nonsynonymous Substitutions

To test the hypothesis that *B. aphidicola* (Bp) evolves more rapidly than the other two species, we estimated the relative rates of sequence evolution for a set of orthologous genes present in all three *B. aphidicola* species. The tests were performed as implemented in RRTree (Robinson-Rechavi and Huchon 2000), using *W. glossinidia* (330 orthologs) and *B. floridanus* (337 orthologs) as outgroups. The motivation for using these species as outgroups is that they, despite being closely related to *B. aphidicola* and having similar genomic GC contents (*W. glossinidia* G + C = 22.5%; *B. floridanus* G + C = 27.4%), display different levels of GC skew. The magnitude of the strand-specific DNA asymmetry in *B. floridanus* is even stronger than in *B. aphidicola* (Bp) (table 1), whereas there are no signs of such a bias in *W. glossinidia* and the origin of replication has as yet not been identified in this species.

The relative rate tests indicated that only three genes (*hflC*, *hflK*, and *rne*) evolve with a significantly ($P < 0.05$ with Bonferroni-Holm correction) elevated frequency of nonsynonymous nucleotide substitutions in *B. aphidicola* (Bp) relative to *B. aphidicola* (Sg) and (Ap). All three of these were found to evolve significantly faster in all of the four possible tests ($P < 0.05$). No gene was identified as evolving faster in *B. aphidicola* (Sg) or *B. aphidicola* (Ap).

Patterns of Strand-Specific Nucleotide Substitutions

To explore other reasons for the different magnitudes of the GC skew in the three *B. aphidicola* lineages, we recorded the spectrum of nucleotide substitutions for each codon position individually in each lineage, again using *W. glossinidia* or *B. floridanus* as the outgroups. As expected, we found that the overall distributions of the substitution frequencies differed significantly for the leading and lagging strand genes within each genome (multinomial test, $P < 0.0001$). Most importantly, we noted that the distributions differed significantly for the three *B. aphidicola* genomes on both the leading and lagging strands (multinomial test, $P < 0.0001$). The variance and the SD were estimated for genes longer than 750 bp (supplementary table 1).

Transitions were observed to be the most frequent type of substitutions in *B. aphidicola* (Bp), with a relative predominance of C → T (C:G → T:A) on the leading strand genes (which might arise through C → T on the leading strand as well as through G → A substitutions on the lagging strand) (figs. 2 and 3). The bias was consistently observed at second codon positions, which are the least likely to be saturated (fig. 2A) as well as for all codon positions combined (fig. 2B). The strand bias of the C:G → T:A substitutions in *B. aphidicola* (Bp) was equal or slightly higher than the strand bias of the A:T → G:C substitutions in the same species (fig. 2A).

The substitution patterns in *B. aphidicola* (Ap) and (Sg) was different, with C → T substitutions being equally frequent on both strands or slightly more frequent on lagging strand genes (figs. 2 and 3). However, this pattern is reversed if *B. aphidicola* (Bp) is excluded from the analysis, but even so the predominance of C → T substitutions on the leading strand is much lower in *B. aphidicola* (Ap)

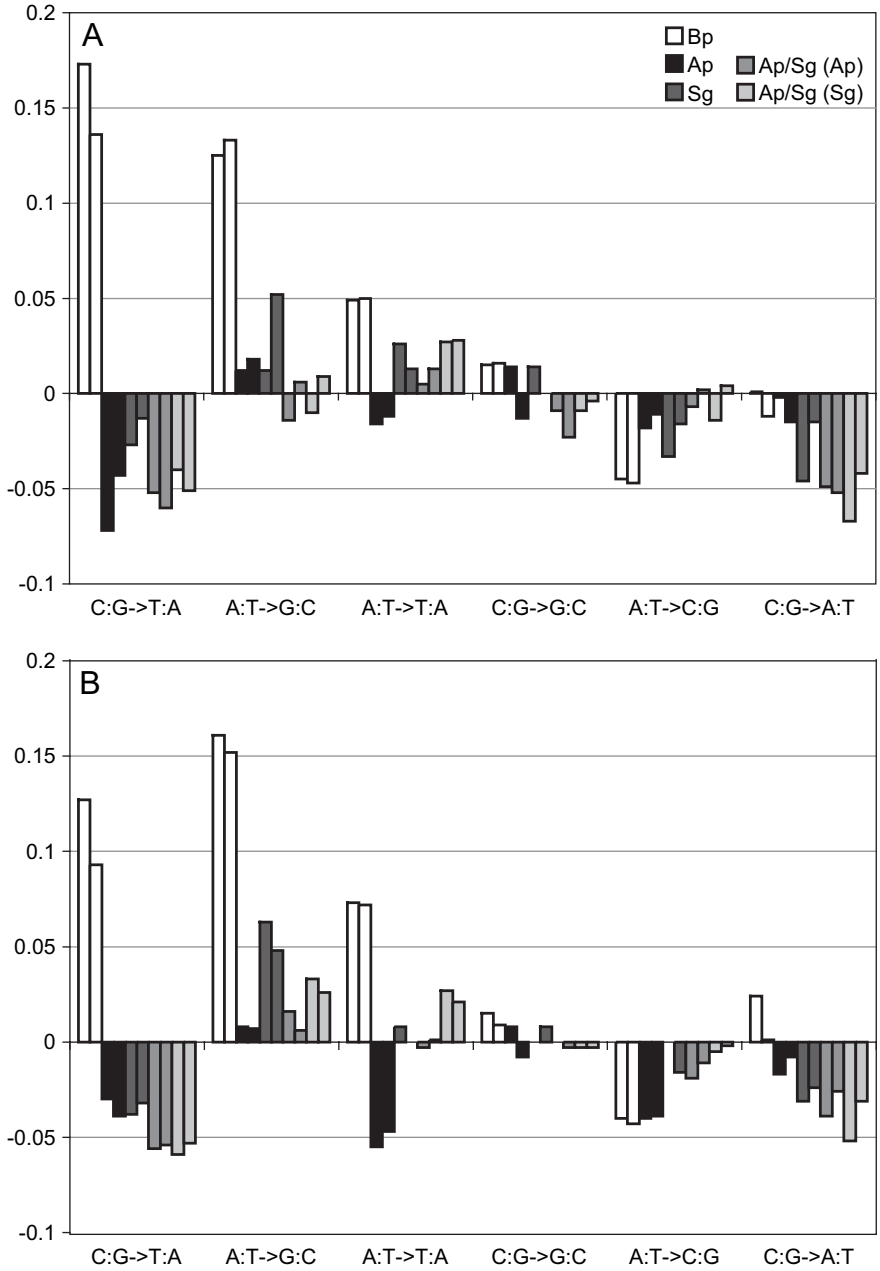

FIG. 2.—Differences in substitution patterns for genes located on the leading and lagging strands of *B. aphidicola* (Bp), (Ap), and (Sg). The values refer to the strand-specific difference in substitution frequencies for (*A*) only second codon positions and (*B*) all three codon positions. C:G → T:A = (C → T(Leading) + G → A(Lagging))−(C → T(Lagging) + G → A(Leading)) and so on. Values above and below 0 refer to the substitution frequencies that are higher on the leading and the lagging strands, respectively. Estimates including the whole branch from the divergence of *B. aphidicola* (Ap) or (Sg) from *B. aphidicola* (Bp) are indicated as Ap/Sg followed by either (Ap) or (Sg), respectively. Each species is represented by two bars of the same color, representing the two outgroups used for the analysis, *B. floridanus* and *W. glossinidia* in that order.

and (Sg) compared to *B. aphidicola* (Bp) (data not shown). A similar substitution pattern was also observed for the ancestral lineage to *B. aphidicola* (Sg) and (Ap) (figs. 2 and 3). Thus, to the extent that C → T deaminations contribute to the strand bias of the C:G → T:A substitutions at second codon positions, we infer that these occur at a higher frequency on the leading strand in *B. aphidicola* (Bp) as compared to the lagging strand (Mann-Whitney, $P < 0.001$). No such relative increase on the leading strand was observed in either of the other two species (figs. 2 and 3).

Even for divergences as low as 10%, the use of parsimony may infer an excess of common to rare changes if the sequences are highly biased (Eyre-Walker 1998). In our comparisons, the substitution frequencies were in the range of 0.15 (for *B. aphidicola* (Ap) and (Sg)) to 0.30 (for *B. aphidicola* (Bp) relative to (Ap) or (Sg)) nonsynonymous substitutions per site and the frequencies of T/(T + C) at second codon positions were 67% on the leading strand and 64% on the lagging strand. Hence, it is conceivable that we have slightly underestimated the total amount of

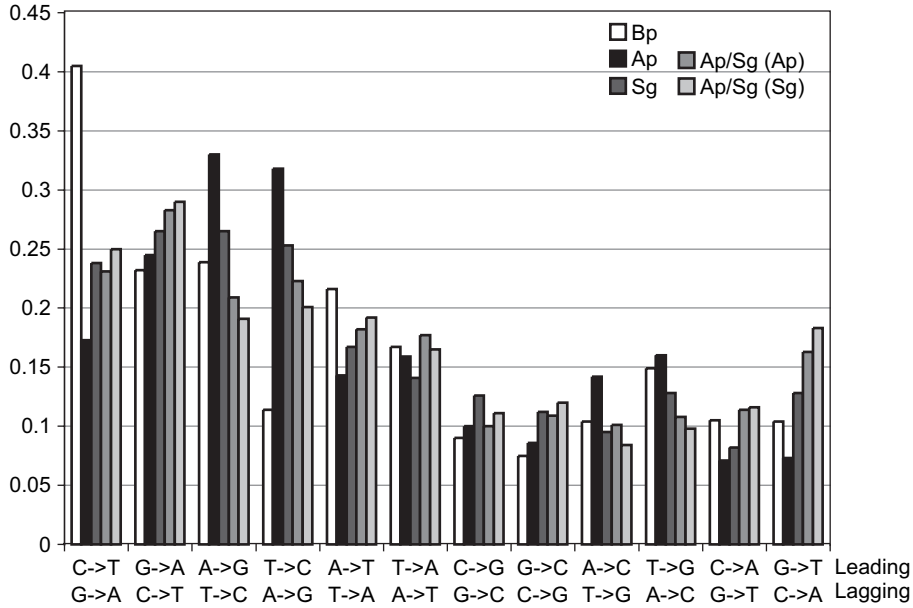

FIG. 3.—The estimated substitution frequencies for genes located on the leading and lagging strands of *B. aphidicola* (Bp), (Ap), and (Sg). The values refer to the frequency of substitutions at second codon positions. Each bar shows the sum of the equivalent substitutions at the leading and lagging strands. Estimates including the whole branch leading to *B. aphidicola* (Ap) or (Sg) are indicated as Ap/Sg followed by either (Ap) or (Sg), respectively. *B. floridanus* was used as the outgroup in the analysis.

substitutions from C → T. However, this is expected to neither influence the inferred relative differences between strands nor between species.

Discussion

Understanding why genomes of obligate intracellular bacteria tend to have a stronger DNA strand bias than those of free-living bacteria (Rocha 2004) may help unravel the mechanisms and genes involved in generating strand-composition biases in bacteria. Strong GC skew in host-associated bacteria has previously been attributed to their extreme gene order stability, which allows mutations to accumulate in the same direction over a long period of time (Rocha 2004). This cannot, however, be the whole story because the three sequenced genomes of *B. aphidicola* show rather different levels of GC-strand bias, despite conserved gene order structures (Tamas et al. 2002; van Ham et al. 2003). The close phylogenetic relationships and the near-identity in gene order offer an excellent opportunity to test how the different levels of GC skew relate to species-specific genome features.

The strand bias is most pronounced in *B. aphidicola* (Bp) (table 1); with a genome size of only 618 kb, this is also the species that has suffered the strongest reductive syndrome of the three (van Ham et al. 2003; Rispe et al. 2004). It is intuitively attractive to think that these two features are related such that the stronger strand asymmetry in this species is due to the loss of one or more components involved in the replication and repair processes, which may either cause a general mutation rate enhancement or result in a different spectrum of mutations.

Our results reject the hypothesis that the different levels of GC skew are due to relative rate differences. We observed no general enhancement of nonsynonymous substitutions in the *B. aphidicola* (Bp) genome indicating that it is not simply a higher amount of substitutions that has given rise to the stronger strand bias in *B. aphidicola* (Bp).

Instead, we suggest that the higher GC skew in *B. aphidicola* (Bp) is due to a different spectrum of mutations. Previous studies have shown a predominance of AT-enriching substitutions in deep branches of the *B. aphidicola* tree, followed by equilibrium or even inversion of this trend in the shallow branches (Clark, Moran, and Baumann 1999). In our analysis, the most striking difference between the three species is the combined higher frequency of C → T mutations on the leading strand and G → A mutations on the lagging strand in *B. aphidicola* (Bp) (fig. 3). This is consistent with the hypothesis that the higher strand bias in *B. aphidicola* (Bp) is due to an increased frequency of cytosine deaminations (Lobry and Sueoka 2002; Rocha 2004). We may ask whether this increase is associated with the loss of specific genes in DNA replication and repair processes, and if so, whether this tells us something about the mechanisms involved in generating DNA strand asymmetry in bacterial genomes.

To address this question, we searched for differences in the sets of replication and repair genes among the three *B. aphidicola* genomes and compared these to the corresponding gene sets in other γ-proteobacterial genomes (table 2). One gene, *mutH*, was identified as solely present in *B. aphidicola* (Bp) but not in the other two *Buchnera* strains. The *mutH* gene codes for the MutH endonuclease, which is a subunit of the mismatch repair (MMR) system. We have considered the possibility that the higher bias in *B. aphidicola* (Bp) is associated with the retention of the *mutH* gene, but for the reasons given below, we argue that the presence

Table 2
Presence and Absence of Replication and Repair Associated Genes in γ-proteobacteria

Species	B_I%[a]	mutH	priA	topA	dnaT	himD	himA	fis	recA
E. coli K12	33.8	+	+	+	+	+	+	+	+
X. campestri ATCC	37.0	–	+	+	–	+	+	+	+
S. typhi CT18	37.5	+	+	+	+	+	+	+	+
X. citri	38.2	–	+	+	–	+	+	+	+
E. coli EDL933	38.9	+	+	+	+	+	+	+	+
S. typhimurium LT2	39.4	+	+	+	+	+	+	+	+
E. coli O157:H7 RIMD	39.6	+	+	+	+	+	+	+	+
Y. pestis KIM	45.0	+	+	+	–	+	+	+	+
V. cholerae chrI	52.9	+	+	+	–	+	+	+	+
V. cholerae chrII	55.1	+	+	+	–	+	+	+	+
H. influenzae	55.2	+	+	+	–	+	+	+	+
B. aphidicola (Ap)	55.5	–	+	+	+	+	+	+	–
C. burnetti	57.4	–	+	+	–	+	+	+	+
P. aeruginosa PAO1	59.0	–	+	+	–	+	+	+	+
B. aphidicola (Sg)	62.9	–	+	+	+	+	+	+	–
P. luminicens	65.4	+	+	+	–	+	+	+	+
P. multocida PM70	67.5	+	+	+	–	+	+	+	+
F. tularensis	69.2	–	+	+	–	–	–	–	+
B. pennsylvanicus	71.0	–	–	–	–	–	–	–	–
B. aphidicola (Bp)	89.0	+	–	–	–	–	–	–	–
B. floridanus	90.6	–	–	–	–	–	–	–	–

[a] Calculated from the values B_I and B_{II} in table 1 as $B_{II}/(B_I + B_{II}) \times 100$.

of this gene is insufficient to explain the higher strand bias in this species.

First, the recognition signal for the MMR system is DNA methylation, which in *Escherichia coli* is accomplished by a DNA methylase. In the absence of a methylated DNA strand, the MMR system will excise nucleotides randomly, which in *E. coli* gives rise to nucleotide substitutions at high frequencies (Lobner-Olesen, Skovgaard, and Marinus 2005). Because no methylation pathway has as yet been identified in *B. aphidicola* and because we do not see a general rate enhancement in *B. aphidicola* (Bp), it is questionable whether an MMR system is operating in this species. Also, the *mutH* gene is not universally conserved among the γ-proteobacterial species, and there seems to be no correlation between its presence/absence and weak/strong DNA strand asymmetries. Finally, deamination mutations, which may account for most of the difference in bias, occur on the parental strand when it is single stranded and hence will not induce mismatches that can be eliminated by the MMR system. For all of these reasons, we consider it unlikely that the presence or absence of an MMR system explains the different levels of GC skew in the three *B. aphidicola* strains.

Another few repair genes have accumulated frameshift mutations in one or more species. Of particular interest here is the *ung* gene, which is involved in the repair of cytosine deaminations. However, also this is an unlikely candidate because the frameshift mutation was detected in *B. aphidicola* (Sg) and there is no evidence of an increased accumulation of cytosine deaminations in this or its sister species *B. aphidicola* (Ap).

More interesting in this context is the absence of the genes *priA*, *dnaT*, *topA*, *himA*, *himD*, and *fis* genes in the *B. aphidicola* (Bp) genome, all of which are involved in DNA replication initiation and re-initiation pathways (Rispe et al. 2004). The helicase protein PriA and the primosomal protein DnaT are crucial for resuming replication fork arrest (Sandler and Marians 2000) and the nucleoproteins HimA and HimD may help to initiate *dnaA*-dependent DNA replication by unwinding the DNA at the replication origin (Von Freiesleben et al. 2000; Ryan et al. 2004).

Two replication restart pathways are operating in *E. coli*; the PriA/PriB DnaT/C complex and the PriC DnaC system (Heller and Marians 2005; Lovett 2005). These differ with respect to the different structures of the arrested fork that they recognize. Whereas the PriA system restores replication after mechanical disruption and works most efficiently when a 3′-end of the leading strand is near to the gap, the PriC system restores replication following arrest of the leading strand biosynthesis. In the latter case, the lagging strand synthesis continues despite the leading strand arrest, resulting in a large gap on the leading strand that serves as the recognition target for PriC. The PriA replication restart pathway has also been studied extensively in *Neisseria gonorrheae* (Kline and Seifert 2005), which differs from *E. coli* in that it does not have a homolog of PriC, and thereby no apparent PriC restart pathway.

Both the PriA and PriC systems require recombination via the RecABCD system (which is not present in either of the *B. aphidicola* strains). Although *priA* mutants are alive in *E. coli*, they are highly defective in recombination and show signs of DNA damage as well as an increased sensitivity to DNA-damaging enzymes. Likewise, mutants in *priA* in *N. gonorrheae* show increased sensitivity to UV irradiation and oxidative-damaging agents (Kline and Seifert 2005). Additionally, they have highly reduced transformation levels, similar to those of *recA* mutants. It is interesting to note that double mutants, *priA-priC*, are not viable in *E. coli*, which suggests that a majority of replication forks are blocked and needs to be restarted (Lovett 2005). Indeed, even if only counting restart events that rely on the DnaC protein in *E. coli*, it was estimated that ca. 18% of replication rounds are arrested (Maisnier-Patin, Nordström, and Dasgupta 2001). Taken together, this provides evidence

for a strong selection pressure on the maintenance of replication restart mechanisms.

A broader analysis revealed that the genes *priA*, *topA*, *himA*, *himD*, and *fis* are conservatively present in the γ-proteobacterial subdivision (table 2). Strikingly, all five genes are absent from *B. floridanus* (Gil et al. 2003) and *Blochmannia pennsylvanicus* (Degnan, Lazarus, and Wernegreen 2005), species that like *B. aphidicola* (Bp) also show atypically strong replication-associated strand asymmetries (table 2). Additionally, *B. floridanus* and *B. pennsylvanicus* have lost the *dnaA* gene. In these species, proteins other than DnaA, such as the nucleoproteins HlpA and Hns or host factors may initiate DNA replication (Gil et al. 2003; Degnan, Lazarus, and Wernegreen 2005). Likewise, the *dnaA* gene is absent from the *W. glossinidia* genome, as is also the *priA* gene (Akman et al. 2002). However, because this species shows no strand bias and the origin of replication has not yet been identified, it was not included in table 2. In contrast to the other endosymbiont genomes, *W. glossinidia* contains the *recA* gene (Akman et al. 2002).

B. aphidicola (Bp), *B. floridanus*, and *B. pennsylvanicus* are unique among γ-proteobacterial genomes in that they encode neither the RecA nor the PriA protein (table 2). Hence, the ability to restart stalled replication forks and repair breaks that arise during DNA synthesis must be severely impaired. Combined with the additional loss of other genes involved in DNA unwinding and restart pathways, one or both DNA strands may spend a longer time in the single-stranded state, and thereby be subjected to a more extensive exposure to deamination mutations.

Thus, based on the results of our analysis, we hypothesize that the exceptionally strong DNA strand asymmetries in the genomes of *B. aphidicola* (Bp) and *B. floridanus* may be due to defective restart mechanisms of stalled replication forks. If strand bias in bacterial genomes is primarily due to cytosine deaminations, and if these accumulate preferentially on one strand at stalled replication forks, the extent of strand bias may reflect the time spent repairing such lesions. Future experimental work is clearly in place in order to determine whether cytosine deaminations accumulate at arrested replication forks, and if so, whether this provides an explanation for the strong strand bias in *B. aphidicola* (Bp).

Finally, deficient or inefficient replication restart mechanisms may provide selection for both small genome size (van Ham et al. 2003) and high chromosomal copy number (Komaki and Ishikawa 2000). This is because the smaller the genome size, the lower the probability that the replication fork is arrested during each replication cycle. Cells with higher genome copy numbers may have a selective advantage because cell division can proceed despite replication fork failure on one or a few genome copies. Thus, inefficient replication restart processes may help explain small genome sizes, high chromosomal copy numbers (Komaki and Ishikawa 2000), as well as strong GC skew in some endosymbiont genomes.

Supplementary Material

Supplementary table 1 is available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

We thank Carolin Frank for helpful discussions, Paul Sharp for sharing results during an initial phase of the study, and Mikael Thollesson for helpful suggestions on the methodology. This work was supported by the Swedish Agricultural Research Council (Formas) and the European Union (EU).

Literature Cited

Akman, L., A. Yamashita, H. Watanabe, K. Oshima, T. Shiba, M. Hattori, and S. Aksoy. 2002. Genome sequence of the endocellular obligate symbiont of tsetse flies, Wigglesworthia glossinidia. Nat. Genet. **32**:402–407.

Altschul, S. F., W. Gish, W. Miller, E. W. Myers, and D. J. Lipman. 1990. Basic local alignment search tool. J. Mol. Biol. **215**:403–410.

Andersson, G. E., and C. G. Kurland. 1991. An extreme codon preference strategy: codon reassignment. Mol. Biol. Evol. **8**: 530–544.

Andersson, G. E., and P. M. Sharp. 1996*a*. Codon usage in the Mycobacterium tuberculosis complex. Microbiology **142**: 915–925.

Andersson, S. G., and P. M. Sharp. 1996*b*. Codon usage and base composition in Rickettsia prowazekii. J. Mol. Evol. **42**: 525–536.

Asakawa, S., Y. Kumazawa, T. Araki, H. Himeno, K. Miura, and K. Watanabe. 1991. Strand-specific nucleotide composition bias in echinoderm and vertebrate mitochondrial genomes. J. Mol. Evol. **32**:511–520.

Clark, M. A., N. A. Moran, and P. Baumann. 1999. Sequence evolution in bacterial endosymbionts having extreme base compositions. Mol. Biol. Evol. **16**:1586–1598.

Degnan, P. H., A. B. Lazarus, and J. J. Wernegreen. 2005. Genome sequence of Blochmannia pennsylvanicus indicates parallel evolutionary trends among bacterial mutualists of insects. Genome Res. **15**:1023–1033.

Eyre-Walker, A. 1998. Problems with parsimony in sequences of biased base composition. J. Mol. Evol. **47**:686–690.

Frank, A. C., and J. R. Lobry. 1999. Asymmetric substitution patterns: a review of possible underlying mutational or selective mechanisms. Gene **238**:65–77.

Frederico, L. A., T. A. Kunkel, and B. R. Shaw. 1990. A sensitive genetic assay for the detection of cytosine deamination: determination of rate constants and the activation energy. Biochemistry **29**:2532–2537.

Gil, R., F. J. Silva, E. Zientz et al. (13 co-authors). 2003. The genome sequence of Blochmannia floridanus: comparative analysis of reduced genomes. Proc. Natl. Acad. Sci. USA **100**: 9388–9393.

Heller, R. C., and K. J. Marians. 2005. The disposition of nascent strands at stalled replication forks dictates the pathway of replisome loading during restart. Mol. Cell **17**:733–743.

Karran, P., and T. Lindahl. 1980. Hypoxanthine in deoxyribonucleic acid: generation by heat-induced hydrolysis of adenine residues and release in free form by a deoxyribonucleic acid glycosylase from calf thymus. Biochemistry **19**:6005–6011.

Klasson, L., and S. G. Andersson. 2004. Evolution of minimal-gene-sets in host-dependent bacteria. Trends Microbiol. **12**:37–43.

Kline, K. A., and H. S. Seifert. 2005. Mutation of the priA gene of Neisseria gonorrhoeae affects DNA transformation and DNA repair. J. Bacteriol. **187**:5347–5355.

Komaki, K., and H. Ishikawa. 2000. Genomic copy number of intracellular bacterial symbionts of aphids varies in response

to developmental stage and morph of their host. Insect Biochem. Mol. Biol. **30**:253–258.
Li, W. H. 1993. Unbiased estimation of the rates of synonymous and nonsynonymous substitution. J. Mol. Evol. **36**:96–99.
Li, W. H., C. I. Wu, and C. C. Luo. 1984. Nonrandomness of point mutation as reflected in nucleotide substitutions in pseudogenes and its evolutionary implications. J. Mol. Evol. **21**:58–71.
Lobner-Olesen, A., O. Skovgaard, and M. G. Marinus. 2005. Dam methylation: coordinating cellular processes. Curr. Opin. Microbiol. **8**:154–160.
Lobry, J. R. 1996. Asymmetric substitution patterns in the two DNA strands of bacteria. Mol. Biol. Evol. **13**:660–665.
Lobry, J. R., and N. Sueoka. 2002. Asymmetric directional mutation pressures in bacteria. Genome Biol. **3**:RESEARCH0058.
Lovett, S. T. 2005. Filling the gaps in replication restart pathways. Mol. Cell **17**:751–752.
Marians, K. J. 1992. Prokaryotic DNA replication. Annu. Rev. Biochem. **61**:673–719.
Maisnier-Patin, S., K. Nordström, and S. Dasgupta. 2001. Replication arrests during a single round of replication of the Escherichia coli chromosome in the absence of the DnaC activity. Mol. Microbiol. **42**:1371–1382.
Moran, N. A., M. A. Munson, P. Baumann, and H. Ishikawa. 1993. A molecular clock in endosymbiotic bacteria is calibrated using the insect host. Proc. R. Soc. Lond. B **253**: 167–171.
Morton, B. R. 1999. Strand asymmetry and codon usage bias in the chloroplast genome of Euglena gracilis. Proc. Natl. Acad. Sci. USA **96**:5123–5128.
Perna, N. T., and T. D. Kocher. 1995. Patterns of nucleotide composition at fourfold degenerate sites of animal mitochondrial genomes. J. Mol. Evol. **41**:353–358.
Rispe, C., F. Delmotte, R. C. van Ham, and A. Moya. 2004. Mutational and selective pressures on codon and amino acid usage in Buchnera, endosymbiotic bacteria of aphids. Genome Res. **14**:44–53.
Robinson-Rechavi, M., and D. Huchon. 2000. RRTree: relative-rate tests between groups of sequences on a phylogenetic tree. Bioinformatics **16**:296–297.
Rocha, E. P. 2004. The replication-related organization of bacterial genomes. Microbiology **150**:1609–1627.
Rocha, E. P., A. Danchin, and A. Viari. 1999. Universal replication biases in bacteria. Mol. Microbiol. **32**:11–16.
Ryan, V. T., J. E. Grimwade, J. E. Camara, E. Crooke, and A. C. Leonard. 2004. Escherichia coli prereplication complex assembly is regulated by dynamic interplay among Fis, IHF and DnaA. Mol. Microbiol. **51**:1347–1359.
Sandler, S. J., and K. J. Marians. 2000. Role of PriA in replication fork reactivation in Escherichia coli. J. Bacteriol. **182**:9–13.
Sharp, P. M., E. Bailes, R. J. Grocock, J. F. Peden, and R. E. Sockett. 2005. Variation in the strength of selected codon usage bias among bacteria. Nucleic Acids Res. **33**:1141–1153.
Shigenobu, S., H. Watanabe, M. Hattori, Y. Sakaki, and H. Ishikawa. 2000. Genome sequence of the endocellular bacterial symbiont of aphids Buchnera sp. APS. Nature **407**:81–86.
Tamas, I., L. Klasson, B. Canback, A. K. Naslund, A. S. Eriksson, J. J. Wernegreen, J. P. Sandstrom, N. A. Moran, and S. G. Andersson. 2002. 50 million years of genomic stasis in endosymbiotic bacteria. Science **296**:2376–2379.
Tanaka, M., and T. Ozawa. 1994. Strand asymmetry in human mitochondrial DNA mutations. Genomics **22**:327–335.
Thompson, J. D., D. G. Higgins, and T. J. Gibson. 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res. **22**:4673–4680.
van Ham, R. C., J. Kamerbeek, C. Palacios et al. (16 co-authors). 2003. Reductive genome evolution in Buchnera aphidicola. Proc. Natl. Acad. Sci. USA **100**:581–586.
Von Freiesleben, U., K. V. Rasmussen, T. Atlung, and F. G. Hansen. 2000. Rifampicin-resistant initiation of chromosome replication from oriC in ihf mutants. Mol. Microbiol. **37**:1087–1093.
Wright, S. P. 1992. Adjusted P-values for simultaneous inference. Biometrics **48**:1005–1013.

Jennifer Wernegreen, Associate Editor

Accepted February 7, 2006