Similar compositional biases are caused by very different mutational effects

Eduardo P.C. Rocha, Marie Touchon and Edward J. Feil

Genome Res. 2006 16: 1537-1547 originally published online October 26, 2006
Access the most recent version at doi:10.1101/gr.5525106

Letter

Similar compositional biases are caused by very different mutational effects

Eduardo P.C. Rocha,[1,2,4] Marie Touchon,[1,2] and Edward J. Feil[3]

[1]*Unité Génétique des Génomes Bactériens, URA 2171, Institut Pasteur, 75015 Paris, France;* [2]*Atelier de BioInformatique, Université Pierre et Marie Curie-Paris 6, 75005 Paris, France;* [3]*Department of Biology and Biochemistry, University of Bath, Claverton Down, Bath BA2 1AJ, United Kingdom*

Compositional replication strand bias, commonly referred to as GC skew, is present in many genomes of prokaryotes, eukaryotes, and viruses. Although cytosine deamination in ssDNA (resulting in C→T changes on the leading strand) is often invoked as its major cause, the precise contributions of this and other substitution types are currently unknown. It is also unclear if the underlying mutational asymmetries are the same among taxa, are stable over time, or how closely the observed biases are to mutational equilibrium. We analyzed nearly neutral sites of seven taxa each with between three and six complete bacterial genomes, and inferred the substitution spectra of fourfold degenerate positions in nonhighly expressed genes. Using a bootstrap procedure, we extracted compositional biases associated with replication and identified the significant asymmetries. Although all taxa showed an overrepresentation of G relative to C on the leading strand (and imbalances between A and T), widely variable substitution asymmetries are noted. Surprisingly, all substitution types show significant asymmetry in at least one taxon, but none were universally biased in all taxa. Notably, in the two most biased genomes, A→G, rather than C→T, shapes the compositional bias. Given the variability in these biases, we propose that the process is multifactorial. Finally, we also find that most genomes are not at compositional equilibrium, and suggest that mutational-based heterotachy is deeply imprinted in the history of biological macromolecules. This shows that similar compositional biases associated with the same essential well-conserved process, replication, do not reflect similar mutational processes in different genomes, and that caution is required in inferring the roles of specific mutational biases on the basis of contemporary patterns of sequence composition.

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

The study of genome composition made direct and important contributions to our understanding of DNA structure and evolution well before complete genome sequences were available (Chargaff 1950; Sueoka 1962). Since then, many studies have attempted to infer mutational scenarios to account for compositional deviations such as asymmetric nucleotide composition on the two strands of replication. A major difficulty arises from the fact that 12 different substitutions are possible between the four nucleotides in DNA. Because very different sets of mutations may lead to similar compositional biases, it is usually a very speculative exercise to infer the grounds of the relevant mutational asymmetries just by analyzing compositional deviations. Since many different chemical attacks and repair mechanisms affect DNA mutations (Friedberg et al. 1995), it is even more difficult to unravel the biological basis of the mutational biases. The usual processes of inference also implicitly assume that compositional deviations reflect the current mutational biases affecting genomes. However, compositional deviations accumulate through millions of years, and genome rearrangements, repair, or replication gene losses or gains can lead to rapid shifts in the mutation spectra. In order to detect contemporary mutation biases, it is therefore necessary to examine de novo mutational spectra, which requires a massive analysis of orthologous genes between multiple very closely related genomes.

[4]Corresponding author.
E-mail erocha@pasteur.fr; fax 33-1-44-27-63-12.
Article published online before print. Article and publication date are at http://www.genome.org/cgi/doi/10.1101/gr.5525106.

Here, we use such an approach to investigate how the asymmetric replication of DNA into a leading and a lagging strand (Kitani et al. 1985) shapes the relative frequency of each individual mutation type on each strand in widely divergent taxa. We then tried to understand how these mutational biases can explain compositional differences between the two strands. Typically, compositional strand bias is evident by the relative enrichment of G (and T) over C (and A) in the replicating leading strand, although curiously, the enrichment of G tends to be more pronounced than the enrichment of T, and the latter is sometimes inversed (i.e., A is enriched in the leading strand). Replication strand bias has been noted in the genomes of bacteria (Lobry 1996; McInerney 1998; Mrázek and Karlin 1998), archaea (Lopez et al. 1999; Lopez and Philippe 2001; Worning et al. 2006), viruses (Mrázek and Karlin 1998; Grigoriev 1999), organelles (Andersson and Kurland 1991; Pesole et al. 1999; Bielawski and Gold 2002), and humans (Touchon et al. 2005). The intensity of the bias in some genomes is extremely strong, being the most important factor shaping heterogeneities in intragenomic amino acid composition (Lafay et al. 1999; Mackiewicz et al. 1999; Rocha et al. 1999; Tillier and Collins 2000a).

The customary explanation for compositional strand bias is that it arises from a longer exposure in the ssDNA state of the template serving to synthesize the lagging strand. Cytosine deamination, because it occurs so much more frequently in ssDNA (Coulondre et al. 1978), has been proposed to be the most important cause for the bias (Reyes et al. 1998; Frank and Lobry 1999). This hypothesis is attractive because it explains a nearly universal bias by a fundamental chemical property of DNA while

being corroborated by some genomic data (see Frank and Lobry 1999; Rocha 2004b and references therein). However, although more frequent C→T mutations in the leading strand can help to explain the typical enrichment of both G and T in the leading strand, it does not explain why this enrichment is more pronounced for G than for T (Rocha and Danchin 2001). Also, there are some exceptions to the systematic association of G and T enrichment in the leading strand. In the low G+C firmicutes, (e.g., *Bacillus* and *Staphylococcus*), the frequency of T relative to A is smaller in the leading than in the lagging strand genes (Lobry 1996; Worning et al. 2006), but leading strand genes are still G rich. Another notable exception is the high G+C species *Streptomyces coelicolor*, in which the leading strand is slightly C richer (Bentley et al. 2002). Thus C→T changes cannot act alone to create the observed biases, and other substitution types must also play a role. The asymmetric deamination of adenine has also been proposed to produce preferentially A→G mutations in the leading strand (Reyes et al. 1998; Mackiewicz et al. 2003), but this also does not solve the problem as the stronger enrichment of G than T on the leading strand implicates the secondary role of a substitution type that enriches for G without a commensurate enrichment of T (e.g., C→G, C→A) on the leading strand. Alternatively (or in addition), a loss of T without a commensurate loss of G (e.g., T→A, T→G) will also result in stronger GC biases. Finally, in many cyanobacteria and mollicutes, there are no apparent strand compositional biases, suggesting that mutational biases are either absent or cancel each other out in these taxa.

A more detailed analysis of genes that have suffered a strand switch since speciation in *Chlamydiacea* and *Bacillii* was compatible, although not indicative, with C→T deamination being responsible for a maximum of two-thirds of the bias, complemented by important C→G asymmetries (Rocha and Danchin 2001). A major problem with this and other subsequent studies aiming at identifying the mutational basis of compositional strand bias is that they relied on sequences saturated with synonymous substitutions or sites highly affected by selection, and it was often not possible to orientate the changes, that is, separating a change X→Y from a change Y→X (Rocha and Danchin 2001; Szczepanik et al. 2001; Klasson and Andersson 2006). The inference of mutational biases is very hazardous when substitutions are near saturation (Eyre-Walker 1998), especially when some nucleotides are much more frequent than others. It is also unclear if the statistics used to assign statistical significance are robust to deviations of normality in the data. Another problem with earlier analyses is that they focused on one or two clades, which may not capture the diversity of the processes leading to compositional strand bias. Finally, although the response of gene composition to strand switch has been studied (Rocha and Danchin 2001; Szczepanik et al. 2001; Tillier and Collins 2000b), the possibility that genes are not at compositional equilibrium even when they have not engaged in strand switch is rarely, if ever, considered.

Here, we use an extensive set of multigenomic comparisons and statistical procedures to directly identify those mutation types occurring at significantly different frequencies on each strand. Very closely related genomes have few, if any, multiple substitutions, and also provide a more reliable mutational footprint, as the changes observed will have arisen recently and will not have been filtered to any large degree by purifying selection (Rocha et al. 2006). This is especially true if one analyzes fourfold degenerate third codon positions in non-highly expressed genes, as these positions are expected to be under very weak, if any, selection. We use more than three genomes in most data sets, thus increasing the confidence in the assignment of directionality to changes (Table 1). In all cases, we analyze >100,000 sites for each strand and use nonparametric bootstrap procedures to evaluate the statistical robustness of the results. The analysis of substitution frequencies shows a much more diverse picture of the asymmetric replicative processes shaping the composition of genomes than expected, and suggests that the asymmetry of mutation spectra differs markedly between taxa and may be subject to frequent change.

Results

Checking subsaturation levels and that genes evolve at the same rates irrespective of strands

We divided the orthologs within each clade according to the strand upon which they are coded, and removed genes not consistently found in the same strand within a given taxon. We restricted our analysis to fourfold degenerate sites to avoid the effect of selection on nonsynonymous changes and removed the 10% putatively most highly expressed genes to decrease the problems associated with selection on codon usage (see Methods). For the clade with highest codon usage bias, *Escherichia coli*, we repeated the entire analysis removing the 15% most highly ex-

Table 1. Bacterial genomes used in the study

Genomes	No. of strains	Group	G+C[a]	B_l[b]	ΔGC[c]	ΔAT[d]	Max d_S[e]	Alignment's length	
								Lead	Lag
Escherichia	6	γ-Proteobacteria	51%	0.061	0.10	−0.06	0.077	1,099,425	762,978
Streptococcus	6	Firmicute	38%	0.147	0.29	0.08	0.048[f]	932,679	210,291
Bacillus	5	Firmicutes	35%	0.314	0.61	0.12	0.36	1,359,318	395,040
Rickettsia	4	α-Proteobacteria	29%–32%	0.126	0.24	−0.08	0.22[f]	316,977	222,648
Staphylococcus	4	Firmicutes	33%	0.159	0.32	0.03	0.08	1,164,927	427,185
Bordetella	3	β-Proteobacteria	68%	0.136	0.09	−0.25	0.041	1,176,738	705,645
Neisseria	5	β-Proteobacteria	52%	0.142	0.22	−0.17	0.15	525,744	493,308

[a]G+C content.
[b]B_l is the level of strand bias (Methods).
[c]Difference in GC skew between genes in the leading and in the lagging strand.
[d]Difference in AT skew between genes in the leading and in the lagging strand.
[e]Maximal d_S among the taxa and number of sites in the alignments. d_S between the outgroup and the most distant ingroup.
[f]Only ingroups analyzed.

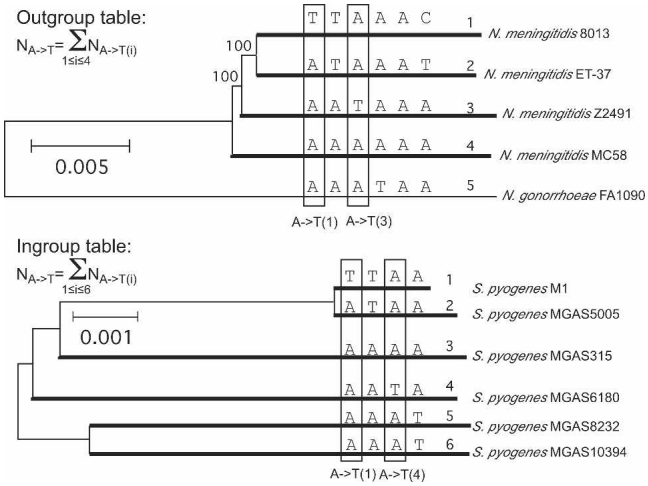

Figure 1. Scheme of the analysis. For most taxa we can reliably predict the outgroup (*top*). In this case, only the changes taking place in each of the terminal leaves of the ingroups are counted (thick lines), and only if there were no changes in the same position elsewhere in the tree, that is, if all other nucleotides in the same position are strictly identical. The table contains the sum of all substitutions of each type. When there is no single outgroup, or it cannot be reliably predicted, the analysis is done on all terminal branches of the tree (*bottom*). A substitution is marked if it only occurred in that branch of the tree.

pressed genes, without any significant changes in the results (Supplemental Fig. 1). We then estimated the number of multiple substitutions in the ingroups and compared it with the expected number both in the *Bacillus* (distant comparisons) and in the *Staphylococcus* (close comparisons). We found single substitutions in an excess of multiple substitutions by a factor of 10 in *Bacillus* and 100 in *Staphylococcus*. We then checked that these values are within the order of magnitude of the expected number if substitutions accumulate randomly. In *Bacillus* we obtained slightly more multiple substitutions than expected given the frequency of single substitutions (16% more), whereas in *Staphylococcus* we obtained 35% less. The larger difference in the latter is most likely due to the very low number of multiple substitutions observed (<50); thus the difference from the expected value can be accounted for by the small sample size. Finally, we computed the rates of synonymous (d_S) substitutions for each gene (Yang and Nielsen 2000). The highest d_S is associated with the comparisons involving the outgroup in the bacilli, *Bacillus cereus* ATCC14579, which shows a d_S of ~0.36 with the ingroups. Although the substitution spectrum of outgroups is not used, it may influence the inference of substitutions in the ingroups. We thus computed the substitution table of this genome by assuming that a position for which it differed from all the others, when these are identical among them, is a substitution in the outgroup. The corresponding substitution table is very strongly correlated to the one obtained for *Bacillus anthracis*, the one with the smallest terminal branch in the group (Pearson correlation coefficient = 0.97). This shows that multiple substitutions are few, within the expected bounds, and that they are not seriously biasing the results even in the most distant elements of the clade.

In our tests we examine the asymmetry of specific substitution frequencies (Fig. 1). If genes in the two types of replicating strands evolve at significantly different rates, then the observed differences in relative substitution frequencies could potentially reflect distinct selection pressures on codon usage in the two populations of genes. Hence, we tested if there were systematic differences in d_S between the strands. A paired *t*-test on the log-transformed d_S values for the leading and lagging strands shows no significant difference ($P > 0.3$) (Fig. 2). An exhaustive analysis shows that d_S values in leading and lagging strands are only significantly distinct in one comparison out of 61 ($P = 0.02$, after a sequential Bonferroni correction for multiple tests). Naturally, statistical tests are limited by sample sizes; but these are very large in the present analysis. This strongly suggests that previous observations that lagging strand genes evolve faster (Rocha and Danchin 2001; Mackiewicz et al. 2003) are probably caused by purifying selection on genes that are more frequent in one replicating strand, such as highly expressed genes. The differences all but disappear when this effect is controlled by removing the highly expressed genes.

Analysis of substitution frequencies

After having proceeded through all the checks described in the previous section, we computed the relative substitution frequency tables from the data. First, we opposed the substitution frequencies of complementary changes. A bootstrap procedure tests the differences between, for example, $C\rightarrow T_{leading} + G\rightarrow A_{lagging}$ versus $G\rightarrow A_{leading} + C\rightarrow T_{lagging}$. If there is no strand bias, then one would expect these summed substitution frequencies to be the same. Because it allows pooling the data sets, this is statistically more powerful than analyzing separately all changes in the two strands. However, we also did the complementary analysis, which is described in the next section.

Before considering strand asymmetries, it is clear that the overall mutational spectra are very diverse among different bacteria (Table 2). This was expected because bacteria exhibit widely different nucleotide compositions, with G+C contents varying from 25% to 75% (Sueoka 1962; Muto and Osawa 1987) and at fourfold degenerate sites between <10% and >90%. It is therefore not surprising that changes leading to G+C enrichment are much more frequent in G+C-rich bacteria such as *Bordetella*, than, say, among the A+T-rich *Streptococcus pyogenes* (Table 2. As expected,

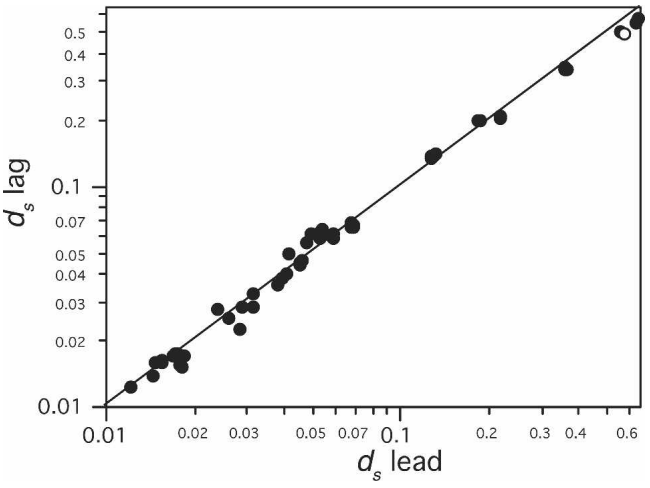

Figure 2. Average d_S values of genes in the leading and lagging strands. The open circle indicates the only comparison between (*R. conorii* and *R. prowazekii*) genes in the leading and lagging strand that is significantly different after applying a Bonferroni correction for multiple tests ($P < 0.05$).

Table 2. Normalized mutation frequencies per 1000 positions and transitions/transversions ratio in the pooled data analysis

Change	*Bacillus*	*Bordetella*	*Escherichia*	*Neisseria*	*Rickettsia*	*Staphylococcus*	*Streptococcus*
A→C	2.252	1.451	1.753	2.032	9.469	0.262	0.645
T→G	3.703	1.162	1.987	2.428	10.175	0.348	0.643
P	− −	NS	−	NS	NS	−	NS
A→G	9.068	4.539	5.299	5.763	20.763	1.036	1.859
T→C	6.130	5.057	5.206	5.573	20.331	0.832	1.917
P	++	NS	NS	NS	NS	++	NS
A→T	2.327	0.323	2.196	1.041	5.814	0.751	0.953
T→A	3.165	0.262	2.011	0.986	5.862	0.802	0.916
P	− −	NS	NS	NS	NS	NS	NS
C→A	2.842	0.203	1.735	1.717	8.696	1.341	1.383
G→T	2.740	0.178	1.868	1.451	13.101	1.038	1.696
P	NS	NS	NS	++	− −	+	−
C→G	1.353	0.558	1.177	1.869	1.693	0.366	0.698
G→C	1.022	0.459	1.089	1.412	1.889	0.280	0.736
P	+	+	NS	++	NS	NS	NS
C→T	15.570	1.590	9.228	8.070	34.975	3.848	6.633
G→A	15.722	0.736	6.719	5.784	35.528	3.655	4.628
P	NS	++	++	++	NS	NS	++
ts/tv	4.79	5.19	3.83	3.89	3.94	3.61	3.92

P is the *P*-value associated with the nonparametric bootstrap test (see Methods) that genes have a different nucleotide frequency in a given mutation than in its complement as seen from the leading strand (i.e., A→G includes $A\rightarrow G_{leading}$ and $G\rightarrow A_{lagging}$). (+/−) $P < 0.05$; (++/− −) $P < 0.01$.
(NS) Nonsignificant difference.
(+/++) Gain or (−/− −) loss in the top change (e.g., A→G in the A→G and T→C comparison).

transitions are much more frequent than transversions, but the range is rather large, from 3.61 to 5.16 times more frequent (Table 2). The frequency of some mutations is more surprising. For example, G→C and C→G transversions are often found to be extremely rare (Hudson et al. 2003), but in our data they are not always the rarest, for example, in *Bordetella* or *Neisseria* (note that frequencies are normalized by nucleotide composition, thus this is not a trivial association with G+C content).

The number of complementary substitution types showing significant strand asymmetries is strikingly variable among genomes, with a minimum of one in six in *Rickettsia* to four in six in *Bacillus* (Table 2; Fig. 3). This is not a trivial consequence of the varying statistical power of the comparisons owing to the differing numbers of substitutions, as *Rickettsia* and *Bacillus* represent comparisons involving more changes than many of the other taxa. Hence, one must conclude that the number of asymmetric substitutions is highly variable among bacteria. In *Bacillus* the two pairs that show no significant asymmetry are C→A (G→T) and, surprisingly, C→T (G→A). In fact, and quite unexpectedly, in *Bacillus*, as in *Staphylococcus* and *Rickettsia*, there are no more C→T changes than G→A changes in the leading strand. Hence, in these genomes, compositional strand bias cannot result from preferential cytosine deamination in the leading strand. This set includes two of the three firmicutes and one of the four proteobacteria, showing that the absence of asymmetric cytosine deamination is not clade specific. Also, it should be pointed out that the low G+C firmicutes, *Bacillus* and *Staphylococcus*, show the two largest compositional strand biases (Table 1). Hence, these results downplay the role of cytosine deamination in generating strand asymmetry in general, and in particular within the two most strongly biased genomes. In contrast, A→G substitutions are significantly associated with the leading strand only in *Bacillus* and *Staphylococcus*. This substitution type, and not C→T changes, is therefore accounting for leading strand G enrichment in these species.

In four of the seven genomes, we did find an asymmetry between C→T and G→A changes that is compatible with, although not demonstrative of, preferential cytosine deamination in the leading strand. However, it should be emphasized that all substitutions show significant asymmetry in at least one group of bacteria. Some transversions are almost as consistently biased as C→T versus G→A. For example, A→C versus T→G and C→G versus G→C are significantly asymmetric in three of the seven groups and always show the same sign. The significant preference for C→G changes over G→C changes in *Bacillus*, *Bordetella*, and *Neisseria* can help to explain why GC skews tend to be stronger than AT skews. Furthermore, in the firmicutes, we note a significant leading strand preference for T→G (*Bacillus* and *Staphylococcus*), T→A (*Bacillus* only), and C→A (*Staphylococcus* only), all of which may contribute to the unusual leading strand enrichment of A in these species. The latter case of C→A change is the only example of inconsistent asymmetry throughout the data sets. The contrast of C→A with G→T shows four cases of significant asymmetries, two with each sign, that is, in *Neisseria* and *Staphylococcus*, the frequency of C→A is higher than G→T in the leading strand, whereas the inverse is found in *Rickettsia* and *Streptococcus*. Furthermore C→A versus G→T is the only significant asymmetry noted in *Rickettsia* but in this case appears to be particularly strong. Overall, these results indicate that: (1) All substitution types show significant strand asymmetry in at least one genome; (2) different genomes show very different numbers of significantly asymmetric changes; (3) no single type of change is systematically associated with compositional strand bias; (4) most types of change are consistent, in the sense that either they

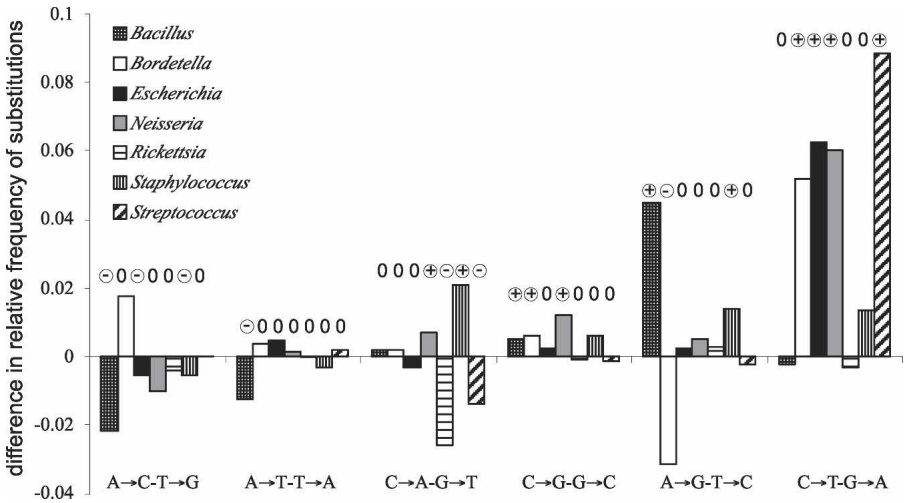

Figure 3. Difference between the pairs of symmetric substitutions in different genomes. We took the data in Supplemental Table 1 and normalized for each genome so that the sum of the frequencies of each type of substitution is 1. Hence, high absolute values reveal unbalance within the clade, but they differ in terms of statistical robustness because of the different number of variable sites in each clade. (+/−) Significantly different from 0 ($P < 0.05$).

are not significant or they are of the same sign in this set of genomes.

Direct comparison of changes between strands

The previous approach has the advantage that pooling the data increases statistical signal. However, when we compute $C{\rightarrow}T_{leading} + G{\rightarrow}A_{lagging}$ against $G{\rightarrow}A_{leading} + C{\rightarrow}T_{lagging}$, we are not separately testing $C{\rightarrow}T_{leading}$ versus $C{\rightarrow}T_{lagging}$ and $G{\rightarrow}A_{leading}$ versus $G{\rightarrow}A_{lagging}$. To verify these results, we also carried out the nonpooled analysis (see Methods). In the majority of the cases (32/42), the results of the tests are strictly identical (Supplemental Table 2). In the remaining cases, the results of the tests are not identical but can be trivially explained by the lower power of the tests caused by the smaller sample sizes. For example, in the pooled analysis, C→G is more frequent in the leading strand of *Neisseria* than G→C (Table 2). In the nonpooled analysis of *Neisseria*, the frequency of $C{\rightarrow}G_{leading}$ is, indeed, higher than that of $C{\rightarrow}G_{lagging}$, but the difference is not significant. Similarly, in *Streptococcus*, C→A is less frequent in the leading strand than G→T in Table 2, whereas in Supplemental Table 2 none of them is significantly asymmetric. It is therefore likely that some substitutions are nonsignificant in this analysis simply because the power of the test is lower for the analysis where data sets are smaller, given the smaller number of changes available for comparison. This is especially true for groups of genomes with few genes in the lagging strand, such as the Firmicutes, and for very recently diverged genomes with fewer substitutions. This analysis therefore depicts and confirms the previous one: Many distinct substitutions are responsible for replication strand bias, and these differ between genomes. Notably, the results concerning the cytosine deamination theory are strictly identical, showing an equal frequency of C→T changes in three out of the seven clades.

Are genomes generally close to equilibrium?

The previous sections dealt with the observed substitution frequencies, that is, the likelihood that a given nucleotide will change into another. Such values are a proxy of the mutational biases operating in the genome and can be used to estimate the nucleotide composition at equilibrium. However, the actual composition at fourfold degenerate sites of the genome may not correspond to the equilibrium values, because of recent changes in the mutational spectra or because of preferential selection for certain nucleotides. Hence, we sought to determine if genomes were close to equilibrium relative to compositional strand bias.

For this, we took the mutation spectra of each replicating strand and computed the nucleotide composition changes until equilibrium (Fig. 4A,B). In the majority of cases, we found a gap between the expected compositions of fourfold synonymous positions given the mutational spectra and their actual composition. To examine how closely the predicted equilibria fit the extant genome compositions, we first subtracted the observed skews in the lagging strand, (G − C)/(G+C) and (A − T)/(T+A), from those observed in the leading strand, thus providing an index of strand asymmetry, as in Rocha and Danchin (2001). We then computed the expected genome composition at equilibrium. Finally, we computed the differences in skews between the leading and lagging strands from the predicted composition and plotted the net difference in strand asymmetry between the actual genome and the expected genome at equilibrium (Fig. 4C,D,E). For some genomes, the differences are very pronounced. For example, in *Bacillus*, C increases relative to G in the leading strand, and both nucleotides remain in the same relative frequency in the lagging strand, resulting in a net loss of GC skew (Fig. 4C). On the other hand, A increases relative to T in the leading strand and decreases in the lagging strand, resulting in a small net gain of AT skew (Fig. 4D). At equilibrium, *Bacilli* remain among the most biased genomes, equivalent only to *Bordetella*. In the latter, G is increasing relative to C in the leading strand and decreasing in the lagging strand, whereas the inverse happens to A relative to T. The overall skew (B_1) is expected to increase significantly in this group (Fig. 4E). In *Staphylococcus*, the changes are very small, and the observed skews closely fit the substitution spectra. A dramatic change is observed in the data set for *Rickettsia*. These genomes, which according to current composition show average biases, have a composition at equilibrium that leads to slightly lower AT skews and to negative GC skews. Hence, if genomes evolve according to the computed mutational spectra, *Bordetella*, *Neisseria*, *Escherichia*, and *Streptococcus* will become more biased, *Bacillus* will have a lower bias, and the bias in *Rickettsia* will decrease concomitantly with an inversion in GC skews. Overall, four genomes will become more skewed and two less so. Only *Staphylococcus* will remain nearly unchanged. Naturally, this does not mean that more genomes are gaining compositional strand bias than losing it, since both inversions of genes from one strand to the other and horizontal gene transfer are expected to decrease the genome overall biases.

It is important at this stage to note that even though two of the three genomes lacking C→T asymmetry are losing composi-

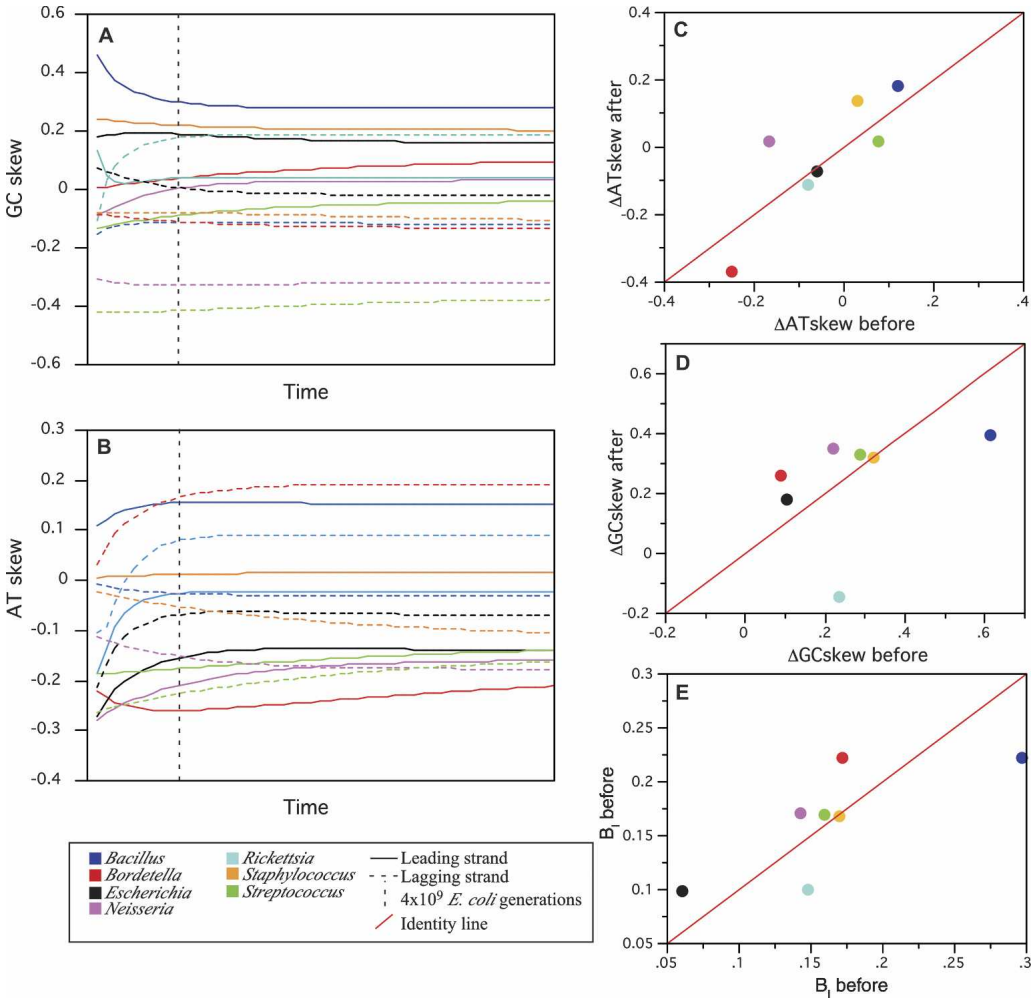

Figure 4. Evolution of (*A*) GC skews and (*B*) AT skews from actual values to the values close to equilibrium in each replicative strand of each clade using the substitution frequencies computed in Supplemental Table 2. The results for the leading strand are in solid lines, and the results for lagging strands are in dashed lines. The units in the *x*-axes are arbitrary and represent the iteration of the transition matrix in the simulations. To provide an order of magnitude of the number of generations required to attain equilibrium, we indicate by a dashed vertical line the point corresponding to 4×10^9 *E. coli* generations (~20 million yr). This calculation is just indicative and assumes a mutation rate of 10^{-9} per generation and 200 generations per year. (*C*) ΔAT skew, (*D*) ΔGC skew, and (*E*) B_I change from the actual genome (before) to the values predicted at equilibrium (after).

tional strand bias (*Staphylococcus* and *Bacillus*), at equilibrium these still remain among the most skewed genomes. More strikingly, this analysis suggests a large discrepancy between the actual composition and the one at equilibrium. To test that changes depicted in Figure 4 are significant, we computed the substitution spectra for each of the 1000 bootstrap experiments and then inferred the composition at equilibrium. This was done for each taxa and for each strand. Using the distribution of nucleotide compositions at equilibrium for each 1000 bootstrap analyses, we tested if >95% of the values were lower or higher than the one of the actual data set, that is, than the current composition of the genome. A large majority of tests (42/56) showed a significant ($P < 0.05$) deviation between the composition of the actual genome and the expected values at equilibrium (Fig. 5). This reinforces the previous analyses and shows that compositions away from equilibrium are the rule, not the exception, even in positions expected to be under very weak, if any, selection. A remarkable exception is found in *Staphylococcus*, which is at equilibrium. This may partly be caused by the low number of substitutions in the set (Supplemental Table 2). However, we included 2254 substitutions in the leading strand of this group, which is more than in other sets that show systematic deviations from equilibrium, for example, *Streptococcus* and *Bordetella*. Hence, the number of substitutions should be sufficient

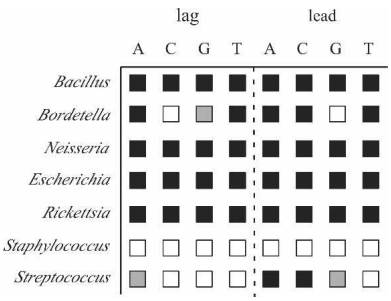

Figure 5. Results of the tests that composition at equilibrium is significantly different from the current composition. For each nucleotide the test was done on the lagging/leading strand. (Gray) $P < 0.05$; (black) $P < 0.01$; (white) NS.

to allow the detection of an important bias, and it is reasonable to assume that the *Staphylococcus* genome composition is very close to equilibrium.

Discussion

The availability of multiple complete genome sequences for single species or genera provides the means to statistically compare the patterns of polymorphism between closely related genomes, the large amounts of sequence data compensating for the rarity of nucleotide changes. However, such an approach requires special care. Sequences must have very low error rates, and orthology assignment must be conservative. The use of more than three genomes provides a more stringent approach to assigning directionality of changes by parsimony. The sequence data we have used here, in particular for the *Staphylococcus* and *Bacillus* genera, have passed extreme tests of accuracy (see Rocha et al. 2006 and references therein). We conservatively assigned orthology by using the information on reciprocal best hits, followed by two filters to minimize the problems of paralogy or xenology, one removing highly divergent genes and the second removing genes outside syntons. We also took care to minimize the effect of selection by using synonymous fourfold degenerate positions and by removing the most highly expressed genes to counter the effects of codon bias. The similarities of the rates of synonymous change on the leading and lagging strands show that changes are nearly neutral or at least that selection is not stronger in one strand than in the other. The major advantage of the method is that it is possible to directly examine substitutions creating asymmetrical compositional biases and to evaluate their significance by a nonparametric bootstrap procedure. As a result, we find that below an apparent uniformity of compositional biases, there are very different mutational asymmetries.

If compositional strand bias were caused solely by the inherent chemical lability of ssDNA and the extended ssDNA state of one strand relative to the other as in the cytosine deamination hypothesis, then we should have found the same substitution asymmetries in the different sets of genomes. Since we did not, one must question if one single cause contributes to the majority of the effect. In fact, some careful calculations suggest that the ssDNA differential exposure may not suffice to explain such an extensive compositional bias. Okazaki fragments in *E. coli* are ~1 kb long (Kitani et al. 1985), and the fork advances in the chromosome at ~1 kb/sec (Bipatnath et al. 1998). Hence, the template to the lagging strand is left half a second more time in the ssDNA state per replication round. For *E. coli*, which has ~200 generations per year, the template to the lagging strand is thus left 100 sec in the ssDNA state per year and 3×10^7 sec in the dsDNA state. The effect could be even smaller in bacteria having fewer generations per year, that is, slow-growers. Incidentally, the latter include the genomes with the highest compositional strand bias, for example, *Borrelia*, *Chlamydia*, and *Buchnera*. Even accepting that cytosine deaminates in ssDNA 140 times faster than in dsDNA, and speculating that U is inefficiently repaired, this seems a very small mutational cause for such a large compositional effect.

Many hypotheses have been put forward to explain replication-associated compositional strand bias. They have been extensively reviewed (Francino and Ochman 1997; Frank and Lobry 1999; Karlin 1999; Rocha 2004b) and are summarized in Table 3. All these hypotheses have the potential to explain part of the available data, but none seems entirely satisfactory. In the light of our results, the simplest explanation is that the bias is multifactorial. One should note that the most frequently cited reason for compositional strand bias, cytosine deamination in ssDNA, could explain a large fraction of strand bias in four out of seven genomes if it accounts for all or a large fraction of C→T substitution asymmetries. Yet, it totally fails to explain the bias in the other three genomes, and this is most significant because two of them are the most biased. Our results suggest that in the latter, G enrichment on the leading strand may predominantly originate from A→G bias rather than C→T bias. Among many other possible hypotheses, this could indicate more frequent deamination of A, not C, in these genomes. The seemingly inevitable conclusion is that an apparently homogeneous compositional bias (GC skew), grounded on a fundamental and highly conserved cellular process (replication), can still have a multifactorial origin in which each factor has a very different relevance in different genomes. A puzzling remaining question is then, why do all these different biases lead to higher GC skew in the leading than in the lagging strand in so many diverse genomes?

We found that most genomes are compositionally away from equilibrium. It has been suggested that this is the case in some regions of the human genome (Lander et al. 2001), particularly in the G+C-rich isochores (Duret et al. 2002). Our data suggest that this may be a general property of genomic sequences. There are two different ways to interpret such a deviation from equilibrium, one based on selection of compositional strand bias and the other on shifting mutational spectra. Selection for nucleotide composition has been proposed in a variety of cases: varying availability of nucleotides in different ecological niches (Rocha and Danchin 2002; Foerstner et al. 2005) and differences in metabolism (Naya et al. 2002; Rocha and Danchin 2002) and temperature (Musto et al. 2004). If G was more adaptive than C in the leading strand (or the reverse on the lagging strand), this would have the advantage of explaining why GC skews are always of the same type, independently of the substitution spectra. Yet, for selection to modulate GC skews and this be revealed in the shift of genome composition from equilibrium, this should involve a biased selection of polymorphisms. The latter necessitates an unlikely large selection coefficient associated with compositional strand bias. This would also require selection for GC skew in some genomes, the ones where the composition is more skewed than expected given the mutation spectra, and against GC skews in the others. Finally, there are other difficulties with this hypothesis because of (1) all the checks we made to remove the effect of selection; (2) the lack of a theory substantiating selection on compositional strand bias; and (3) the results indicating that substitution types causing a qualitatively similar bias are so variable. In the more orthodox neutralist perspective, our results could be explained simply by extensive and frequent variation in rates of the different types of mutations. Such a variation is likely to occur by several mechanisms. Horizontal gene transfer or xenologous replacement of genes related with replication, repair, or simply elements interacting physically with DNA can shift the equilibrium between the different mutations, leading to compositional shifts. The frequent loss, lateral transfer, and recombination of repair genes have been well documented within strains of *E. coli* (Denamur et al. 2000). Shifts in ecological niches can also explain changes in the relative rates of each type of mutation, if they involve a change in the environment, for example, related with temperature, chemical composition, or even nucleotide availability (for the genomes not producing their own nucleotides). *Rickettsia* is an example of a taxon

Table 3. Hypotheses that have been put forward to explain compositional strand bias and some arguments for and against them

Cause	For	Against
Cytosine deamination	C→U is a mutational hotspot in ssDNA (Coulondre et al. 1978), which would lead to C impoverishment in the leading strand (Reyes et al. 1998; Frank and Lobry 1999).	GC skews are higher than expected given AT skews (Rocha and Danchin 2001). Some AT skews are positive (i.e., the leading strand is richer in A, not T) (McLean et al. 1998; Worning et al. 2006).
Mismatch repair	Gapped DNA is more prone to be corrected by mismatch repair, and lagging strand is more subject to it (Radman 1998).	Differential repair is not sufficient or necessary for the appearance of strand bias. There is no strong association between presence of MMR genes and the bias.
DNA polymerase processivity	Lagging strand DNAP could be more prone to dislocate from DNA if it went faster to compensate for the translocation between each Okazaki fragment (Fijalkowska et al. 1998). This could be mutagenic.	Recent data indicate that in T7 the two DNAP are equally processive, with the primase slowing the pace of leading strand synthesis (Lee et al. 2006).
Recombination repair	In γ-proteobacteria the higher biases are in genomes lacking elements (RecA, PriA) associated with the repair of stalled replication forks (Klasson and Andersson 2006).	Genomes without PriA and RecA, e.g., among Mollicutes (Rocha et al. 2005), lack compositional strand bias, and genomes with both proteins have strand bias. Hence, the effect is not sufficient.
Genome rearrangements	Gene switching between replicating strands decreases the bias (Tillier and Collins 2000b; Rocha and Danchin 2001; Szczepanik et al. 2001).	It does not explain how the bias comes about in the first place.
Length of Okazaki fragments	ssDNA is chemically more labile and its differential exposure is proportional to the size of the Okazaki fragments (Rocha 2004b).	It is impossible to test with the available data, but the effects of ssDNA exposure may be too small to be sufficient (see text).
Signals	Many oligonucleotides (Salzberg et al. 1998), among which are recombination signals (e.g., chi sequences) (El Karoui et al. 1999), and segregation signals (e.g., KOPS) (Bigot et al. 2005; Levy et al. 2005) are overrepresented in the leading strand.	The signals are compositionally biased as the leading strand (i.e., they are G rich), but they are not nearly enough to explain the extent of the bias (Frank and Lobry 1999; Tillier and Collins 2000a).
DNA polymerase composition	Genomes with both PolC and DnaE (α-subunits of DNA polymerases) are enriched in A in the leading strand, whereas the others are enriched in T (Worning et al. 2006).	The effect is not valid for the opposition between G and C within replicating strands (Rocha 2002), i.e., it does not explain GC skews.
Frameshift mutagenesis	+1 G frameshifts occur fivefold more frequently during synthesis of series of Gs in the leading strand, which may favor runs of Gs in the leading strand (Gawel et al. 2002).	Long runs of Gs are rare in most bacterial genomes (Rocha 2003), which suggests that this cannot explain the extent of the bias.

recently adapted to an intracellular lifestyle, and has undergone extensive pseudogenization. Infection of human cells by *Rickettsia* is associated with oxidative damage (Santucci et al. 1992), which, through 8-hydroxyguanine lesions, induces G→T mutations. This may help account for the unusually strong C→A/G→T biases as well as the large gap between the genome composition and the one expected at equilibrium. Finally, even if a genome is not acquiring or losing genes and if the environment is stable, changes in the mutation spectra may occur by the evolution of proteins leading to the increase in some types of mutations. These may eventually be compensated by the decrease in other types of mutations, and thus be neutrally fixed. Further work will be necessary to better understand the evolution of mutation spectra through time, although some of our preliminary observations suggest little variation within the species domain. This is consistent with mutation spectra fluctuating around an average behavior within closely related genomes, but also with this average behavior shifting apart with time within lineages because of changes in the repair and replication machinery, but possibly also metabolic changes and nucleotide availability. Still, one is left with the puzzling observation that qualitatively similar GC skews result from different mutation spectra. In any case, these results clearly show the importance of considering heterotachy when analyzing sequence evolution (Lopez et al. 2002).

One must be extremely careful when incriminating specific mutation types to the compositional deviation of DNA sequences from an average value. For every type of deviation there are multiple repair genes whose presence or absence could potentially explain the effect. However, if one does not know which are the asymmetric mutations, one can totally fail to pinpoint the relevant gene(s) and understand its fundamental cause. Furthermore, our data suggest that most frequently there may not even be such a single gene or level of analysis, as slight changes in the biochemical characteristics of the proteins involved in cellular processes affecting mutation types could suffice to change sequence composition dramatically. Challenging Ockham's razor, even simple, nearly ubiquitous compositional biases, caused by the essential and highly conserved process of replication, can be underlined by a large complexity of biological phenomena.

Methods

Data

We used seven groups of complete genomes of bacterial strains or species (Table 1; Supplemental Table 1). These include six strains of *E. coli* or *Shigella*, six strains of *Streptococcus pyogenes*, five strains of the *B. cereus* group including *B. anthracis* and *Bacillus thurigiensis*, four species of *Rickettsia*, four strains of *Staphylococcus aureus*, three species of *Bordetella*, and five strains of *Neisseria*, one *Neisseria gonorrhoeae* and four *Neisseria meningitidis*. The complete list of strains, accession numbers, and the phylogenetic trees for each group are presented as Supplemental material. One should note that the separation between named strains and named species in bacteria is highly controversial (Gevers et al. 2005). Hence, some of these genomes (e.g., *Escherichia* and *Shigella* or the *Bor-*

detella) are classed as representing different species or genera but, in fact, correspond to highly similar core genomes.

Definition of orthology

A preliminary set of orthologs was defined by identifying unique pairwise reciprocal best hits, with at least 40% similarity in protein sequence and <20% difference in length. This list was then refined by combining the information on the distribution of similarity of these putative orthologs and the data on gene order conservation (as in Rocha et al. 2006). Because few rearrangements are observed at these short evolutionary distances, genes outside conserved blocks of synteny are likely to be xenologs or paralogs. Hence, we conservatively used the distribution of sequence similarity within reciprocal best hits, together with the classification of these genes as either syntenic or nonsyntenic, to set appropriate lower thresholds of protein sequence similarity between orthologs within each group: *Bacillus* (90%, mean >99%), *E. coli* (90%, mean >99%), *Streptococcus* (95%, mean >99%), *Staphylococcus* (95%, mean >99%), *Bordetella* (90%, mean >99%), *Rickettsia* (80%, mean >93%), *Neisseria* (90%, mean >97%). The definitive list of orthologs for each group was defined as the intersection of pairwise lists.

Alignments and inference of substitution tables

The protein sequences of the orthologs were aligned using CLUSTALW (Thomson et al. 1994) and back-translated into DNA sequences. This produced multiple alignments with a very large number of positions, which compensates for the relatively low density of substitutions (Table 1). For each set of multiple alignments, we counted the number of each type of directed change at third codon positions corresponding to amino acids fourfold degenerated (quartets). When we could reliably identify an outgroup in the taxa, we built an outgroup mutation table (Fig. 1). This is the sum of all substitutions of each type observed within the ingroups. A substitution is only counted if it occurs in one single terminal branch of the tree and if all the other elements strictly respect the consensus (defined by the outgroup). For *Streptococcus* and *Rickettsia*, there was no single outgroup, and in these cases we compute an ingroup mutation table (Fig. 1). In this table, we compute all types of substitution that take place in one single terminal branch of the tree and sum the occurrence of each mutation type across all terminal branches. An individual mutation is included only if there is a consensus in all the other genomes.

Determination of substitution frequencies and statistical tests

The absolute values of substitutions from i to j (e.g., A→T) at fourfold degenerate positions were converted to relative substitution frequencies f_{ij} by dividing them by the average number of nucleotides i in all the sequences for a given taxon (i.e., A in the precedent example) (Gojobori et al. 1982). Since we only analyze the fourfold degenerate codon positions, we normalized according to the frequencies of nucleotides at these positions. This allows comparing directly the frequencies between different types of substitutions in a data set. Because we are interested in computing the asymmetries in the frequencies of complementary changes, we cumulated the biases between the leading and the lagging strands. For example, the relative frequency of C→T changes was computed taking into account C→T changes in leading strand genes and G→A changes in lagging strand genes. Similarly, the relative frequency of G→A changes accounts for G→A changes in the leading strand genes and C→T changes in the lagging strand genes. To check that this is not in any way biasing our analysis, we also did the analysis of leading and lagging strands separately, which revealed concordant results in the vast majority of cases. The assessment of significant asymmetry was done by a nonparametric bootstrap procedure in which we sample 1000 times each set of multiple alignments and compute at each time the normalized relative frequencies. We consider that f(X→Y) is significantly more (less) frequent than the complementary change at a given P-value if no more than $P/2$ ($1 - P/2$) percent of the pairwise comparisons of the bootstrapped relative frequencies show lower (higher) f(X→Y) in the leading strand. For the analysis separating leading and lagging strand genes, the same bootstrap analysis is done between X→Y in the genes in one strand and the same change in the genes of the other strand. Hence, for the previous example, in the first analysis we test by bootstrap if

$$H0{:}[f(C \rightarrow T_{leading}) + f(G \rightarrow A_{lagging})] - [f(G \rightarrow A_{leading}) + f(C \rightarrow T_{lagging})] = 0$$

whereas in the second we test separately if

$$H0{:}f(C \rightarrow T_{leading}) - f(C \rightarrow T_{lagging}) = 0$$

and

$$H0{:}f(G \rightarrow A_{leading}) - f(G \rightarrow A_{lagging}) = 0$$

GC and AT skews

We identified the origin and terminus of replication in each genome using cumulative GC skews and AT skews analysis in 10-kb sliding windows (Grigoriev 1998), where

$$\text{GCskew} = [\text{G} - \text{C}]/[\text{G}+\text{C}]$$

$$\text{ATskew} = [\text{A} - \text{T}]/[\text{T}+\text{A}]$$

We also used the positioning of genes that tend to be close to the origin of replication such as *dnaA* (Mackiewicz et al. 2004) to define the origin. We then classed all genes according to their presence on the leading or lagging strand in the respective genome and only kept the ones always present on the same replicating strand in all genomes of the taxa. The shift of genes from one replicating strand to the other is a rare event and very unlikely to be a source of error as this analysis encompasses very short time scales. To support this further, no significant strand shifts were observed in any of the genomes of *Bacillus*, *Staphylococcus*, or *Escherichia*, with the exception of *Shigella flexneri*. Hence all genes consistently present in the same strand were accepted. The level of compositional asymmetry between strands was computed using the composition of the third codon position of fourfold degenerate codons (q) in genes of the leading and the lagging strand and following (Lobry and Sueoka 2002).

$$B_I = \sqrt{\left(\frac{G_{q,lead}}{G_{q,lead}+C_{q,lead}} - \frac{G_{q,lag}}{G_{q,lag}+C_{q,lag}}\right)^2 + \left(\frac{T_{q,lead}}{T_{q,lead}+A_{q,lead}} - \frac{T_{q,lag}}{T_{q,lag}+A_{q,lag}}\right)^2}$$

Higher B_I values indicate higher bias, although not necessarily G and T richness in the leading strand. To account for this, we also computed the difference in GC and AT skews for genes in the leading and lagging strand.

$$\Delta GCskew = GCskew_{q,leading} - GCskew_{q,lagging}$$

$$\Delta ATskew = ATskew_{q,leading} - ATskew_{q,lagging}$$

Removing highly expressed genes

Although we restrict our analysis to fourfold degenerate sites, this does not guarantee that such substitutions are nearly neutral or that they are immune to other mutational biases. Highly expressed genes are much more conserved than other genes (Sharp 1991; Rocha and Danchin 2004; Drummond et al. 2005). This includes nonsynonymous but also synonymous substitutions, because codon usage is under strong selection in highly expressed genes (Grantham et al. 1981), especially among fast-growing bacteria (Rocha 2004a). If such genes were not removed from the analysis, then the identified substitutions might not reflect the mutation spectrum associated to replication. Highly expressed genes may also suffer mutational biases because of transcription-coupled repair (Francino et al. 1996; Lopez and Philippe 2001; Hudson et al. 2003). The problem of highly expressed genes is especially important for the analysis of compositional strand bias since highly expressed genes are more likely to be essential and essential genes accumulate in the leading strand (Rocha and Danchin 2003). Hence, we used the CAI index (Sharp and Li 1987) to remove the top 10% genes with most biased codon usage, expected to be the most highly expressed. We checked that removing even more genes (15%) from the genome with the highest codon usage bias (*E. coli*) led to the same results (Supplemental Fig. 1). In fact, one should note that this is rather conservative, as slow-growing bacteria such as *Rickettsia* are likely to have little, if any, codon usage bias.

Acknowledgments

We thank the Sanger Centre and the Institut Pasteur for kindly sharing sequence data before publication. The sequence data of *E. coli* O42 and *N. meningitidis* C ET-37 were produced by the Pathogen Sequencing Unit at the Sanger Institute and can be obtained from http://www.sanger.ac.uk/Projects/Pathogens. The data for *N. meningitidis* C 8013 were provided by the unit "Génomique des microorganismes pathogènes" from the Institut Pasteur. M.T. is funded by the Conseil Régional de l'Ile de France.

References

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

Bentley, S.D., Chater, K.F., Cerdeno-Tarraga, A.M., Challis, G.L., Thomson, N.R., James, K.D., Harris, D.E., Quail, M.A., Kieser, H., Harper, D., et al. 2002. Complete genome sequence of the model actinomycete *Streptomyces coelicolor* A3(2). *Nature* **417:** 141–147.

Bielawski, J.P. and Gold, J.R. 2002. Mutation patterns of mitochondrial H- and L-strand DNA in closely related Cyprinid fishes. *Genetics* **161:** 1589–1597.

Bigot, S., Saleh, O.A., Lesterlin, C., Pages, C., El Karoui, M., Dennis, C., Grigoriev, M., Allemand, J.F., Barre, F.X., and Cornet, F. 2005. KOPS: DNA motifs that control *E. coli* chromosome segregation by orienting the FtsK translocase. *EMBO J.* **24:** 3770–3780.

Bipatnath, M., Dennis, P.P., and Bremer, H. 1998. Initiation and velocity of chromosome replication in *Escherichia coli* B/r and K-12. *J. Bacteriol.* **180:** 265–273.

Chargaff, E. 1950. Chemical specificity of nucleic acids and mechanism of their enzymatic degradation. *Experientia* **6:** 201–240.

Coulondre, C., Miller, J.H., Farabaugh, P.J., and Gilbert, W. 1978. Molecular basis of base substitution hotspots in *Escherichia coli*. *Nature* **274:** 775–780.

Denamur, E., Lecointre, G., Darlu, P., Tenaillon, O., Acquaviva, C., Sayada, C., Sunjevaric, I., Rothstein, R., Elion, J., Taddei, F., et al. 2000. Evolutionary implications of the frequent horizontal transfer of mismatch repair genes. *Cell* **103:** 711–721.

Drummond, D.A., Bloom, J.D., Adami, C., Wilke, C.O., and Arnold, F.H. 2005. Why highly expressed proteins evolve slowly. *Proc. Natl. Acad. Sci.* **102:** 14338–14343.

Duret, L., Semon, M., Piganeau, G., Mouchiroud, D., and Galtier, N. 2002. Vanishing GC-rich isochores in mammalian genomes. *Genetics* **162:** 1837–1847.

El Karoui, M., Biaudet, V., Schbath, S., and Gruss, A. 1999. Characteristics of χ distribution on different bacterial genomes. *Res. Microbiol.* **150:** 579–587.

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

Fijalkowska, I.J., Jonczyk, P., Tkaczyk, M.M., Bialokorska, M., and Schaaper, R.M. 1998. Unequal fidelity of leading strand and lagging strand DNA replication on the *Escherichia coli* genome. *Proc. Natl. Acad. Sci.* **95:** 10020–10025.

Foerstner, K.U., von Mering, C., Hooper, S.D., and Bork, P. 2005. Environments shape the nucleotide composition of genomes. *EMBO Rep.* **6:** 1208–1213.

Francino, M.P. and Ochman, H. 1997. Strand asymmetries in DNA evolution. *Trends Genet.* **13:** 240–245.

Francino, M.P., Chao, L., Riley, M.A., and Ochman, H. 1996. Asymmetries generated by transcription-coupled repair in enterobacterial genes. *Science* **272:** 107–109.

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

Friedberg, E.C., Walker, G.C., and Siede, W. 1995. *DNA repair and mutagenesis*. ASM Press, Washington, DC.

Gawel, D., Jonczyk, P., Bialoskorska, M., Schaaper, R.M., and Fijalkowska, I.J. 2002. Asymmetry of frameshift mutagenesis during leading and lagging-strand replication in *Escherichia coli*. *Mutat. Res.* **501:** 129–136.

Gevers, D., Cohan, F.M., Lawrence, J.G., Spratt, B.G., Coenye, T., Feil, E.J., Stackebrandt, E., Van de Peer, Y., Vandamme, P., Thompson, F.L., et al. 2005. Re-evaluating prokaryotic species. *Nat. Rev. Microbiol.* **3:** 733–739.

Gojobori, T., Li, W.H., and Graur, D. 1982. Patterns of nucleotide substitution in pseudogenes and functional genes. *J. Mol. Evol.* **18:** 360–369.

Grantham, R., Gautier, C., Gouy, M., Jacobzone, M., and Mercier, R. 1981. Codon catalog usage is a genome strategy modulated for gene expressivity. *Nucleic Acids Res.* **9:** r43–r74.

Grigoriev, A. 1998. Analyzing genomes with cumulative skew diagrams. *Nucleic Acids Res.* **26:** 2286–2290.

Grigoriev, A. 1999. Strand-specific compositional asymmetries in double-stranded DNA viruses. *Virus Res.* **60:** 1–19.

Hudson, R.E., Bergthorsson, U., and Ochman, H. 2003. Transcription increases multiple spontaneous point mutations in *Salmonella enterica*. *Nucleic Acids Res.* **31:** 4517–4522.

Karlin, S. 1999. Bacterial DNA strand compositional asymmetry. *Trends Microbiol.* **7:** 305–308.

Kitani, T., Yoda, K., Ogawa, T., and Okazaki, T. 1985. Evidence that discontinuous DNA replication in *Escherichia coli* is primed by approximately 10 to 12 residues of RNA starting with a purine. *J. Mol. Biol.* **184:** 45–52.

Klasson, L. and Andersson, S.G. 2006. Strong asymmetric mutation bias in endosymbiont genomes coincide with loss of genes for replication restart pathways. *Mol. Biol. Evol.* **23:** 1031–1039.

Lafay, B., Lloyd, A.T., McLean, M.J., Devine, K.M., Sharp, P.M., and Wolfe, K.H. 1999. Proteome composition and codon usage in spirochaetes: Species-specific and DNA strand-specific mutational biases. *Nucleic Acids Res.* **27:** 1642–1649.

Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., et al. 2001. Initial sequencing and analysis of the human genome. *Nature* **409:** 860–921.

Lee, J.B., Hite, R.K., Hamdan, S.M., Xie, X.S., Richardson, C.C., and van Oijen, A.M. 2006. DNA primase acts as a molecular brake in DNA replication. *Nature* **439:** 621–624.

Levy, O., Ptacin, J.L., Pease, P.J., Gore, J., Eisen, M.B., Bustamante, C., and Cozzarelli, N.R. 2005. Identification of oligonucleotide sequences that direct the movement of the *Escherichia coli* FtsK translocase. *Proc. Natl. Acad. Sci.* **102:** 17618–17623.

Lobry, J.R. 1996. Asymmetric substitution patterns in the two DNA strands of bacteria. *Mol. Biol. Evol.* **13:** 660–665.

Lobry, J. and Sueoka, N. 2002. Asymmetric directional mutation pressures in bacteria. *Genome Biol.* **3:** research0058.

Lopez, P. and Philippe, H. 2001. Composition strand asymmetries in prokaryotic genomes: Mutational bias and biased gene orientation. *C. R. Acad. Sci. III* **324:** 201–208.

Lopez, P., Philippe, H., Myllykallio, H., and Forterre, P. 1999. Identification of putative chromosomal origins of replication in Archaea. *Mol. Microbiol.* **32:** 883–886.

Lopez, P., Casane, D., and Philippe, H. 2002. Heterotachy, an important process of protein evolution. *Mol. Biol. Evol.* **19:** 1–7.

Mackiewicz, P., Gierlik, A., Kowalczuk, M., Dudek, M.R., and Cebrat, S.

1999. How does replication-associated mutational pressure influence amino acid composition of proteins? *Genome Res.* **9:** 409–416.
Mackiewicz, P., Mackiewicz, D., Kowalczuk, M., Dudkiewicz, M., Dudek, M.R., and Cebrat, S. 2003. High divergence rate of sequences located on different DNA strands in closely related bacterial genomes. *J. Appl. Genet.* **44:** 561–584.
Mackiewicz, P., Zakrzewska-Czerwinska, J., Zawilak, A., Dudek, M.R., and Cebrat, S. 2004. Where does bacterial replication start? Rules for predicting the oriC region. *Nucleic Acids Res.* **32:** 3781–3791.
McInerney, J.O. 1998. Replicational and transcriptional selection on codon usage in *Borrelia burgdorferi*. *Proc. Natl. Acad. Sci.* **95:** 10698–10703.
McLean, M.J., Wolfe, K.H., and Devine, K.M. 1998. Base composition skews, replication orientation and gene orientation in 12 prokaryote genomes. *J. Mol. Evol.* **47:** 691–696.
Mrázek, J. and Karlin, S. 1998. Strand compositional asymmetry in bacterial and large viral genomes. *Proc. Natl. Acad. Sci.* **95:** 3720–3725.
Musto, H., Naya, H., Zavala, A., Romero, H., Alvarez-Valin, F., and Bernardi, G. 2004. Correlations between genomic GC levels and optimal growth temperatures in prokaryotes. *FEBS Lett.* **573:** 73–77.
Muto, A. and Osawa, S. 1987. The guanine and cytosine content of genomic DNA and bacterial evolution. *Proc. Natl. Acad. Sci.* **84:** 166–169.
Naya, H., Romero, H., Zavala, A., Alvarez, B., and Musto, H. 2002. Aerobiosis increases the genomic guanine plus cytosine content (GC%) in prokaryotes. *J. Mol. Evol.* **55:** 260–264.
Pesole, G., Gissi, C., De Chirico, A., and Saccone, C. 1999. Nucleotide substitution rate of mammalian mitochondrial genomes. *J. Mol. Evol.* **48:** 427–434.
Radman, M. 1998. DNA replication: One strand may be more equal. *Proc. Natl. Acad. Sci.* **95:** 9718–9719.
Reyes, A., Gissi, C., Pesole, G., and Saccone, C. 1998. Asymmetrical directional mutation pressure in the mitochondrial genome of mammals. *Mol. Biol. Evol.* **15:** 957–966.
Rocha, E.P.C. 2002. Is there a role for replication fork asymmetry in the distribution of genes in bacterial genomes? *Trends Microbiol.* **10:** 393–396.
Rocha, E.P.C. 2003. An appraisal of the potential for illegitimate recombination in bacterial genomes and its consequences: From duplications to genome reduction. *Genome Res.* **13:** 1123–1132.
Rocha, E.P. 2004a. Codon usage bias from tRNA's point of view: Redundancy, specialization, and efficient decoding for translation optimization. *Genome Res.* **14:** 2279–2286.
Rocha, E.P.C. 2004b. The replication-related organisation of the bacterial chromosome. *Microbiology* **150:** 1609–1627.
Rocha, E.P.C. and Danchin, A. 2001. Ongoing evolution of strand composition in bacterial genomes. *Mol. Biol. Evol.* **18:** 1789–1799.
Rocha, E.P.C. and Danchin, A. 2002. Competition for scarce resources might bias bacterial genome composition. *Trends Genet.* **18:** 291–294.
Rocha, E.P.C. and Danchin, A. 2003. Essentiality, not expressiveness, drives gene strand bias in bacteria. *Nat. Genet.* **34:** 377–378.
Rocha, E.P.C. and Danchin, A. 2004. An analysis of determinants of protein substitution rates in Bacteria. *Mol. Biol. Evol.* **21:** 108–116.
Rocha, E.P.C., Danchin, A., and Viari, A. 1999. Universal replication bias in bacteria. *Mol. Microbiol.* **32:** 11–16.
Rocha, E.P.C., Cornet, E., and Michel, B. 2005. Comparative and evolutionary analysis of the bacterial homologous recombination systems. *PLoS Genet.* **1:** e15.
Rocha, E.P.C., Maynard Smith, J., Hurst, L.D., Holden, M.T., Cooper, J.E., Smith, N.H., and Feil, E. 2006. Comparisons of d_N/d_S are time-dependent for closely related bacterial genomes. *J. Theor. Biol.* **239:** 226–235.
Salzberg, S.L., Salzberg, A.J., Kerlavage, A.R., and Tomb, J.-F. 1998. Skewed oligomers and origins of replication. *Gene* **217:** 57–67.
Santucci, L.A., Gutierrez, P.L., and Silverman, D.J. 1992. *Rickettsia rickettsii* induces superoxide radical and superoxide dismutase in human endothelial cells. *Infect. Immun.* **60:** 5113–5118.
Sharp, P.M. 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.
Sharp, P.M. and Li, W.H. 1987. The codon Adaptation Index—A measure of directional synonymous codon usage bias, and its potential applications. *Nucleic Acids Res.* **15:** 1281–1295.
Sueoka, N. 1962. On the genetic basis of variation and heterogeneity of DNA base composition. *Proc. Natl. Acad. Sci.* **48:** 582–591.
Szczepanik, D., Mackiewicz, P., Kowalczuk, M., Gierlik, A., Nowicka, A., Dudek, M.R., and Cebrat, S. 2001. Evolution rates of genes on leading and lagging DNA strands. *J. Mol. Evol.* **52:** 426–433.
Thomson, J.D., Higgins, D.G., and Gibson, T.J. 1994. Clustal W: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, positions-specific gap penalties and weight matrix choice. *Nucleic Acids Res.* **22:** 4673–4680.
Tillier, E.R. and Collins, R.A. 2000a. The contributions of replication orientation, gene direction, and signal sequences to base-composition asymmetries in bacterial genomes. *J. Mol. Evol.* **50:** 249–257.
Tillier, E.R. and Collins, R.A. 2000b. Replication orientation affects the rate and direction of bacterial gene evolution. *J. Mol. Evol.* **51:** 459–463.
Touchon, M., Nicolay, S., Audit, B., Brodie Of Brodie, E.B., d'Aubenton-Carafa, Y., Arneodo, A., and Thermes, C. 2005. Replication-associated strand asymmetries in mammalian genomes: Toward detection of replication origins. *Proc. Natl. Acad. Sci.* **102:** 9836–9841.
Worning, P., Jensen, L.J., Hallin, P.F., Staerfeldt, H.H., and Ussery, D.W. 2006. Origin of replication in circular prokaryotic chromosomes. *Environ. Microbiol.* **8:** 353–361.
Yang, Z. and Nielsen, R. 2000. Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models. *Mol. Biol. Evol.* **17:** 32–43.

Received May 19, 2006; accepted in revised form August 30, 2006.