A Comparative Genomics Approach to DNA Asymmetry

M. PILAR FRANCINO[a] AND HOWARD OCHMAN[b,c]

[a]*Department of Biology, University of Rochester, Rochester, New York 14627, USA*

[b]*Department of Ecology and Evolutionary Biology, 233 Life Sciences South, University of Arizona, Tucson, Arizona 85721, USA*

With the availability of complete sequences for several bacterial genomes, it is now possible to uncover patterns of genome organization common to very divergent taxa. Most bacterial chromosomes are polarized around their origins of replication in terms of both base composition[1–4] and gene distribution.[3–8] For nine completely sequenced prokaryotic species, Freeman *et al.*[9] recently showed that features of base composition, such as purine excess, reflect the partitioning of genes between the complementary strands of DNA. Here, we apply a comparative approach across species to examine the relation between strand-biased gene distribution and the most common compositional feature of bacterial chromosomes, GC skew.[1]

GC skew refers to an observed excess of G over C in the leading strand of replication. Because we expect equal frequencies of G and C under no strand bias for either mutation or selection, the excess of G over C might indicate that different spectra of mutations occur during leading and lagging strand synthesis.[1–3] However, strand-biased gene distribution could also produce compositional asymmetry between the strands due to (1) selection on amino acid content and/or synonymous codon usage and (2) strand-specific mutational biases introduced by transcription (through transcription-coupled repair and the preferential deamination of the coding strand).[10,11]

We analyzed gene distribution and GC skew in the completely sequenced bacterial chromosomes for which the origin of replication has been located. Origins of replication have been determined experimentally for only two of the six species analyzed, *Escherichia coli* and *Bacillus subtilis*, whereas for *Haemophilus influenzae*, *Borrelia burgdorferi*, *Mycoplasma pneumoniae*, and *M. genitalium*, assignment of replication origins is based on the presence of conserved genes and/or nucleotide motifs and is confirmed by the associated switches in base composition.[1–9] For our analysis, the terminus was assigned halfway through the chromosome from the origin of replication, except for *E. coli*, where the known location of *terC* was used, and for the linear chromosome of *B. burgdorferi*, where replication was assumed to proceed from the center to the ends of the chromosome. In all six genomes, the majority of genes are encoded on the leading strand, presumably to avoid head-on collisions between replication and transcription complexes.[12,13] However, there is substantial variation in the degree of coding asymmetry between the strands, which is highest in gram-positives (*B. subtilis* and mycoplasmas), intermediate in the spirochete (*B. burgdorferi*), and lowest in proteobacteria (*E. coli* and *H. influenzae*; TABLE 1). We first asked if strand bias in overall gene distribution is related to GC skew across these bacterial

[c]To whom correspondence should be addressed. Phone, 520-626-8355; fax, 520-621-3709; e-mail, hochman@u.arizona.edu

TABLE 1. Strand Assignment of Genes and Codon Bias in Bacterial Genomes

Species	Proportion Coded on Leading Strand		X^2 Leading/X^2 Lagging[a]	Mean X^2[b]
	Total Genes	High Bias Genes		
Escherichia coli	0.55	0.60	1.06	0.73 ± 0.60
Haemophilus influenzae	0.55	0.67	1.25	0.78 ± 0.61
Borrelia burgdorferi	0.66	0.81	1.34	1.04 ± 0.60
Bacillus subtilis	0.74	0.74	0.98	0.56 ± 0.47
Mycoplasma pneumoniae	0.78	0.62	0.89	0.64 ± 0.34
Mycoplasma genitalium	0.80	0.67	0.90	1.01 ± 0.39

[a]Ratio of the mean X^2 for all genes on the leading strand over the mean X^2 for all genes on the lagging strand. Note that this ratio is >1.0 in genomes with moderate to intermediate levels of total coding asymmetry but <1.0 in other genomes.

[b]Mean X^2 and standard deviation for all genes in the genome.

genomes. FIGURE 1A shows that there is clearly no association between these two genomic properties.

However, genes can be biased between the strands not only in quantity, but also in quality. If selection acts to reduce conflict between transcription and replication, as suggested by the pervasive strand bias in total gene numbers, genes may well be partitioned between the strands according to their levels of expression. In unicellular organisms, levels of gene expression often correlate with degree of bias in synonymous codon usage.[14–17] This, in addition to the increased mutational biases due to frequent transcription, is likely to result in highly expressed genes having a more skewed base composition.[10] Therefore, we asked whether the distribution of genes with the highest codon biases could explain the observed levels of GC skew.

For all genes in each of the six species, we calculated the deviation from random usage of codons by means of a X^2 index. X^2 is computed as the average of the squared deviations of relative synonymous codon usage values (RSCU) from unity, weighted by the frequency of the encoded amino acid. Values of RSCU are obtained by dividing the observed frequency of each codon by the frequency expected under equal usage of synonymous codons. For each species, we defined high bias genes as those with X^2 values above 1.3, or 1 standard deviation above the mean X^2 across all genomes (0.74 ± 0.56), ensuring that the selected sets of genes have comparable levels of codon bias in all species. Over all genomes, the majority of highly biased (and, presumably, highly expressed) genes are encoded on the leading strand, and there is a strong correlation between the proportion of highly biased genes on this strand and the GC skew ($r = 0.916$, $p = 0.01$; FIG. 1B). The ranking of species according to their proportions of high bias genes on the leading strand is markedly different from that based on total genes. The reason is that high bias genes are overrepresented in the leading strands of species with moderate and intermediate levels of

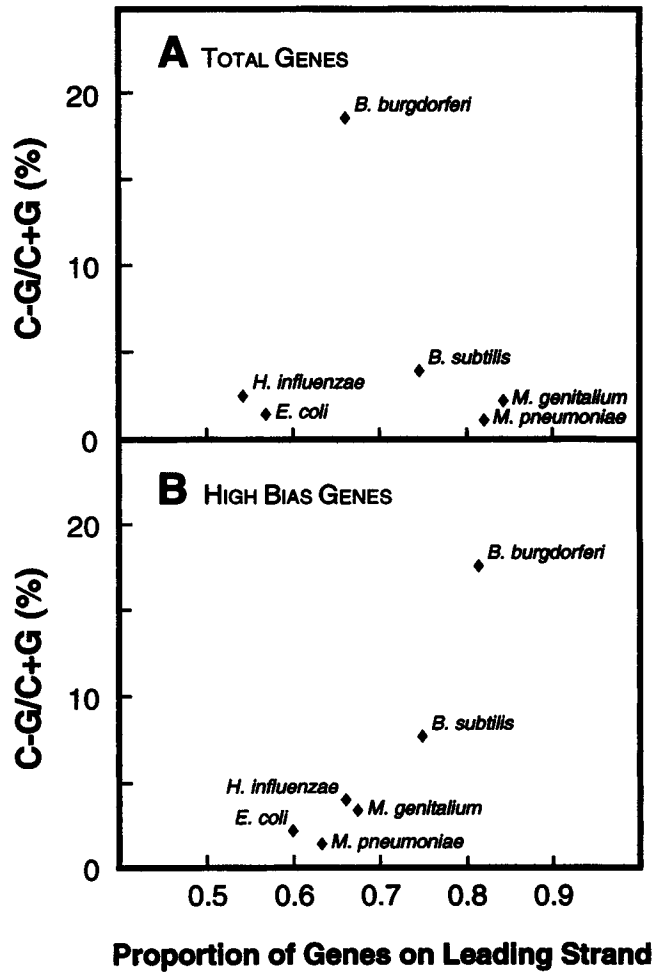

FIGURE 1. GC skew versus proportion of genes encoded on the leading strand for six bacterial genomes. C–G/C+G (%) indices were computed for each origin-to-terminus half of the chromosome. Plotted C–G/C+G percentages correspond to the counter-clockwise half of the chromosome and, therefore, to the lagging strand of replication, as manifest in their positive values; values for the other half of the chromosome are of the same magnitude, but of opposite sign. Leading strand genes were counted over the entire chromosome. (**A**) The lack of association between GC skew and the proportion of total genes on the leading strand. (**B**) The strong correlation between GC skew and the proportion of high bias genes ($X^2 > 1.3$) on the leading strand ($r = 0.916$, $p = 0.01$). Base composition, gene distribution, and codon bias were analyzed with the computer program DNA Master (DNA Master is available from Jeffrey G. Lawrence at the following site: http://cobamide2.bio.pitt.edu//computer.htm).

total coding asymmetry, but underrepresented in the leading strands of species with high levels of total coding asymmetry, relative to the expectations based on the distribution of the totality of genes (TABLE 1).

Hence, the degree of GC skew in a genome is related to the manner in which highly biased genes are partitioned between the strands, not to the overall distribution of genes. This presumably reflects a larger skew in base composition for this subset of genes, because of the limited set of codons that they use and the increased effects of mutational biases induced during frequent transcription.[10] Moreover, because GC skew is also apparent in transcribed but untranslated regions,[1] codon usage cannot be its only cause, emphasizing the role of transcription itself.

REFERENCES

1. LOBRY, J.R. 1996. Asymmetric substitution patterns in the two DNA strands of bacteria. Mol. Biol. Evol. **13:** 660–665.
2. LOBRY, J.R. 1996. Origin of replication of *Mycoplasma genitalium*. Science **272:** 745–746.
3. BLATTNER, F.R. *et al.* 1997. The complete nucleotide sequence of *Escherichia coli* K-12. Science **277:** 1453–1462.
4. FRASER, C.M. *et al.* 1997. Genomic sequence of a Lyme disease spirochaete, *Borrelia burgdorferi*. Nature **390:** 580–586.
5. FLEISCHMANN, R.D. *et al.* 1995. Whole-genome random sequencing and assembly of Haemophilus influenzae Rd. Science **269:** 496–512.
6. FRASER, C.M. *et al.* 1995. The minimal gene complement of *Mycoplasma genitalium*. Science **270:** 397–403.
7. HIMMELREICH, R. *et al.* 1996. Complete sequence analysis of the genome of the bacterium *Mycoplasma pneumoniae*. Nucleic Acids Res. **24:** 4420–4449.
8. KUNST, F. *et al.* 1997. The complete genome sequence of the Gram-positive bacterium *Bacillus subtilis*. Nature **390:** 249–256.
9. FREEMAN, J.M. *et al.* 1998. Patterns of genome organization in bacteria. Science **279:** 1827.
10. FRANCINO, M.P. & H. OCHMAN. 1997. Strand asymmetries in DNA evolution. Trends Genet. **13:** 240–245.
11. MRAZEK, J. & S. KARLIN. 1998. Strand compositional asymmetry in bacterial and large viral genomes. Proc. Natl. Acad. Sci. USA **95:** 3720–3725.
12. BREWER, B. 1988. When polymerases collide: Replication and the transcriptional organization of the *E. coli* chromosome. Cell **53:** 679–686.
13. FRENCH, S. 1992. Consequences of replication fork movement through transcription units *in vivo*. Science **258:** 1362–1365.
14. SHARP, P.M. & W.-H. LI. 1987. The codon adaptation index—a measure of directional synonymous codon usage bias, and its potential applications. Nucleic Acids Res. **15:** 1281–1284.
15. SHIELDS, D.C. & P.M. SHARP. 1987. Synonymous codon usage in *Bacillus subtilis* reflects both translational selection and mutational biases. Nucleic Acids Res. **15:** 8023–8040.
16. SHARP, P.M. & G. MATASSI. 1994. Codon usage and genome evolution. Curr. Opin. Genet. Dev. **4:** 851–860.
17. ANDERSSON, G.E. & P.M. SHARP. 1996. Codon usage in the *Mycobacterium tuberculosis* complex. Microbiology **142:** 915–925.