Gene Location and Bacterial Sequence Divergence

Alex Mira[*1] *and Howard Ochman*[*†]

[*]Department of Ecology and Evolutionary Biology, University of Arizona; [†]Department of Biochemistry and Molecular Biophysics, University of Arizona

Previous comparison of a relatively small set of homologous genes from *Escherichia coli* and *Salmonella typhimurium* revealed that genes nearer to the origin of replication had substitution rates lower than genes closer to the replication terminus. The recently completed sequences of numerous bacterial genomes have allowed us to test whether this effect of distance from the replication origin on substitution rates, as observed for the *E. coli–S. typhimurium* comparison, is a general feature of bacterial genomes. Extending the analysis to all 3,000 *E. coli–S. typhimurium* homologs confirmed the significant association between chromosomal position and synonymous site divergence. However, the effect, though still significant, is not as dramatic as originally thought. A similar association between relative chromosomal location and synonymous substitution rate was detected in the majority of other bacterial species comparisons within α- and γ- Proteobacteria, and Firmicutes but was absent in Chlamydiales. The opposite trend, i.e., a *decrease* in synonymous divergence with distance from the replication origin, was detected in Mycobacteria. Analysis of the patterns of nucleotide substitutions revealed that the distance effect is not affected by gene orientation and is mainly caused by an increase in rates of transversions, suggesting that this effect may not be caused by recombinational repair or biased gene conversion, as originally suggested.

Introduction

Several factors, including codon usage bias, gene expression level, strand location and orientation, protein hydrophobicity and mutational bias, affect substitution rates at synonymous sites (Ikemura 1981; Sharp and Li 1987; Lobry 1996; de Miranda et al. 2000; Moran and Wernegreen 2000; Francino and Ochman 2001). In addition to these factors, Sharp et al. (1989) detected a tendency for genes located near the replication origin to undergo lower rates of synonymous substitutions than genes situated closer to the terminus. After accounting for differences in codon bias, genes near the replication origin were estimated to have a substitution rate about half that of genes closer to the terminus (Sharp et al. 1989). These results were based on the analysis of the relatively few ($n = 67$) pairs of homologous gene sequences that were then available for the enteric bacteria *Escherichia coli* and *Salmonella typhimurium*, and the significant association was largely attributable to the limited set of low-divergence genes near the replication origin.

Recent completion of *E. coli* and *Salmonella* genome sequencing (Blattner et al. 1997; McClelland et al. 2001; Parkhill et al. 2001; Perna et al. 2001) allows a reexamination of this distance effect on synonymous substitution rates based on the entire complement of homologous genes in these organisms. In addition, the full sequences of many bacterial genomes have been completed, including many pairs of closely related species, allowing a test of whether the effect of chromosome position on sequence divergence is a general feature of bacterial genomes. Because substitution rates at synonymous sites are less influenced by selection than at nonsynonymous positions, changes at these sites can provide substantial information about the underlying rates of mutations. We have examined the relationship between synonymous substitution rates and chromosomal position in 14 bacterial species pairs, each sharing a large number of gene homologs, in order to study differences in sequence divergence along the chromosome.

The distance effect could be attributable to increased mutation rates or decreased repair capabilities because genes are situated further from the replication origin. Although the molecular basis of these differences in mutation rates has not been addressed experimentally, it was originally hypothesized to be the outcome of more frequent recombinational repair or biased gene conversion (Sharp et al. 1989; Sharp 1991; Birky and Walsh 1992), which might arise from higher gene dosage near the origin, as achieved by multiple replication forks. Because the growth conditions and the number of coincident replication forks per cell are variable among species, the strength of the distance effect in different taxa could lend support to this explanation. In addition, we have determined the patterns of individual substitutions at synonymous positions in order to elucidate the potential causes of differences in substitution rates at different positions of the chromosome.

Materials and Methods

Selection of Species

Pairs of bacteria were selected based on the following criteria: (1) a majority of genes had an unambiguous homolog in the two genomes and (2) average synonymous substitution rates were not in saturation. Fully annotated genome sequences for the species *E. coli* (strains MG1655 and O157:H7), *Rickettsia prowazekii*, *R. conorii*, *Mycobacterium leprae*, *M. tuberculosis* H37Rv, *Pseudomonas aeruginosa*, *Chlamydia muridarum*, *C. trachomatis*, *Salmonella enterica* serovar Typhi, *S. en-*

[1] Present address: Department of Molecular Evolution, Evolutionary Biology Centre, Uppsala University, Norbyvägen 18C, SE-752 36 Uppsala, Sweden.

Key words: substitution rates, replication origin, genome evolution, *Escherichia coli*.

Address for correspondence and reprints: Howard Ochman, Department of Biochemistry and Molecular Biophysics, University of Arizona, P.O. Box 210088, Tucson, Arizona 85721.
E-mail: hochman@email.arizona.edu.

Mol. Biol. Evol. 19(8):1350–1358. 2002
© 2002 by the Society for Molecular Biology and Evolution. ISSN: 0737-4038

terica serovar Typhimurium, *Listeria innocua*, and *L. monocytogenes*, were obtained from NCBI (http://www.ncbi.nlm.nih.gov/). Complete but unannotated genome sequences were obtained from the Unfinished Genomic Sequenced Data section at the TIGR website (http://www.tigr.org/) for *Pseudomonas putida* KT2400, and from the Gonococcal Genome Sequencing Project at the University of Oklahoma for *Neisseria gonorrhoeae* (http://www.genome.ou.edu/gono.html).

Determining Positions of Replication Origin and Terminus

The location of the origin of replication in each genome is based on information present in the databases as derived by experimental evidence (Weigel et al. 1997; Barekzi et al. 2001), the presence of the *dnaA* box sequence (Salazar et al. 1996; Gasc et al. 1998), shifts in G+C-skew at third codon positions (Lobry 1996; Read et al. 2000; Kuroda et al. 2001; Ogata et al. 2001) and shifts in skewed oligonucleotides (or both) (Salzberg et al. 1998). The replication terminus for each genome was located at the position most distant from the origin and usually coincided with a second shift in G+C-skew (Lobry 1996; Salzberg et al. 1998). Gene positions were estimated as the distance from the origin of replication to the start of the open reading frame, regardless of coding strand. Homologous genes were excluded from the analysis if they occupied genome positions that differed by more than one tenth the length of the genome in either species. This elimination of homologs was especially relevant to the analyses of the Pseudomonads, which have undergone high levels of genome rearrangements, and to the analysis of the two Mycobacteria, which differ by one megabase in genome size (Cole et al. 2001).

Sequence Analysis and Codon Usage

The genes of a reference species were searched for sequence homology using BLAST similarity searches (Altschul et al. 1997) against the full sequence of a subject genome. A gene in the subject species was considered a homolog when it shared at least 60% sequence identity over at least 80% of the length of the reference gene. Genes shorter than 150 bp were excluded from the analysis, and those genes with different orientations in the two species from each compared pair were also eliminated. Homologous sequences were aligned using the Gap command of the GCG package (Devereux, Haeberli, and Smithies 1984). Divergence rates at silent sites (K_s) were obtained through Diverge in GCG, which applies the algorithm by Li (1993) and Pamilo and Bianchi (1993). The accuracy of using this measure of K_s for estimating synonymous divergence has been validated in the pair *E. coli-S. typhimurium* (Smith and Eyre-Walker 2001), but the assumptions in the use of this distance statistic might be violated in some genomes with extreme G + C compositions. Substitutions were identified as one of six types: A ↔ T, A ↔ G, A ↔ C, C ↔ G, C ↔ T, and C ↔ A. Because the ancestral state of sequences is unknown in pairwise comparisons, directionality of nucleotide substitutions was not determined. Codon usage bias was estimated by the χ^2 measure (Shields et al. 1998) using the publicly available DNA Master program from J. G. Lawrence (http://cobamide2.bio.pitt.edu/computer.htm).

Results

Results from simple regression analyses (using distance from replication origin as the predictor) and multiple regression analyses (using both distance from the origin and codon usage bias as predictors) are presented in table 1. In the majority of species pairs considered, there is a significant effect of distance from the replication origin on the synonymous substitution rate (K_s), with genes closer to the terminus having higher substitution rates (fig. 1). However, this distance effect accounts for, at most, 7% of the variation in a species pair (table 1); and, expectedly, a large proportion of the variation in K_s is explained by codon usage bias, which is known to be inversely related to K_s (Sharp and Li 1987; Smith and Eyre-Walker 2001). For the *Escherichia* O157-*Salmonella* comparisons, genes close to the replication terminus have, on average, 50% higher divergence at synonymous sites than genes near the replication origin. (The same result was obtained when comparing the K-12 strain of *E. coli* with *Salmonella*.) No distance effect was detected in *C. muridarum-C. trachomatis*, or for species pairs with very low levels of synonymous site divergence (table 1).

The most curious result was detected in the species pair *M. leprae-M. tuberculosis*, which shows a significant *negative* correlation between K_s and distance from the origin of replication. The negative association is detected at silent sites of functional homologous genes (fig. 1) and was further examined in the large number of pseudogenes in *M. leprae* (Cole et al. 2001). Because pseudogenes are nonfunctional, they might more accurately reflect the actual pattern of mutation than do synonymous codon positions (Gojobori, Ishii, and Nei 1982; Andersson and Andersson 2001). Regression analysis of substitutions in the pseudogenes not in saturation against distance from the origin revealed a negative slope that was only marginally significant ($t = -2.02$, $P = 0.044$).

Although silent sites may not be entirely neutral because many species show a nonrandom choice of codons (Gouy and Gautier 1982; Ikemura 1985), the distance effect observed in the species pairs presented in figure 1 is not affected by codon usage bias. This is supported by two results. First, individual regression analyses of distance from the origin as a predictor of K_s performed on different codon usage bias categories (low, intermediate, and high) yielded consistent results for each species comparison. Figure 2 shows this result for homologs from *E. coli* and *S. typhimurium*. The strength of the distance effect did not statistically differ among the three codon bias categories ($F = 0.97$, $P = 0.38$, ANCOVA). In addition, there was no effect of distance from the origin on codon usage bias for all species pairs (simple regression analysis, data not

Table 1
Effect of Distance from Replication Origin on Synonymous Site Divergence (K_s)

	No. of Ribosomal		Distance Effect[b]				Distance + Codon Bias Effects		
Species Pair	Operons	Avg. K_s[a]	n[c]	r^2	t Value	P	r^2	t Value[d]	P[d]
Escherichia coli O157[e]–*Salmonella typhimurium*	7	0.99	3,046	6.7	14.81	<0.0001	0.337	14.03 −35.24	<0.0001 <0.0001
Escherichia coli O157[e]–*Salmonella typhi*	7	0.97	2,415	0.041	10.23	<0.0001	0.288	9.74 −29.01	<0.0001 <0.0001
Chlamydia muridarum–*Chlamydia trachomatis*	1–2	0.79	823	<0.0001	−0.52	0.606	0.305	−1.11 −17.58	0.27 <0.0001
Pseudomonas aeruginosa–*Pseudomonas putida*[f]	4	0.78	1,861	0.019	6.00	<0.0001	0.073	5.19 −10.44	<0.0001 <0.0001
Mycobacterium leprae–*Mycobacterium tuberculosis*	1	0.74	338	0.085	−5.70	<0.0001	0.145	−5.55 −4.95	<0.0001 <0.0001
Listeria innocua–*Listeria monocytogenes*	6	0.62	2,500	0.036	9.48	<0.0001	0.163	9.44 −19.21	<0.0001 <0.0001
Rickettsia prowazekii–*Rickettsia conorii*	1	0.34	700	0.036	5.19	<0.0001	0.094	4.76 −6.79	<0.0001 <0.0001
Neisseria meningitidis B–*Neisseria gonorrhoeae*	4	0.14	1,619	<0.01	−0.26	0.796	0.66	−0.31 −3.26	0.76 0.001
Helicobacter pylori 26695–*Helicobacter pylori* J99	2–3	0.13	1,381	<0.01	0.82	0.409	0.16	1.03 −1.01	0.30 0.31
Escherichia coli K12–*Escherichia coli* O157	7	0.07	3,826	0.05	−1.36	0.175	0.40	−1.69 −3.64	0.09 <0.001
Neisseria meningitidis A–*Neisseria meningitidis B*	4	0.07	1,792	0.03	−0.76	0.450	0.25	−0.76 −1.97	0.45 0.048

[a] Calculated by the method of Li (1993).
[b] r^2, t value and P columns represent reduction of variance in the data, t value and P value of the linear regression analysis.
[c] Number of homologous genes compared in species pair.
[d] Data in the first row represent t and P values of the distance effect, data in the second row represent t and P values of codon bias effect.
[e] Comparisons using the strain K-12 of *E. coli* produced almost identical results.
[f] Data include all genes, regardless of strand location.

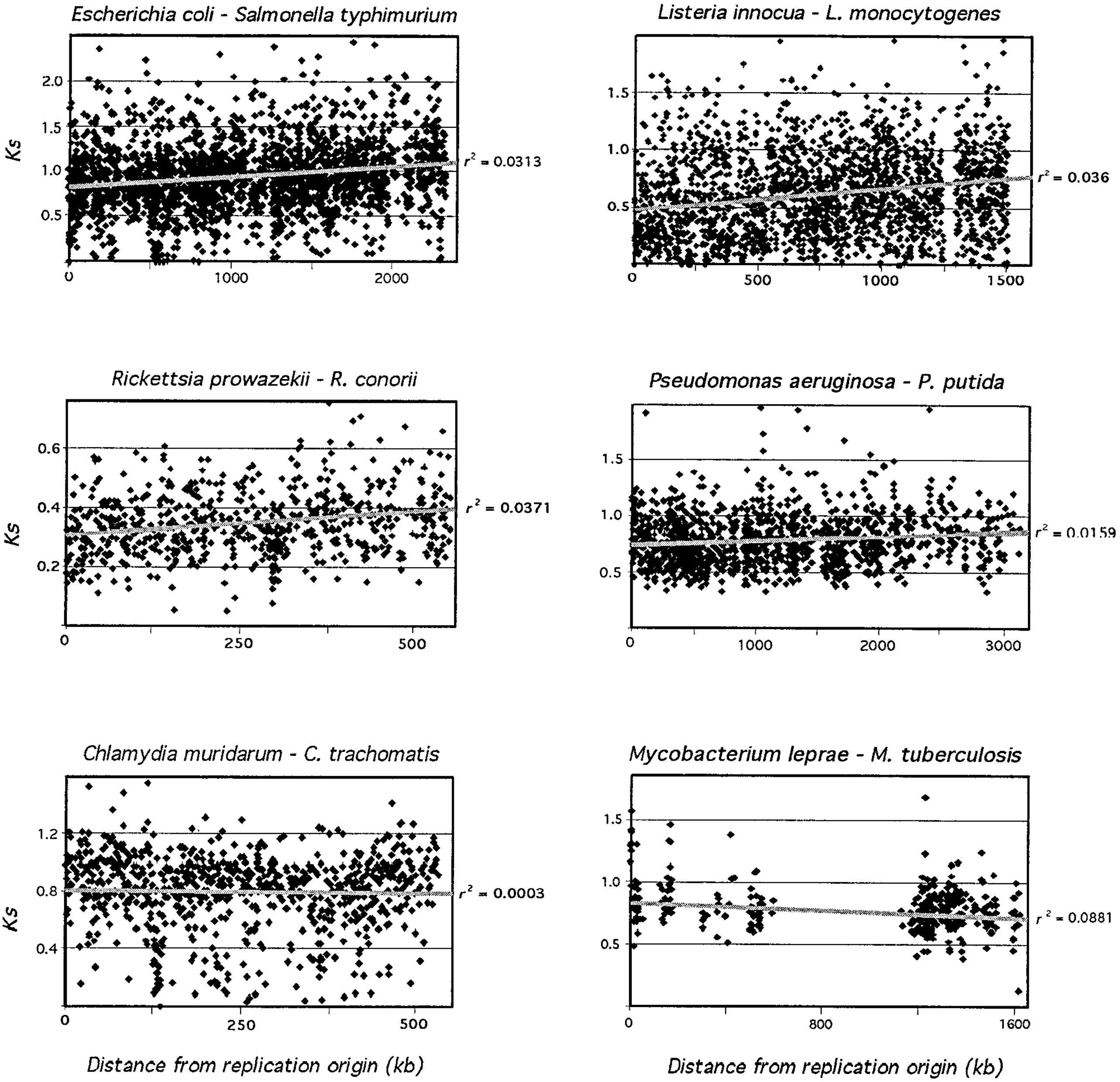

FIG. 1.—Effect of distance from the replication origin on synonymous substitution rates per synonymous site (K_S) in pairs of bacterial species. Each panel displays the regression for homologous genes in the species listed, with r^2 values indicated.

shown); in other words, highly biased genes were distributed equally throughout the chromosome (fig. 2), and the distance effect was not a by-product of low-biased genes being clustered near the terminus.

Analyses of the different types of nucleotide substitutions against chromosome position revealed that most of the distance effect is attributable to transversions (fig. 3). Although transitions also show a positive effect, they are less affected by distance from the origin and, in all species pairs in which a positive distance effect was detected, there are significantly lower slope coefficients and r^2 values for transitions than for transversions ($P < 0.01$ in all cases). Transitions and transversions were further separated into individual substitution types: C ↔ T, A ↔ G, G ↔ T, A ↔ C, A ↔ T and C ↔ G (table 2). In both the *Escherichia-Salmonella* and *R. prowazeki-R. conorii* comparisons G ↔ T and A ↔ C transversions were influenced most by distance from the replication origin, whereas in *L. innocua-L. monocytogenes* the largest effect was detected for A ↔ T transversions. The significant distance effect in the *Pseudomonas* comparison depends upon all substitution types, but no individual changes displayed a significant distance effect. In the *C. muridarum-C. trachomatis* comparison, C ↔ T transitions and A ↔ T transversions showed positive distance effects, and C ↔ G transversions a negative effect, producing the overall nonsignificant effect of distance on K_S. Finally, the comparison

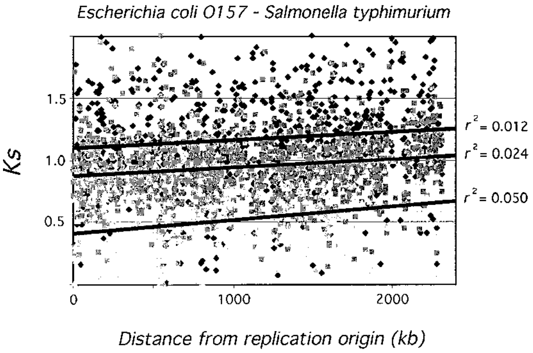

FIG. 2.—Effect of distance from the replication origin on synonymous divergence for genes having low (<0.30; black diamonds), intermediate (0.30–0.45; dark squares), or high (>0.45; pale triangles) levels of codon bias, as measured by the codon adaptation index (CAI) of Sharp and Li (1987), in the pair *E. coli* strain O157 and *S. enterica* serovar Typhimurium. The regression lines of the three CAI categories are indicated.

of the two *Mycobacteria* presents a weak negative distance effect that was significant for A ↔ G, A ↔ C, and A ↔ T substitutions. Thus, the distance effect in most species is primarily caused by changes in the rates of transversions; however, no specific transversion contributes to the effect in all species. Because certain transversions modify the GC content of a sequence, it is also notable that the difference between the GC contents of homologs increased with distance from the replication origin in all bacterial pairs except the Chlamydiales and Pseudomonads (table 2).

To further examine the potential factors influencing the increase in substitution rates with distance from the replication origin, simple regression analyses of distance as a predictor of K_s were performed separately for genes having opposite orientations with respect to their replication direction (table 3). Genes oriented in the same direction as replication (forward genes) and in the opposite direction to replication (reverse genes) both showed similar distance effects analogous to the one observed when all genes were considered. There was no significant difference in the strength of the distance effect for forward and reverse genes, indicating that the

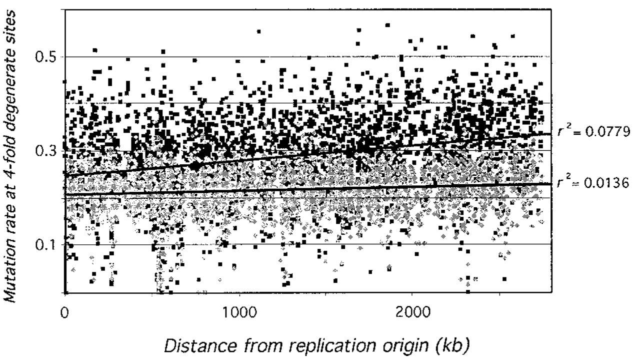

FIG. 3.—The effect of distance from the replication origin on transitions (gray diamonds) and transversions (black squares) rates at fourfold degenerate sites for comparisons of homologs from *E. coli* strain O157:H7 and *S. enterica* serovar Typhimurium.

Table 2
Effect of Distance from the Replication Origin on GC Content, and Rates of Transitions and Transversions

			TRANSITIONS[a]		TRANSVERSIONS[a]			
	%GC[b]	GC DIFF.[c]	C ↔ T	A ↔ G	G ↔ T	A ↔ C	A ↔ T	C ↔ G
Escherichia coli–Salmonella typhi	52.3–53.8	5.21***	5.00***	6.54***	9.17***	8.08***	6.95***	6.70***
			0.83	1.41	2.73	2.13	1.59	1.48
Escherichia coli–Salmonella typhimurium	52.2–53.9	3.34**	3.71***	6.90***	10.38***	10.78***	8.52***	8.05***
			0.44	1.50	3.33	3.58	2.26	2.03
Listeria innocua–Listeria monocytogenes	38.0–38.5	−4.69***	9.15***	3.77**	6.15***	8.33***	12.6***	5.05***
			3.18	0.55	1.46	2.65	5.87	0.99
Rickettsia prowazeki–Rickettsia conorii	30.3–32.8	4.55***	1.38	0.24	4.79***	4.11***	3.77**	1.66
			0.26	0.01	3.06	2.27	1.92	0.38
Pseudomonas aeruginosa–Pseudomonas putida	66.7–63.0	−1.48	−0.75	−2.15	−0.27	−1.86	−1.75	1.92
			0.02	0.18	0.003	0.14	0.12	0.14
Chlamydia muridarum–Chlamydia trachomatis	40.8–41.6	−0.11	2.90*	−0.19	−1.77	−0.56	3.07*	−3.57**
			1.01	0.004	0.38	0.04	1.13	1.52
Mycobacterium leprae–Mycobacterium tuberculosis	59.7–65.5	−5.33***	−2.51	−2.69*	−2.63*	−2.77*	−3.02*	−2.56
			0.45	0.52	0.50	0.55	0.65	0.47

[a] Data presented as t values (top row for each species pair) and 100 × r^2 values (underneath) for the regression analyses of distance from replication origin as a predictor of substitutions per site at third codon positions.

[b] Average GC content of homologous genes.

[c] Data represent t values for the regression analyses of distance from replication origin as a predictor of difference in GC contents of homologous genes.

* $P < 0.01$; ** $P < 0.001$; *** $P < 0.0001$.

Table 3
Effect of Gene Orientation (GO) on the Distance Effect

	Regression Analysis[a]					
	GO[b]	n	r^2	*t* value	Slope[c]	Gene Orientation Effect[d]
Escherichia coli O157–*Salmonella typhimurium*	F	1,485	0.0384	7.69***	0.19 ± 0.025	$F_{1,2509}$ = 1.46, P = 0.227
	R	1,028	0.0226	4.87***	0.15 ± 0.031	
Escherichia coli O157–*Salmonella typhi*	F	1,323	0.0356	6.99***	0.19 ± 0.027	$F_{1,2320}$ = 1.10, P = 0.293
	R	1,001	0.0587	7.89***	0.24 ± 0.031	
Chlamydia muridarum–Chlamydia trachomatis	F	406	0.0011	−0.65 NS	−0.03 ± 0.044	$F_{1,720}$ = 0.01, P = 0.910
	R	318	0.0008	−0.51 NS	−0.03 ± 0.056	
Pseudomonas aeruginosa–Pseudomonas putida	F	720	0.0145	3.25**	0.12 ± 0.037	$F_{1,1300}$ < 0.01, P = 0.972
	R	584	0.0188	3.34***	0.14 ± 0.041	
Mycobactericum leprae–Mycobactericum tuberculosis	F	221	0.0945	−4.78***	−0.31 ± 0.064	$F_{1,335}$ = 0.29, P = 0.593
	R	118	0.1012	−3.61***	−0.32 ± 0.088	
Listeria innocua–Listeria monocytogenes	F	1,960	0.0340	8.30***	0.18 ± 0.022	$F_{1,2524}$ = 0.01, P = 0.903
	R	568	0.0423	4.99***	0.20 ± 0.041	
Rickettsia prowazekii–Listeria conorii	F	437	0.0341	3.92***	0.18 ± 0.047	$F_{1,697}$ = 0.14, P = 0.709
	R	264	0.0466	3.58***	0.21 ± 0.060	

[a] Simple regression analysis of distance from the replication origin as a predictor of K_s.

[b] Symbols indicate gene orientation (GO). F = forward genes (orientated in the same direction as DNA replication); R = reverse genes (oriented opposite to DNA replication).

[c] Data represent coefficients of the regression line ± standard errors.

[d] *F* ratios and *P* values of ANCOVA with distance from replication origin as covariate and gene orientation as main factor.

**$P < 0.01$; ***$P < 0.001$; NS = not significant.

observed distance effect is not influenced by gene orientation.

Discussion

General Features

In the majority of bacterial species considered, there is a significant increase in synonymous site divergence with distance from the replication origin. Other bacterial pairs such as *Streptococcus pneumoniae-S. pyogenes* and *Staphylococcus aureus-S. epidermis* also show a significant distance effect but were not included in the analysis because their high divergence produced many genes with saturated K_s values. The only comparisons for which the distant effect was not observed were: (1) bacterial species that were very closely related, (2) the *Chlamydias*, and (3) the *Mycobacteria*, where a negative association was found. In the *E. coli-Salmonella* comparisons, a gene closer to the replication terminus undergoes about a 50% increase in synonymous divergence when compared with genes nearer to the origin. Although the distance effect is less dramatic than the twofold difference originally proposed (Sharp et al. 1989; Sharp 1991), it is profound and pervasive in phylogenetically distant bacterial clades.

We examined whether the distance effect is also present in archeabacteria. Although the mechanisms of DNA replication have not been fully elucidated in archaebacteria, recent research demonstrates that replication occurs from a single origin in some species such as *Pyrococcus* (Myllykallio et al. 2001; Smith et al. 1997; Salzberg et al. 1998). However, in a comparison of homologs between *Pyrococcus abysii* and *P. horikoshi*, there is no significant relationship between K_s and the distance from the putative replication origin ($r^2 = 0.001$, $P > 0.1$).

Possible Causes

Changes in substitution rates with distance from the replication origin could result from either differences in mutation rates or differences in repair rates at different positions of the chromosome. The distance effect was originally hypothesized to result from more frequent recombinational repair or biased gene conversion near the origin (Sharp et al. 1989; Sharp 1991; Birky and Walsh 1992) as achieved from the presence of multiple replication forks which produce multiple copies of sequences closer to the origin. Because the distance effect preferentially acts on transversions, it is difficult to see how such a substitutional pattern could arise from a nondiscriminating repair process, such as gene conversion or homologous exchange. The number of replication forks within a cell is largely influenced by the growth rate, which is highly variable among the bacteria analyzed in the present study (Mira, Moran, and Ochman 2001). Although growth rates are sometimes difficult to estimate, there is an association between the number of ribosomal RNA operons and growth rate across bacterial species (Asai et al. 1999; Klappenbach, Dunbar, and Schmidt 2000). When the strength of the distance effect is compared with the number of ribosomal operons across species (including the *Chlamydia*, *Listeria*, *Rickettsia*, *Mycobacterium*, and *Pseudomonas* pairs, together with the comparisons *S. pyogenes-S. pneumoniae*, *S. aureus-S. epidermis*, and *E. coli-S. typhimurium*), there is a positive relationship ($r^2 = 0.66$): the enteric bacteria have the strongest distance effect and also the highest number of ribosomal RNA operons (seven), whereas *Chlamydia* and *Mycobacterium*, which display no distance effect, have only one or two ribosomal operons. However, this result is equivocal: *Rickettsia* contains but a single ribosomal operon but shows a strong distance effect. In addition, this relationship between rRNA operons and the strength of the distance effect is based on a small set of phylogenetically distant, but not completely independent, comparisons.

To gain additional insights into the potential mechanisms involved in the distance effect, we examined the frequency of individual substitutions at different parts of the chromosome. In the cases where a positive distance effect was detected, transitions generally increased with distance from the origin, but to a significantly lower extent than transversions. When the different transitions and transversions were evaluated, the substitutions contributing most to the distance effect varied according to the specific bacterial pair. For example, G ↔ T and A ↔ C transversions are most prevalent in *Escherichia*, *Salmonella*, and *Rickettsia*, in contrast to A ↔ T transversions in *Listeria*.

Gene orientation does not influence the distance effect: genes of forward and of reverse orientation show a similar increase in K_s values with distance from replication origin. In species pairs for which there is a significant distance effect, the GC content of homologs differs most in genes situated away from the origin of replication. The extent to which this is a cause or result of the distant effect is not known. It is possible that some of the mechanisms affecting GC composition, such as mutational bias, are intensified when a gene is located closer to the replication terminus. Bacterial chromosomes are thought to move through a stationary machinery for replicating DNA (Lemon and Grossman 1998), and the newly formed replication origins appear to move toward the pole of cells (Webb et al. 1998). It is possible that this replication process creates differences in enzyme activity (and mutation rates) along different parts of the chromosome. For example, the DNA polymerase may tend to fall off the replicating DNA strand as replication progresses and the reassembling of the polymerase can be error-prone (Goodman 2000; Courcelle and Hanawalt 2001).

Seeming Exceptions

Focusing on the cases where a distance effect on K_s was not detected offers additional insights into its possible causes. For example, when homologs from strains within a single species were compared, as possible in *E. coli*, *Neisseria meningitidis*, and *Helicobacter pylori*, no significant distance effects were observed, and the same was true for the pair *N. meningitidis-N. gonorrhoeae*. Thus, in comparisons where there are low levels of sequence divergence between homologs, we detected no effect of distance from the replication origin on substitution rates (table 1). In these cases, there is probably insufficient variation to detect a change in substitution frequencies across the chromosome, particularly if rarer transversions are responsible for the phenomenon. In addition, recombination between such closely related strains might diminish the overall amount of detected divergence.

Another case in which there is no significant association between distance from the replication origin and K_s is in the *C. muridarum-C. trachomatis* comparison. The relatively small chromosome of these parasitic bacteria (Read et al. 2000) might contribute to the absence of a distance effect. In these genomes, a gene can be, at most, 500 kilobases (kb) from the replication origin, a distance that may not be sufficient to produce a significant effect in this species. For example, when analyzing only the genes in the initial 500 kb of *E. coli* and *S. typhi* chromosomes, there is no significant distance effect ($t = -0.67$, $P = 0.55$). However, in *Rickettsia*, which has approximately the same genome size as *Chlamydia*, a distance effect is apparent.

During the process of genome reduction, both *Rickettsia* and *Chlamydia* have lost several DNA repair genes (Stephens et al. 1998; Andersson and Andersson 2001), and if any are uniquely involved in the preferential repair of close-to-the-origin genes, their absence might eliminate a distance effect.

Whatever mechanism underlies the distance effect, the increase in synonymous divergence with distance from the replication origin should be apparent in spontaneous mutation or substitution rates measured under experimental conditions. However, Hudson et al. (2002) failed to detect an effect of distance from the replication origin on the mutation rate of *lacZ* alleles inserted at four sites in the *Salmonella* genome. In contrast, they found the highest mutation rate at a locus of intermediate position between the replication origin and terminus. The basis for this discrepancy could be that laboratory conditions produce a different mutational spectrum than that under natural conditions (Hudson et al. 2002). Although the distance effect was not apparent in this experimental setting, it has influenced the rates and patterns of molecular evolution across a wide range of bacterial genomes.

Acknowledgments

We thank Susan Miller and the Biotechnology Computing Facility of the University of Arizona for computer programs and Jeff Lawrence for distributing the DNA Master sequence analysis program. This work was supported by grants from the NIH and DOE.

LITERATURE CITED

ALTSCHUL, S. F., T. L. MADDEN, A. A. SCHAFFER, J. ZHANG, Z. ZHANG, W. MILLER, and D. J. LIPMAN. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. **25**:3389–3402.

ANDERSSON, J. O., and S. G. E. ANDERSSON. 2001. Pseudogenes, junk DNA, and the dynamics of *Rickettsia* genomes. Mol. Biol. Evol. **18**:829–839.

ASAI, T., C. CONDON, J. VOULGARIS, D. ZAPOROJETS, B. SHEN, M. AL-OMAR, C. SQUIRES, and C. L. SQUIRES. 1999. Construction and initial characterization of *Escherichia coli* strains with few or no intact chromosomal rRNA operons. J. Bacteriol. **181**:3803–3809.

BAREKZI, N., K. BEINLICH, T. T. HOANG, X. Q. PHAM, R. KARKHOFF-SCHWEIZER, and H. P. SCHWEIZER. 2001. High-frequency flp recombinase-mediated inversions of the oriCcontaining region of the *Pseudomonas aeruginosa* genome. J. Bacteriol. **182**:7070–7074.

Birky, C. W. Jr., and J. B. Walsh. 1992. Biased gene conversion, copy number, and apparent mutation rate differences within chloroplast and bacterial genomes. Genetics **130**:677–783.

Blattner, F. R., G. Plunkett, III, C. A. Bloch et al. (17 co-authors). 1997. The complete genome sequence of *Escherichia coli* K-12. Science **277**:1453–1474.

Cole, S. T., K. Eiglmeier, J. Parkhill, K. D. James, N. R. Thomson, P. R. Wheeler, N. Honore, T. Garnier, C. Churcher, and D. Harris. 2001. Massive gene decay in the leprosy bacillus. Nature **409**:1007–1011.

Courcelle, J., and P. C. Hanawalt. 2001. Participation of recombination proteins in rescue of arrested replication forks in UV-irradiated *Escherichia coli* need not involve recombination. Proc. Natl. Acad. Sci. USA **98**:8196–8202.

Devereux, J., P. Haeberli, and O. Smithies. 1984. A comprehensive set of sequence analysis programs for the VAX. Nucleic Acids Res. **12**:387–395.

Francino, M. P., and H. Ochman. 2001. Deamination as the basis of strand-asymmetric evolution in transcribed *Escherichia coli* sequences. Mol. Biol. Evol. **18**:1147–1150.

Gasc, A. M., P. Giammarinaro, S. Richter, and M. Sicard. 1998. Organization around the dnaA gene of *Streptococcus pneumoniae*. Microbiology **144**:433–439.

Gojobori, T., K. Ishii, and M. Nei. 1982. Patterns of nucleotide substitution in pseudogenes and functional genes. J. Mol. Evol. **18**:360–369.

Goodman, M. F. 2000. Coping with replication ‘train wrecks’ in *Escherichia coli* using Pol V, Pol II and RecA proteins. Trends Biochem. Sci. **25**:189–195.

Gouy, M., and Gautier, C. 1982. Codon usage in bacteria: correlation with gene expressivity. Nucleic Acids Res. **10**: 7055–7074.

Hudson, R. E., U. Bergthorsson, J. R. Roth, and H. Ochman. 2002. Effect of chromosome location on bacterial mutation rates. Mol. Biol. Evol. **19**:85–92.

Ikemura, T. 1981. Correlation between the abundance of *Escherichia coli* transfer RNAs and the occurrence of the respective codons in its protein genes: a proposal for a synonymous codon choice that is optimal for the *E. coli* translational system. J. Mol. Biol. **151**:389–409.

Ikemura, T. 1985. Codon usage and tRNA content in unicellular and multicellular organisms. Mol. Biol. Evol. **2**:13–34.

Klappenbach, J. A., J. M. Dunbar, and T. M. Schmidt. 2000. rRNA operon copy number reflects ecological strategies of bacteria. Appl. Environ. Microbiol. **66**:1328–1333.

Kuroda, M., T. Ohta, I. Uchiyama et al. (37 co-authors). 2001. Whole genome sequencing of meticillin-resistant *Staphylococcus aureus*. Lancet **357**:1225–1240.

Lemon, K. P., and A. D. Grossman. 1998. Localization of bacterial DNA polymerase: evidence for a factory model of replication. Science **282**:1516–1519.

Li, W. H. 1993. Unbiased estimation of the rates of synonymous and non-synonymous substitution. J. Mol. Evol. **36**: 96–99.

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

McClelland, M., K. E. Sanderson, J. Spieth et al. (26 co-authors). 2001. Complete genome sequence of *Salmonella enterica* serovar Typhimurium LT2. Nature **413**:852–856.

Mira, A., N. A. Moran, and H. Ochman. 2001. Deletional bias and the evolution of bacterial genomes. Trends Genet. **17**:589–596.

de Miranda, A. B., F. Alvarez-Valin, K. Jabbari, W. M. Degrave, and G. Bernardi. 2000. Gene expression, amino acid conservation, and hydrophobicity are the main factors shaping codon preferences in *Mycobacterium tuberculosis* and *Mycobacterium leprae*. J. Mol. Evol. **50**:45–55.

Moran, N. A., and J. J. Wernegreen. 2000. Lifestyle evolution in symbiotic bacteria: insights from genomics. Trends Ecol. Evol. **15**:321–326.

Myllykallio, H., P. Lopez, P. Lopez-Garcia, R. Heilig, W. Saurin, Y. Zivanovic, H. Philippe, P. Forterre. 2001. Bacterial mode of replication with eukaryotic-like machinery in a hyperthermophilic archaeon. Science **288**:2212–2215.

Ogata, H., S. Audic, P. Renesto-Audiffren et al. (11 co-authors). 2001. Mechanisms of evolution in *Rickettsia conorii* and *R. prowazekii*. Science **293**:2093–2098.

Pamilo, P., and N. O. Bianchi. 1993. Evolution of the *zfx* and *zfy* genes—rates and interdependence between the genes. Mol. Biol. Evol. **10**:271–281.

Parkhill, J., G. Dougan, K. D. James et al. (41 co-authors). 2001. Complete genome sequence of a multiple drug resistant *Salmonella enterica* serovar Typhi CT18. Nature **413**: 848–852.

Perna, N. T., G. Plunkett III, V. Burland et al. (28 co-authors). 2001. Genome sequence of enterohaemorrhagic *Escherichia coli* O157:H7. Nature **409**:529–533.

Read, T. D., R. C. Brunham, C. Shen et al. (25 coauthors). 2000. Genome sequences of *Chlamydia trachomatis* MoPn and *Chlamydia pneumoniae* AR39 Nucleic Acids Res. **28**: 1397–1406.

Salazar, L., H. Fsihi, E. de Rossi, G. Riccardi, C. Rios, S. T. Cole, and H. E. Takiff. 1996. Organization of the origins of replication of the chromosomes of *Mycobacterium smegmatis*, *Mycobacterium leprae* and *Mycobacterium tuberculosis* and isolation of a functional origin from *M. smegmatis*. Mol. Microbiol. **20**:283–293.

Salzberg, S. L., A. J. Salzberg, A. R. Kerlavage, and J. F. Tomb. 1998. Skewed oligomers and origins of replication. Gene **217**:57–67.

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 W. H. Li. 1987. The rate of synonymous substitution in enterobacterial genes is inversely related to codon usage bias. Mol. Biol. Evol. **4**:222–230.

Sharp, P. M., D. C. Shields, K. H. Wolfe, and W. H. Li. 1989. Chromosomal location and evolutionary rate variation in enterobacterial genes. Science **246**:808–810.

Shields, D. C., P. M. Sharp, D. G. Higgins, and F. Wright. 1998. ‘‘Silent’’ sites in *Drosophila* genes are not neutral: evidence of selection among synonymous codons. Mol. Biol. Evol. **5**:704–716.

Smith, D. R., L. A. Doucette-Stamm, C. Deloughery et al. (37 co-authors). 1997. Complete genome sequence of *Methanobacterium thermoautotrophicum* ΔH: functional analysis and comparative genomics. J. Bacteriol. **179**: 7135–7155.

SMITH, N. G., and A. EYRE-WALKER. 2001. Nucleotide substitution rate estimation enterobacteria: approximate and maximum-likelihood methods lead to similar conclusions. Mol. Biol. Evol. **18**:2124–2126.

STEPHENS, R. S., S. KALMAN, C. LAMMEL et al. (12 co-authors). 1998. Genome sequence of an obligate intracellular pathogen of humans: *Chlamydia trachomatis*. Science **282**: 754–759.

WEBB, C. D., P. L. GRAUMANN, J. A. KAHANA, A. A. TELEMAN, P. A. SILVER, and R. LOSICK. 1998. Use of time-lapse microscopy to visualize rapid movement of the replication origin region of the chromosome during the cell cycle in *Bacillus subtillus*. Mol. Microbiol. **28**:883–892.

WEIGEL, C., A. SCHMIDT, B. RUCKERT, R. LURZ, and W. MESSER. 1997. DnaA protein binding to individual DnaA boxes in the *Escherichia coli* replication origin, oriC. EMBO J. **16**:6574–6583.

KENNETH WOLFE, reviewing editor

Accepted April 12, 2002