62 Pelet, A. *et al.* (1998) Various mechanisms cause RET-mediated signaling defects in Hirschsprung disease. *J. Clin. Invest.* 101, 1415–1423
63 Takahashi, M. *et al.* (1999) Co-segregation of MEN2 and Hirschsprung disease: the same mutation of RET with both gain and loss-of-function? *Hum. Mutat.* 13, 331–336
64 Mulligan, L.M. *et al.* (1994) Diverse phenotypes associated with exon 10 mutations of the RET proto-oncogene. *Hum. Mol. Genet.* 3, 2163–2167
65 Geneste, O. *et al.* (1999) Two distinct mutations of the RET receptor causing Hirschsprung disease impair the binding of signalling effectors to a multifunctional docking site. *Hum. Mol. Genet.* 8, 1989–1999
66 Pasini, B. *et al.* (1995) Loss of function effect of RET mutations causing Hirschsprung disease. *Nat. Genet.* 10, 35–40
67 Cosma, M.P. *et al.* (1998) Mutations in the extracellular domain cause RET loss of function by a dominant negative mechanism. *Mol. Cell. Biol.* 18, 3321–3329
68 Angrist, M. *et al.* (1998) Human GFRA1: cloning, mapping, genomic structure, and evaluation as a candidate gene for Hirschsprung disease susceptibility. *Genomics* 48, 354–362
69 Myers, S.M. *et al.* (1999) Investigation of germline GFR alpha-1 mutations in Hirschsprung disease. *J. Med. Genet.* 36, 217–220
70 Onochie, C.I. *et al.* (2000) Characterisation of the human GFRalpha-3 locus and investigation of the gene in Hirschsprung disease. *J. Med. Genet.* 37, 669–673
71 Vanhorne, J.B. *et al.* Cloning and characterization of the GFRA2 locus and investigation of the gene in Hirschsprung disease. *Hum. Genet.* (in press)
72 Bolk, S. *et al.* (2000) A human model for multigenic inheritance: phenotypic expression in Hirschsprung disease requires both the RET gene and a new 9q31 locus. *Proc. Natl. Acad. Sci. U. S. A.* 97, 268–273
73 Chakravarti, A. and Lyonnet, S. (2001) Hirschsprung disease. In *Molecular Bases of Hereditary Diseases* (Scriver *et al.*, eds), pp. 6231–6255,.
74 Lang, D. *et al.* (2000) Pax3 is required for enteric ganglia formation and functions with Sox10 to modulate expression of c-ret. *J. Clin. Invest.* 106 (8), 963–971
75 Grimm, J. *et al.* (2001) Novel p62dok family members, dok-4 and dok-5, are substrates of the c-Ret receptor tyrosine kinase and mediate neuronal differentiation. *J. Cell Biol.* 154, 345–354
76 Burton, M.D. *et al.* (1997) RET proto-oncogene is important for the development of respiratory CO_2 sensitivity. *J. Auton. Nerv. Syst.* 63, 137–143

Deletional bias and the evolution of bacterial genomes

Alex Mira, Howard Ochman and Nancy A. Moran

Although bacteria increase their DNA content through horizontal transfer and gene duplication, their genomes remain small and, in particular, lack nonfunctional sequences. This pattern is most readily explained by a pervasive bias towards higher numbers of deletions than insertions. When selection is not strong enough to maintain them, genes are lost in large deletions or inactivated and subsequently eroded. Gene inactivation and loss are particularly apparent in obligate parasites and symbionts, in which dramatic reductions in genome size can result not from selection to lose DNA, but from decreased selection to maintain gene functionality. Here we discuss the evidence showing that deletional bias is a major force that shapes bacterial genomes.

Alex Mira
Howard Ochman*
Nancy A. Moran
Dept of Ecology and Evolutionary Biology, University of Arizona, Tucson, Arizona 85721, USA.
*e-mail: hochman@email.arizona.edu

When compared with eukaryotes, bacteria, including eubacteria and archaebacteria, accommodate a rather narrow range of variation in genome size. Whereas eukaryotic genomes vary in size by four orders of magnitude (from about 10^7–10^{11} basepairs), there is only about one order-of-magnitude difference across bacterial genome sizes[1–3]. However, the difference in the ranges of genome size in eukaryotes and bacteria is not reflected in corresponding differences in gene number. Unlike eukaryotes, the genome size variation in bacteria translates almost directly into biochemical, physiological and organismal complexity because the majority of sequences are functional protein-coding regions (Fig. 1). Among bacteria for which complete genomic sequences are available, a tenfold variation in genome size is reflected by a similar difference in total gene number[4,5] (Table 1). By contrast, yeast and humans have genomes that differ by almost 300-fold in size, yet they have only a sixfold difference in gene number[6–8].

What is the source of variation in genome size in bacteria? On the basis of the distribution of genome sizes and the orientation of apparently duplicated genes, it was once thought that new bacterial genomes evolved by repeated events of genome doubling[1,9,10]. However, subsequent analyses of additional genomes provided several lines of evidence against this hypothesis. First, related bacteria having genomes of similar sizes often contain very different complements of genes, and arrangements of duplicated genes are not consistent across taxa[11,12]. Second, the variation in

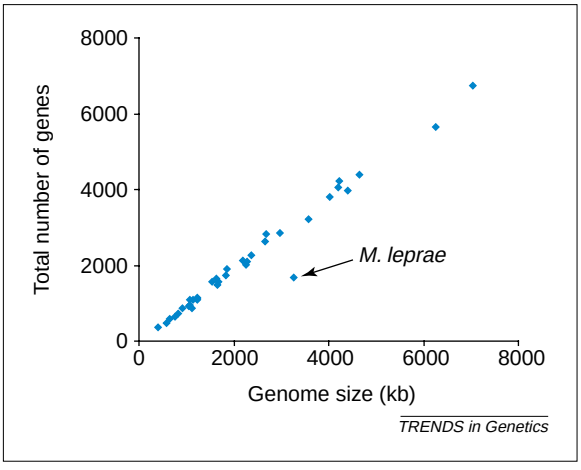


Fig. 1. Association between genome size and gene number in bacteria. Numbers include protein-coding and RNA genes ($R^2 = 0.945$). When the number of annotated pseudogenes is added to the number of functional genes, *Mycobacterium leprae* falls on the regression line. Taxa are listed in Table 1.

 0168-9525/01/$ – see front matter PII: S0168-9525(01)02447-7

Table 1. Characteristics of sequenced prokaryotic genomes

	Taxonomic group	Genome size (bp)	% occupied by genes[a]	Total gene number	Average gene length (bp)	Pseudogene number[b]	Mean spacer length[c] (bp)	Median spacer length[c] (bp)
Eubacteria								
Mycoplasma genitalium	Low GC Gram +	580 074	90.4	481	1094	–	116.0	4
Buchnera aphidicola	γ-Proteobacteria	640 681	87.9	601	937	8	127.5	74
Ureaplasma urealyticum	Low GC Gram +	751 719	92.5	652	1069	–	86.9	37
Mycoplasma pneumoniae	Low GC Gram +	816 394	88.8	732	812	–	124.9	21
Borrelia burgdorferi	Spirochaetales	910 724	94.2	873	986	–[e]	60.1	19
Chlamydia trachomatis	Chlamydiales	1 042 519	90.9	937	1013	–	101.5	53
Chlamydia murinarum	Chlamydiales	1 069 411	90.9	949	1029	–	102.6	43
Rickettsia prowazekii	α-Proteobacteria	1 111 523	76.0	871	971	12	306.2	123
Treponema pallidum	Spirochaetales	1 138 011	93.1	1082	986	–	72.9	35
Chlamydophila pneumoniae J138	Chlamydiales	1 228 267	89.6	1110	997	–	115.5	52
Chlamydophila pneumoniae AR39	Chlamydiales	1 229 858	89.2	1152	956	–	115.7	49
Chlamydophila pneum. CWCL029	Chlamydiales	1 230 230	88.8	1097	1000	–	126.2	55
Aquifex aeolicus	Aquificales	1 551 335	93.7	1574	930	–	62.5	10
Campylobacter jejuni	ε-Proteobacteria	1 641 481	94.3	1654	939	20	56.4	8
Helicobacter pylori J99	ε-Proteobacteria	1 643 831	90.1	1491	996	–[f]	108.9	23
Helicobacter pylori 26695	ε-Proteobacteria	1 667 867	88.4	1553	952	–[f]	124.9	25
Haemophilus influenzae	γ-Proteobacteria	1 830 138	87.0	1745	914	–	136.5	68
Thermotoga maritima	Thermotogales	1 860 725	93.8	1895	927	–	60.7	6
Neisseria meningitidis strain A	β-Proteobacteria	2 184 406	82.7	2121	853	56	178.0	85
Pasteurella multocida	γ-Proteobacteria	2 257 487	88.9	2015	998	–	124.7	64
Neisseria meningitidis strain B	β-Proteobacteria	2 272 351	78.6	2096	853	–[g]	231.9	94
Lactococcus lactis	Low GC Gram +	2 365 589	84.7	2268	882	–	159.7	96
Xylella fastidiosa	γ-Proteobacteria	2 679 306	83.8	2822	797	–	154.4	76
Deinococcus radiodurans[d]	Thermus/ Deinococcus	3 060 986	88.7	2996	960	–	122.1	44
Mycobacterium leprae	Actinobacteria	3 268 203	76.8	2770	908	1081	546.4	106
Synechocystis PCC6803	Cyanobacteria	3 573 470	87.0	3219	969	–	143.9	103
Caulobacter crescentus	α-Proteobacteria	4 016 947	90.5	3794	961	–	201.6	62
Vibrio cholerae[d]	γ-Proteobacteria	4 033 464	86.4	3949	876	–	136.6	78
Bacillus halodurans	Low GC Gram +	4 202 353	84.9	4066	879	–	156.4	87
Bacillus subtilis	Low GC Gram +	4 214 814	87.8	4221	880	–	121.9	72
Mycobacterium tuberculosis	Actinobacteria	4 411 529	90.5	3970	1009	9	105.2	49
Escherichia coli K12	γ-Proteobacteria	4 639 221	87.9	4405	1027	–	128.8	63
Pseudomonas aeruginosa	γ-Proteobacteria	6 264 403	89.5	5640	995	–	116.6	68
Mesorhizobium loti	α-Proteobacteria	7 036 074	86.4	6752	905	–	142.2	73
Averages			**88.0**	**1860**	**947**		**147.0**	**57**
Archaea								
Thermoplasma acidophilum	Thermoplasmales	1 564 906	86.0	1526	897	–	134.3	65
Thermoplasma volcanium	Thermoplasmales	1 584 804	85.7	1548	880	–	146.1	53
Methanococcus jannaschii	Methanococcales	1 664 970	87.8	1758	834	–	115.6	49
Pyrococcus horikoshii	Thermococcales	1 738 505	84.4	2113	814	–	128.8	13
Methanobacterium thermoautotrophicum	Methanobacteriales	1 751 377	90.8	1916	832	–	83.8	37
Pyrococcus abysii	Thermococcales	1 765 118	81.8	1765	912	–	93.5	14
Halobacterium sp.	Halobacteriales	2 014 239	88.0	2110	841	–	115.3	56
Archaeoglobus fulgidus	Archaeoglobales	2 178 400	91.7	2420	832	–	74.5	13
Averages			**87.0**	**1983**	**855**		**112.0**	**37**

[a]Includes RNA genes.
[b]Includes only annotated chromosomal pseudogenes.
[c]Spacers between overlapping genes are scored as zero.
[d]Chromosomes 1 and 2 combined.
[e]Although no chromosomally encoded pseudogenes have been identified, *Borrelia* contains many plasmid-borne pseudogenes.
[f]Sequenced genome contains no annotated pseudogenes, but at least six were identified in other strains.
[g]Sequenced genome contains no annotated pseudogenes, but at least five were identified in other strains.

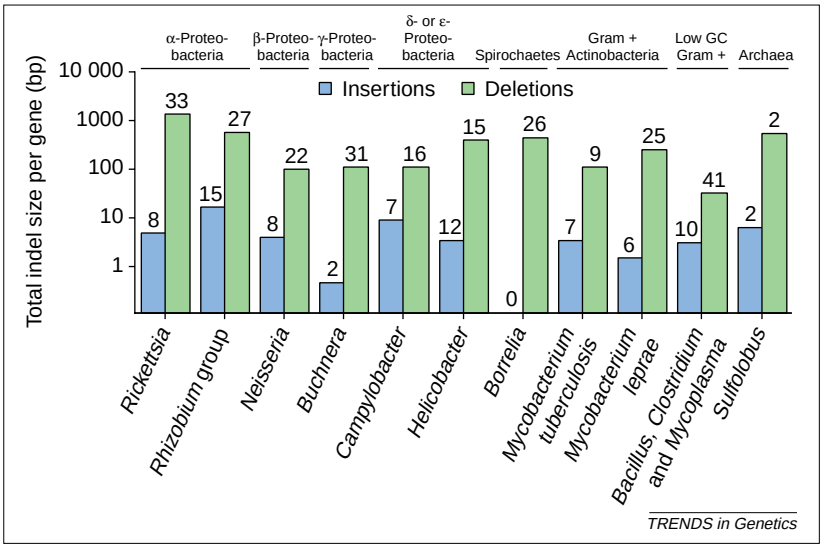


Fig. 2. Frequency of deletions and insertions in bacterial genomes. Frequencies based on comparative analyses of pseudogenes with their functional counterparts from a closely related species, generally from the same genus, and with at least one functional gene in a closely related outside reference species. Bars represent the average total size of deletions and insertions per pseudogene. Numbers at tops of bars represent the numbers of each type of event. Analyzed pseudogenes were: *hmbR*, *opaA*, *opaB*, *yedI**, *fhuA*, *porA*, *pilC2*, *opcB* (*Neisseria meningitidis, Neisseria gonorrhoeae*); *kdpA*, *kdpC*, *ast*, *glpT'**, cj0565, cj794** (*Campylobacter jejuni*); *fhuA*, cytochrome P-450, IS1380, *egl*, groEL (*Rhizobium* sp.); *vacA*, *oipA*, *iceA*, *rfaJ*, OMP29, HP1589 (*Helicobacter pylori*); *ahcY* (*Sulfolobus solfataricus*); groES2 (*Rhodobacter sphaeroides*); *msp1b1pg* (*Anaplasma marginale*); ORF3, *aatA* (*Agrobacterium rhizogenes*); *epsD* (*Azospirillum brasilense*); *oxyR*, Rv1503c-Rv1504c, *hypB*, Rv3349c** (*Mycobacterium tuberculosis*); *bfrB*, *csp*, *ackA-pta*, *fadE7*, *cysM* (*Mycobacterium leprae*); *vmp*, *vlp*, BBG20**, BBQ71**, BBQ55 (*Borrelia* sp.); *hmw2*, *lppA* (*Mycoplasma* sp.); *treP*, *hblB*, transposase, *s14*, *fhuC* (*Bacillus* sp.); *cpe*, p-21 (*Clostridium* sp.); recombinase, *sat4* (*Staphylococcus* sp.); *hisG*, *hisC* (*Lactococcus lactis*); *ace* (*Enterococcus faecalis*). Genes marked with an asterisk are those in which a functional equivalent in the same bacterial group could not be found; two asterisks indicate that no suitable homolog could be found in an outside reference species.

Box 1. Replication advantage: organellar and bacterial genomes compared

The compactness of bacterial and organellar genomes is often attributed to selection for rapid replication of DNA (Refs a–g). It is known, for example, that mitochondria with chromosomal deletions accumulate with time in animal tissues[h] and that plant plastids with deletions can accumulate *in vitro*[i]. Although organelles are derived from bacteria, the latter have a complex cell structure, and chromosomal replication rate has little effect on cell division rate. Comparing across diverse bacterial taxa, doubling times show no relationship with genome size (see Fig. I). Strains of *E. coli* can vary by as much as 25% in chromosome length, but growth rates are not correlated with genome size[j], under either rich or poor nutrient conditions[k]. Moreover, the translational efficiency of ribosomes and tRNA abundance, not replication rate, are primary determinants of cell division rates in bacterial isolates[l–n]. It is unclear what selective forces act on growth rates under natural conditions: wild strains of *E. coli* are highly variable in their doubling times[o], and certain intracellular bacteria with reduced genomes display the slowest growth rates among prokaryotes[p]. The translational machinery comprises about half of the dry weight of a bacterial cell, and the process consumes up to 80% of the cell's energy[q]. It is therefore probable that the profound effects on growth rates of ribosome kinetics and the translational process mask any influences of genome size on replication speed.

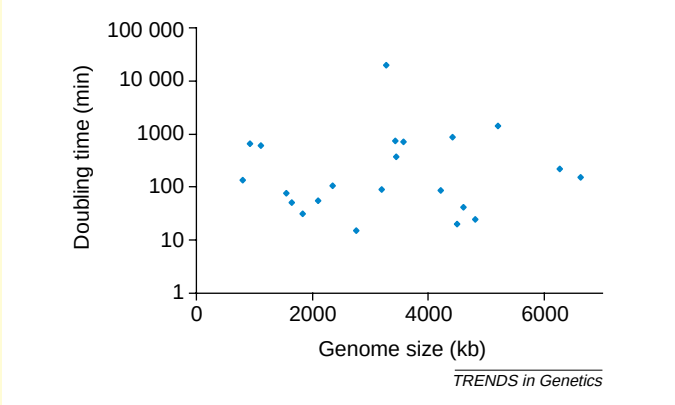


Fig. I. Doubling times of bacteria under laboratory conditions do not correlate with genome size. Data are for 22 species for which doubling times were available in the literature, and include bacteria from ten major taxonomic divisions.

References

a Selosse, M-A. *et al.* (2001) Reducing the genome size of organelles favours gene transfer to the nucleus. *Trends Ecol. and Evol.* 16, 135–141

b Andersson, S.G.E. and Kurland, C.G. (1998) Reductive evolution of resident genomes. *Trends Microbiol.* 6, 263–268

c Cavalier-Smith, T. (1987) The simultaneous symbiotic origin of mitochondria, chloroplasts and microbodies. *Ann. New York Acad. Sci.* 503, 53–71

d Samuelsson, T. and Borén, T. (1992) Evolution of macromolecule synthesis. In *Mycoplasmas: Molecular Biology and Pathogenesis* (Maniloff, J. *et al.*, eds), pp. 575–591, ASM Press

e Andersson, S.G.E. and Kurland, C.G. (1995) Genomic evolution drives the evolution of the translation system. *Biochem. Cell Biol.* 73, 775–787

f Maniloff, J. (1996) The minimal cell genome: 'On being the right size'. *Proc. Natl. Acad. Sci. U. S. A.* 93, 10004–10006

g Vellai, T. and Vida, G. (1999) The origin of eukaryotes: the difference between prokaryotic and eukaryotic cells. *Proc. R. Soc. London Ser. B* 266, 1571–1577

h Cortopassi, G.A. *et al.* (1992) A pattern of accumulation of a somatic deletion of mitochondria in ageing human tissues. *Proc. Natl. Acad. Sci. U. S. A.* 89, 7370–7374

i Day, A. and Ellis, T.H.N. (1984) Chloroplast DNA deletions associated with wheat plants regenerated from pollen: possible basis for maternal inheritance of chloroplasts. *Cell* 39, 359–368

j Bergthorsson, U. and Ochman, H. (1998) Distribution of chromosome length variation in natural isolates of *Escherichia coli*. *Mol. Biol. Evol.* 15, 6–16

k Bergthorsson, U. and Ochman, H. (1995) Heterogeneity of genome sizes among natural isolates of *Escherichia coli*. *J. Bacteriol.* 177, 5784–5789

l Kurland, C.G. and Ehrenberg, M. (1987) Growth-optimizing accuracy of gene expression. *Annu. Rev. Biophys.* 16, 291–317

m Emilsson, V. and Kurland, C.G. (1990) Growth rate dependence of transfer RNA abundance in *Escherichia coli*. *EMBO J.* 9, 4359–4366

n Mikkola, R. and Kurland, C.G. (1991) Is there a unique ribosome phenotype for naturally occurring *Escherichia coli*? *Biochimie* 73, 1061–1066

o Mikkola, R. and Kurland, C.G. (1992) Selection of laboratory wild-type phenotype from natural isolates of *Escherichia coli* in chemostats. *Mol. Biol. Evol.* 9, 394–402

p Cole, S.T. *et al.* (2001) Massive gene decay in the leprosy bacillus. *Nature* 409, 1007–1011

q Maaloe, O. (1979) Regulation of the protein-synthesizing machinery in ribosomes, tRNA, factors and so on. In *Biological Regulation and Development* (Vol. 1) (Goldberger, R.F., ed.), pp. 487–542, Plenum Press

Box 2. Testing reductive evolution in intracellular bacteria

Bacterial lineages that form obligate, intimate associations with hosts show a repeated pattern of extreme reduction in genome size[a–d]. Obligate pathogens, such as *Mycoplasma*, *Rickettsia* and *Chlamydia*, have genome sizes of approximately one megabase (Mb) or less[e], as do the endosymbiotic bacteria of aphids and tsetse flies (the bacteria *Buchnera* and *Wigglesworthia*, respectively)[f,g]. By contrast, free-living bacteria, including enteric bacteria, which are closely related to *Buchnera*, have genomes of 3–6 Mb (Ref. e). Because bacterial genomes consist primarily of genes encoding proteins (Table 1), this large-scale genome reduction implies a massive loss of genes. Genome reduction could result from several processes:

Hypothesis 1: Selection favoring small genome size drives the reduction. This selection could potentially be due to faster replication or energy savings connected with a small genome size[h–k]. This hypothesis is often presumed to be true without direct examination (Box 1).

Hypothesis 2: Increase in the rate of deletions or in the degree of deletional bias of novel mutations drives the reduction.

Hypothesis 3: Decrease in selection across many loci results in a large proportion of the genome that is effectively neutral, and these regions are eliminated by deletional bias in the mutation pattern. Effective neutrality of genes in newly symbiotic or pathogenic lineages could occur for several reasons. First, a portion of this reduction in genome size could be attributable to the loss of genes that are not required in the intracellular habitat. For example, genes underlying the biosynthesis of nutrients that can be assimilated from the host cell are missing in obligate symbionts and pathogens[l–p], and genes underlying cell motility have been eliminated from the genome of *Buchnera*, which is nonmotile[c]. Second, increased levels of genetic drift, resulting from the population structure imposed by obligate association with hosts, could result in the fixation of missense mutations within genes that are beneficial but unnecessary. For example, numerous genes underlying DNA repair and genes enhancing transcription and translation are lacking in bacteria with small genomes. High levels of genetic drift in endosymbionts and pathogens are supported by a number of other features including rapid polypeptide evolution[r], lack of codon preferences[r], and low levels of polymorphism within species[s]. Third, the habitats within hosts can be relatively benign, resulting in less stringent selection on the efficiency of basic cellular processes. Under any of these changes, fixation of mutations destroying gene functionality increases due to insufficient purifying selection.

Under Hypotheses 1 and 2, small genomes are expected to be more streamlined than large-genome counterparts. Whether driven by greater selection or by increased mutational pressure, functionless regions should be more effectively eliminated in the smaller genomes. By contrast, Hypothesis 3 does not predict that smaller genomes will also be more tightly packed. Under this hypothesis, deletional bias is ubiquitous in larger and smaller genomes, and smaller genomes differ in the extent to which selection counters mutation. Indeed, at some stages one might predict that genomes undergoing reduction will have more regions with no function, because they can retain pseudogenes sequences that have been rendered functionless but have not been entirely eliminated by selection.

As an initial test of these predictions, we examined the lengths of spacer regions across the sequenced bacterial genomes. We consider that a large proportion of these spacers are functionless, an assumption that is supported by the fact that their lengths are highly variable, their sequences are not conserved across strains or species, and some are formed by remnants of eroded pseudogenes (Table 2). Across the sequenced genomes, there is no correlation between spacer length and genome size (Fig. 3), suggesting that genome reduction is not the result of direct selection favoring elimination of functionless DNA.

In many cases, fully sequenced small-genome organisms are phylogenetically distant from fully sequenced large-genome organisms, making direct comparisons of spacers less compelling. However, *Buchnera aphidicola*, the obligate endosymbiont of aphids, has a very small genome (0.64 Mb) and is closely related to *E. coli*, which has a relatively large genome (4.6–5.5 Mb) (Refs c,f,t). Many regions of synteny are maintained between *E. coli* and *Buchnera* and, within these, are instances of spacers flanked by the same genes in both species. We compared the lengths of corresponding spacers in *Buchnera* and *E. coli* to determine the pattern of change in orthologous regions. Although segments of the *Buchnera* genome are

genome sizes within a bacterial species is often large enough to obscure placement into a discrete size class[13]. Third, phylogenetic analyses reveal that bacteria with the smallest genomes are derived from bacteria with larger genomes[14].

Deletional bias in bacterial genomes

Although duplications and resulting paralogous genes are apparent in many bacterial genomes[12], there is growing evidence that the primary route through which bacterial species obtain new genes is by lateral transfer[15–17]. If lateral gene transfer is an ongoing process, why are bacterial genomes compact and not ever-expanding because of the influx of foreign sequences? The obvious answer is that lineages must undergo the inactivation and loss of genes, and the elimination of the corresponding DNA that made up the genes. This could result if the mutational process driving the structural evolution of chromosomes is biased towards DNA loss. This bias could be of two types. Large deletions might remove one or more genes in a single event. Alternatively, a gene could first lose function through point or frameshift mutations to generate a pseudogene that is eroded and eliminated by subsequent small deletions. For deletional bias to lead to the loss of genes and the corresponding DNA, selection on those genes must be ineffective owing to low selection coefficients or to small population size.

Table I. Lengths of spacers in *Buchnera aphidicola* (*A. pisum*) and *Escherichia coli* MG1655

		Number	Mean length (bp)	Range (bp)	Median (bp)
Orthologous spacers[a]	*E. coli*	194	30.8 ± 3.6	0–315	12
	Buchnera	194	31.7 ± 3.1	0–319	15
All spacers	*E. coli*	4403	120.4 ± 2.0	0–1730	63
	Buchnera[b]	586	117.6 ± 6.0	0–1318	74

[a]Spacers flanked by the same coding genes, excluding spacers with annotated regulatory regions.
[b]Excluding spacers flanking annotated pseudogenes.

rearranged on a chromosomal scale, local gene order and orientation is maintained for many genome fragments. We identified all syntenic fragments and determined the spacer lengths wherever both flanking genes were present in both species (spacers containing an annotated functional region were not included). Although spacers can contain regulatory regions that could be under selection, portions of spacers consist of nonfunctional DNA, as supported by their extreme AT bias, lack of sequence conservation and loss of promoter regions in *Buchnera*[c,u].

The average length and overall size range of orthologous spacers are virtually identical in the two species and show no trend towards reduction in *Buchnera* (Table I; T value = –0.78, P = 0.82; n = 194). This test could be biased if spacers contain regions affecting gene expression, raising the possibility that spacer size is preserved by selection in one or both lineages. However, because preservation of promoter sequences appears to be stronger in *E. coli*, non-neutrality of spacers would bias the comparison towards smaller spacers in *Buchnera*, a bias that would have produced an opposite trend and weighed against Hypothesis 3. *Buchnera* lacks promoters in more than 35% of cases in which the orthologous *E. coli* region is known to contain them, and many ribosome-binding Shine–Delgarno sequences are deteriorated as well. Despite stronger selective constraints acting on *E. coli* spacers, the orthologous spacers in *Buchnera* have not decreased in size. Thus, the test is conservative and not biased toward smaller spacers in *Buchnera*, providing evidence against direct selection as the basis for reduction in *Buchnera* genome size.

References

a Andersson, S.G.E. and Kurland, C.G. (1995) Genomic evolution drives the evolution of the translation system. *Biochem. Cell. Biol.* 73, 775–787
b Maniloff, J. (1996) The minimal cell genome: 'On being the right size'. *Proc. Natl. Acad. Sci. U. S. A.* 93, 10004–10006
c Shigenobu, S. *et al.* (2000) Genome sequence of the endocellular bacterial symbiont of aphids *Buchnera* sp. APS. *Nature* 407, 81–86
d Wernegreen, J.J. *et al.* (2000) Decoupling of genome size and sequence divergence in a symbiotic bacterium. *J. Bacteriol.* 182, 3867–3869
e Casjens, S. (1998) The diverse and dynamic structure of bacterial genomes. *Annu. Rev. Genet.* 32, 339–377
f Akman, L. and Aksoy, S. (2001) A novel application of gene arrays: *Escherichia coli* array provides insight into the biology of the obligate endosymbiont of tsetse flies. *Proc. Natl. Acad. Sci. U. S. A.* 98, 7546–7551
g Wernegreen, J.J. *et al.* (2000) Decoupling of genome size and sequence divergence in a symbiotic bacterium. *J. Bacteriol.* 182, 3867–3869
h Cavalier-Smith, T. (1987) The simultaneous symbiotic origin of mitochondria, chloroplasts and microbodies. *Ann. New York Acad. Sci.* 503, 53–71
i Samuelsson, T. and Borén, T. (1992) Evolution of macromolecule synthesis. In *Mycoplasmas: Molecular Biology and Pathogenesis* (Maniloff, J. *et al.* eds), pp. 575–591, ASM Press
j Maniloff, J. (1996) The minimal cell genome: 'On being the right size'. *Proc. Natl. Acad. Sci. U. S. A.* 93, 10004–10006
k Vellai, T. and Vida, G. (1999) The origin of eukaryotes: the difference between prokaryotic and eukaryotic cells. *Proc. R. Soc. Lond. Ser. B* 266, 1571–1577
l Fraser, C.M. *et al.* (1995) The minimal gene complement of *Mycoplasma genitalium*. *Science* 270, 397–403
m Casjens, S. *et al.* (2000) A bacterial genome in flux: the twelve linear and nine circular extrachromosomal DNAs in an infectious isolate of the Lyme disease spirochete *Borrelia burgdorferi*. *Mol. Microbiol.* 35, 490–516
n Parkhill, J. *et al.* (2000) Complete DNA sequence of a serogroup A strain of *Neisseria meningitidis* Z2491. *Nature* 404, 502–506
o Fraser, C.M. *et al.* (1997) Genomic sequence of a Lyme disease spirochaete, *Borrelia burgdorferi*. *Nature* 390, 580–586
p Fraser, C.M. *et al.* (1998) Complete genome sequence of *Treponema pallidum*, the syphilis spirochete. *Science* 281, 375–388
q Moran, N.A. (1996) Accelerated evolution and Muller's ratchet in endosymbiotic bacteria. *Proc. Natl. Acad. Sci. U. S. A.* 93, 2873–2878
r Wernegreen, J.J. and Moran, N.A. (1999) Evidence for genetic drift in endosymbionts (*Buchnera*): analyses of protein-coding genes. *Mol. Biol. Evol.* 16, 83–97
s Funk, D.J. *et al.* (2001) Intraspecific variation in symbiont genomes: bottlenecks and the aphid-buchnera association. *Genetics* 157, 477–489
t Charles, H. and Ishikawa, I. (1999) Physical and genetic map of the genome of *Buchnera*, the primary endosymbiont of the pea aphid, *Acyrthosiphon pisum*. *J. Mol. Evol.* 16, 1586–1598
u Clark, M.A. *et al.* (1998) Sequence analysis of a 34.7-kb DNA segment from the genome of *Buchnera aphidicola* (endosymbiont of aphids) containing *groEL*, *dnaA*, the *atp* operon, *gidA*, and *rho*. *Curr. Microbiol.* 36, 158–163

The existence of a deletional bias can be evaluated by comparing known pseudogenes with their functional counterparts in the same or other taxa. Because pseudogenes are functionless and under no selective constraints, the profile of mutations in these regions reflects the actual frequency of different mutational classes[18–20]. In *Rickettsia* pseudogenes, deletions are more common and longer than insertions[20,21]. To assess the extent of deletional bias in bacteria, we surveyed cases of known pseudogenes in a broad taxonomic range of bacteria (Fig. 2). In every case, deletions are more frequent than insertions and have a much greater net effect on genome size. Here, we are assuming that the observed deletional frequency in pseudogenes reflects the pattern of spontaneous mutation and is not elevated by selection on genome size itself. This assumption is based on the observation that a single deletion of one or a few basepairs will alter the size of the genome by a factor of ~10^{-6}. Thus, even if DNA replication were to limit the rate of cell division (Box 1), the selection coefficient on small deletions would be so small as to approximate zero[19].

The comparisons in Fig. 2 are conservative estimates of the amounts of deleted DNA because they do not take into account large deletions that remove all or most of a gene, such that it is absent, nor genes that have mutated to a degree such that they

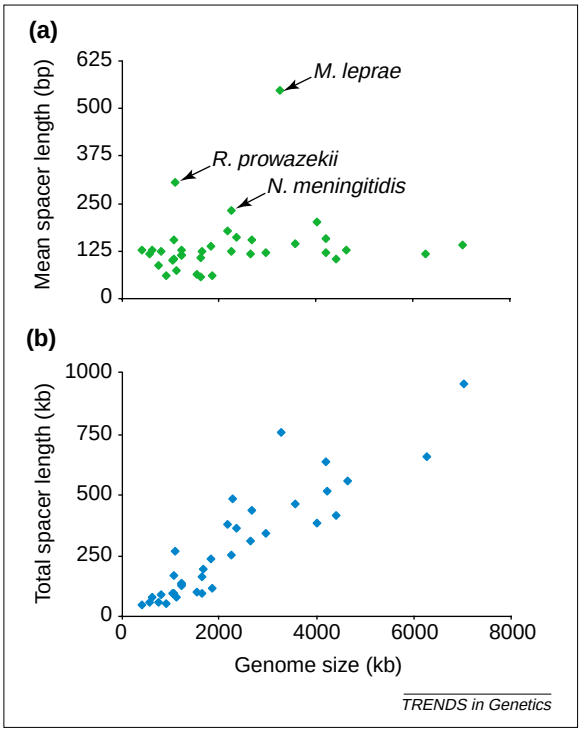


Fig. 3. The relationship between genome size and noncoding DNA in eubacteria. (a) Relationship between genome size and mean spacer length. The three species with exceptionally long spacers are bacteria with high numbers of pseudogenes: *Mycobacterium leprae*, *Rickettsia prowazekii* and *Neisseria meningitidis*. (b) Relationship between genome size and summed spacer length in eubacteria ($R^2 = 0.762$). Taxa are listed in Table 1.

are unrecognizable. For example, intergenic spacers in *Mycobacterium leprae* are, on average, more than five times longer than those in its sister species, *Mycobacterium tuberculosis* (Table 1), suggesting the presence of additional pseudogenes eroded beyond recognition. Thus, considering that our analysis is limited to the more conservative comparisons, the propensity for deletion is profound. Analyses of the

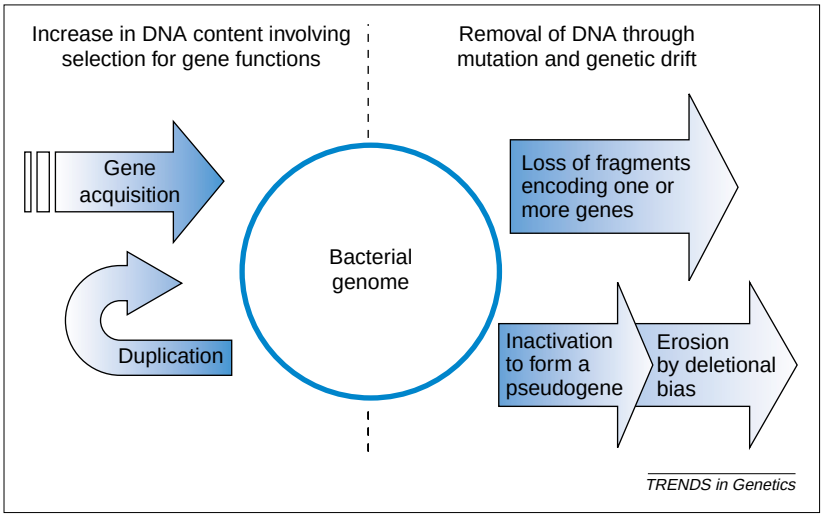


Fig. 4. Processes involved in the evolution of genome size in bacteria. New sequences are acquired by DNA transfer and gene duplication, the former being the predominant mode of DNA increase within most species. DNA loss can be produced by large deletions eliminating one or more genes in a single event, or by loss of function followed by subsequent deletions of the resulting pseudogenes.

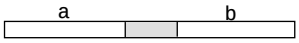
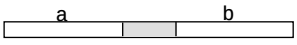
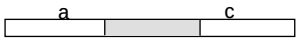
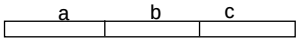
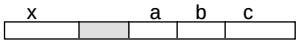
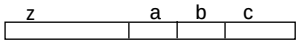
total gene contents in natural isolates of *Escherichia coli* show that large deletions also are sometimes fixed within lineages, removing several genes at once[22]. Similarly, at least 25 long deletions were identified in different strains of *M. tuberculosis*, with one event removing as many as 16 open reading frames (ORFs)[36]. Large deletions have also been inferred for intracellular bacterial symbionts[23].

These observations suggest that there is a constant eroding force of deletion that must be counterbalanced by selection on gene function. This requirement for selection to maintain sequences offers a general explanation for the paucity of nonfunctional DNA in bacteria. In addition, it has often been proposed that genome size, especially that of bacteria in intracellular habitats, is itself the object of selection; that is, that the tight packing and small size of bacterial genomes is an adaptation to promote efficiency or competitiveness during replication[14,24–28]. This view is consistent with any mutational pattern, including one biased towards deletion, but requires that fixation of individual deletions is driven by selection favoring smaller genomes. As noted above, small deletions would not confer substantial reduction in chromosome replication rate[19], and the median size of the deletions in the pseudogene dataset is only 3 bp (Fig. 2). Analyses of sequenced genomes of different sizes suggest that genome reduction in intracellular symbionts does not result from selection (Box 2).

Degraded genes and persistent pseudogenes

The overall proportion of noncoding DNA is fairly similar (around 12%) among the fully sequenced genomes of different taxonomic groups (Fig. 3). These noncoding sequences provide a glimpse of the effects of deletional bias. For example, the large number of gene regions conserved between *Buchnera* and *E. coli* allows a direct assessment of how the extreme reduction of genomes occurs[29]. Within regions that show synteny with *E. coli*, spacers from *Buchnera* can be divided into two categories: 'ancient' spacers, which are those flanked on both sides by the same genes in *E. coli* and *Buchnera*, and 'amended' spacers, which are those in which one or more genes are absent from *Buchnera*, but with flanking genes that have the same gene order as in *E. coli* (Table 2). The ancient spacers, which are presumed to be orthologous sequences, are of similar average length in both species, despite the fact that *Buchnera* has undergone massive genome reduction (Box 2). By contrast, the amended spacers are over three times longer than the ancient spacers (Table 2). The most likely explanation for the additional length is that the amended spacers contain the residue of ancestral genes and represent highly eroded pseudogenes. The same length increase is found in *Buchnera* spacers located between regions of synteny with *E. coli*. These could

Table 2. Spacer lengths in *Buchnera* (*A. pisum*)

Spacers within *Buchnera*			Number	Mean length (bp)
Ancient spacers[a]		*Buchnera*	270	55.1
		*E. coli*		
Amended spacers[b]		*Buchnera*	165	188.0
		*E. coli*		
Spacers at termini of syntenic regions[c]		*Buchnera*	162	188.4
		*E. coli*		

[a]Spacers with the same flanking genes in both *Buchnera* and *Escherichia coli*.
[b]Spacers with flanking genes that conserve ancestral order and that occur where one or more additional genes are present in *E. coli*.
[c]Spacers at the ends of rearranged fragments.

be degraded pseudogenes formed during chromosome rearrangements. Similar instances of eroded pseudogenes can be observed in *M. leprae*, based on comparisons with *M. tuberculosis*[30]. The retention of such augmented functionless regions in a highly reduced genome suggests that the loss of gene function is not strictly coupled with the loss of DNA.

Persistence of DNA following loss of gene function is also evident from analyses of bacterial genomes. For example, the eight pseudogenes annotated in *Buchnera* of *Acyrthosiphon pisum* are not appreciably shorter than their functional counterparts in *E. coli* (T-value = –0.447, $P = 0.669$). In one case, the *cmk* pseudogene shares an identical 16-bp deletion in *Buchnera* from *A. pisum* and in *Buchnera* from *Schizaphis graminum*, which diverged at least 50 million years ago[31]. This implies that *cmk* has been a pseudogene for at least that long, yet there is little other reduction in the length of this pseudogene in either species. Other examples of ancient pseudogenes occur in *Buchnera* from *Diuraphis noxia*, in which *trpEG* pseudogenes are shared among isolates from diverse geographic locations[32]. In *R. prowazekii*, only 76% of the genome encodes proteins (Table 1), and at least 12 pseudogenes have been retained for long periods within this clade of intracellular pathogens[21,33]. Other examples of persistent pseudogenes are known from the spirochete *Borrelia burgdorferi*[34], *Neisseria meningitidis*[35] and *M. leprae*[30].

A model for genome size evolution in bacteria

These observations, from analyses of complete genomic sequences, suggest a simple model for the evolution of bacterial genome size based on the outcome of several opposing forces (Fig. 4). Deletional bias and genetic drift cause genomes to contract, whereas selection on gene function causes genomes to maintain DNA. Accretions in genome size depend on either duplications or the acquisition of exogenous DNA, but these events are only effective if the new genes confer some benefit (or if the DNA is somehow parasitic). Small-genome bacteria, by being sequestered in hosts, might have a reduced opportunity for gene uptake, and they might lose pathways required to incorporate exogenous DNA. Thus, the larger genome size of free-living bacteria could reflect more-frequent acquisition of new genes, greater need for metabolic versatility or more-effective selection for the maintenance of weakly beneficial genes.

The loss of DNA occurs both through large deletions spanning multiple loci[23,36] and through small deletions of one or a few nucleotides (Fig. 2)[20,21]. The relative importance of these two routes will vary among bacterial lineages. On the basis of comparisons of *Buchnera* to free-living relatives, early stages in genome reduction involved large deletions spanning as many as 50 genes[23]. By contrast, modern *Buchnera* have stable genome sizes[37], presumably because large deletions, involving loss of genes, are lethal. The final reduction of spacers will be slow, because most large deletions will overlap coding or promoter regions and thus be selected against, and small deletions remove DNA only gradually.

Not all of the processes depicted in Fig. 4 occur at the same rates, and some bacterial genomes might be in the process of increasing or decreasing in size. In particular, the transition to intracellular life probably imposes an abrupt decrease in the effectiveness of selection on many genes, whereas drift and deletional bias remove nonfunctional DNA at low rates. The differences among organisms in spacer length and the persistence of recognizable pseudogenes will reflect the time elapsed since the lifestyle shifts occurred and the frequency of deletion events. Thus, immediately following a change to a lifestyle in which selection is less effective in preserving gene function, many pseudogenes will appear. The unusually large proportion of pseudogenes in *M. leprae*[30] could represent an early stage in the process of genome reduction. *Buchnera*, *Mycoplasma genitalium* and *R. prowazekii*, which have relatively short spacers and fewer pseudogenes, might be at more-advanced stages in this process. Thus, at any point through evolutionary time, deletional bias appears to be a major force shaping bacterial genomes, perhaps explaining their small size and tight packing of their genes.

Acknowledgements
We thank Jacob Russell, Alex Kondrashov and two anonymous reviewers for helpful comments. The work was supported by grants GM55535 and GM56120 from NIH and grant DEB-9978518 from NSF.

References

1 Herdman, M. (1985) The evolution of bacterial genomes. In *The Evolution of Genome Size* (Cavalier-Smith, T., ed.), pp. 37–68, John Wiley & Sons
2 Casjens, S. (1998) The diverse and dynamic structure of bacterial genomes. *Annu. Rev. Genet.* 32, 339–377
3 Gregory, T.R. and Hebert, P.D.N. (1999) The modulation of DNA content: proximate causes and ultimate consequences. *Genome Res.* 9, 317–324
4 Fraser, C.M. *et al.* (1995) The minimal gene complement of *Mycoplasma genitalium*. *Science* 270, 397–403
5 Kaneko, T. *et al.* (2000) Complete genome structure of the nitrogen-fixing symbiotic bacterium *Mesorhizobium loti*. *DNA Res.* 7, 331–338
6 Goffeau, A. (1996) Life with 6000 genes. *Science* 274, 546
7 Lander, E.S. *et al.* (2001) A physical map of the human genome. *Nature* 409, 934–941
8 Venter, J.C. *et al.* (2001) The sequence of the human genome. *Science* 291, 1304–1351
9 Wallace, D.C. and Morowitz, H.J. (1973) Genome size and evolution. *Chromosoma* 40, 121–126
10 Riley, M. and Anilionis, A. (1978) Evolution of the bacterial genome. *Annu. Rev. Microbiol.* 32, 519–560
11 Huynen, M. and Bork, P. (1998) Measuring genome evolution. *Proc. Natl. Acad. Sci. U. S. A.* 95, 5849–5856
12 Jordan, I.K. *et al.* (2001) Lineage-specific gene expansions in bacterial and archaeal genomes. *Genome Res.* 11, 555–565
13 Bergthorsson, U. and Ochman, H. (1995) Heterogeneity of genome sizes among natural isolates of *Escherichia coli*. *J. Bacteriol.* 177, 5784–5789
14 Andersson, S.G.E. and Kurland, C.G. (1998) Reductive evolution of resident genomes. *Trends Microbiol.* 6, 263–268
15 Ochman, H. *et al.* (2000) Lateral gene transfer and the nature of bacterial innovation. *Nature* 405, 299–304
16 Nelson, K.E. *et al.* (1999) Evidence for lateral gene transfer between Archaea and Bacteria from genome sequence of *Thermotoga maritima*. *Nature* 399, 323–329
17 Lawrence, J.G. and Ochman, H. (1998) Molecular archaeology of the *Escherichia coli* genome. *Proc. Natl. Acad. Sci. U. S. A.* 95, 9413–9417
18 Gojobori, T. *et al.* (1982) Patterns of nucleotide substitution in pseudogenes and functional genes. *J. Mol. Evol.* 18, 360–369
19 Petrov, D.A. and Hartl, D.L. (1998) High rate of DNA loss in the *Drosophila melanogaster* and *Drosophila virilis* species groups. *Mol. Biol. Evol.* 15, 293–302
20 Andersson, J.O. and Andersson, S.G.E. (2001) Pseudogenes, junk DNA, and the dynamics of *Rickettsia* genomes. *Mol. Biol. Evol.* 18, 829–839
21 Andersson, J.O. and Andersson, S.G.E. (1999) Genome degradation is an ongoing process in *Rickettsia*. *Mol. Biol. Evol.* 16, 1178–1191
22 Ochman, H. and Jones, I.B. (2000) Evolutionary dynamics of full genome content in *Escherichia coli*. *EMBO J.* 19, 6637–6643
23 Ochman, H. and Moran, N.A. (2001) Genes lost and genes found: evolution of bacterial pathogenesis and symbiosis. *Science* 292, 1096–1098
24 Cavalier-Smith, T. (1987) The simultaneous symbiotic origin of mitochondria, chloroplasts and microbodies. *Ann. New York Acad. Sci.* 503, 53–71
25 Samuelsson, T. and Borén, T. (1992) Evolution of macromolecule synthesis. In *Mycoplasmas: Molecular Biology and Pathogenesis* (Maniloff, J. *et al.*, eds), pp. 575–591, ASM Press
26 Andersson, S.G.E. and Kurland, C.G. (1995) Genomic evolution drives the evolution of the translation system. *Biochem. Cell. Biol.* 73, 775–787
27 Maniloff, J. (1996) The minimal cell genome: 'On being the right size'. *Proc. Natl. Acad. Sci. U. S. A.* 93, 10004–10006
28 Vellai, T. and Vida, G. (1999) The origin of eukaryotes: the difference between prokaryotic and eukaryotic cells. *Proc. R. Soc. London Ser. B* 266, 1571–1577
29 Shigenobu, S. *et al.* (2000) Genome sequence of the endocellular bacterial symbiont of aphids *Buchnera* sp. APS. *Nature* 407, 81–86
30 Cole, S.T. *et al.* (2001) Massive gene decay in the leprosy bacillus. *Nature* 409, 1007–1011
31 Clark, M.A. *et al.* (1999) Sequence evolution in bacterial endosymbionts having extreme base composition. *Mol. Biol. Evol.* 16, 1586–1598
32 Wernegreen, J.J. and Moran, N.A. (2000) Decay of mutualistic potential in aphid endosymbionts through silencing of biosynthetic loci: *Buchnera* of *Diuraphis*. *Proc. R. Soc. London Ser. B* 267, 1423–1431
33 Andersson, S.G.E. *et al.* (1998) The genome sequence of *Rickettsia prowazekii* and the origin of mitochondria. *Nature* 396, 133–140
34 Casjens, S. *et al.* (2000) A bacterial genome in flux: the twelve linear and nine circular extrachromosomal DNAs in an infectious isolate of the Lyme disease spirochete *Borrelia burgdorferi*. *Mol. Microbiol.* 35, 490–516
35 Parkhill, J. *et al.* (2000) Complete DNA sequence of a serogroup A strain of *Neisseria meningitidis* Z2491. *Nature* 404, 502–506
36 Kato-Maeda, M. *et al.* (2001) Comparing genomes within the species *Mycobacterium tuberculosis*. *Genome Res.* 11, 547–554
37 Wernegreen, J.J. *et al.* (2000) Decoupling of genome size and sequence divergence in a symbiotic bacterium. *J. Bacteriol.* 182, 3867–3869