A selective force favoring increased G+C content in bacterial genes

Rahul Raghavan, Yogeshwar D. Kelkar, and Howard Ochman[1]

Department of Ecology and Evolutionary Biology, Yale University, New Haven, CT 06520

Edited by Eviatar Nevo, Institute of Evolution, Haifa, Israel, and approved July 31, 2012 (received for review April 3, 2012)

Bacteria display considerable variation in their overall base compositions, which range from 13% to over 75% G+C. This variation in genomic base compositions has long been considered to be a strictly neutral character, due solely to differences in the mutational process; however, recent sequence comparisons indicate that mutational input alone cannot produce the observed base compositions, implying a role for natural selection. Because bacterial genomes have high gene content, forces that operate on the base composition of individual genes could help shape the overall genomic base composition. To explore this possibility, we tested whether genes that encode the same protein but vary only in their base compositions at synonymous sites have effects on bacterial fitness. *Escherichia coli* strains harboring G+C-rich versions of genes display higher growth rates, indicating that despite a pervasive mutational bias toward A+T, a selective force, independent of adaptive codon use, is driving genes toward higher G+C contents.

bacterial adaptation | genome evolution | mutational patterns

Bacterial genomes are highly variable in their overall base compositions, with sequenced genomes ranging from 13% to 75% G+C (1, 2). Genomic base composition in bacteria was the first genetic character to be considered selectively neutral, in the sense that the variation is not adaptive but instead due solely to interspecies differences in the mutational process (3, 4). This view was bolstered by sequence information showing that the base compositional differences among bacteria were most pronounced at synonymous and noncoding positions, sites that are thought to be under the least selective constraints (5). Furthermore, repeated attempts to link genomic base composition with an environmental factor or other selective processes have met with limited success (6–10).

Recently, however, studies based on the comparative analyses of gene sequences have challenged the notion that base composition is driven solely by mutational biases. These studies revealed that mutation is universally biased toward A+T, suggesting selection as an agent that maintains the contemporary base compositions in bacterial genomes (11–13). A difficulty with ascribing a role for selection in determining genomic base composition is that the selective force must be sufficiently strong to operate on a single-nucleotide change despite the fact that each change makes only a tiny contribution toward the overall G+C content of the genome. However, because bacterial genomes are composed primarily of protein-coding genes (14), a selective force that acts on each gene to increase its G+C content can cumulatively influence the overall genomic base composition. To test this possibility, we examined the fitness effects of protein-coding gene variants that differed only in their base composition at synonymous codon positions. For two genes, both of which encode proteins that are neither native nor physiologically relevant to *Escherichia coli*, we detected a strong and significant association between G+C contents and bacterial fitness. G+C-enrichment of mRNAs is observed in a majority of bacterial species, and the G+C content of functionless, nontranscribed regions in the *E. coli* genome are decidedly lower than that of fourfold degenerate sites in protein-coding genes. Taken together, these data indicate that selection operating on the base composition of individual coding regions guides genomic base composition in bacteria.

Results

Base Composition of an Expressed Gene Impacts *E. coli* Growth Rate. *E. coli* strains containing variants of a plasmid-borne green fluorescent protein (GFP) gene that differ in base composition (40.4–53.7% G+C; averaging 126 substitutions between gene pairs) were tested for growth at various time points after GFP induction. At 2 h postinduction, there is a significant association between the G+C composition of the expressed GFP gene and bacterial generation time, with strains expressing genes of higher G+C contents exhibiting significantly higher growth rates; the effect becomes more pronounced at later time points (Fig. 1). Performing the same experiment without IPTG induction of the GFP genes abolished the association between bacterial growth rate and the G+C contents of the GFP genes (Fig. S1), indicating that the effect of base composition on growth rate requires gene expression. Additionally, there is no association between the base composition of a GFP gene and its level of protein production (as measured by GFP fluorescence) at any point during the growth experiment (Fig. S2), nor is there an association between codon adaptation index (CAI) of the GFP gene variants and growth rates (Fig. 2).

Association Between Base Composition and Growth Rate Is Observed for Multiple Genes. To determine whether the association between base composition and *E. coli* growth rate was in some way limited to the GFP constructs, we tested a set of *Bacillus* phage ϕ29 DNA polymerase genes that varied over a more confined range of G+C contents (43.7–47.2% G+C; averaging 213 substitutions between gene pairs). Again, the sequence variants of this gene have similar CAI values, and each encodes the identical protein. As observed with the GFP constructs, there is a significant association between the base compositions of the ϕ29 DNA polymerase gene and bacterial growth rate (Fig. 3), and the association is only apparent when the gene is induced (Fig. S3).

Obstruction of Translation Diminishes the Effects of Base Composition on Growth Rate. To establish whether the effect of transcript base composition on bacterial growth rate requires translation, we tested the growth rates of *E. coli* strains that harbor constructs in which translation was prevented through the removal of the ribosome-binding site and start codon, and the introduction of stop codons at the 5′ ends, of the GFP gene. When transformed with these constructs, an association between the base composition of induced genes and bacterial growth rate remains, but it is no longer significant (Fig. 4). This trend suggests that there are

Author contributions: R.R. and H.O. designed research; R.R. and Y.D.K. performed research; R.R., Y.D.K., and H.O. analyzed data; and R.R., Y.D.K., and H.O. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

[1]To whom correspondence should be addressed. E-mail: howard.ochman@yale.edu.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1205683109/-/DCSupplemental.

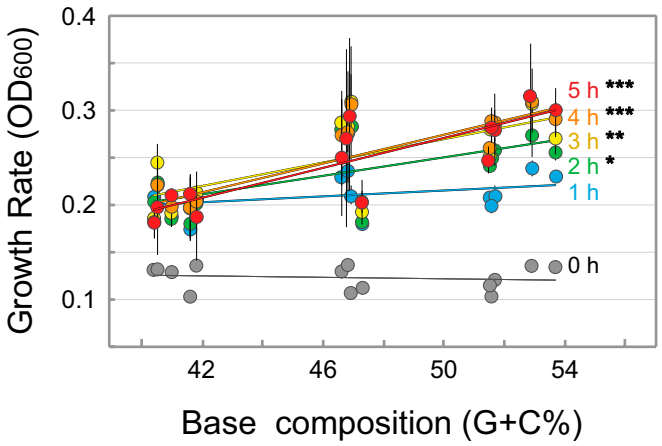

Fig. 1. Growth rates of isogenic strains expressing GFP genes of different G+C contents. Growth rate (OD600) was measured hourly for 5 h after induction with 1 mM IPTG. A significant association between %G+C and growth rate is observed at 2 h after induction and is evident at all later time points (0 h, $r^2 = 0.03$; 1 h, $r^2 = 0.15$; 2 h, $r^2 = 0.40$; 3 h, $r^2 = 0.47$; 4 h, $r^2 = 0.67$; 5 h, $r^2 = 0.72$). Asterisks denote level of significance: $*P < 0.05$, $**P < 0.01$, $***P < 0.001$. Values represent the mean ± SD of three replicates. For detailed results of regressions, see Table S1.

other unrecognized selective pressures that counteract the A+T-biased mutational process.

G+C Enrichment of mRNAs Occurs in the Majority of Bacterial Species. If there is widespread selection across bacterial species for G+C-rich transcripts, then synonymous sites can no longer be considered as evolving in a neutral manner and their G+C contents are expected to be higher than those of noncoding regions within the corresponding genome. We examined the association between nucleotide compositions of fourfold degenerate sites (GC4) and noncoding genomic regions (GCnc) in a phylogenetically diverse sample of sequenced bacterial genomes and found a positive nonlinear relationship in which GC4 was higher than GCnc in the vast majority of genomes with base compositions

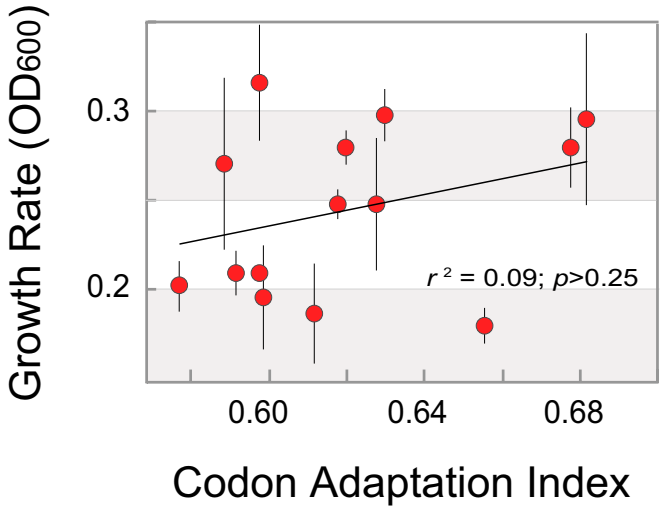

Fig. 2. No significant association between codon adaptation index (CAI) and growth rate. Growth of *E. coli* strains expressing GFP genes with CAI values ranging from 0.58 to 0.68 (calculated using CAIcal; ref. 15) were monitored by measuring OD600 every hour. Values shown were measured at 5 h after gene induction and represent the mean ± SD of three replicates.

over 40% G+C (Fig. 5), suggesting that selection acts to enrich transcripts with G/C nucleotides. The converse was detected (i.e., GC4 < GCnc) in most bacterial genomes with base compositions less than 40% G+C.

To further understand the influence of transcription on genomic nucleotide composition, we measured the G+C content of nontranscribed intergenic regions situated between the transcription start sites (TSSs) of 222 divergently transcribed adjacent genes in *E. coli*. We found that the base composition of these intergenic regions averaged only 44.1% G+C, even after accounting for (and removing) the A+T-rich −10 and −35 promoters contained within these regions (Fig. S4). This base composition is significantly lower than GC4 ($P < 10^{-16}$, two-sample t test; Fig. S4) but higher than the equilibrium G+C content of 32% (0.28–0.37; 95% CI) predicted for *E. coli* (16).

Discussion

We demonstrate that the overall base composition of highly expressed genes affects bacterial fitness: Strains expressing equivalent genes of higher G+C contents have significantly faster doubling times. These findings shed light on the evidence that G+C content of many bacterial genomes is higher than that predicted on the basis of mutational patterns, but only to the extent that selection is operating at the level of individual protein-coding sequences. The interplay between mRNA phenotype and the translation machinery partially explains the observed phenomenon based on several observations: (*i*) The association between increasing G+C contents and growth rates is observed only in genes that are both transcribed and translated (Fig. 4); (*ii*) there is no correspondence between the base composition of a GFP gene and its level of protein production (Fig. S2); (*iii*) mRNAs with higher G+C contents exhibited higher stability (Fig. S5); and (*iv*) mRNAs are G+C-enriched at fourfold degenerate sites relative to noncoding sites in the majority of bacteria (Fig. 5) and to regions shown experimentally to be nontranscribed in the *E. coli* genome (Fig. S4).

The degree of secondary structure near the 5′-end of mRNAs has been shown to negatively influence the levels of protein expression (17–19), which is also observed for the GFP sequence variants that we tested (Fig. S6*A*). However, the stability of mRNA secondary structures near their 5′ ends (-4 to +37) was not correlated with bacterial growth rates (Fig. S6*B*), ruling out the possibility that the higher fitness of G+C-rich GFP variants results from suppressing the translation of wasteful proteins. Other features of a coding region, aside from those that specify mRNA stability and the amino acid sequence of its encoded protein, may influence its nucleotide sequence, as shown recently for the tendency of bacterial genes to avoid internal Shine–Delgarno sequences (20).

Examined across bacterial genomes, it has been recorded by us and by others (e.g., refs. 5, 13, 21, and 22) that the G+C contents of synonymous sites are often higher than those of noncoding regions (Fig. 5). Because synonymous sites were previously considered to be evolving neutrally in all but the most highly expressed bacterial genes, this pattern has been taken as evidence that noncoding regions are under some form of selection for lower G+C contents. However, our results indicate the converse: Selection instead serves to increase the G+C contents of synonymous sites. Because this selective force for higher G+C contents is operating on expressed sequences, we examined the base composition of spacers situated between divergent TSSs, regions with very low probability of expression (Fig. S4). The G+C content of this set of nontranscribed regions is substantially lower than GC4 but higher than the predicted equilibrium G+C% of 0.32 (16), suggesting that although the region may be evolving in a predominately neutral manner, it might still contain regulatory sequences that are under selection.

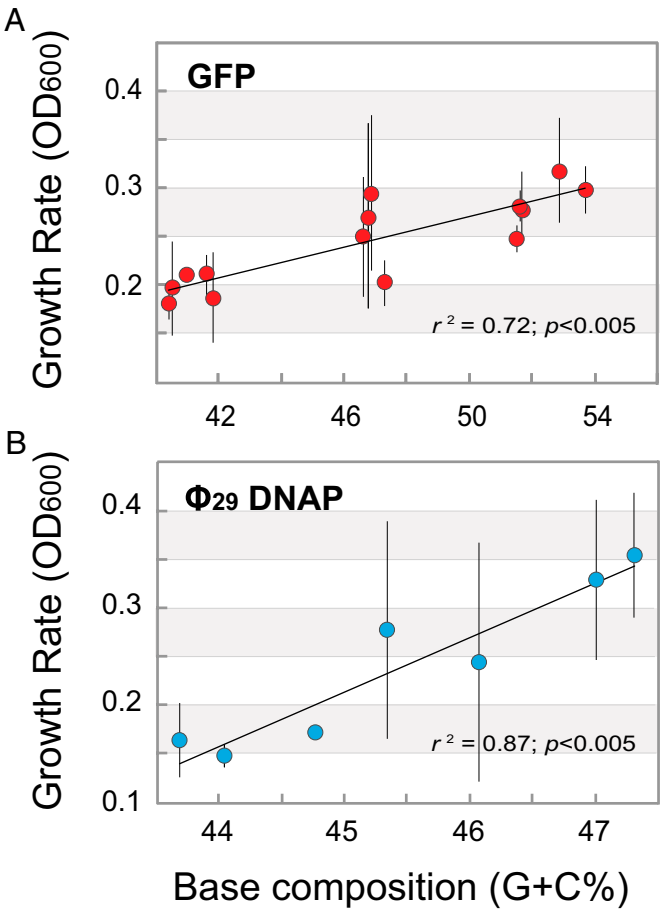

Fig. 3. Significant association between bacterial growth rate and base composition for genes spanning different ranges of G+C contents. Growth rates for *E. coli* strains expressing GFP gene variants with base compositions ranging from 40.4–53.7% G+C (*A*) and *Bacillus* phage ϕ29 DNA polymerase gene variants with base compositions ranging from 43.7–47.2% G+C (*B*), measured at 5 h after induction with 1mM IPTG. Values represent the mean ± SD of three replicates.

It is curious that the G+C content of noncoding sequences (GCnc) is higher than that of synonymous sites (GC4) in the most A+T-rich genomes (Fig. 5). Because A/T-to-G/C mutations at synonymous sites are rare in A+T-rich genomes (12, 13, 23), it is possible that the enrichment of A and U in mRNA transcripts is caused by the formation of secondary structures that are stabilized by A:U pairings, which occur much more readily than mutations leading to G:C pairings. Although the association between GC4 and GCnc is expected to become nonlinear at the limits of the distribution, a logistic curve best fits the relationship between GC4 and GCnc suggesting that there is cooperativity in the occurrence of complementary A:T or G:C pairings, as might be expected if there were compensatory changes that serve to stabilize mRNAs (24).

In sum, for two genes, neither of which encodes a physiologically relevant protein, we detected a strong and significant association between G+C content and bacterial fitness, indicating that a selective force within genes compensates for mutations that are naturally A+T-biased in bacteria. The elevated fitness of strains expressing mRNAs with high G+C content was most pronounced when genes are expressed at very high levels, suggesting that the selective forces responsible for this effect are low (12, 13), perhaps on par with those operating on adaptive codon bias (25). However, because bacterial genomes consist primarily of protein-coding regions, even a small selective effect that operates to change the base composition of individual genes in a common direction will have the effect of altering overall genomic base composition. Under this view, the extreme A+T richness of highly reduced genomes of host-restricted bacteria would result from the reduced efficacy of selection in these species (11, 26–28) rather than from changes in mutational patterns. Thus, the broad variation in genomic base composition among bacterial species reflects, in part, differences in population-level parameters that affect the efficacy of selection.

Materials and Methods

A set of 14 clones containing GFP genes that varied in their base compositions from 40.4% to 53.7% G+C was selected from those reported in Kudla et al. (18). A second set of plasmids containing the *Bacillus* phage ϕ29 DNA polymerase gene with base compositions ranging from 43.7% to 47.2% G+C were selected from those reported in Welch et al. (29). Within each set, the amino acid sequence of the protein encoded by all constructs was identical, and clones in each set were selected to represent a narrow range of CAI values for the *E. coli* host (CAI of GFP constructs = 0.58–0.68; CAI of ϕ29 constructs = 0.53–0.59) to abate the effects of selection on translational optimization (30). To further analyze the sole effect of base composition on bacterial fitness, we selected constructs with very different nucleotide sequences, even among those of similar base composition. For example, among the GFP constructs of low G+C content (40.4–41.8%), there are, on average, 58 substitutions; among those of medium G+C (46.6–47.3%), 121 substitutions; and among those of high G+C (51.5–53.7%), 99 substitutions. There are, on average, 126 synonymous substitutions between pairs of GFP constructs (240 codons), and 213 differences between the ϕ29 DNA polymerase constructs (575 codons). Genes were supplied in similar pET vectors (Novagen), with the gene of interest linked to a T7 promoter. To test constructs in which translation is prevented, we designed GFP genes in which the start codon was removed and stop codons introduced into their 5′ ends. These amplicons were cloned into a pET11a vector (Novagen) that was engineered, by XbaI and BamHI digestion, to lack its ribosome-binding site.

To express the plasmid libraries and monitor growth rates, we transformed constructs into *E. coli* BL21(DE3) (NEB), which has a chromosomal copy of T7 RNA polymerase under the control of a *lac*UV5 promoter. *E. coli* strain Lemo21(DE3) (NEB), which is similar to BL21(DE3) but can express T7 lysozyme under a rhamnose promoter, was used to measure mRNA stability.

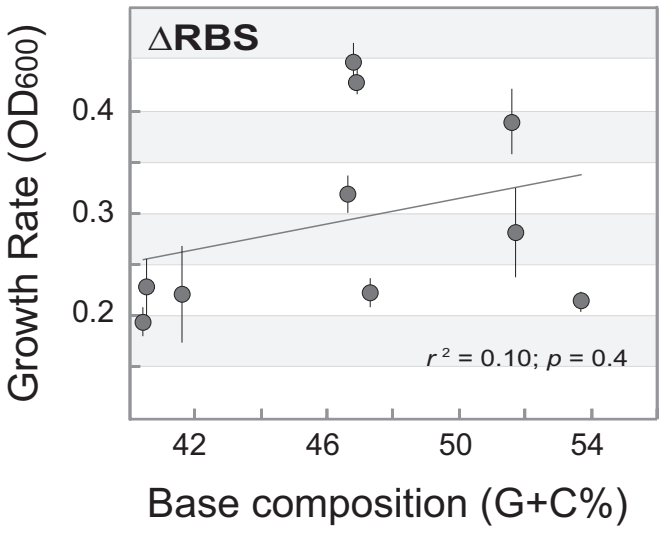

Fig. 4. Effect of translation on the association between base composition and bacterial growth rates. Growth of *E. coli* strains expressing GFP gene variants of different base compositions in which translation of the GFP gene was prevented by removing the ribosome binding sites (ΔRBS) and start codons. Growth rates measured at 5 h after induction and represent the mean ± SD of three replicates.

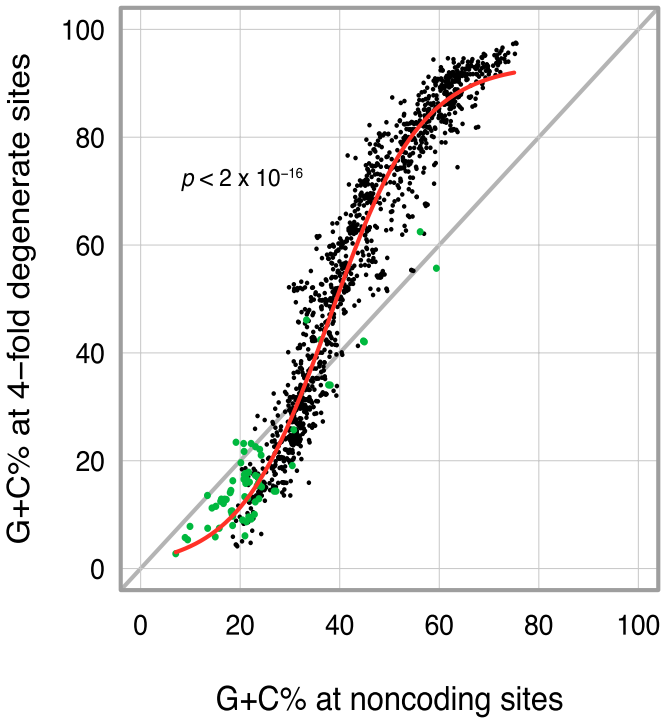

Fig. 5. Relationship between the average G+C content at fourfold degenerate sites of all protein-coding genes and the average G+C content of noncoding intergenic regions in fully sequenced bacterial genomes (n = 1430). Green circles denote bacteria with genome size of less than one megabase (n = 67), and the red line indicates the logistic regression model fitted to the data.

For growth rate experiments, 10 μL of an overnight culture were inoculated into 140 μL of LB medium containing 100 μg/mL ampicillin into 96-well plates and grown at 37 °C with constant shaking. After 1 h, 1 mM IPTG was added, and GFP fluorescence and optical density (OD) were measured each hour for 5 h on a Victor3 microplate reader (PerkinElmer). Regression analyses of bacterial growth rates and GFP fluorescence were performed on values averaged from three independent experiments.

For measurements of mRNA stability, Lemo21(DE3) cells were grown as above but induced with 20 μM IPTG for 2 h, followed by the induction of T7 lysozyme with 2 mM L-rhamnose. After 15 min, samples were collected and processed for quantification as described above. Quantitative PCR was performed with primers that target the 3′-end common to all of the GFP mRNAs, and transcript abundance was calculated from threshold cycle (Ct) after normalizing to the Ct obtained with 16S rRNA primers (dCt).

Genome sequences and annotations of 1,430 fully sequenced bacterial genomes were obtained from NCBI FTP server (ftp.ncbi.nih.gov). For each genome, noncoding regions were identified as those not encoding proteins, ribosomal RNA, and transfer RNA, and the fourfold degenerate sites of protein-coding genes were identified based on the standard genetic code. The G+C content of fourfold degenerate sites (GC4) was calculated for each genome as the proportion of all fourfold degenerate sites in the entire genome that are either G or C, and the G+C content of noncoding regions (GCnc) for each genomes was similarly calculated. The relationship between GC4 and GCnc (Fig. 5 and Fig. S7) was fitted with a logistic function, and the goodness of fit of this regression was measured using analysis of deviance with the nls package in R.

To examine nontranscribed regions that are putatively free of selective constraints, we located intergenic regions situated between experimentally verified, divergent transcription start sites in the *E. coli* MG1655 genome (31, 32). Any of these intergenic regions that contained noncoding RNAs (33) were removed before analysis. Because many promoter sequences are AT-rich and could bias nucleotide counts in short intergenic regions, the AT-rich σ70 promoter sequences (TATAAT, -10 region and TTGACA, -35 region) were not included in the calculations of G+C contents. RNA free energy was calculated using the mfold web server (34).

ACKNOWLEDGMENTS. We thank Joshua Plotkin and Grzegorz Kudla for providing the GFP constructs and Mark Welch for the *Bacillus* phage ϕ29 DNA polymerase plasmids. We also thank Nancy Moran and Adam Eyre-Walker for helpful discussions and comments on the manuscript, Kim Hammond for assistance with the figures, and the staff at the Yale University Faculty of Arts and Sciences High Performance Computing Center. This work was supported in part by National Institutes of Health Grant GM74738 (to H.O.).

1. McCutcheon JP, Moran NA (2010) Functional convergence in reduced genomes of bacterial symbionts spanning 200 My of evolution. *Genome Biol Evol* 2:708–718.
2. Thomas SH, et al. (2008) The mosaic genome of *Anaeromyxobacter dehalogenans* strain 2CP-C suggests an aerobic common ancestor to the delta-proteobacteria. *PLoS ONE* 3:e2103.
3. Sueoka N (1962) On the genetic basis of variation and heterogeneity of DNA base composition. *Proc Natl Acad Sci USA* 48:582–592.
4. Freese E (1962) On the evolution of base composition of DNA. *J Theor Biol* 3:82–101.
5. Muto A, Osawa S (1987) The guanine and cytosine content of genomic DNA and bacterial evolution. *Proc Natl Acad Sci USA* 84:166–169.
6. Hurst LD, Merchant AR (2001) High guanine-cytosine content is not an adaptation to high temperature: a comparative analysis amongst prokaryotes. *Proc Biol Sci* 268:493–497.
7. Naya H, Romero H, Zavala A, Alvarez B, Musto H (2002) Aerobiosis increases the genomic guanine plus cytosine content (GC%) in prokaryotes. *J Mol Evol* 55:260–264.
8. Foerstner KU, von Mering C, Hooper SD, Bork P (2005) Environments shape the nucleotide composition of genomes. *EMBO Rep* 6:1208–1213.
9. Musto H, et al. (2006) Genomic GC level, optimal growth temperature, and genome size in prokaryotes. *Biochem Biophys Res Commun* 347:1–3.
10. Wang HC, Susko E, Roger AJ (2006) On the correlation between genomic G+C content and optimal growth temperature in prokaryotes: Data quality and confounding factors. *Biochem Biophys Res Commun* 342:681–684.
11. Balbi KJ, Rocha EPC, Feil EJ (2009) The temporal dynamics of slightly deleterious mutations in *Escherichia coli* and *Shigella* spp. *Mol Biol Evol* 26:345–355.
12. Hildebrand F, Meyer A, Eyre-Walker A (2010) Evidence of selection upon genomic GC-content in bacteria. *PLoS Genet* 6:e1001107.
13. Hershberg R, Petrov DA (2010) Evidence that mutation is universally biased towards AT in bacteria. *PLoS Genet* 6:e1001115.
14. Mira A, Ochman H, Moran NA (2001) Deletional bias and the evolution of bacterial genomes. *Trends Genet* 17:589–596.
15. Puigbò P, Bravo IG, Garcia-Vallve S (2008) CAIcal: A combined set of tools to assess codon usage adaptation. *Biol Direct* 3:38.
16. Lynch M (2007) *The origins of genome architecture* (Sinauer, Sunderland).
17. Eyre-Walker A, Bulmer M (1993) Reduced synonymous substitution rate at the start of enterobacterial genes. *Nucleic Acids Res* 21:4599–4603.
18. Kudla G, Murray AW, Tollervey D, Plotkin JB (2009) Coding-sequence determinants of gene expression in *Escherichia coli*. *Science* 324:255–258.
19. Allert M, Cox JC, Hellinga HW (2010) Multifactorial determinants of protein expression in prokaryotic open reading frames. *J Mol Biol* 402:905–918.
20. Li GW, Oh E, Weissman JS (2012) The anti-Shine-Dalgarno sequence drives translational pausing and codon choice in bacteria. *Nature* 484:538–541.
21. Bernardi G, Bernardi G (1986) Compositional constraints and genome evolution. *J Mol Evol* 24:1–11.
22. Osawa S, et al. (1988) Directional mutation pressure and transfer RNA in choice of the third nucleotide of synonymous two-codon sets. *Proc Natl Acad Sci USA* 85:1124–1128.
23. McCutcheon JP, Moran NA (2012) Extreme genome reduction in symbiotic bacteria. *Nat Rev Microbiol* 10:13–26.
24. Chamary JV, Hurst LD (2005) Evidence for selection on synonymous mutations affecting stability of mRNA secondary structure in mammals. *Genome Biol* 6:R75.
25. Sharp PM, Emery LR, Zeng K (2010) Forces that influence the evolution of codon bias. *Philos Trans R Soc Lond B Biol Sci* 365:1203–1212.
26. Moran NA (1996) Accelerated evolution and Muller's rachet in endosymbiotic bacteria. *Proc Natl Acad Sci USA* 93:2873–2878.
27. Andersson SGE, Kurland CG (1998) Reductive evolution of resident genomes. *Trends Microbiol* 6:263–268.
28. Moran NA, McCutcheon JP, Nakabachi A (2008) Genomics and evolution of heritable bacterial symbionts. *Annu Rev Genet* 42:165–190.
29. Welch M, et al. (2009) Design parameters to control synthetic gene expression in *Escherichia coli*. *PLoS ONE* 4:e7002.
30. Sharp PM, Li W-H (1987) The Codon Adaptation Index—a measure of directional synonymous codon usage bias, and its potential applications. *Nucleic Acids Res* 15:1281–1295.
31. Keseler IM, et al. (2011) EcoCyc: A comprehensive database of *Escherichia coli* biology. *Nucleic Acids Res* 39(Database issue):D583–D590.
32. Raghavan R, Sage A, Ochman H (2011) Genome-wide identification of transcription start sites yields a novel thermosensing RNA and new cyclic AMP receptor protein-regulated genes in *Escherichia coli*. *J Bacteriol* 193:2871–2874.
33. Raghavan R, Groisman EA, Ochman H (2011) Genome-wide detection of novel regulatory RNAs in *E. coli*. *Genome Res* 21:1487–1497.
34. Zuker M (2003) Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Res* 31:3406–3415.