Available online at www.sciencedirect.com
ScienceDirect
Biochemical and Biophysical Research Communications 356 (2007) 20–25
BBRC
www.elsevier.com/locate/ybbrc

GC content variability of eubacteria is governed by the pol III α subunit

Xiaoqian Zhao [a,c,1], Zhang Zhang [a,b,c,1], Jiangwei Yan [a,c,1], Jun Yu [a,b,d,*]

[a] Key Laboratory of Genome Sciences and Information, Beijing Institute of Genomics, Chinese Academy of Sciences, Beijing 101300, China
[b] Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100080, China
[c] Graduate School of Chinese Academy of Sciences, Beijing 100039, China
[d] James D. Watson Institute of Genome Sciences, Zhejiang University, Hangzhou 310007, China

Received 11 February 2007
Available online 28 February 2007

Abstract

Eubacterial genomes have highly variable GC content (0.17–0.75) and the primary mechanism of such variability remains unknown. The place to look for is what actually catalyzes the synthesis of DNA, where DNA polymerase III is at the center stage, particularly one of its 10 subunits—the α subunit. According to the dimeric combination of α subunits, GC contents of eubacterial genomes were partitioned into three groups with distinct GC content variation spectra: dnaE1 (full-spectrum), dnaE2/dnaE1 (high-GC), and polC/dnaE3 (low-GC). Therefore, genomic GC content variability is believed to be governed primarily by the α subunit grouping of DNA polymerase III; it is of essence in genome composition analysis to take full account of such a grouping principle. Since horizontal gene transfer is very frequent among bacterial genomes, exceptions of the grouping scheme, a few percents of the total, are readily identifiable and should be excluded from in-depth analyses on nucleotide compositions.

Keywords: Genomic GC content; DNA polymerase III α subunit; dnaE; polC

The compositional dynamics of prokaryotic genomes, such as their guanidine and cytosine contents, varies broadly and its mechanistic basis at molecular level has not been revealed [1–3]. It has been proposed that the genomic GC (gGC) content variation is influenced by multiple factors, including both extrinsic factors and intrinsic mutational bias [4,5]. For instance, it was proposed that the gGC content increase was advantageous for eubacteria exposed to sunlight ultraviolet [6], higher in nitrogen-fixing aerobic eubacteria than in non-nitrogen-fixing species within the same genus [6], and increasing significantly in aerobic bacteria as compared to anaerobic ones [7]. Some of the authors even pointed out that eubacteria relying on their hosts for survival, such as obligatory pathogens and symbionts, tend to have low gGC content [8], and others argued for a correlation between gGC content and the optimal growth temperature not only among thermopiles but also mesophiles [5,9–12]. It is clear that a simple inclusive theory or many correlations between gGC and other biotic or abiotic factors are not going to reveal the causative details of the gGC variation, even in the simplest cellular organisms—eubacteria.

The most important intrinsic factor is mutational bias that may come from two basic sources: DNA replication or repair errors and horizontal gene transfer [13]. However, the transferred DNA segments may ameliorate to reflect the base composition of the recipient genome over time and may not alter gGC content in dramatic ways [14]. As we refer to nucleotide mutation, it is impossible to decouple DNA replication and repair from gGC dynamics or divergence among bacterial genomes, since the chemical reactions to polymerize DNA in replication and repair are catalyzed by the same enzymes or their complex [15]. Nevertheless, these enzymes should have left sequence

* Corresponding author. Address: Key Laboratory of Genome Sciences and Information, Beijing Institute of Genomics, Chinese Academy of Sciences, Beijing 101300, China. Fax: +86 10 80498676.
E-mail address: junyu@genomics.org.cn (J. Yu).
[1] These authors contributed equally to this work.

0006-291X/$ - see front matter © 2007 Elsevier Inc. All rights reserved.
doi:10.1016/j.bbrc.2007.02.109

signatures from accumulative errors in the genomes over evolutionary time scale, especially for prokaryotes that are most ancient and evolving in a fast pace, especially for prokaryotes that have the broadest gGC content variation [16].

Toward this end, we focused our analysis on the major prokaryotic DNA polymerases involving DNA replication and repair, in particular the α subunit of DNA polymerase III, to search for correlations between gGC content and the α subunit variants. In our previous study, we have classified these variants into four major isoforms: dnaE1, dnaE2, dnaE3, and polC [17] (Figure S1). This grouping scheme is consistent with two previous notions: biochemical characterization revealed that polC possesses an epsilon domain absent in dnaE and dnaE has at least two other isoforms among bacteria taxa [17,18]. Our analysis clearly divided eubacteria into three groups with distinct gGC content variation spectra according to the dimeric combination of α subunits: dnaE1 (homodimer, full-spectrum), dnaE2/dnaE1 (heterodimer, high-GC), and polC/dnaE3 (heterodimer, low-GC).

Materials and methods

We collected 381 complete genome sequences from eubacterial data available in GenBank (November 26, 2006) [19] and extracted the sequences of all α subunits from the genome annotations and subsequent manual identification by using blast-based tools. With a length cut-off of 625 amino acids, we had 380 eubacterial genomes containing 596 α subunits (Figure S2). Only one bacterium, *Saccharophagus degradans* 2–40, was excluded from further analyses since its two α subunits are too short (471 and 571 amino acids, respectively) to pass the threshold. The numbers of α subunits in these eubacterial genomes ranged from one to four and 199 genomes have more than two α subunits. We also classified 380 eubacterial genomes according to 16S rRNA-based phylogeny and NCBI's taxonomy database [19] (Table S1). We did *t*-test to evaluate gGC differences between the dnaE1 and the other two groups, dnaE2/dnaE1 and polC/dnaE3 (see Results and discussion). *P* values were calculated by *t*-test with one tail and estimated by equal variance.

Results and discussion

DNA polymerase III α subunit governs the gGC variability

According to the dimeric combination of α subunits [17], we plotted 380 eubacterial genomes in our dataset into three groups, namely dnaE1 (homodimer), dnaE2/dnaE1 (heterodimer), and polC/dnaE3 (heterodimer) (Fig. 1). Unlike eukaryotic genomes whose average gGC contents are rather constant around 0.40 [3,4], the GC content of eubacterial genomes varies widely from 0.17 to 0.75 (Fig. 2), and the extremities are represented by *Candidatus Carsonella ruddii* for the minimum [20] and *Anaeromyxobacter dehalogenans* 2CP-C (unpublished) for the maximum [19]. We categorized the gGC content according to possession of the α subunit for all the genomes, resulted in three groups containing 181, 103, and 96 genomes, respectively. Based on this grouping scheme, we found that the gGC content of the dnaE1 group has the broadest distribution, from 0.17 to 0.70 with a mean of 0.45 and a median of 0.44; whereas the ranges of gGC contents are 0.45–0.75 with a mean of 0.63 and a median of 0.64 for the dnaE2/dnaE1 group, and 0.24–0.56 with its mean and median both of 0.36 for the polC/dnaE3 group. Detailed information on gGC contents of each phylum in the three groups are summarized in Table 1. We also evaluated our grouping scheme with *t*-test based on gGC differences between the dnaE1 group and the other two groups, and found the differences are extremely significant with *P* values smaller than 0.001.

The most striking result is the fact that the gGC content of the dnaE2/dnaE1 group is always equal to or higher than 0.45. This "magic" number seems very stringent but there has not been a mechanistic explanation as to why these bacteria across five different phyla follow this rule in such an inflexible way (Fig. 1). In contrast, the gGC content of the polC/dnaE3 group are almost exclusively below 0.45 with merely ten exceptions (*Bacillus licheniformis* ATCC 14580, *Geobacillus kaustophilus* HTA426, *B. licheniformis* DSM 13, *Lactobacillus brevis* ATCC 367, *Lactobacillus casei* ATCC 334, *Lactobacillus delbrueckii bulgaricus* ATCC BAA-365, *L. delbrueckii bulgaricus*, *Desulfitobacterium hafniense* Y51, *Moorella thermoacetica* ATCC 39073 and *Thermotoga maritim*) but they are stayed as largely lower gGC. The most GC-flexible group, containing exclusively dnaE1 that is believed to function as a homodimer, not only has most of the bacteria in this dataset but also covers a broad gGC range.

We are facing an interesting but rather fundamental paradox where dnaE1 and its other two derivatives, dnaE2 and dnaE3, may actually have different GC content preferences when they are involved in DNA synthesis. In other words, dnaE1 and its two other isoforms may be involved in either DNA replication or DNA repair, and may be involved in both. DnaE1 is responsible for chromosomal replication as we know [21], and it is also demonstrated to be required for UV mutagenesis and optimal repair of methyl methanesulfonate (MMS) and hydrogen peroxide damages [22–24]. Generally, dnaE1 has dual function concerning DNA replication and repair, and accordingly the gGC content varies broadly from 0.17 to 0.70 in dnaE1 group.

If we assume that dnaE1 is involved in both DNA replication and repair, in the dnaE2/dnaE1 group how dnaE1 and dnaE2 divide up their work and the differentiated emphasis resulted in higher gGC content? Indeed, dnaE2 as the DNA polymerase III catalytic subunit was demonstrated to be involved in damage-inducible mutagenesis, together with imuA and imuB, forming a damage-inducible operon that is upregulated in *Caulobacter crescentus* [25]. It is also suggested dnaE2 might participate in error-prone DNA repair synthesis and point mutations during translesion synthesis (TLS) in *Mycobacterium tuberculosis* [26]. Since dnaE2 is the unique contributor for damage repair and mutagenesis, it may play a dominant role in maintaining the high gGC content for these genomes (Fig. 1).

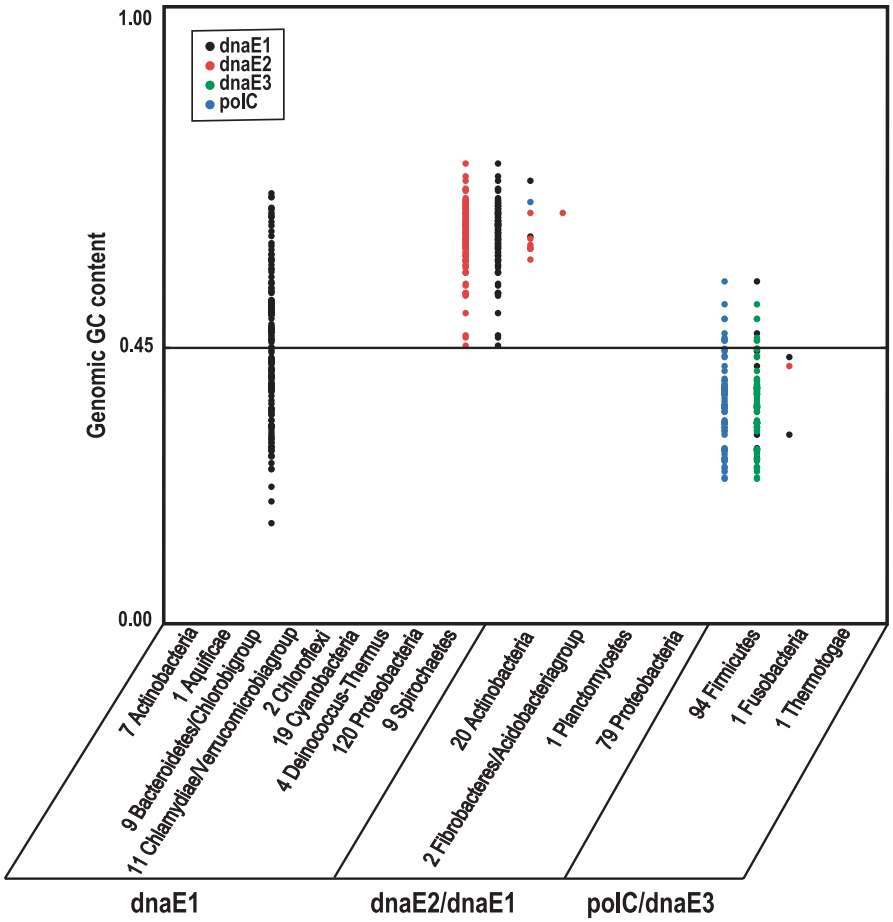

Fig. 1. Three eubacterial groups according to the possession of α subunit isoforms. The dnaE1 group only harbors dnaE1 that forms homodimers for its active form and has the most variable gGC contents. The dnaE2/dnaE1 group possesses both dnaE2 and dnaE1, which form heterodimers for catalytic activities, and the gGC content of this bacterial group is always higher than 0.45 in our current dataset. The polC/dnaE3 group is composed of polC and dnaE3, and a greater majority of the bacteria in this group have gGC contents lower than 0.45. Since horizontal gene transfer is very frequent among bacterial genomes exceptions do exist as some of the bacteria actually have more than what they suppose to have.

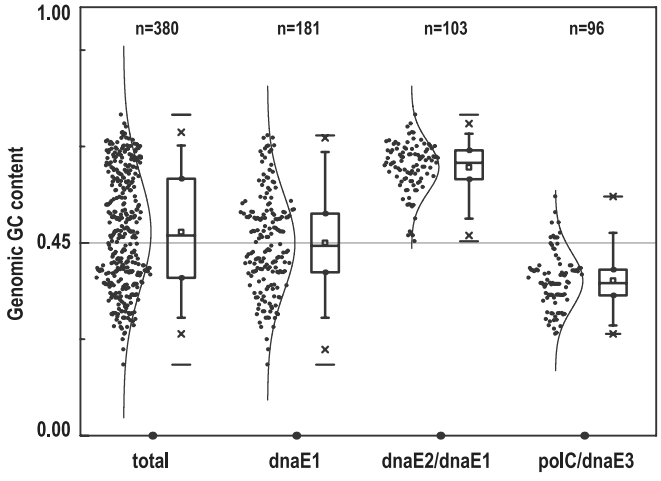

Fig. 2. gGC distribution among eubacterial genomes sequenced thus far. The actual data points representing each genomes (left) and statistical analysis (right) are both shown with a putative threshold of 0.45 (horizontal grey line). The extremities are represented by *C. ruddii* for the minimum [20] and *A. dehalogenans* strain 2CP-C (unpublished) for the maximum. Symbols are: top horizontal bars, maximum; bottom horizontal bars, minimum; top *x*, 99%; bottom *x*, 1%; box, 25–75%; □, median; *n*, number of organisms.

Again, in *C. crescentus*, a member of the dnaE2/dnaE1 group, it was observed that a high proportion of G:C to C:G transversions [25], as opposed to G:C to A:T transitions that is a hallmark of UV mutagenesis described for *Escherichia coli*, the most known and typical member of the dnaE1 group. Therefore, we speculate that dnaE2 may be responsible for the higher gGC content in this group.

The third group, similar to the dnaE2/dnaE1 group, also possesses two α subunit isoforms, polC and dnaE3, presumably forming a functional heterodimer at the replication fork [27]. Since an epsilon domain is identified in polC for a 3′–5′ proofreading exonuclease activity, it is unquestionable that polC is involved in repair and mutagenesis [28]. Moreover, dnaE3 is also recognized as an error-prone polymerase, acting as a TLS polymerase, shown clearly in *Streptococcus pyogenes* [29,30]. In addition, it was found that depletion of either polC or dnaE3 (although it was not distinguished in the report) in *Bacillus subtilis* prevents UV-induced mutagenesis; this result suggests that polC and dnaE3 definitely function in DNA repair and mutagenesis in *B. subtilis*, aside from their major roles in DNA replication [31]. In contrast to the dnaE2/

Table 1
gGC content distributions in different dnaE groups in eubacterial phyla

Group	Phylum	Number of isolates	gGC content	*P* value
dnaE1		181	0.17–0.70	
	Actinobacteria	7	0.46–0.70	
	Aquificae	1	0.43	
	Bacteroidetes/Chlorobi group	9	0.39–0.66	
	Chlamydiae/Verrucomicrobia group	11	0.35–0.41	
	Chloroflexi	2	0.47–0.49	
	Cyanobacteria	19	0.31–0.62	
	Deinococcus-Thermus	4	0.67–0.70	
	Proteobacteria	119	0.17–0.68	
	Spirochaetes	9	0.28–0.53	
dnaE2/dnaE1		103	gGC > 0.45 (0.45–0.75)	$P = 1.65 \times 10^{-37}$
	Actinobacteria[a]	20	0.53–0.73	
	Fibrobacteres/Acidobacteria group	2	0.58–0.62	
	Planctomycetes	1	0.55	
	Proteobacteria	80	0.45–0.75	
polC/dnaE3		96	gGC < 0.45[c] (0.24–0.56)	$P = 7.13 \times 10^{-12}$
	Firmicutes[b]	94	0.24–0.56	
	Fusobacteria	1	0.27	
	Thermotogae	1	0.46	

[a] *Symbiobacterium thermophilum* IAM14863 possesses a polC in addition to dnaE2 and dnaE1.
[b] One isolate (*Carboxydothermus hydrogenoformans* Z-2901) has dnaE2 other than polC and dnaE1.
[c] The group has 10 genomes that have gGC contents above 0.45.

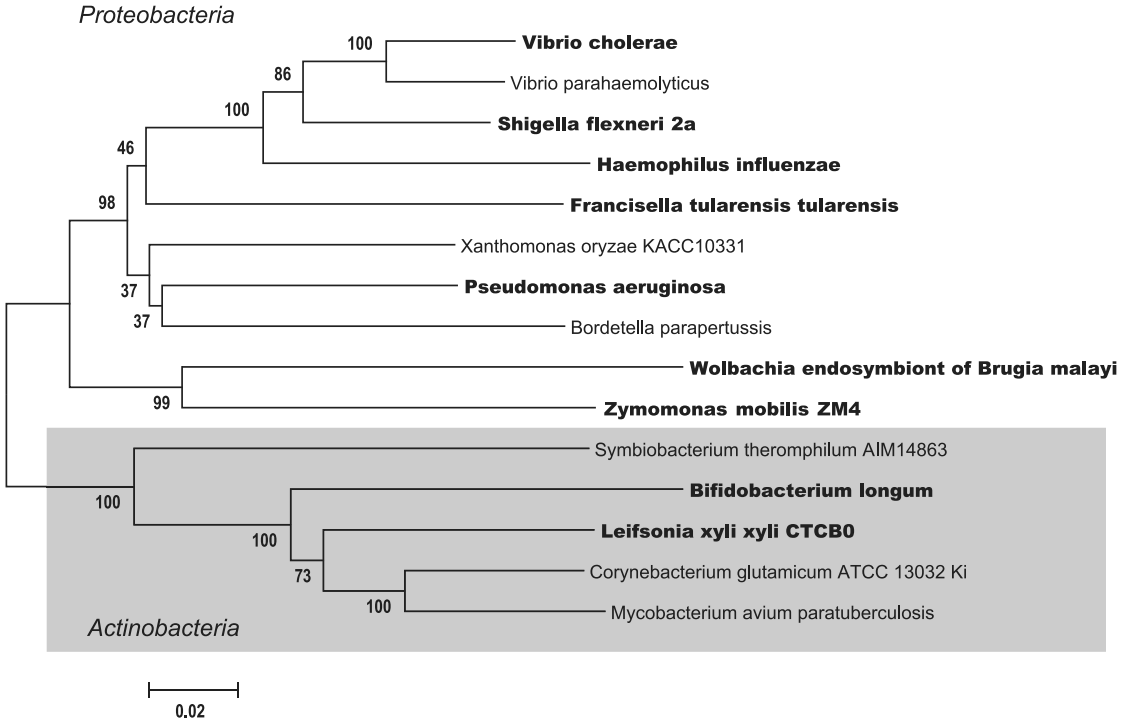

Fig. 3. 16S rRNA-based phylogeny of selected species from *Proteobacteria* and *Actinobacteria*. Shading differentiates *Actinobacteria* (shaded) from *Proteobacteria*. Bold letters are used to highlight the dnaE1 group. The tree was constructed with the program MEGA (version 3.1) by neighbor-joining with Kimura 2-parameter distance (scale bar). The reliability of the branching orders was estimated by bootstrapping (1000 bootstrap replications, 64238 random seed), and bootstrap values (percentage after 1000 iterations) for major branches are shown.

dnaE1 group, the gGC content in polC/dnaE3 group preferentially stay low, with great majority below 0.45; since polC and dnaE3 almost exclusively belong to this group (see the supplementary material for details on exceptional cases) and responsible for both replication and repair, the lower gGC content should be a collective contribution from both isoforms.

The gGC grouping scheme breaks taxonomic boundaries

In the process of correlating gGC contents with the possession of α subunit isoforms, we realized that this scheme does not entirely follow the classical taxonomy, albeit most of the bacteria in our collection do. Two phyla, *Proteobacteria* and *Actinobacteria*, have their

members split into the dnaE1 and dnaE2/dnaE1 groups. Phyla *Proteobacteria* and *Actinobacteria* have total of 199 and 27 genomes in our dataset, respectively; between the two groups, the former has a split of 120 and 79, whereas the latter has a split of 7 and 20 (Fig. 1). Therefore, the grouping scheme is definitely not a rare event. In addition, the genomes from the two phyla in the two groups all share dnaE1; this fact suggests that the higher gGC content in the dnaE2/dnaE1 group is determined by the existence of dnaE2 not dnaE1 that allows a broad gGC variation in its own group. Finally, we only found the gGC content grouping conflict in these two rather ancient phyla but not in any other phyla in our entire dataset so that such a conflict is quite limited among all eubacteria. We at present time are unable to exclude the possibility that the dnaE2 subunit is capable of transferring between the genomes of the two groups within each phylum. To further illustrate this cross-phylum phenomenon, we constructed a phylogenetic tree that included all 380 genomes by using 16S rRNAs (data not show). For a better display, we randomly selected 10 species from *Proteobacteria* and 5 species from *Actinobacteria* (Fig. 3), and the 15 bacteria were distinctly divided into each phylum, whereas the dnaE1 and dnaE2/dnaE1 groups were mixed.

Three scenarios are worthy of bearing in mind. First, since gGC content variations are not strictly taxonomic, the α subunits divergence scheme (equal to gGC content variation) should occur earlier than taxonomic divergence. An alternative hypothesis is that dnaE2 may evolve to be functional in the two phyla independently, assuming these two phyla are the oldest. Second, if the molecular mechanism is simplistically, as proposed here, α subunit-dependent, horizontal gene transfer might play a major role to change the α-subunit-possession groups, especially for dnaE1 group that has a functional form of dnaE1 homodimer. Third, if horizontal gene transfer is so prevalent among prokaryotic genomes, either compatibility or dominance scheme has to be introduced to explain why some of the eubacterial types are even possible to correlate with taxonomy. The short answer for these questions is that these possibilities may all hold truth based on the current dataset.

Acknowledgments

We thank our colleague Mr. Chen Chen for valuable discussions. This work was supported by a grant from the Chinese Academy of Sciences awarded to J.Y. (KSCX2-SW-331).

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.bbrc.2007.02.109.

References

[1] S.D. Bentley, J. Parkhill, Comparative genomic structure of prokaryotes, Annu. Rev. Genet. 38 (2004) 771–792.
[2] E. Barbu, K.Y. Lee, R. Wahl, Content of purine and pyrimidine base in desoxyribonucleic acid of bacteria, Ann. Inst. Pasteur. (Paris) 91 (1956) 212–224.
[3] A.N. Belozersky, A.S. Spirin, A correlation between the compositions of deoxyribonucleic and ribonucleic acids, Nature 182 (1958) 111–112.
[4] N. Sueoka, On the genetic basis of variation and heterogeneity of DNA base composition, Proc. Natl. Acad. Sci. USA 48 (1962) 582–592.
[5] H.C. Wang, E. Susko, A.J. Roger, On the correlation between genomic G+C content and optimal growth temperature in prokaryotes: data quality and confounding factors, Biochem. Biophys. Res. Commun. 342 (2006) 681–684.
[6] C.E.A. McEwan, D. Gatherer, N.R. McEwan, Nitrogen-fixing aerobic bacteria have higher genomic GC content than non-fixing species within the same genus, Hereditas 128 (1998) 173–178.
[7] H. Naya, H. Romero, A. Zavala, B. Alvarez, H. Musto, Aerobiosis increases the genomic guanine plus cytosine content (GC%) in prokaryotes, J. Mol. Evol. 55 (2002) 260–264.
[8] E.P. Rocha, A. Danchin, Base composition bias might result from competition for metabolic resources, Trends Genet. 18 (2002) 291–294.
[9] S.A. Marashi, Z. Ghalanbor, Correlations between genomic GC levels and optimal growth temperatures are not 'robust', Biochem. Biophys. Res. Commun. 325 (2004) 381–383.
[10] H. Musto, H. Naya, A. Zavala, H. Romero, F. Alvarez-Valin, G. Bernardi, The correlation between genomic G+C and optimal growth temperature of prokaryotes is robust: a reply to Marashi and Ghalanbor, Biochem. Biophys. Res. Commun. 330 (2005) 357–360.
[11] S. Basak, S. Mandal, T.C. Ghosh, Correlations between genomic GC levels and optimal growth temperatures: some comments, Biochem. Biophys. Res. Commun. 327 (2005) 969–970.
[12] H. Naya, D. Gianola, H. Romero, J.I. Urioste, H. Musto, Inferring parameters shaping amino acid usage in prokaryotic genomes via Bayesian MCMC methods, Mol. Biol. Evol. 23 (2006) 203–211.
[13] J.G. Lawrence, H. Ochman, Molecular archaeology of the *Escherichia coli* genome, Proc. Natl. Acad. Sci. USA 95 (1998) 9413–9417.
[14] J.G. Lawrence, H. Ochman, Amelioration of bacterial genomes: rates of change and exchange, J. Mol. Evol. V44 (1997) 383–397.
[15] J.Q. Svejstrup, Mechanisms of transcription-coupled DNA repair, Nat. Rev. Mol. Cell Biol. 3 (2002) 21–29.
[16] A. Muto, S. Osawa, The guanine and cytosine content of genomic DNA and bacterial evolution, Proc. Natl. Acad. Sci. USA 84 (1987) 166–169.
[17] X. Zhao, J. Hu, J. Yu, Comparative analysis of eubacterial DNA polymerase III alpha subunits, Genomics Proteomics Bioinf. 4 (2006) 203–211.
[18] I. Bruck, M. O'Donnell, The DNA replication machine of a gram-positive organism, J. Biol. Chem. 275 (2000) 28971–28983.
[19] NCBI.
[20] A. Nakabachi, A. Yamashita, H. Toh, H. Ishikawa, H.E. Dunbar, N.A. Moran, M. Hattori, The 160-kilobase genome of the bacterial endosymbiont Carsonella, Science 314 (2006) 267.
[21] C.S. McHenry, DNA polymerase III holoenzyme of *Escherichia coli*, Annu. Rev. Biochem. 57 (1988) 519–550.
[22] B.A. Bridges, R.P. Mottershead, Mutagenic DNA repair in *Escherichia coli*. III. Requirement for a function of DNA polymerase III in ultraviolet-light mutagenesis, Mol. Gen. Genet. 144 (1976) 53–58.
[23] B.A. Bridges, H. Bates, Mutagenic DNA repair in *Escherichia coli*. XVIII. Involvement of DNA polymerase III alpha-subunit (DnaE protein) in mutagenesis after exposure to UV light, Mutagenesis 5 (1990) 35–38.

[24] M.E. Hagensee, S.K. Bryan, R.E. Moses, DNA polymerase III requirement for repair of DNA damage caused by methyl methanesulfonate and hydrogen peroxide, J. Bacteriol. 169 (1987) 4608–4613.
[25] R.S. Galhardo, R.P. Rocha, M.V. Marques, C.F. Menck, An SOS-regulated operon involved in damage-inducible mutagenesis in *Caulobacter crescentus*, Nucleic Acids Res. 33 (2005) 2603–2614.
[26] H.I. Boshoff, M.B. Reed, C.E. Barry III, V. Mizrahi, DnaE2 polymerase contributes to in vivo survival and the emergence of drug resistance in *Mycobacterium tuberculosis*, Cell 113 (2003) 183–193.
[27] E. Dervyn, C. Suski, R. Daniel, C. Bruand, J. Chapuis, J. Errington, L. Janniere, S.D. Ehrlich, Two essential DNA polymerases at the bacterial replication fork, Science 294 (2001) 1716–1719.
[28] A. Johnson, M. O'Donnell, Cellular DNA replicases: components and dynamics at the replication fork, Annu. Rev. Biochem. 74 (2005) 283–315.
[29] I. Bruck, M.F. Goodman, M. O'Donnell, The essential C family DnaE polymerase is error-prone and efficient at lesion bypass, J. Biol. Chem. 278 (2003) 44361–44368.
[30] B. Tippin, P. Pham, M.F. Goodman, Error-prone replication for better or worse, Trends Microbiol. 12 (2004) 288–295.
[31] E. Le Chatelier, O.J. Becherel, E. d'Alencon, D. Canceill, S.D. Ehrlich, R.P. Fuchs, L. Janniere, Involvement of DnaE, the second replicative DNA polymerase from *Bacillus subtilis*, in DNA mutagenesis, J. Biol. Chem. 279 (2004) 1757–1767.