AMERICAN SOCIETY FOR MICROBIOLOGY | mBio®

RESEARCH ARTICLE
Ecological and Evolutionary Science

Codon Usage Optimization in the Prokaryotic Tree of Life: How Synonymous Codons Are Differentially Selected in Sequence Domains with Different Expression Levels and Degrees of Conservation

José Luis López,[a] Mauricio Javier Lozano,[a] María Laura Fabre,[a] Antonio Lagares[a]

[a]Instituto de Biotecnología y Biología Molecular, CONICET, CCT-La Plata, Departamento de Ciencias Biológicas, Facultad de Ciencias Exactas, Universidad Nacional de La Plata, La Plata, Argentina

José Luis López and Mauricio Javier Lozano contributed equally to this work. Author order was determined alphabetically.

ABSTRACT Prokaryote genomes exhibit a wide range of GC contents and codon usages, both resulting from an interaction between mutational bias and natural selection. In order to investigate the basis underlying specific codon changes, we performed a comprehensive analysis of 29 different prokaryote families. The analysis of core gene sets with increasing ancestries in each family lineage revealed that the codon usages became progressively more adapted to the tRNA pools. While, as previously reported, highly expressed genes presented the most optimized codon usage, the singletons contained the less selectively favored codons. The results showed that usually codons with the highest translational adaptation were preferentially enriched. In agreement with previous reports, a C bias in 2- to 3-fold pyrimidine-ending codons, and a U bias in 4-fold codons occurred in all families, irrespective of the global genomic GC content. Furthermore, the U biases suggested that U_3-mRNA–U_{34}-tRNA interactions were responsible for a prominent codon optimization in both the most ancestral core and the highly expressed genes. A comparative analysis of sequences that encode conserved (*cr*) or variable (*vr*) translated products, with each one being under high (HEP) and low (LEP) expression levels, demonstrated that the efficiency was more relevant (by a factor of 2) than accuracy to modeling codon usage. Finally, analysis of the third position of codons (GC3) revealed that in genomes with global GC contents higher than 35 to 40%, selection favored a GC3 increase, whereas in genomes with very low GC contents, a decrease in GC3 occurred. A comprehensive final model is presented in which all patterns of codon usage variations are condensed in four distinct behavioral groups.

IMPORTANCE The prokaryotic genomes—the current heritage of the most ancient life forms on earth—are comprised of diverse gene sets, all characterized by varied origins, ancestries, and spatial-temporal expression patterns. Such genetic diversity has for a long time raised the question of how cells shape their coding strategies to optimize protein demands (i.e., product abundance) and accuracy (i.e., translation fidelity) through the use of the same genetic code in genomes with GC contents that range from less than 20 to more than 80%. Here, we present evidence on how codon usage is adjusted in the prokaryotic tree of life and on how specific biases have operated to improve translation. Through the use of proteome data, we characterized conserved and variable sequence domains in genes of either high or low expression level and quantitated the relative weight of efficiency and accuracy—as well as their interaction—in shaping codon usage in prokaryotes.

KEYWORDS codon usage selection, mutational bias, genome evolution, core genes, singletons, translation efficiency, translation accuracy

Citation López JL, Lozano MJ, Fabre ML, Lagares A. 2020. Codon usage optimization in the prokaryotic tree of life: how synonymous codons are differentially selected in sequence domains with different expression levels and degrees of conservation. mBio 11:e00766-20. https://doi.org/10.1128/mBio.00766-20.

Invited Editor Marc Bailly-Bechet, Université Claude Bernard Lyon 1

Editor Christa M. Schleper, University of Vienna

Address correspondence to Antonio Lagares, lagares@biol.unlp.edu.ar.

Received 2 April 2020
Accepted 16 June 2020
Published 21 July 2020

The wide range of GC contents exhibited by prokaryote genomes—i.e., from less than 20% to 80%—is believed to be primarily caused by interspecies differences in mutational processes that operate on both the coding and the noncoding regions (1–6). Since prokaryote genomes consist mainly of coding regions that tightly reflect the genomic GC content, mutational bias is a main force that shapes the codon usage of the majority of the genes (7, 8). Thus, understanding how selection is coupled to mutational processes to model codon usage under such diverse GC contents is an essential issue (9–11). Recent evidence suggests that prokaryotic genomes with intermediate to high GC contents are affected by mutations that are universally biased in favor of AT replacements (12, 13). That process is counterbalanced by selection-based constraints that, in turn, increase the GC content and fine-tune codon usage—i.e., the so-called mutation-selection-drift model (14–16). Intragenomic codon usage heterogeneities, however, are always present among different gene sets—i.e., between core genes that are shared throughout a given lineage and singletons (unique accessory genes) that are taxon and/or strain specific (17, 18). Furthermore, in a multipartite genome, the linkage between the physical patterns of heterogeneity in codon usage and the replicon location of the different core genes has also been recently demonstrated (19). The analysis of intragenomic codon usage heterogeneities by different authors (20, 21) has served to identify at least the following three distinctive gene groups. The first comprises the majority of the coding sequences that are associated with the so-called typical codon usage, while the second consists of the putative highly expressed (PHE) genes involving codon usages that are the best adapted to the translational machinery (20, 22–26). The third contains genes that encode the accessory information, including the singletons (unique genes) that are present in mobile genetic elements as well as in the most stable replicons (21, 27–30). The intracellular variations in codon usage can be explained on the basis of selective pressures that operate with different strengths depending on gene function and the resulting impact on cellular fitness (31). A search for the biochemical basis associated with the heterogeneity in codon usage among different gene sets has been the focus of numerous studies. Several lines of evidence have indicated that the biased codon usage in PHE genes correlates with the copy number of the specific tRNA species that decode the preferred codons (23, 32, 33) and with an optimal codon-anticodon interaction (34). The latter includes both the classical Watson-Crick interactions (WCIs) and a wobble base pairing with the corresponding cognate tRNAs. All these interactions have been taken into consideration in order to define different numerical indices (35, 36) as estimators of the codon adaptation to the existing tRNA pool. Though not considered in currently used translation-adaptation indices, evidence has also been found for other nonstandard codon-anticodon interactions which, by improving the decoding capacity, are also relevant to codon usage evolution (37–40).

The analysis of an extensive number of genes with different functions, degrees of ubiquitousness, and degrees of phylogenetic conservation has demonstrated that codon usage is related to gene expression level (32, 41, 42), the degree of conservation (18, 31, 43, 44), the genomic location—i.e., chromosome, chromid, or plasmidome (19, 45, 46)—and different features such as codon ramps and mRNA secondary structure, among others (47–49). Current evidence indicates that accessory genes involve atypical codon usages (21, 28, 46, 50) compared to the most conserved (ancestral) core genes in a given lineage. The latter genes, for their part, exhibit adaptational variations in codon usages ranging from typical to more biased, as the one observed in genes that correspond to highly abundant proteins which are coded by PHE genes (51). Moreover, that core genes may also exhibit remarkable codon usage heterogeneities has been recently demonstrated (19).

In the work reported here, after examining 29 different prokaryote families, we performed a consolidated analysis aimed at characterizing the specific intragenomic codon variations that lead to differences in codon usage between gene sets with diverse expression levels and degrees of conservation in a given lineage. The evaluation of intragenic regions with different coding characteristics—compared to strategies based on the global analyses of complete genes—enabled the recognition of different patterns of codon usages within a message to be translated. Thus, the questions emerged of (i) whether the codon usage patterns associated with highly expressed amino acid sequences (i.e., affecting efficiency) were the same as those associated with genes encoding highly conserved sequences (i.e., affecting accuracy) and (ii) whether the requirements for translation efficiency and accuracy were fully independent or whether those two types of demands interacted. The results have indicated how, even in organisms with quite different GC contents, alterations in specific codons are associated with a selective adaptation of the most ancestral genes compared to the adaptation of those genes that are newer in the phylogeny. Through an independent analysis of sequences associated with variable or conserved regions having different expression levels (i.e., low versus high), we were able to identify the specific codon usages associated with translation efficiency and accuracy as well as quantitatively estimate their relative relevance to codon usage.

RESULTS

Ancestry-dependent codon usage bias as revealed by the analysis of core genes from diverse prokaryotic families. López et al. (19) have recently demonstrated that in a model proteobacterium, the more ancestral the core genes were, the better adapted their codon usages were to the translational machinery. In order to investigate if such a correlation was associated with a general phenomenon in different prokaryote taxa, we assembled different core gene sets that progressed deeper into the phylogenies of 27 Gram-negative and -positive eubacterial families spanning the phyla *Proteobacteria*, *Actinobacteria*, *Firmicutes*, and *Bacteroidetes* along with 2 archaeal families from the phylum *Euryarchaeota*. Table S1a (tab 1) in the supplemental material itemizes for each taxon the number of genes in each gene set from the most recent core 1 (C1) to the most ancestral core n (Cn). In each prokaryote family, the most ancestral core gene set (Cn) consisted of 100 to 500 orthologs. The codon usage variation with gene ancestry within a given prokaryote family was evaluated through a correspondence analysis (CA) that used as variables the raw codon counts (RCC) of the individual genes in each genome analyzed (see Materials and Methods). The average values of the first two components for the core gene sets C1 to Cn were projected on the CA plots. Figure 1 (left graphs) depicts the CA for four genomes specifically selected to represent groups of organisms with different types of CA plots and GC3 contents in their core genes, PHE genes, and singletons, namely, groups A to D (see Materials and Methods). CA were also calculated using relative synonymous-codon usages (RSCUs) as input variables instead of RCC as presented in Fig. S1A. In agreement with a recent study with *Sinorhizobium meliloti* (19), in all instances a directional shift in the codon usage positions was evident from the most recent toward the most ancestral core gene set. That this ancestry-dependent pattern of codon usage variation was observed in even quite distant prokaryote families among those analyzed in this study was remarkable (cf. the CA plots for all other species in Fig. S1B, left graphs). In the evolution of core codon usages, however, the extent of the observed shifts and the type of synonymous codons enriched in each taxon (i.e., the direction of change) varied markedly among different families (Fig. 1, Fig. S1A, and Fig. S1B, right graphs).

The general features that characterized the bias in codon usages can be summarized as follows. First, a general pattern indicated that in bacteria from groups B, C, and D, the PHE genes are enriched in codons with higher GC3 than in singletons (Fig. 1 and Fig. S1, right graphs). Conversely, an AU enrichment in the third position of codons was observed in the ancestral core fractions of organisms from group A which have extremely low GC contents. Second, from C1 to Cn in the CA plot, the codon usages gradually shifted away from the position of the singletons (the unique genes) to approach the region where the PHE genes were located (Fig. 1 and Fig. S1, left graphs). Similar results were obtained when PHE genes were subtracted from the different Cn cores (i.e., Cn-woPHE [Fig. S2]). Thus, the overall evidence suggested that gene ancestry correlated with a codon usage optimization that resembled the one observed in the

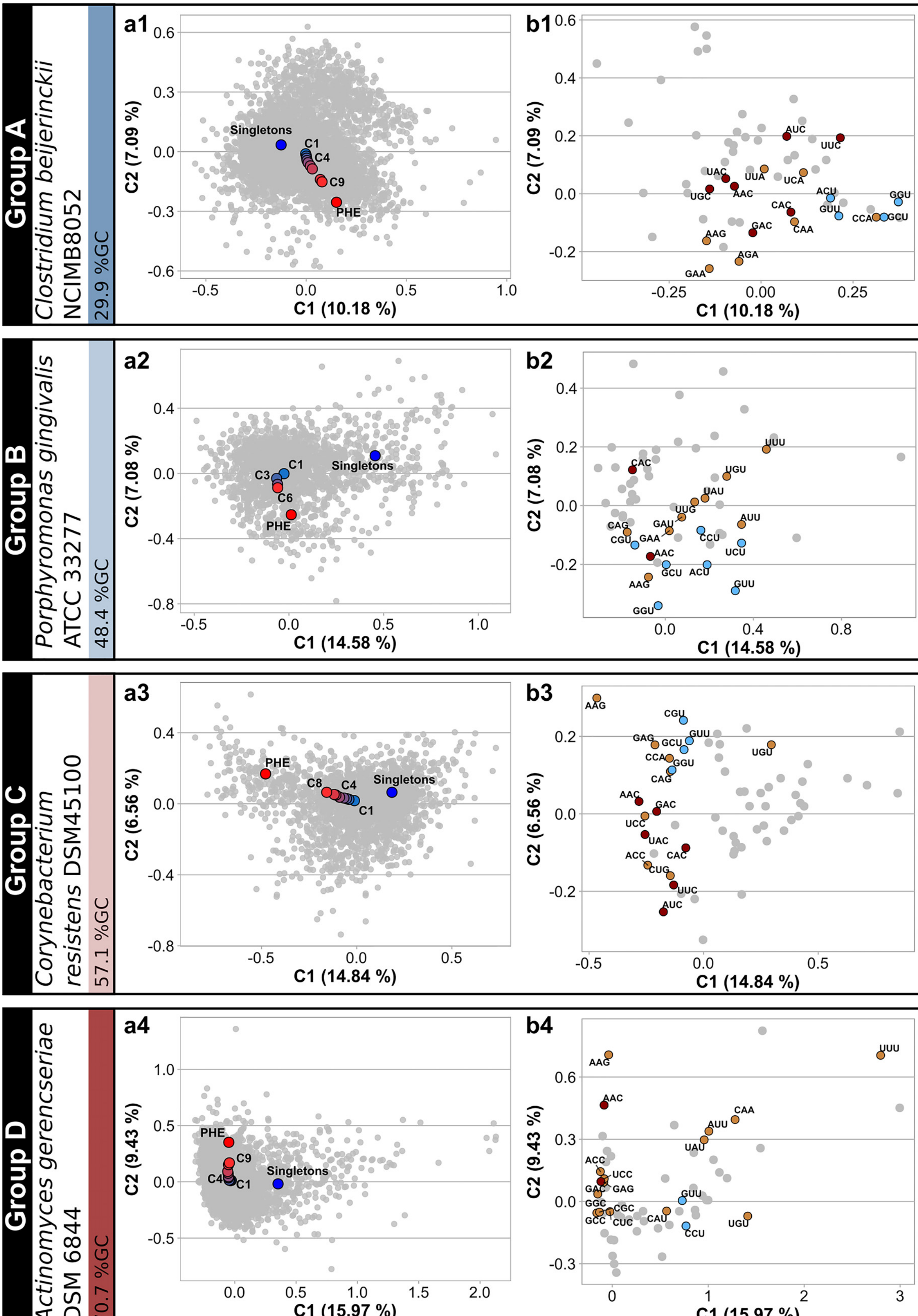

FIG 1 Raw-codon-count-based correspondence analysis (CA) plots of core-gene sets with different degrees of conservation throughout the phylogeny of selected prokaryote families (groups A to D). (a1 to a4) In 4 reference strains with different GC contents, individual genes (in

(Continued on next page)

PHE genes. Nonetheless, the most ancestral core genes (i.e., the Cn gene sets) never overlapped with the position of the PHE genes in the CA plots. In most prokaryote species, the order of positions in the CA plot followed the sequence singletons-C1-Cn, which series was associated with both an enrichment in C-ending 2-/3-fold degenerate codon families (i.e., a C bias in the 2-/3-fold degenerate pyrimidine-ending codons) (Fig. S3A, panel a, shows a significant C bias from the most recent to the most ancestral core gene set—both without PHE genes—with a P value of <0.02 [t test]) and an additional enrichment in U-ending 4-fold degenerate codon families (Fig. S3A, panel b, shows a significant U bias associated with gene ancestry, with a P value of <0.05 [t test]). Such C and U biases were found to be even more intense when comparing the most ancestral core gene sets against PHE genes (Fig. S3B). Each of the previous effects varied in relative intensity among the different prokaryote families and was more intense in microorganisms from groups B and C (central blue and orange bars in Fig. S3 [all panels]). In agreement with previous reports (52), no specific C bias with increasing core gene ancestry was observed in the TGC codon (Cys) irrespective of the group under consideration. Comparable codon enrichments also were found when comparing Cn-woPHE genes (i.e., Cn without PHE genes) to PHE genes (Fig. S3B, panels a and b, where a significant C bias [$P < 0.002$, t test], except for group D, and a significant U bias [$P < 0.03$, t test] were observed). Wald et al. (40) have previously reported that the C and the U biases are associated with an improved codon usage correspondence to the anticodons of the tRNA pool. The combined effects of the C and U biases are the basis for the "rabbit head" distribution of genes that can be observed in most of the CA plots (gray dots), an effect that was originally described for *Escherichia coli* (21). Contrasting with the codon usage of core and PHE genes, the singleton genes tend to be enriched in A/U-ending codons.

Indication from m-tAI values that the codon usages of the most ancestral genes are better adapted to the cellular translational machinery. In order to explore how extensive the correlations between codon usage, gene ancestry, and translation efficiency were, we calculated the modal species-specific tRNA-adaptation index (m-tAI) values for the C1 to the Cn genes for a given strain and used those indices to estimate the adaptation of each gene set to the tRNA pool. Each m-tAI takes into consideration both the copy number of each tRNA structural gene as an estimation of that tRNA's cellular concentration and the codon-anticodon interactions, including the classical Watson-Crick interactions (WCIs) along with the wobble rules (see Materials and Methods). Figure 2 and Fig. S4, left graphs, illustrate how with progressive gene ancestry the m-tAI generally increases to often approach that of the PHE genes, thus evidencing that the most ancestral cores are enriched in genes that displayed adaptive—i.e., selection-dependent—changes in their codon usage. That such m-tAI increases with progressive ancestry had been observed in strains from group A (average Spearman coefficient = 0.99 and P value = 0.002), group B (average Spearman coefficient = 0.66 and P value = 0.02), and group C (average Spearman coefficient 0.90 and P value = 0.08) was indeed remarkable (cf. Fig. 2 and Fig. S4, left graphs). Unfortunately, nonstandard forms of base pairing, such as U:U interactions and others, are not included in the m-tAI calculations, and this fact might negatively impact the way that the m-tAi varies with ancestry, in particular in organisms from the GC-rich group D. In the reference strains from these prokaryote families, the PHE genes (red dashed lines) were always associated with higher m-tAI values than those of the core gene sets from

FIG 1 Legend (Continued)
gray) are represented in the space of the first-two CA components, with the percent variation of components 1 and 2 being indicated on the axes. CAs were computed using raw codon counts (RCC) as the input variables. Average coordinates (centroids) for different gene sets (i.e., singletons in blue, C1 to Cn in a gradient from blue to red, and putative highly expressed [PHE] genes in red) were projected on the CA space. In C1 to Cn, the higher number, the more ancestral the core gene set within the phylogeny. Table S1a (tab 1) lists the prokaryote species that were used to construct each Ci gene set by means of the EDGAR software (57, 58). (b1 to b4) Plots describing codon relative weight in the first two principal-component positions of the CA. Codons with the highest codon usage frequency (CUF) enrichment for each amino acid from C1 to PHE (i.e., those codons that better represent translational adaptation) are colored light brown, except when those same codons corresponded also to a 2-/3-fold C or to a 4-fold U bias, in which case they are colored dark brown and light blue, respectively.

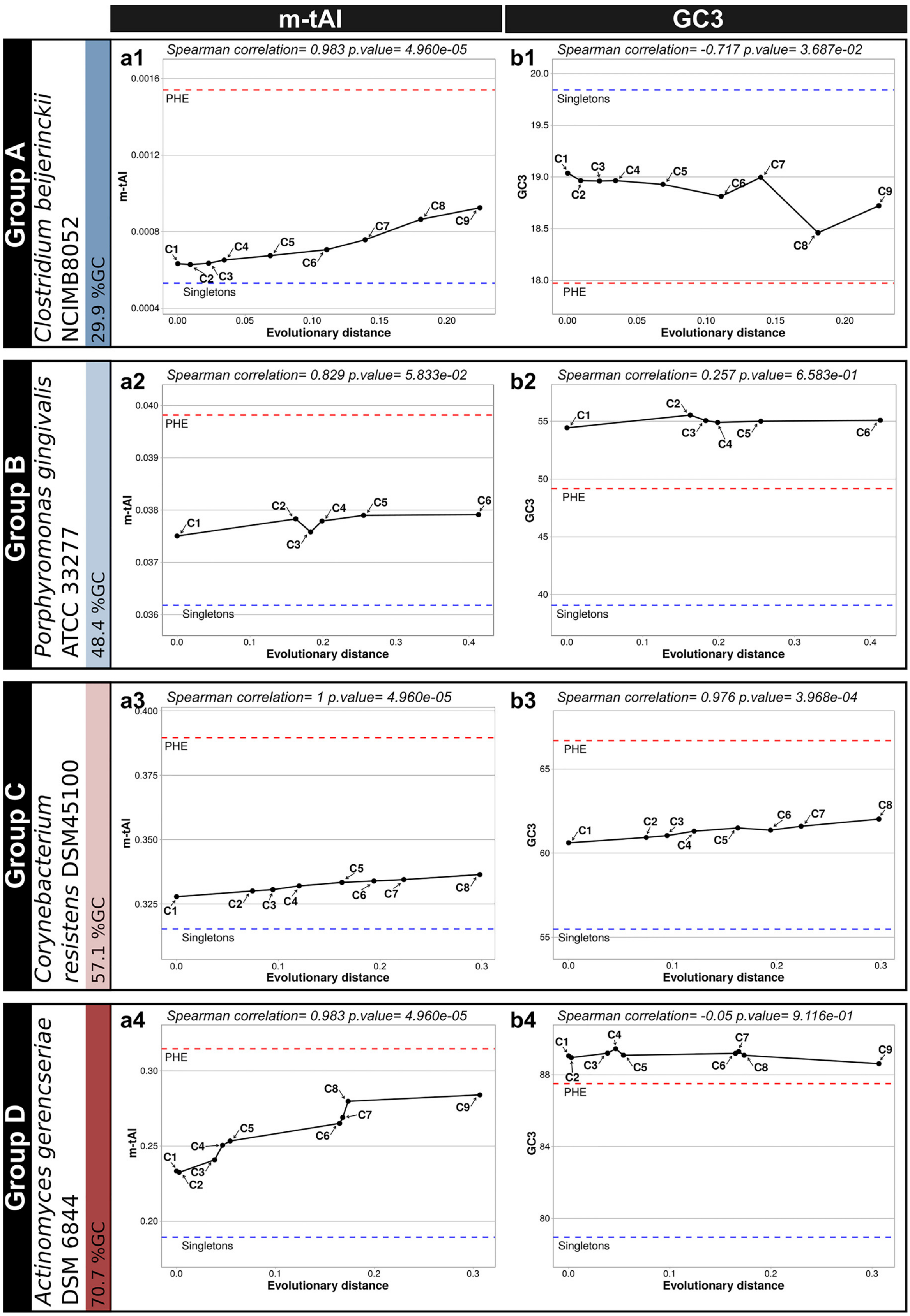

FIG 2 Codon usage adaptations to the cellular tRNA pool, and changes in the GC3 content of different prokaryote core genes. The reference strains represented are the same five as in Fig. 1. (a1 to a4) In each panel, the modal tRNA adaptation index (m-tAI) calculated

(Continued on next page)

the same genome. Conversely, singletons (blue dashed lines) were always the gene sets with the lower m-tAIs, suggesting that accessory genes (i.e., those present in plasmids and phages and the unique genes in chromosomes) involve codon usages that—most likely due to their nonessential character—are far from being optimized with respect to the host translation machinery. Strains with the characteristics described above have genomes with quite diverse GC contents, ranging from ca. 30% to more than 70%. Exceptions to the general increase in the m-tAI values with ancestry are likely due to m-tAI deficiencies to quantitate nonstandard codon-anticodon interactions (i.e., those different from WCIs, along with wobble base pairing) (35).

Effect of codon optimization on the GC content. An analysis of the prokaryote genomes with different GC contents enabled us to explore how the GC composition at the third base of codons (i.e., the GC3) changed in the core gene sets over ancestry and to compare the results with the GC3 in PHE genes and singletons. Since the first two positions in codons are constrained by the protein-coding information, most of the GC changes result in variations in synonymous codons (2). As we have seen in the two previous sections, core genes adjust their codon usages in the direction of the PHE genes (Fig. 1 and Fig. S1, left graphs) in order to improve translation (Fig. 2 and Fig. S4, left graphs). The question thus was raised as to how bacteria with different GC contents changed their GC3 compositions in the process of adapting their codon usages. The results presented in Fig. 2 and Fig. S4 (right graphs) show that changes in GC3 in genomes from groups A to D each follow a distinctive pattern as determined by comparing singletons to Ci-Cn to PHE genes. Whereas in genomes that belong to group A (overall GC content lower than ca. 35%) the GC3 decreases from singletons to Ci to PHE genes (cf. Fig. 2, panel b1), in the genomes included in group C the GC3 either increases from singletons to Ci to PHE (cf. Fig. 2, panel b3) or plateaus in Ci to PHE genes at a high level (cf. Fig. S4, panel b17). In contrast, genomes pertaining to group B exhibited a biphasic pattern with an initial GC3 increase from the level of the singletons up to the contents of the Ci series (with i varying from 1 to n) followed by a later decrease from the Cn values down to those of the PHE genes (cf. Fig. 2, panel b2). Those changes in the group B genomes were reflected in pronounced forward and backward movements in the position of the core genes in the CA plots, first from singletons to Ci and then from Cn to the PHE genes (cf. Fig. S1, organisms in group B). A similar biphasic pattern in the CA plots could also be recognized, though softened, in certain species that were included in group C or even group D, in which the PHE genes did not evidence a decrease in GC3 levels compared to those of the core genes. The genomes in group D had extremely high global GC contents and had GC3 values in all their core gene sets (C1 to Cn) that were comparable to—though slightly higher—than the corresponding values in their PHE genes. Next, we describe how individual codons for a given amino acid are selected in the most ancestral core gene sets.

Characterization of codons that improve adaptation to the tRNA pool. The variations in the use of individual codons when progressing from the C1 to the Cn gene sets were analyzed in the different prokaryote genomes, together with the tRNA gene copy numbers and the absolute adaptiveness values (*Wi*s; see Materials and Methods). Figure 3 and Fig. S5A illustrate the codon usage frequencies (CUFs; see Materials and Methods) for singletons, PHE genes, and core genes with increasing ancestry together with the tRNA gene copies and the *Wi*s (Fig. S5B summarizes the *Wi*s, ΔCn-C1, and ΔPHE-Cn in the different genomes studied). In agreement with previous reports (10),

FIG 2 Legend (Continued)
for each of the Ci gene sets as described in Materials and Methods is plotted on the ordinate as a function of the evolutionary distance indicated on the abscissa (Table S1a, tab 2) as inferred from the corresponding phylogenetic trees included in Table S1a to c. Higher values of m-tAI indicate an enrichment in the codon usage frequencies of those synonymous codons better adapted to the host cell tRNA pool. The C1 to Cn gene sets plotted are the same as those presented in Fig. 1. The red and blue horizontal dashed lines correspond to the respective m-tAI values calculated for the PHE genes and the singletons. (b1 to b4) In each panel, the average GC3 content in each core gene set of increasing ancestry is plotted on the ordinate as a function of the evolutionary distance indicated on the abscissa as in panels a1 to a4. The PHE genes and the singletons are represented as red and blue horizontal dashed lines, respectively.

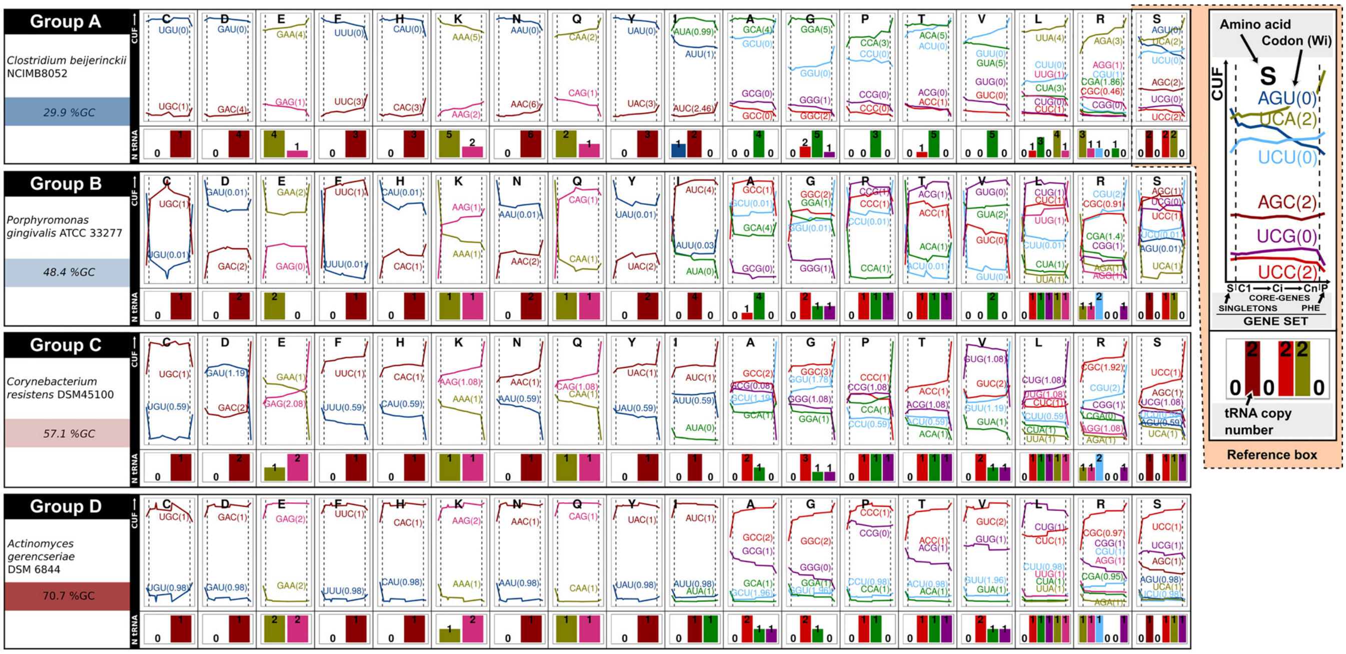

FIG 3 Codon usage frequencies and absolute adaptiveness values (*Wi*s) of the gene sets analyzed in this work, together with the tRNA gene copy numbers for strains of the four reference groups A to D. For the amino acid indicated by the corresponding single-letter identification code located above each panel, the change in the modal CUFs (see Materials and Methods) of the core gene sets with increasing ancestries (left to right, C1 to Cn), the PHE genes, and the singletons are plotted in the upper portions as solid horizontal curves for each of the indicated codon triplets between the two vertical broken lines, for the singletons to the left of the first of those lines, and for the PHE genes to the right of the second (with singletons and PHE genes being located at the beginning and the end of the curves, respectively). The CUFs are represented by different colors, with the associated absolute adaptiveness value (*Wi* [35]) being indicated within parentheses beside each triplet. Finally, the presence and gene copy number of the cognate tRNA species of a given synonymous codon bearing the exact complementary anticodon is depicted with a number and a bar of proportional height in the lower panel in the same color as the corresponding triplet and curve in the upper portion.

our results demonstrated that the CUFs among synonymous codons were strongly influenced by the global GC content in each genome—i.e., codons with G and C in the 3′ position (N_3) were the most abundant synonymous codons in the GC-rich genomes, whereas A and U were predominant in that position in the genomes with low GC contents (Fig. 3 and Fig. S5A). An inspection of the proportion of codon usage for each amino acid in ancestral cores compared to the most recent ones (curves in Fig. 3 and Fig. S5A and B) revealed that in most genomes a C-bias enrichment occurred with increased ancestry at the 3′ position of the 2-fold pyrimidine-ending codons—for Asp (GAC), Phe (UUC), His (CAC), Asn (AAC), and Tyr (UAC)—as well as in the unique 3-fold codons for Ile (AUC). Corresponding to the observed C bias, in all these examples high *Wi*s (shown in parentheses in the figure) were observed for the C-ending codons, which triplets were decoded through exact WCIs with the cognate tRNA species (i.e., with the anticodon $G_{34}N_{35}N_{36}$). Because of the absence of tRNA species bearing anticodons $A_{34}N_{35}N_{36}$ for these five amino acids, lower *Wi*s were obtained for the U-ending codons as the consequence of a weaker wobble codon-anticodon non-WCI recognition. Especially noteworthy was the observation that, though to a lesser extent, the bacteria with extremely low GC contents likewise exhibited a C bias in the 2- to 3-fold codon family, irrespective of a global decrease in the GC3 value, as in the example of *Clostridium beijerinckii* (cf. Fig. 2 and 3).

In the instance of the 2-fold purine-ending codons—that is, GAA and GAG for Glu, AAA and AAG for Lys, and CAA and CAG for Gln—we observed that the codons with G or A in the 3′ position were enriched from C1 to Cn and from Cn to PHE genes (i.e., ΔCn-C1 and ΔPHE-Cn in Fig. S5B, respectively) depending upon which tRNA species (anticodons) were present. In those examples where only the tRNAs bearing the $U_{34}N_{35}N_{36}$ anticodons were present, the cognate A-ending codons recognized by WCIs were the ones that became enriched in the most ancestral core and/or PHE genes (cf.

in Fig. S5B the GA$\underline{A}$ triplet for Glu in *Chromobacterium violaceum*, *Paenibacillus graminis*, *Bacillus subtilis*, *Bordetella holmesii*, and *Leisingera methylohalidivorans*, the AA$\underline{A}$ for Lys in *Methanobrevibacter smithii* and *Bacillus subtilis*, and the CA$\underline{A}$ for Gln in *M. smithii*, *Streptococcus equi*, and *B. subtilis*). Accordingly, these 3′ A-ending codons were associated with higher *Wi*s than the corresponding codons ending in G, as the latter were recognized only by wobble-base pairing (i.e., G_3-U_{34} interaction). In a second circumstance, where both tRNA species for the same amino acid (i.e., those bearing anticodon $U_{34}N_{35}N_{36}$ or $C_{34}N_{35}N_{36}$) were present, a more frequent enrichment in G-ending codons was observed (with few exceptions) since such codons can be decoded by either Watson-Crick or wobble interactions with the tRNA anticodon $C_{34}N_{35}N_{36}$ or $U_{34}N_{35}N_{36}$, respectively. In those few examples where the A-ending codons were more enriched than the G-ending codons, a higher copy number of the tRNA genes was always observed with anticodon $U_{34}N_{35}N_{36}$ than that obtained with the anticodon $C_{34}N_{35}N_{36}$ (cf. in Fig. 3 and Fig. S5A and B the GA$\underline{A}$ triplets for Glu in *Bacteroides vulgatus* and *C. beijerinckii*, the AA$\underline{A}$ triplets for Lys in *Sulfurospirillum multivorans*, and the CA$\underline{A}$ triplets for Gln in *C. beijerinckii* and *S. multivorans*).

A different codon usage bias—in a pattern not found in the 2-/3-fold degenerate amino acids—was observed in codons encoded by 4-fold degenerate amino acids (Val, Thr, Pro, Gly, and Ala) or by the 4-fold boxes of the 6-fold degenerate amino acids (Ser, Leu, and Arg). In these 4-fold groups, a U-bias enrichment (i.e., an NNU codon enrichment) was observed in the PHE genes from most of the genomes irrespective of their GC contents (Fig. 3 and Fig. S5A and B). This enrichment in U-ending codons, previously reported as a U bias (40), could not be explained by WCIs with $A_{34}N_{35}N_{36}$ tRNAs because the latter species are not present in prokaryotes, except in the case of Arg. The observed U bias likely occurred through the previously proposed nonconventional codon-U_3–anticodon-U_{34} interaction that was known to exist in bacteria (53). The presence of $U_{34}N_{35}N_{36}$ tRNA species might, then, lead to an increase in both NNA and NNU codons as a consequence of positive WCIs and U_3-U_{34} interactions, respectively.

All the codon adaptations that we have described in this section referring to core genes proved to be more prominent in the PHE genes, whose triplets were even better adapted to the translational machinery. Contrasting with such a strong pattern of selection-associated codon bias, the singletons displayed codon usages that were in general the most distant from those observed in the PHE genes (as exemplified in the CUFs in Fig. 3 and Fig. S5A and in the CA plots from Fig. 1 and Fig. S1). These observations are also in agreement with variations in the m-tAIs for the different gene sets presented in the previous section.

Search for coding signatures for translation efficiency and accuracy: codon usage profiles associated with sequences encoding HEP_*vr* and HEP_*cr* translated domains. Expression level and amino acid sequence conservation are both parameters that positively correlate with codon usage optimization (54). Nevertheless, the relative relevance of efficiency and accuracy to translation and the way in which either one of those parameters affects the other have not yet been investigated in detail. A central limitation that made such studies difficult was associated with the natural genomic heterogeneity in gene ancestry along with the expression level and the sequence conservation (structural constraints) in the translated products. In order to reduce the degrees of freedom in the analysis, for each of six different bacterial species, we created two distinct gene sets based on the experimental proteome data. One of those gene sets consisted of genes encoding proteins with the highest expression levels in the cell (i.e., the HEP), while the other was associated with proteins with low cellular abundance (i.e., the LEP). Then, the conserved (*cr*) and variable (*vr*) sequences among the orthologs were collected from each individual gene (see Materials and Methods), and the corresponding highly expressed conserved (HEP_*cr*), highly expressed variable (HEP_*vr*), lowly expressed conserved (LEP_*cr*), and lowly expressed variable (LEP_*vr)* modal codon usages were used to calculate the relative distances illustrated in the neighbor-joining tree presented in Fig. 4. In five out of the six species present in the trees (all except *Mycobacterium fortuitum*), the HEP_*cr* and HEP_*vr* sequences separated from

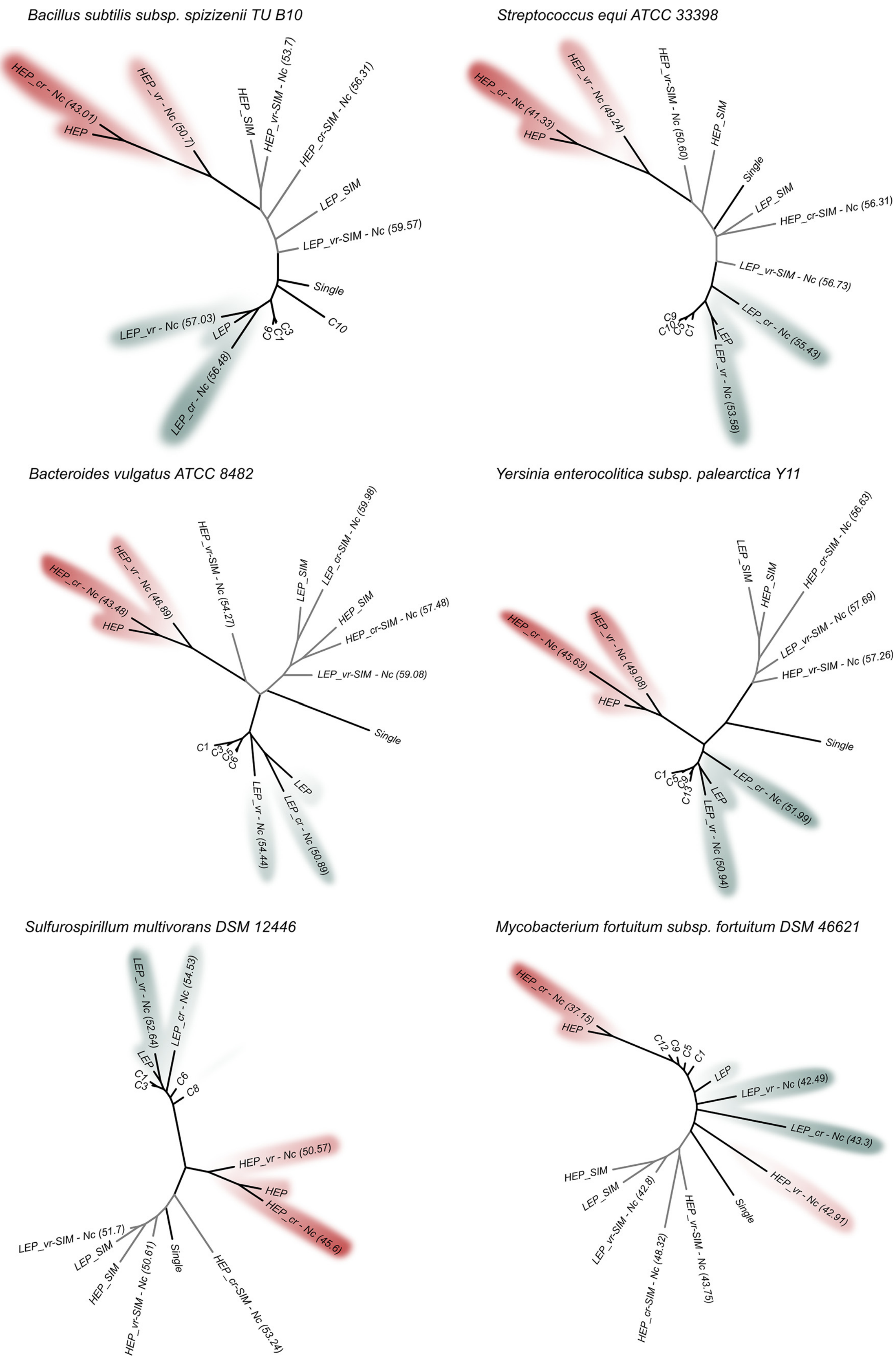

FIG 4 Neighbor-joining distance trees of different gene sets encoding HEP, LEP, and their associated conserved (*cr*) and variable (*vr*) regions based on the corresponding modal codon usage. Modal codon usage-based neighbor-joining trees with black branches were

(Continued on next page)

those of the singletons, the core genes, and all the LEPs as a result of a strong codon usage adaptation (also reflected in the low effective number of codons [Nc_s] associated with the HEPs, indicated in parentheses following labels in the tree). Furthermore, the large distance in the tree between HEP_*cr* and LEP_*cr* (where both sequences encode regions with conserved amino acids but with different expression levels) compared to the much shorter distance between HEP_*cr* and HEP_*vr* (where both encode highly expressed products with different degrees of conservation) pointed to the quantitatively stronger effect of efficiency over accuracy in shaping codon usage bias. Control data sets were incorporated into the trees in Fig. 4 (branches in gray) using artificially evolved sequences with no pressure for codon selection (see Materials and Methods). The results show that, as expected, for the six analyzed genera the distance between the most and the least selected gene sets (i.e., distance from HEP_*cr* to LEP_*vr*) was always larger in natural genes than in the simulated sequences without selection (i.e., [distance from HEP_*cr* to LEP_*vr*]$^{\text{natural}}$/[distance from HEP_*cr* to LEP_*vr*]$^{\text{simulated}}$ $>$ 1, with an average value $\pm$ standard deviation [SD] = 1,69 $\pm$ 0.20). Thus, natural sequences display more divergent (positively adapted) codon usages. It is noteworthy that SIM sequences tended to group with singletons—the least adapted gene set—and had in general higher Nc values than their corresponding natural sequences.

Codons that were optimized as a result of accuracy under high and under low expression—i.e., HEP_*cr*–HEP_*vr* and LEP_*cr*–LEP_*vr*, respectively, labeled "*A*" for "accuracy" at the bottom of Fig. 5—were highly coincident with the codons that were optimized through efficiency—i.e., HEP_*cr*–LEP_*cr* and HEP_*vr*–LEP_*vr*, labeled "*E*" for "efficiency." For some organisms, the greater distance between HEP_*cr* and HEP_*vr* than between LEP_*cr* and LEP_*vr* (Fig. 4) indicates a stronger influence of accuracy in codon usage optimization when operating under high-expression conditions, thus pointing to an interaction between the simultaneous requirements of high fidelity and efficiency. The most relevant contributions to the global difference in codon usage between HEP and LEP were efficiency (both in conserved and in variable regions) (*E* columns in Fig. 5) followed by accuracy under high expression (first *A* column in Fig. 5) (the stronger the contribution of each factor, either *E* or *A*, the shorter the distance in brackets to HEP-LEP in the figure). The heat maps display the complete profiles of preferred codons for sequences requiring high translational accuracy and/or efficiency (protein demands). As expected, the preferred codon for each amino was in agreement with the C and U bias and the tRNA copy number described in the previous sections. In light of these results, the highly expressed variable and conserved domains constitute the basis for explaining the observed codon usage optimization in the most ancestral core gene sets (Cn), which concentrate HEPs (Table S3). Figure 6 illustrates that HEP sequences (red dots) are those under the highest selective pressure to optimize codon usage because of both their expression level and their degree of conservation.

DISCUSSION

Since gene adaptation to a host cell is expected to be associated with an improved codon selection for translation efficiency and accuracy (42, 55), we investigated corre-

FIG 4 Legend (Continued)
constructed for the indicated natural gene sets and their intragenic regions (*cr* and *vr*) following the method described by Karberg et al. (17) along with the neighbor-joining program of the Phylip package (62). Artificially simulated sequences were used as controls in the neighbor-joining tree (SIM labeled data and gray branches in the tree). Such artificially generated sequences were evolved under a model with no pressure for codon selection and preserving the same K_A/K_S ratio as that corresponding to each of their natural HEP/LEP set of homologs (see Materials and Methods). LEP_*cr*-SIM sequences are not included since, on average, fewer than 53 conserved amino acid positions/protein were collected in the simulation. Phylogenetic trees were drawn through the use of the Figtree application (59). Abbreviations: C1 to Ci, core gene sets with increasing ancestry; single, singletons; HEP, genes encoding proteins with the highest expression level; LEP, genes encoding proteins with the lowest expression level; HEP_*cr*, conserved HEP sequences (dark red); HEP_*vr*, variable HEP sequences (light red); LEP_*cr*, conserved LEP sequences (dark blue); and LEP_*vr*, variable LEP sequences (light blue). HEP and LEP *cr* and *vr* subfractions were recovered as indicated in Materials and Methods through the use of the polypeptide sequences included in C13 for *Yersinia enterocolitica* subsp. *palearctica* Y11, C10 for *Streptococcus equi* ATCC 33398, C8 for *Sulfurospirillum multivorans* DSM 12446, C9 for *Bacillus subtilis* subsp. *spizizenii* TU B 10, C6 for *Bacteroides vulgatus* ATCC 8482, and C12 for *Mycobacterium fortuitum* subsp. *fortuitum* DSM 46621 (ATCC 6841). The effective number of codons (Nc_s) as previously defined by Wright (71) are indicated in parentheses for the *cr* and *vr* subset of sequences.

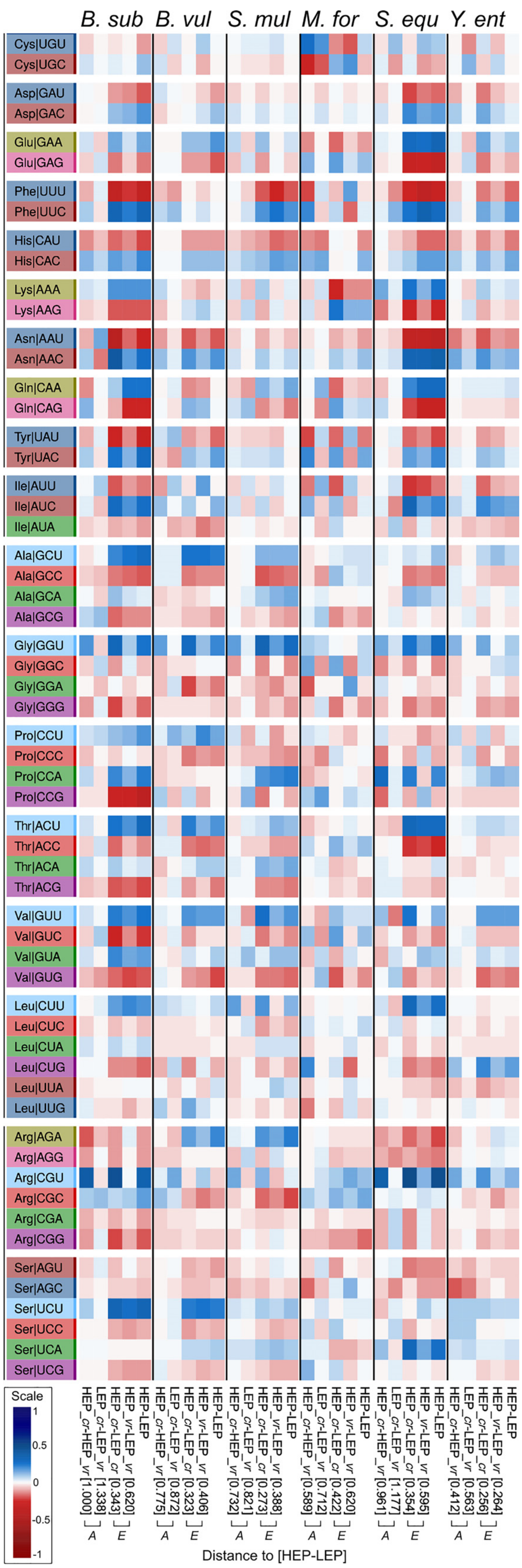

FIG 5 Heat map representation expressing differences in modal codon usage profiles between the indicated gene sets. The color scale from red to blue indicates the relative level of use of each particular

(Continued on next page)

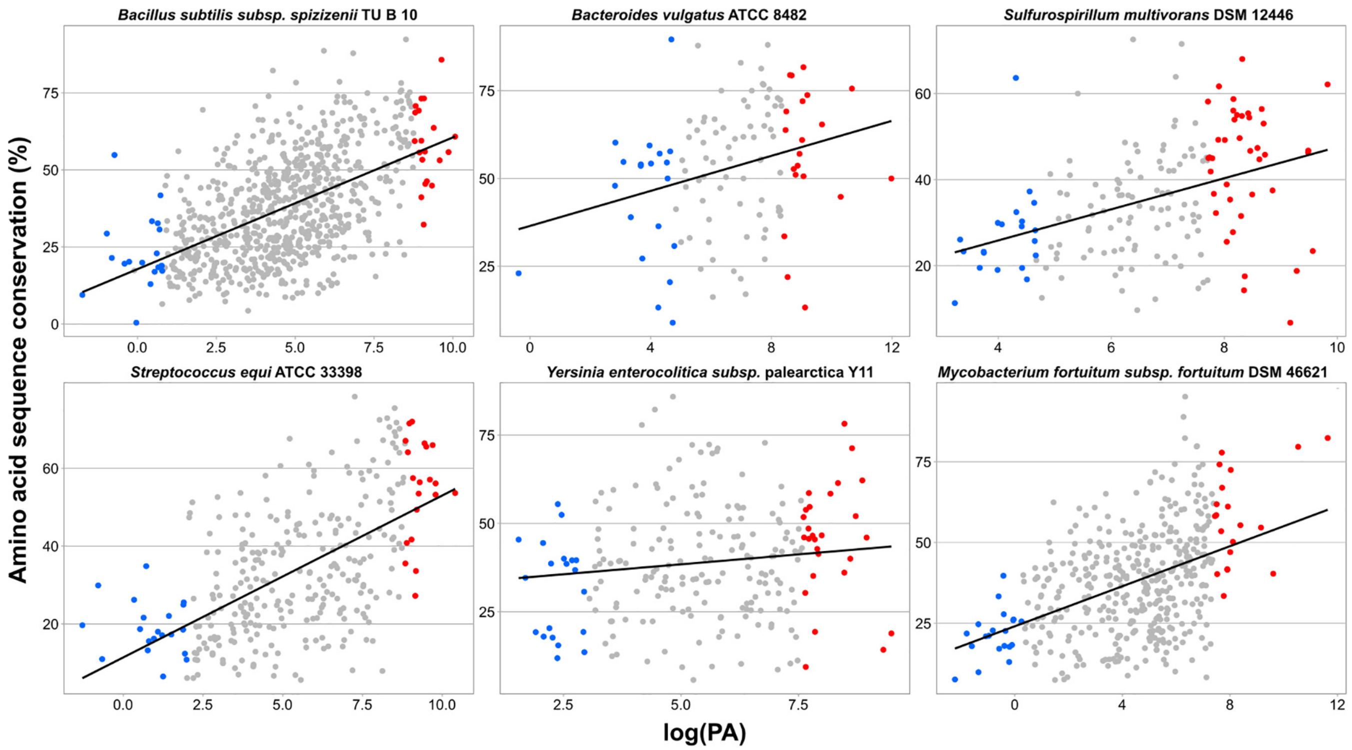

FIG 6 Amino acid sequence conservation in proteins with different cellular abundances. The amino acid sequence conservation calculated for proteins of the indicated bacterial species and core fractions (see Materials and Methods) are plotted on the ordinate as a function of the logarithm of the corresponding protein abundance (logPA) on the abscissa. The red and blue dots correspond to HEP and LEP, respectively, with all the other proteins of the same core represented in gray. The linear regressions and graphs were all made with the ggplot2 library from the R package.

lations between core gene ancestry and their modal codon usage within a given prokaryote family. In order to ascertain if the adaptation of the most ancestral core genes was an extensive phenomenon among prokaryotes, we analyzed core modal codon usages in 27 different species of *Bacteria* and 2 of *Archaea*. That in the CA plots the most ancestral core genes had been the ones with the closest location to the PHE genes in all families was remarkable and strongly indicated a core codon usage adaptation that likely operated to improve translation. In agreement with the position of the different gene sets in the CA plots, the m-tAI values served to confirm that the PHE genes were the best-adapted gene set, followed by the Cn to C1 core genes, in that order, and finally by the singletons, with those being the least adapted genes with the lowest m-tAIs in the genome. Studies by others have previously demonstrated that the level of gene expression together with the need to preserve accuracy during the translation of conserved amino acid regions are both among the main parameters that

FIG 5 Legend (Continued)
codon in a gene set compared to that of another (i.e., gene set 1 versus gene set 2). The blue color corresponds to the dominant use of a particular codon in gene set 1 over the use of the same codon in gene set 2 (and vice versa for the red color). Amino acids are indicated in the standard three-letter code. The heat map was constructed through the use of the phytools R package (72). Distance between gene sets was determined using their corresponding modal codon usages as previously reported (46). "HEP (gene set 1)–LEP (gene set 2)" represents the profile of the optimized codons when comparing the coding strategies in high- versus low-expression genes (i.e., reflecting differences in their modal codon usages). The columns indicated by "*A*" correspond to the profiles of codons optimized as a result of accuracy (i.e., differences between HEP_*cr*–HEP_*vr* and LEP_*cr*–LEP_*vr*). The columns indicated by "*E*" correspond to the profiles of optimized codons through high expression (i.e., reflecting differences in efficiency between HEP_cr–LEP_*cr* and HEP_*vr*–LEP_*vr*). The numbers in brackets indicate the extent to which changes induced by either efficiency or accuracy approach the differences in codon usage between HEP and LEP (i.e., distances from each column to the column HEP–LEP). The shorter the distance in brackets the stronger the evolutionary constraint and the contribution of the indicated factor (i.e., accuracy or expression level) to codon optimization.

govern codon usage selection (54). The bioinformatics isolation of conserved (*cr*) and variable (*vr*) coding sequence domains from genes under high-expression (HEP) and low-expression (LEP) regimes served in this work to ascertain quantitatively the relative contribution of efficiency (expression level) versus accuracy during the selection-based codon usage optimization. According to the observed neighbor-joining distances (Table S3 worksheet "distances" and tree in Fig. 4), changes in codon usage derived from differences in gene expression levels—i.e., the efficiency in terms of the distance from the LEP to the HEP—were between 1.25 to 2.35 times greater than the changes in codon usage resulting in increased accuracy—i.e., the distance from *vr* to *cr*. The increasing proportion of highly expressed variable and specially conserved sequences (i.e., HEP_*vr* and HEP_*cr*) in the most ancestral gene sets constituted the basis for explaining the corresponding high degree of codon usage optimization that gradually increased from C1 to Cn.

The central question therefore was how adaptive changes in codon usage—which alterations become reflected in m-tAI values— occurred in prokaryotes with quite diverse GC contents (10). Because of the small amount of intergenic DNA in prokaryotes, genomic differences in base composition must mainly derive from changes in the coding regions. Within the alterations in the open reading frames, changes in GC are preferentially associated with modifications in the GC3, and only to a lower extent with alterations in the GC content of the first two codon positions (2, 4). How mutational bias (12) competes with selection (15) to drive all these changes is not yet fully understood. The codon usage biases described here were associated with the four different patterns of GC3 changes summarized in the schemes presented in Fig. 7 (i.e., the genome groups A, B, C, and D). The group A genomes, those having an extremely low GC content and with their GC3 frequency decreasing from C1 to Cn, proved to have only the tRNA-U_{34} to recognize 4-fold synonymous codons in one or more amino acids. In such instances, the observed core-gene AT enrichment over ancestry appeared to be directly affected by selection (as with the PHE genes), where codons NNA (via WCIs with the tRNA-U_{34}) and NNU (via nonconventional U-U interactions) were preferentially enriched over NNC and NNG codons. Though both of those changes were probably related to improvements in translation efficiency, such increases are not always reflected in the m-tAIs since, as stated earlier, U-U interactions are not considered in the calculation of that index. Unfortunately, when we (data not shown) and others (36) have attempted to improve the tRNA adaptation index by including additional nonstandard base pairings, we obtained no better results. Nonetheless, under the assumption that the PHE genes are the best adapted to the translational machinery, in genomes with extremely low GC content—such as those belonging to group A—the observed AT3 enrichment from C1 to Cn to PHE (Fig. 7, right side) should mainly result from selection. According to Hildebrand et al. (15), the mutational processes in very-low-GC organisms favor a GC3 enrichment. That the core and PHE genes in bacteria that belonged to group A had been selected to bear lower GC3 values than singletons in order to improve translation in view of the previous pattern of increasing GC content was remarkable, with this circumstance being a result of the above-mentioned enrichment in NNA and NNU triplets compared to their proportion in the synonymous codons (Fig. 7, right side). In group B genomes, the biphasic pattern observed from singletons to PHE genes could be explained by an initial increase in GC3-rich codons from singletons to core genes, followed by a later U bias from core genes to PHE genes. That initial GC3 enrichment followed by a U3 increase was sufficient to explain the basis of the previously reported "rabbit head" distribution of codon usages that characterizes most prokaryote genomes (21, 56). What should be also especially noted is that the PHE genes separated from the Cn (in both the CA and the GC3 plots) because of a much more intense U bias likely associated with the difference in expression levels between the two gene sets. In the type C genomes, in which the GC3 always increased, the absence of a strong U bias from the Cn to the PHE genes led to a less pronounced—i.e., more linear—"rabbit-head" distribution of genes in the CA plot. In addition to that general trend, *Yersinia enterocolitica*, *Methanolacinia petrolearia*, and *Sphingomonas*

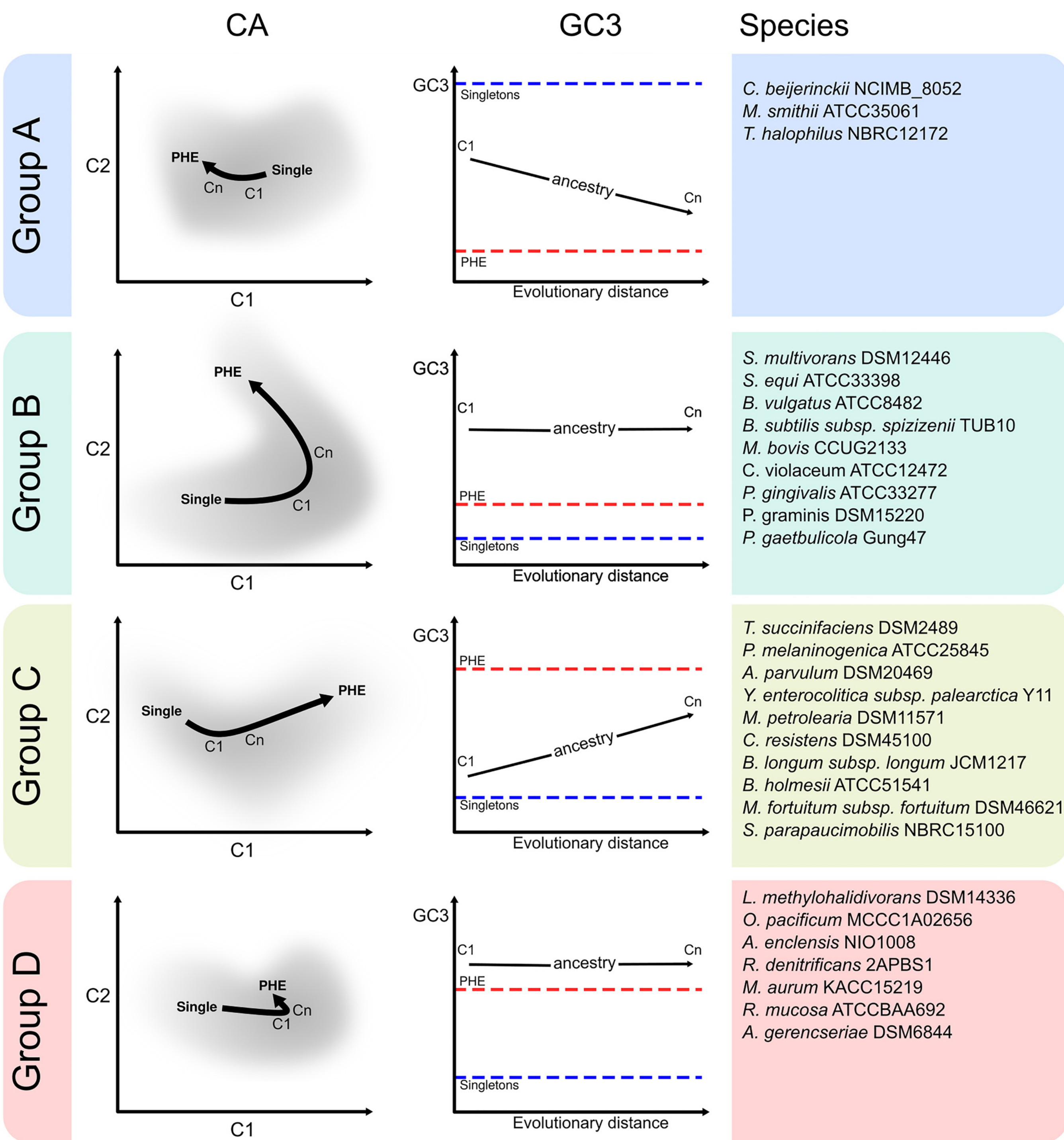

FIG 7 Schematic representation in a cartoon format of general codon-usage patterns observed in different prokaryote families. For the prokaryote strains whose genomes were classified as belonging to groups A, B, C, and D and which are listed to the right of each set of graphs, cartoons with the associated correspondence analysis and GC3 variation pattern among the core gene sets of increasing ancestry (light gray) are presented, along with the corresponding singletons (blue) and PHE genes (red). The light blue arrow indicates the direction of the U bias and the red arrow that of the C bias. The right graph is a plot of the GC3 content on the ordinate as a function of increasing evolutionary distance on the abscissa, with the red horizontal dashed line indicating the PHE genes and the blue the singletons.

parapaucimobilis could be considered as having behavior intermediate between that of the bacteria in group C and that of the bacteria in group B. Finally, the group D genomes, which had extremely high GC contents, were the most restricted with respect to GC3 variations. The quite small compositional variations in that group of genomes became apparent in the compacted location of the different core and PHE genes in the CA plots. What was remarkable is that in group D genomes a U bias (though much less intense than in the genomes of groups A, B, and C) was still a visible variable that differentiated codon usages between the core and the PHE genes. As stated above, the noninclusion of U-U interactions in the m-tAI calculation limited the use of this parameter to expresses the translational adaptation of those gene sets in which a U bias was dominant. Pouyet et al. (11) present a model to predict and separate the relative contribution of mutational bias (N layer), codon selection (C layer), and amino acid composition (A layer) on the global GC content and the GC3 content. Our analysis is fully consistent with the results reported by Pouyet et al. (11) where the C layer (codon selection/translational selection) has a stronger influence on the GC3 of genes with low effective number of codons (Nc) (such as Cn and PHE) compared to the influence on genes with the highest Nc (such as C1).

The results presented here together with previous evidence from other authors have enabled a comprehensive analysis of the principal basis underlying the changes associated with the optimization of codon usage in prokaryotes in different gene sets and in organisms with different GC contents. As stated previously, the overall codon usage is known to be constrained by genome-wide mutational processes (7, 8, 10) that are considered a main force in shaping the global GC content. The intragenomic codon usage will concurrently become accommodated through selection-driven processes, as has also been extensively reported (32, 34, 41, 47). In order to further our knowledge of the relevance of those factors/forces generating intragenomic variations, we investigated the different nucleotide base changes underlying the selection of preferred codons in the core and PHE genes of representative prokaryote species. The analysis of gene sets with different expression levels and degrees of conservation in organisms with diverse global GC contents enabled a description of how core codon usage approaches that existing in the PHE genes and how nucleotide changes correlate with an improved compatibility between the genes and the coexisting tRNA pool. That C- and U-ending codons in 2-/3-fold and in 4-fold degenerate amino acids, respectively, were specifically enriched as a result of selection to improve translation has been previously reported for different prokaryotic genomes (40). Using separate analyses focused on different gene sets, we demonstrated in this study that similar selection-driven adaptations in codon usage have taken place from singletons to core genes to PHE genes. The intensity and relevance of the C and U biases were dependent on the particular genome—and especially on the genomic GC content—as well as on the gene fractions under consideration. In contrast to the codon usage variations occasioned by selection in the core and PHE genes, the singletons constituted the gene set characterized by both the lowest GC3 content as a result of the AT mutation that is universally biased in prokaryotes (12) and a much more relaxed selection than that of the most ancestral genes, with the sole exception of the extremely low-GC-containing genomes of group A. In addition to a description of the basic changes that together conform the intracellular codon usage variations, further investigation should be focused on the analysis of the time course required by the newly acquired information to be properly incorporated into the genetic language of the host cell (codon usage tuning).

MATERIALS AND METHODS

Prokaryote families selected for analysis in this study and identification of core genes and singletons. We screened the EDGAR public project database (57, 58) available at https://edgar.computational.bio.uni-giessen.de/cgi-bin/edgar_login.cgi, chose several prokaryote families that included at least 20 complete genomes each, and finally selected 27 bacterial and 2 archaeal families (Table S1a, tab 1). A specific core gene set was defined as a group of genes whose orthologs are present in a given set of species under investigation. For each of the families selected, sequential core gene sets with increasing ancestry (C1 through Cn) were calculated. To that end, first the phylogenetic tree for each family was extracted from EDGAR and one species per family was chosen as a reference. Next, the different core gene sets were obtained by incorporating into the analysis new species having sequentially increasing phylogenetic distances from the reference strain (accordingly, by following the tree from the branches to the root). Table S1a to c lists the phylogenetic trees used for these calculations as well as the particular species that were included in each core gene set (C1 to Cn) for the different prokaryote families. The phylogenetic trees were edited with the Figtree (59) and Inkskape programs (TEAM-

Inkscape). At least six core gene sets differing in size from ca. 50 to 100 genes each were calculated per family. Table S2, tabs 2 to 30, lists the singletons—those corresponding to genes that were specific to the reference strains with no orthologs within the family—as calculated with EDGAR.

PHE genes. For each of the selected reference genomes, we obtained a set of genes encoding ribosomal proteins and tRNA synthetases (24, 60). Table S2, tab 1, itemizes the PHE (putative highly expressed genes) whose orthologs were obtained and analyzed in each reference genome.

Codon usage diversity groups. The prokaryote species studied in this work were classified into four different groups based on the compositional characteristics of their codon usage. Eubacterial and archaeal species were classified into groups A to D according to their global GC contents and to the relative GC3 contents (i.e., percent GC at the third position of codons) among their core genes (C1 to Cn), PHE, and singletons; as follows: group A, which included species with very low global GC (<36%) and where $GC3^{singletons} > GC3^{C1-Cn} > GC3^{PHE}$; group B, which included species with low to intermediate global GC (48% in average) and where $GC3^{C1-Cn} >>> GC3^{PHE}$ and $GC3^{singletons}$; group C, which included species with intermediate global GC (53% on average) where $GC3^{PHE} > GC3^{C1-Cn} > GC3^{singletons}$; and group D, which included species with very high global GC (68% in average) and where $GC3^{C1-Cn}$ was greater than or comparable to $GC3^{PHE} > GC3^{singletons}$. Groups A and group D were those that included the species with lower and higher global GC contents, respectively.

Highly and lowly expressed proteins within the same core gene set. Integrated expression data from the Protein Abundance Database (PaxDB [61]) were retrieved for the bacterial strains *Yersinia pestis* CO92, *Streptococcus pyogenes* M1 GAS, *Campylobacter jejuni* subsp. *jejuni* NCTC 11168, *Bacillus subtilis* subsp. *subtilis* strain 168, *Bacteroides thetaiotaomicron* VPI 5482, and *Mycobacterium tuberculosis* H37Rv. Assuming that orthologs have comparable expression levels within the same—or closely related—species and using the PaxDB data from the above-indicated 6 strains, we inferred putative expression data for the proteomes of the microorganisms presented in Fig. 4 to 6 and listed in Table S3. Then, for selected core fractions, we obtained one subset of genes encoding highly expressed proteins (HEP) plus another subset codifying lowly expressed proteins (LEP). For 23 out of the 29 prokaryotic genomes that we analyzed, no proteome data were available, nor were any in phylogenetically related microorganisms.

Analysis of codon usage in gene sequence regions that encode either conserved or variable amino acid positions in the HEP and LEP subsets. Individual genes that belonged to the HEP and LEP groups were aligned with the corresponding orthologs. Then codons corresponding to conserved and variable amino acid positions in the HEP genes were separated and each concatenated to generate the HEP_*cr* and HEP_*vr* sequence groups. Through the use of a similar procedure with the LEP genes, the LEP_*cr*, and LEP_*vr* sequences were also generated. Codons categorized as belonging to the *cr* group were those associated with positions with fully conserved amino acids throughout the alignment. Codons categorized as belonging to the *vr* group were those associated with positions where none of the amino acids aligned (at that specific point) reached a proportion higher than 0.5. The modal codon usage (46) of each collection of *cr* and *vr* sequences were calculated and used for further analysis.

Codon composition based HEP/LEP_*cr*/*vr* distance trees: control trees with artificially evolved sequences under no pressure for codon selection. Modal codon usage-based neighbor-joining trees were constructed for the indicated gene sets and intragenic regions (*cr* and *vr*) following the method described by Karberg et al. (17) along with the neighbor-joining program of the Phylip package (62). Artificially simulated sequences were used as controls in the neighbor-joining tree. Such generated sequences were evolved under a model with no pressure for codon selection and preserving the same K_A/K_S ratio (i.e., the ratio between non-synonymous to synonymous substitutions) as that corresponding to each of the natural HEP/LEP set of homologs. Amino acid and codon alignments were generated with TranslatorX using MUSCLE (http://translatorx.co.uk/). For proteins in both HEP and LEP groups, we inferred the most likely evolutionary model using the amino acid alignments and modeltest-ng (63) as well as a maximum likelihood phylogenetic tree using codonphyml (64). Next, we used both the inferred trees and codon alignments to optimize a codon evolutionary model using codeml from the PAML suite (65). For simplicity, an M0 model with F3×4 codon equilibrium frequencies was used. F3×4 frequencies avoided the introduction of compositional biases. The obtained parameters of the model, which included the K_A/K_S value for each set of HEP and LEP orthologs, were used to generate simulated DNA sequences for each protein using PAML evolver software. Such artificially generated sequences were used to recover the *cr*/*vr* simulated data sets (namely, the SIM data sets) using CUBACR and the procedure described above to obtain the natural HEP_*cr*/*vr* and LEP_*cr*/*vr* data sets (see previous section). The scripts used for this analysis are included in CUBACR (https://github.com/maurijlozano/CUBACR).

Correspondence analyses. (i) RCC-based analyses. Raw-codon-count (RCC)-based correspondence analyses (CA) were performed using bash and R-software homemade scripts which can be found at the CUBES software page (this work; available at https://github.com/maurijlozano/CUBES). Briefly, G. Olsen codon usage software was used to count codons on coding sequences (available at http://www.life.illinois.edu/gary/programs.html), data were loaded on R, and the correspondence analyses were run using the FactoMiner (66) and Factoextra (https://CRAN.R-project.org/package=factoextra) packages. Plots were made using the ggplot2, ggrepel, ggthemes, and gtools R packages. For each core gene set, the CA coordinates were calculated as the arithmetic mean of the first and second dimensions of all the genes present in that set (centroids). Then, a plot was generated containing all the coding sequences, together with the projections of the core-gene sets (C1 to Cn), the singletons and PHE genes.

(ii) RSCU-based CAs and calculation of modal codon usages. The CAs based on the use of relative synonymous codon usages (RSCUs) (67) of all individual genes from a given genome were calculated by CodonW (68). The modal codon usages (46) were calculated for the core genes, singletons, and PHE genes. Artificial sequences representing modal codon usages (i.e., modal sequences) and the amino acid composition present in each core fraction (Cn) were generated through the use of a homemade Perl script (calculate_modals2.pl) from the CUBES package and incorporated into the CA. In order to accurately represent the modal codon usage, particularly for synonymous codons from low-abundance amino acids, modal sequences were designed with a length of at least 10,000 codons. RSCU-based CA plots (see the supplemental material) were generated through the use of the Ggplot2 program (69) and edited with Inkskape (TEAM-Inkscape).

tRNA gene copy number and m-tAI. The gene copy number of each tRNA in the different prokaryote species studied was downloaded from the GtRNAdb server (http://gtrnadb.ucsc.edu), which uses predictions made by the program tRNAscan-SE (70). For each reference genome, the copy number for the tRNAs and the sequences of all the open reading frames were used to train the S_{ij} weights as previously reported, with that parameter estimating the efficiency of the interaction between the *i*th codon and the *j*th anticodon in a given organism (35, 36). The procedure is, briefly, as follows. With randomly generated S_{ij} starting points—i.e., having values that range between 0 and 1—the tAI was calculated for each coding sequence by means of the tAI package (35; https://github.com/mariodosreis/tai). Next, the directional codon bias score (DCBS [36]) was calculated through the use of the script seq2DCBS.pl (CUBES package). Finally, the Nelder-Mead optimization algorithm from R project was used (instead of the hill-climbing algorithm) to search for the S_{ij} value that maximized the Spearman rank correlation between the DCBS and the tAI. A script for bulk S_{ij} estimation is available in the CUBES package (https://github.com/maurijlozano/CUBES, calculate_sopt_DCBS_GNM_f.sh). The estimated sets of S_{ij} weights were used to calculate both the *Wi*s—i.e., the absolute adaptiveness values—as originally reported by dos Reis et al. (35) and the modal tRNA-adaptation index (m-tAI) for different species and gene sets (i.e., core and PHE genes plus singletons) as a measure of their efficiency in being recognized by the intracellular tRNA pool. The m-tAIs were calculated from the previously generated modal sequences by means of the tAI_Modal_g.sh script from the CUBES package. The Spearman coefficient was used to characterize how m-tAIs changed, from C1 to Cn, over ancestry.

SUPPLEMENTAL MATERIAL

Supplemental material is available online only.
FIG S1, PDF file, 2.9 MB.
FIG S2, PDF file, 0.5 MB.
FIG S3, PDF file, 0.4 MB.
FIG S4, PDF file, 2.9 MB.
FIG S5, PDF file, 1.8 MB.
TABLE S1a, XLSX file, 2.3 MB.
TABLE S1b, XLSX file, 2.7 MB.
TABLE S1c, XLSX file, 1.5 MB.
TABLE S2, XLSX file, 2.7 MB.
TABLE S3, XLSX file, 1.5 MB.

ACKNOWLEDGMENTS

We are grateful to Paula Giménez, Silvana Tongiani (both members of CPA CONICET at IBBM), and to Ruben Bustos from UNLP for their technical assistance and to Donald F. Haggerty for editing the final version of the manuscript.

This research was supported by the National Science and Technology Research Council (Consejo Nacional de Investigaciones Científicas y Técnicas, Argentina; PIP 2014-0420; and PUE-CONICET-2016-22920160100090CO to the IBBM), the Ministry of Science Technology and Productive Innovation (Ministerio de Ciencia Tecnolología e Innovación Productiva [MinCyT], Argentina; PICT-2012-1719, PICT-2015-2452, and PICT-2017-2022), and Ciencia y Tecnología para el Desarrollo (CYTED; acción 115RT0492). J.L.L., M.J.L., and M.L.F. were supported by CONICET, and A.L. was supported by CONICET and by the UNLP (Universidad Nacional de La Plata).

This paper is dedicated to the memory of the late Gabriel Favelukes (deceased 2020) in heartfelt recognition of his foundational work in the biochemical sciences at the Facultad de Ciencias Exactas, Universidad Nacional de La Plata.

REFERENCES

1. Sueoka N. 1962. On the genetic basis of variation and heterogeneity of DNA base composition. Proc Natl Acad Sci U S A 48:582–592. https://doi.org/10.1073/pnas.48.4.582.
2. Sueoka N. 1988. Directional mutation pressure and neutral molecular evolution. Proc Natl Acad Sci U S A 85:2653–2657. https://doi.org/10.1073/pnas.85.8.2653.
3. Osawa S, Ohama T, Yamao F, Muto A, Jukes TH, Ozeki H, Umesono K. 1988. Directional mutation pressure and transfer RNA in choice of the third nucleotide of synonymous two-codon sets. Proc Natl Acad Sci U S A 85:1124–1128. https://doi.org/10.1073/pnas.85.4.1124.
4. Sueoka N. 1992. Directional mutation pressure, selective constraints, and genetic equilibria. J Mol Evol 34:95–114. https://doi.org/10.1007/BF00182387.
5. Sueoka N. 1995. Intrastrand parity rules of DNA base composition and

usage biases of synonymous codons. J Mol Evol 40:318–325. https://doi.org/10.1007/BF00163236.

6. Sueoka N. 1999. Two aspects of DNA base composition: G+C content and translation- coupled deviation from intra-strand rule of A = T and G = C. J Mol Evol 49:49–62. https://doi.org/10.1007/PL00006534.
7. Knight RD, Friedland SJ, Landweber LF. 2001. A simple model based on mutation and selection explains trends in codon and amino-acid usage and GC composition within and across genomes. Genome Biol 2:RESEARCH0010. https://doi.org/10.1186/gb-2001-2-4-research0010.
8. Chen SL, Lee W, Hottes AK, Shapiro L, Mcadams HH. 2004. Codon usage between genomes is constrained by genome-wide mutational processes. Proc Natl Acad Sci U S A 101:3480–3485. https://doi.org/10.1073/pnas.0307827100.
9. Sharp PM, Stenico M, Peden JF, Lloyd AT. 1993. Codon usage: mutational bias, translational selection, or both? Biochem Soc Trans 21:835–841. https://doi.org/10.1042/bst0210835.
10. Hershberg R, Petrov DA. 2009. General rules for optimal codon choice. PLoS Genet 5:e1000556. https://doi.org/10.1371/journal.pgen.1000556.
11. Pouyet F, Bailly-Bechet M, Mouchiroud D, Guéguen L. 2016. SENCA: a multilayered codon model to study the origins and dynamics of codon usage. Genome Biol Evol 8:2427–2441. https://doi.org/10.1093/gbe/evw165.
12. Hershberg R, Petrov DA. 2010. Evidence that mutation is universally biased towards AT in bacteria. PLoS Genet 6:e1001115. https://doi.org/10.1371/journal.pgen.1001115.
13. Bohlin J, Eldholm V, Brynildsrud O, Petterson J-O, Alfsnes K. 2018. Modeling of the GC content of the substituted bases in bacterial core genomes. BMC Genomics 19:589. https://doi.org/10.1186/s12864-018-4984-3.
14. Bulmer M. 1991. The selection-mutation-drift theory of synonymous codon usage. Genetics 129:897–907.
15. Hildebrand F, Meyer A, Eyre-Walker A. 2010. Evidence of selection upon genomic GC-content in bacteria. PLoS Genet 6:e1001107. https://doi.org/10.1371/journal.pgen.1001107.
16. Raghavan R, Kelkar YD, Ochman H. 2012. A selective force favoring increased G+C content in bacterial genes. Proc Natl Acad Sci U S A 109:14504–14507. https://doi.org/10.1073/pnas.1205683109.
17. Karberg KA, Olsen GJ, Davis JJ. 2011. Similarity of genes horizontally acquired by *Escherichia coli* and *Salmonella enterica* is evidence of a supraspecies pangenome. Proc Natl Acad Sci U S A 108:20154–20159. https://doi.org/10.1073/pnas.1109451108.
18. Bohlin J, Eldholm V, Pettersson JHO, Brynildsrud O, Snipen L. 2017. The nucleotide composition of microbial genomes indicates differential patterns of selection on core and accessory genomes. BMC Genomics 18:151. https://doi.org/10.1186/s12864-017-3543-7.
19. López JL, Lozano MJ, Lagares A, Fabre ML, Draghi WO, Del Papa MF, Pistorio M, Becker A, Wibberg D, Schlüter A, Pühler A, Blom J, Goesmann A, Lagares A. 2019. Codon usage heterogeneity in the multipartite prokaryote genome: selection-based coding bias associated with gene location, expression level, and ancestry. mBio 10:e00505-19. https://doi.org/10.1128/mBio.00505-19.
20. Grantham R, Gautier C, Gouy M, Jacobzone M, Mercier R. 1981. Codon catalog usage is a genome strategy modulated for gene expressivity. Nucleic Acids Res 9:213. https://doi.org/10.1093/nar/9.1.213-b.
21. Médigue C, Rouxel T, Vigier P, Hénaut A, Danchin A. 1991. Evidence for horizontal gene transfer in *Escherichia coli* speciation. J Mol Biol 222:851–856. https://doi.org/10.1016/0022-2836(91)90575-q.
22. Grosjean H, Fiers W. 1982. Preferential codon usage in prokaryotic genes: the optimal codon-anticodon interaction energy and the selective codon usage in efficiently expressed genes. Gene 18:199–209. https://doi.org/10.1016/0378-1119(82)90157-3.
23. Kanaya S, Yamada Y, Kudo Y, Ikemura T. 1999. Studies of codon usage and tRNA genes of 18 unicellular organisms and quantification of *Bacillus subtilis* tRNAs: gene expression level and species-specific diversity of codon usage based on multivariate analysis. Gene 238:143–155. https://doi.org/10.1016/s0378-1119(99)00225-5.
24. Karlin S, Mrazek J. 2000. Predicted highly expressed genes of diverse prokaryotic genomes. J Bacteriol 182:5238–5250. https://doi.org/10.1128/JB.182.18.5238-5250.2000.
25. dos Reis M, Wernisch L, Savva R. 2003. Unexpected correlations between gene expression and codon usage bias from microarray data for the whole Escherichia coli K-12 genome. Nucleic Acids Res 31:6976–6985. https://doi.org/10.1093/nar/gkg897.
26. Supek F, Škunca N, Repar J, Vlahoviček K, Šmuc T. 2010. Translational selection is ubiquitous in prokaryotes. PLoS Genet 6:e1001004. https://doi.org/10.1371/journal.pgen.1001004.
27. Mrázek J, Karlin S. 1999. Detecting alien genes in bacterial genomes. Ann N Y Acad Sci 870:314–329. https://doi.org/10.1111/j.1749-6632.1999.tb08893.x.
28. Daubin V, Lerat E, Perrière G. 2003. The source of laterally transferred genes in bacterial genomes. Genome Biol 4:R57. https://doi.org/10.1186/gb-2003-4-9-r57.
29. Daubin V, Ochman H. 2004. Bacterial genomes as new gene homes: the genealogy of ORFans in E. coli. Genome Res 14:1036–1042. https://doi.org/10.1101/gr.2231904.
30. Ochman H, Lerat E, Daubin V. 2005. Examining bacterial species under the specter of gene transfer and exchange. Proc Natl Acad Sci U S A 102:6595–6599. https://doi.org/10.1073/pnas.0502035102.
31. Yannai A, Katz S, Hershberg R. 2018. The codon usage of lowly expressed genes is subject to natural selection. Genome Biol Evol 10:1237–1246. https://doi.org/10.1093/gbe/evy084.
32. Ikemura T. 1981. Correlation between the abundance of Escherichia coli transfer RNAs and the occurrence of the respective codons in its protein genes: a proposal for a synonymous codon choice that is optimal for the E. coli translational system. J Mol Biol 151:389–409. https://doi.org/10.1016/0022-2836(81)90003-6.
33. Ikemura T. 1982. Correlation between the abundance of yeast transfer RNAs and the occurrence of the respective codons in protein genes. differences in synonymous codon choice patterns of yeast and Escherichia coli with reference to the abundance of isoaccepting transfer RNAs. J Mol Biol 158:573–597. https://doi.org/10.1016/0022-2836(82)90250-9.
34. Bulmer M. 1987. Coevolution of codon usage and transfer RNA abundance. Nature 325:728–730. https://doi.org/10.1038/325728a0.
35. dos Reis M, Savva R, Wernisch L. 2004. Solving the riddle of codon usage preferences: a test for translational selection. Nucleic Acids Res 32:5036–5044. https://doi.org/10.1093/nar/gkh834.
36. Sabi R, Tuller T. 2014. Modelling the efficiency of codon-tRNA interactions based on codon usage bias. DNA Res 21:511–526. https://doi.org/10.1093/dnares/dsu017.
37. Gerber AP, Keller W. 2001. RNA editing by base deamination: more enzymes, more targets, new mysteries. Trends Biochem Sci 26:376–384. https://doi.org/10.1016/s0968-0004(01)01827-8.
38. Agris PF, Vendeix FAP, Graham WD. 2007. tRNA's wobble decoding of the genome: 40 years of modification. J Mol Biol 366:1–13. https://doi.org/10.1016/j.jmb.2006.11.046.
39. Novoa EM, Pavon-Eternod M, Pan T, Ribas de Pouplana L. 2012. A role for tRNA modifications in genome structure and codon usage. Cell 149:202–213. https://doi.org/10.1016/j.cell.2012.01.050.
40. Wald N, Alroy M, Botzman M, Margali H. 2012. Codon usage bias in prokaryotic pyrimidine-ending codons is associated with the degeneracy of the encoded amino acids. Nucleic Acids Res 40:7074–7083. https://doi.org/10.1093/nar/gks348.
41. Sharp PM, Bailes E, Grocock RJ, Peden JF, Sockett RE. 2005. Variation in the strength of selected codon usage bias among bacteria. Nucleic Acids Res 33:1141–1153. https://doi.org/10.1093/nar/gki242.
42. Stoletzki N, Eyre-Walker A. 2007. Synonymous codon usage in *Escherichia coli*: selection for translational accuracy. Mol Biol Evol 24:374–381. https://doi.org/10.1093/molbev/msl166.
43. Ran W, Kristensen DM, Koonin EV. 2014. Coupling between protein level selection and codon usage. mBio 5:e00956-14. https://doi.org/10.1128/mBio.00956-14.
44. Jara E, Morel MA, Lamolle G, Castro-Sowinski S, Simón D, Iriarte A, Musto H. 2018. The complex pattern of codon usage evolution in the family Comamonadaceae. Ecol Genet Genomics 6:1–8. https://doi.org/10.1016/j.egg.2017.11.002.
45. McInerney JO. 1998. Replicational and transcriptional selection on codon usage in Borrelia burgdorferi. Proc Natl Acad Sci U S A 95:10698–10703. https://doi.org/10.1073/pnas.95.18.10698.
46. Davis JJ, Olsen GJ. 2010. Modal codon usage: assessing the typical codon usage of a genome. Mol Biol Evol 27:800–810. https://doi.org/10.1093/molbev/msp281.
47. 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. https://doi.org/10.1098/rstb.2009.0305.
48. Shabalina SA, Spiridonov NA, Kashina A. 2013. Sounds of silence: synonymous nucleotides as a key to biological regulation and complexity. Nucleic Acids Res 41:2073–2094. https://doi.org/10.1093/nar/gks1205.

49. Quax TEF, Claassens NJ, Söll D, van der Oost J. 2015. Codon bias as a means to fine-tune gene expression. Mol Cell 59:149–161. https://doi.org/10.1016/j.molcel.2015.05.035.
50. Lerat E, Daubin V, Ochman H, Moran NA. 2005. Evolutionary origins of genomic repertoires in bacteria. PLoS Biol 3:e130. https://doi.org/10.1371/journal.pbio.0030130.
51. Davis JJ, Olsen GJ. 2011. Characterizing the native codon usages of a genome: an axis projection approach. Mol Biol Evol 28:211–221. https://doi.org/10.1093/molbev/msq185.
52. Supek F, Vlahoviček K. 2005. Comparison of codon usage measures and their applicability in prediction of microbial gene expressivity. BMC Bioinformatics 6:182. https://doi.org/10.1186/1471-2105-11-463.
53. Ran W, Higgs PG. 2010. The influence of anticodon-codon interactions and modified bases on codon usage bias in bacteria. Mol Biol Evol 27:2129–2140. https://doi.org/10.1093/molbev/msq102.
54. Gingold H, Pilpel Y. 2011. Determinants of translation efficiency and accuracy. Mol Syst Biol 7:481. https://doi.org/10.1038/msb.2011.14.
55. Lawrence JG, Ochman H. 1997. Amelioration of bacterial genomes: rates of change and exchange. J Mol Evol 44:383–397. https://doi.org/10.1007/pl00006158.
56. McInerney JO. 1997. Prokaryotic genome evolution as assessed by multivariate analysis of codon usage patterns. Microb Comp Genomics 2:89–97. https://doi.org/10.1089/omi.1.1997.2.89.
57. Blom J, Albaum SP, Doppmeier D, Pühler A, Vorhölter FJ, Zakrzewski M, Goesmann A. 2009. EDGAR: a software framework for the comparative analysis of prokaryotic genomes. BMC Bioinformatics 10:154. https://doi.org/10.1186/1471-2105-10-154.
58. Yu J, Blom J, Glaeser SP, Jaenicke S, Juhre T, Rupp O, Schwengers O, Spänig S, Goesmann A. 2017. A review of bioinformatics platforms for comparative genomics. Recent developments of the EDGAR 2.0 platform and its utility for taxonomic and phylogenetic studies. J Biotechnol 261:2–9. https://doi.org/10.1016/j.jbiotec.2017.07.010.
59. Rambaut A. 2009. FigTree v1.3.1. http://tree.bio.ed.ac.uk/software/figtree/.
60. Karlin S, Barnett MJ, Campbell AM, Fisher RF, Mrazek J, Mrázek J. 2003. Predicting gene expression levels from codon biases in alpha-proteobacterial genomes. Proc Natl Acad Sci U S A 100:7313–7318. https://doi.org/10.1073/pnas.1232298100.
61. Wang M, Herrmann CJ, Simonovic M, Szklarczyk D, von Mering C. 2015. Version 4.0 of PaxDb: protein abundance data, integrated across model organisms, tissues, and cell-lines. Proteomics 15:3163–3168. https://doi.org/10.1002/pmic.201400441.
62. Felsenstein J. 2005. PHYLIP (Phylogeny Inference Package) version 3.6. http://evolution.genetics.washington.edu/phylip.html.
63. Darriba D, Posada D, Kozlov AM, Stamatakis A, Morel B, Flouri T. 2020. ModelTest-NG: a new and scalable tool for the selection of DNA and protein evolutionary models. Mol Biol Evol 37:291–294. https://doi.org/10.1093/molbev/msz189.
64. Gil M, Zanetti MS, Zoller S, Anisimova M. 2013. CodonPhyML: fast maximum likelihood phylogeny estimation under codon substitution models. Mol Biol Evol 30:1270–1280. https://doi.org/10.1093/molbev/mst034.
65. Yang Z. 2007. PAML 4: phylogenetic analysis by maximum likelihood. Mol Biol Evol 24:1586–1591. https://doi.org/10.1093/molbev/msm088.
66. Lê S, Josse J, Husson F. 2008. FactoMineR: an R package for multivariate analysis. J Stat Softw 25:1–18.
67. Sharp PM, Tuohy TMF, Mosurski KR. 1986. Codon usage in yeast: cluster analysis clearly differentiates highly and lowly expressed genes. Nucleic Acids Res 14:5125–5143. https://doi.org/10.1093/nar/14.13.5125.
68. Peden J. 1999. Analysis of codon usage. PhD dissertation. University of Nottingham, Nottingham, England.
69. Wickham H. 2009. ggplot2: elegant graphics for data analysis. Springer-Verlag, New York, NY.
70. Lowe TM, Chan PP. 2016. tRNAscan-SE On-line: integrating search and context for analysis of transfer RNA genes. Nucleic Acids Res 44:W54–W57. https://doi.org/10.1093/nar/gkw413.
71. Wright F. 1990. The "effective number of codons" used in a gene. Gene 87:23–29. https://doi.org/10.1016/0378-1119(90)90491-9.
72. Revell LJ. 2012. phytools: an R package for phylogenetic comparative biology (and other things). Methods Ecol Evol 3:217–223. https://doi.org/10.1111/j.2041-210X.2011.00169.x.