SUPPLEMENTARY INFORMATION

STATISTICAL EVALUATION OF STRUCTURAL CHANGES IN SKEW PATTERNS

SI Table 1. Percent classification of chromosomes according to their skew patterns along the plus strand and the CDS concatenate.

| mono- and di-nucleotide skews | plus strand          | CDS concatenate   |
|-------------------------------|----------------------|------------------|
|                              | sharp | vague | flat   | sharp | vague | flat   |
| $S^{AT}$                      | 57.35| 21.47 | 21.18  | 0.88  | 21.47 | 77.65  |
| $S^{GC}$                      | 85.88| 9.12  | 5.00   | 1.18  | 25.88 | 72.94  |
| $S^{AG-CT}$                   | 56.76| 18.53 | 24.71  | 0.88  | 23.82 | 75.29  |
| $S^{GA-TC}$                   | 71.47| 13.82 | 14.71  | 0.88  | 20.29 | 78.82  |
| $S^{GG-CC}$                   | 80.88| 13.53 | 5.59   | 1.47  | 21.18 | 77.35  |
| $S^{AA-TT}$                   | 48.53| 23.53 | 27.94  | 1.76  | 26.18 | 72.06  |
| $S^{AC-GT}$                   | 81.18| 10.29 | 8.53   | 1.18  | 16.18 | 82.65  |
| $S^{CA-TG}$                   | 85.00| 9.41  | 5.59   | 0.88  | 18.24 | 80.88  |

| di-nucleotide relative abundance skews | plus strand          | CDS concatenate   |
|----------------------------------------|----------------------|------------------|
|                                        | sharp | vague | flat   | sharp | vague | flat   |
| $p^{AG-CT}$                            | 33.24| 21.47 | 45.29  | 1.76  | 15.29 | 82.94  |
| $p^{GA-TC}$                            | 49.12| 22.94 | 27.94  | 0     | 15.29 | 84.71  |
| $p^{GG-CC}$                            | 41.47| 22.06 | 36.47  | 0     | 20.29 | 79.71  |
| $p^{AA-TT}$                            | 53.24| 22.35 | 24.41  | 1.47  | 25.00 | 73.53  |
| $p^{AC-GT}$                            | 44.71| 17.35 | 37.94  | 1.47  | 23.53 | 75.00  |
| $p^{CA-TG}$                            | 42.06| 21.18 | 36.76  | 0.88  | 22.94 | 76.18  |

NOTES.- *sharp*: statistically significant breakpoints are located closer than 5% of the chromosome length to the ori, *vague*: statistically significant breakpoints are detected farther from the ori, and *flat*: no statistically significant breakpoints are detected.

For each chromosome in our collection, we fit the skew values in a linear regression setup and apply a dynamic programming algorithm developed in \(^1,2\), as implemented in the R package *strucchange* \(^3\). This algorithm computes the optimal breakpoints, in which the coefficients of the model shift from one stable regression relationship to another. *Sharp* patterns along the plus strand indicate a statistically significant correlation of skews with replication-associated biases. *Flat* patterns along CDS concatenates suggest that CDS-associated biases are a major determinant of skews.

Regarding relative abundance skews, a significant percentage (up to \(~50\%\)) of the studied genomes exhibit *sharp* patterns along the plus strand, while along CDS concatenates *flat* patterns represent roughly an 80\%. 

1
## STRAND BIASES OF CODING SEQUENCES CAN TRACE SPECIES PHYLOGENY.
### SUPPLEMENTARY TABLES AND FIGURES

**SI Table 2. Topological scores.**

|                          | mononucleotide skews | dinucleotide skews | relative abundance skews | genomic signatures |
|--------------------------|----------------------|--------------------|--------------------------|-------------------|
| Actinobacteria           | 59.5                 | 72.1               | 64.1                     | 67.7              |
| Bacteroidetes            | 77.8                 | 83.3               | 100                      | 100               |
| Chlamydiae               | 90                   | 83.3               | 90                       | 100               |
| Cyanobacteria            | 83.5                 | 80.8               | 81.8                     | 77.6              |
| Firmicutes               | 68.8                 | 77.1               | 80.8                     | 81.1              |
| α-Proteobacteria         | 69.8                 | 79.5               | 72.3                     | 72.5              |
| β-Proteobacteria         | 70.6                 | 72                 | 72.8                     | 79.1              |
| δ-Proteobacteria         | 49.9                 | 53.2               | 62.4                     | 55.8              |
| ε-Proteobacteria         | 100                  | 93.3               | 93.3                     | 93.3              |
| γ-Proteobacteria         | 67.6                 | 75.3               | 74                       | 77.9              |
| Spirochaetes             | 75                   | 75                 | 100                      | 100               |
| Tenericutes              | 71                   | 72.6               | 83.9                     | 74.2              |
| **median**               | **70.8**             | **76.2**           | **81.3**                 | **78.5**          |

**NOTES.-** The percent topological similarity of the skew-based trees to the species-trees. Comparisons are made between species that belong to the same taxonomic rank (phylum or class). Skews are grouped into three classes: mononucleotide skews ($V^{\text{MONO}}$), dinucleotide skews ($V^{\text{DI}}$), relative abundance skews ($V^{\text{RA}}$). In addition, we construct cladograms based on the δ-distance between genomic signatures, using the corresponding topological scores as a benchmark. The median of the topological scores is given for each skew type and for genomic signatures.
SI Figure 1. Cladogram comparison plots. For each phylum or class comparisons are made between the cladograms obtained from NCBI taxonomy (species-trees; red) and the cladograms we construct based on V^{RA} (skew-based trees; black). Bacteria are represented by their NCBI taxonomical ids. For each pair of cladograms, links are drawn between terminal vertices (tips) that correspond to the same organism. The length of the branches is not accounted for. Plots are generated using the function cophyloplot in the R package ape 3.1-2.

Explanatory Comments on SI Figure 1.

1st.- For each pair of bacteria, we apply the symmetric Kullback-Leibler (KL) divergence to quantify the dissimilarity of V^{RA} along their CDS concatenates. In this way we obtain a set of dissimilarities.
between all bacteria within a given phylum or class. We use these sets of dissimilarities to construct the skew-based trees of each phylum or class, via complete-linkage hierarchical clustering. Thus, the branching off points of the skew-based trees are inferred so that all terminal taxa above them share skew patterns that are more similar among them than with any other taxon represented in the cladogram.

2nd.- For each phylum or class the percent topological similarity between the species-tree and the skew-based tree is given by scores, which we obtain using the web-based tool Compare2Trees. For the presented cases, these topological scores are as follows: (a) Firmicutes, 80.8%; (b) Chlamydiae, 90%; (c) Actinobacteria, 64.1%; (d) δ-Proteobacteria, 62.4%. Though Actinobacteria and δ-Proteobacteria have the lowest scores compared to all other cladograms constructed based on VRA (SI Table 2), the structure of their cladogram comparison plots (c,d) is not highly different from the one corresponding to Firmicutes (a) or Chlamydiae (b). Thus, it can be argued that, albeit their relatively low topological scores, Actinobacteria and δ-Proteobacteria have species-specific modes of relative abundance skews, as is also the case for the other bacterial phyla or classes.

To further test our argument, we apply the Grubbs' test for outliers to the full set of scores corresponding to the cladograms we construct based on VRA (SI Table 2). The null-hypothesis (H₀) of the Grubbs' test is that there are no outliers in the data set, while the alternative hypothesis (Hₐ) in the test variant that we use is that the two lowest values (62.4 and 64.1, for δ-Proteobacteria and Actinobacteria, respectively) are outliers. The resulting p-value equals 0.5347 and thus the Hₐ is rejected as non-significant. We conclude that Actinobacteria and δ-Proteobacteria do not constitute exceptional cases with regards to their relative abundance skew profiles.
SI Figure 2. Regression of skews in low-coverage orthologs to the skews of the remaining orthologs. Each point represents a chromosome of our collection. Points that fall on the diagonal (slope=1, y-intercept=0; red dashed line) indicate that skews in low-coverage orthologs are identical to the skews in remaining orthologs. For a given skew, if the regression line tends to fall on to the diagonal (slope close to unity, y-intercept close to zero), there is a general trend of the corresponding strand-asymmerties to be very similar in low-coverage and remaining orthologs, in most of the chromosomes examined.
SI Table 3. Linear regression between skews in the coding strands of *low-coverage* and *remaining* orthologs.

| mono- and di-nucleotide skews | slope  | y-intercept | $r^2$  | $p$-value     | Pearson's $r$ |
|-------------------------------|--------|-------------|--------|---------------|---------------|
| $S^{\text{A-T}}$              | 0.905  | 0.014       | 0.565  | 1.21e-49      | 0.752         |
| $S^{\text{G-C}}$              | 0.939  | 0.00275     | 0.84   | 4.47e-107     | 0.917         |
| $S^{\text{AG-CT}}$            | 0.931  | 0.0149      | 0.834  | 8.39e-105     | 0.913         |
| $S^{\text{GA-TC}}$            | 0.97   | 0.0128      | 0.798  | 1.09e-93      | 0.893         |
| $S^{\text{GG-CC}}$            | 0.929  | 0.0185      | 0.851  | 5.13e-111     | 0.922         |
| $S^{\text{AA-TT}}$            | 0.698  | 0.024       | 0.375  | 9.99e-29      | 0.612         |
| $S^{\text{AC-GT}}$            | 0.832  | 0.0284      | 0.639  | 2.77e-60      | 0.799         |
| $S^{\text{CA-TG}}$            | 0.671  | -0.00135    | 0.424  | 1.7e-33       | 0.651         |
| $S^{\text{AG-CT}}$            | 0.823  | -0.0174     | 0.747  | 7.52e-81      | 0.865         |
| $S^{\text{GA-TC}}$            | 0.719  | 0.0156      | 0.503  | 6.46e-42      | 0.709         |
| $S^{\text{GG-CC}}$            | 0.727  | 0.0274      | 0.358  | 3.07e-27      | 0.599         |
| $S^{\text{AA-TT}}$            | 0.676  | -0.0298     | 0.79   | 2.52e-91      | 0.889         |
| $S^{\text{AC-GT}}$            | 0.915  | 0.0321      | 0.585  | 2.55e-52      | 0.765         |
| $S^{\text{CA-TG}}$            | 0.733  | 0.00153     | 0.747  | 9.38e-81      | 0.864         |

NOTES.- For each skew type, the slope, y-intercept, $r^2$ and $p$-value of the fitted model are given. Also, the Pearson's correlation coefficient ($r$) for each skew in *low-coverage orthologs* and in *remaining orthologs* is computed. When slope is close to 1 and y-intercept is close to 0, *low-coverage* and *remaining* orthologs have highly similar skew patterns.
THE GC BIASED STRUCTURE OF GENETIC CODE IMPOSES SITE-SPECIFIC STRNAD-ASYMMETRIES

**SI Figure 3.** A scatter plot of the GC content entered in the codon probability ($p_i$) function of the GC biased model (Expected GC) versus the resulting GC content measured in the produced artificial sequence (Actual GC). The plotted points fall on or near the diagonal (dotted line), that is, the GC content entered in the $p_i$ function is approximately equal to the resulting GC content ($r^2 = 0.999$), demonstrating that our model is self-consistent. The linear regression function is also provided.
SI Table 4. The mean GC% of each synonymous group and the parameters $R_{i}^{GC}$, $R_{i}^{AT}$ and $n_{i}$ of the GC-biased model for each codon.

| Codon | GC%  | AT%  | $R_{i}^{GC}$ | $R_{i}^{AT}$ | $n_{i}$ |
|-------|------|------|-------------|-------------|--------|
| TTT F | 16.7 | 0.000| 1.200       | 0           |        |
| TCT S | 50.0 | 0.667| 1.333       | 1           |        |
| TAT Y | 16.7 | 0.000| 1.200       | 0           |        |
| TGT C | 50.0 | 0.667| 1.333       | 1           |        |
| TTC F | 16.7 | 2.000| 0.800       | 1           |        |
| TCC S | 50.0 | 1.333| 0.667       | 2           |        |
| TAC Y | 16.7 | 2.000| 0.800       | 1           |        |
| TGC C | 50.0 | 1.333| 0.667       | 2           |        |
| TTA L | 38.9 | 0.000| 1.636       | 0           |        |
| TCA S | 50.0 | 0.667| 1.333       | 1           |        |
| TAA * Ter | | | | | |
| TGA * Ter | | | | | |
| TTG L | 38.9 | 0.857| 1.091       | 1           |
| TGG S | 50.0 | 1.333| 0.667       | 2           |
| TAG * Ter | | | | | |
| TGG W | 66.7 | 1.000| 1.000       | 2           |

CTT L | 38.9 | 0.857| 1.091       | 1           |
| CCT P | 83.3 | 0.800| 2.000       | 2           |
| CAT H | 50.0 | 0.667| 1.333       | 1           |
| CGT R | 72.2 | 0.923| 1.200       | 2           |
| CTC L | 38.9 | 1.714| 0.545       | 2           |
| CCC P | 83.3 | 1.200| 0.000       | 3           |
| CAC H | 50.0 | 1.333| 0.667       | 2           |
| CGC R | 72.2 | 1.385| 0.000       | 3           |
| CTA L | 38.9 | 0.857| 1.091       | 1           |
| CCA P | 83.3 | 0.800| 2.000       | 2           |
| CAA Q | 50.0 | 0.667| 1.333       | 1           |
| CGA R | 72.2 | 0.923| 1.200       | 2           |
| CTG L | 38.9 | 1.714| 0.545       | 2           |
| CCG P | 83.3 | 1.200| 0.000       | 3           |
| CAG Q | 50.0 | 1.333| 0.667       | 2           |
| CGG R | 72.2 | 1.385| 0.000       | 3           |

ATT I | 11.1 | 0.000| 1.125       | 0           |
| ACT T | 50.0 | 0.667| 1.333       | 1           |
| AAT N | 16.7 | 0.000| 1.200       | 0           |
| AGT S | 50.0 | 0.667| 1.333       | 1           |
| ATC I | 11.1 | 3.000| 0.750       | 1           |
| ACC T | 50.0 | 1.333| 0.667       | 2           |
| AAC N | 16.7 | 2.000| 0.800       | 1           |
| AGC S | 50.0 | 1.333| 0.667       | 2           |
| ATA I | 11.1 | 0.000| 1.125       | 0           |
| ACA T | 50.0 | 0.667| 1.333       | 1           |
| AAA K | 16.7 | 0.000| 1.200       | 0           |
| AGA R | 72.2 | 0.462| 2.400       | 1           |
| ATG M | 33.3 | 1.000| 1.000       | 1           |
| ACG T | 50.0 | 1.333| 0.667       | 2           |
| AAG K | 16.7 | 2.000| 0.800       | 1           |
| AGG R | 72.2 | 0.923| 1.200       | 2           |

GTT V | 50.0 | 0.667| 1.333       | 1           |
| GGT A | 83.3 | 0.800| 2.000       | 2           |
| GAT D | 50.0 | 0.667| 1.333       | 1           |
| GGT G | 83.3 | 0.800| 2.000       | 2           |
| GTC V | 50.0 | 1.333| 0.667       | 2           |
| GCC A | 83.3 | 1.200| 0.000       | 3           |
| GAC D | 50.0 | 1.333| 0.667       | 2           |
| GGC G | 83.3 | 1.200| 0.000       | 3           |
| GTA V | 50.0 | 0.667| 1.333       | 1           |
| GCA A | 83.3 | 0.800| 2.000       | 2           |
| GAG E | 50.0 | 1.333| 0.667       | 2           |
| GGA G | 83.3 | 0.800| 2.000       | 2           |
| GTG V | 50.0 | 1.333| 0.667       | 2           |
| GCG A | 83.3 | 1.200| 0.000       | 3           |

NOTES.- To each i\textsuperscript{th} codon corresponds a vector ($x$, $y$, $z$, $n$), where $x$ is the mean GC content of its synonymous codon group, here expressed as a percentage, and $y$, $z$, $n$ are the parameters $R_{i}^{GC}$, $R_{i}^{AT}$, $n_{i}$ of the GC-biased model for the i\textsuperscript{th} codon. $n_{i}$ is the number of G and C of the i\textsuperscript{th} codon. $R_{i}^{GC}$ is the ratio of the i\textsuperscript{th} codon's GC content to the mean GC content of its synonymous group. Similarly, $R_{i}^{AT}$ is the ratio of the i\textsuperscript{th} codon's AT content to the mean AT content of its synonymous group. $R_{i}^{GC}$ and $R_{i}^{AT}$ indicate whether the composition of the i\textsuperscript{th} codon is higher or lower than the mean GC content and the mean AT content of its synonymous codon group, respectively. In this way, $R_{i}^{GC}$ and $R_{i}^{AT}$ capture the GC biases between synonymous codons.
3rd fourfold-degenerate sites

3rd twofold-degenerate sites
SI Figure 4. Heatmaps of Pearson's correlation coefficient ($r$) for codon usage vs. $S_{\text{CDS}}^{\text{A-T}}$ and $S_{\text{CDS}}^{\text{G-C}}$, for both genomic dataset and our GC biased model. Skews are calculated at specific codon sites.
Columns indicate the base composition of the codon site at which skews are computed. Stippled red cells depict negative $r$, while blue cells depict positive $r$. The intensity of the color indicates the degree of correlation, as recited in the color key. Start and stop codons are excluded from our calculations. The six-fold degenerate codons of Leucine and Arginine are labeled in the heatmaps. Codons used in calculations of skews at 3rd fourfold- and 3rd twofold-degenerate sites are marked with “*” and “•”, respectively. “start”: initiation codon, “stop”: termination codons.

**Explanatory Comments on SI Figure 4.**

1st.- Compositional skews respond to the frequencies of certain Leucine and Arginine codons in a way that differentiates them from the other A/T- or G/C-ending codons. Palidwor and co-workers showed that the response of Leucine and Arginine codons to GC mutational bias may lead to atypical usage patterns. These findings were attributed to the fact that Leucine and Arginine are the only amino-acids encoded by codons that allow GC-changing synonymous substitutions in both their 1st and 3rd positions. Along the same lines, the structure of possible synonymous substitutions within the codons of Leucine and Arginine may be responsible for the observed differences in correlation patterns between strand biases and the frequencies of those codons. It is noteworthy that our GC biased model captures these differences (cf. SI Figure 4i vs. j, k vs. l), further supporting our argument that GC mutational bias underlies the correlation of codon usage and site-specific skews.

2nd.- As regards to 2nd codon position, $S_{CDS}^{A-T}$ correlates positively with G/C-ending codons and negatively with A/T-ending codons in our GC biased model, while the inverse holds true in the genomic dataset (SI Figure 4m,n). Thus, other factors than the ones adopted in our model (overall GC content and synonymous-codon GC biases) should operate at these sites to yield the patterns detected in genomic data. The composition of the 2nd codon site is strongly associated with the physicochemical properties of the encoded amino-acids. Codons with purine residues in their 2nd position are assigned to charged or polar amino-acids. In order for integral membrane proteins to maintain their subcellular location, selective pressure is exerted against such codons. Selective pressure may be linked to the peculiarities of $S_{CDS}^{A-T}$ in 2nd codon sites, by imposing specific constraints to GC biased substitutions at these sites. Moreover, selection has been previously implicated in the emergence of atypical $S_{CDS}^{A-T}$ patterns.
## SI Table 5. Molecular network tabulation and assignment to specific molecular phenotypes.

| molecular phenotypes | Molecular network compounds, in terms of the KEGG orthology (KO) groups |
|----------------------|---------------------------------------------------------------------|
| **direct repair**    | phrB  | phrB |
|                      | ogt   | ogt  |
|                      | alkB  | alkB |
|                      | ada   | ada  |
| **BER**              | ung   | ung, [xthA or nfo], polA, [ligATP or ligNAD] |
|                      | mug   | mug, [xthA or nfo], polA, [ligATP or ligNAD] |
|                      | nth   | nth, [xthA or nfo], polA, [ligATP or ligNAD] |
|                      | mutM  | mutM, [xthA or nfo], polA, [ligATP or ligNAD] |
|                      | nei   | nei, [xthA or nfo], polA, [ligATP or ligNAD] |
|                      | tag   | tag, [xthA or nfo], polA, [ligATP or ligNAD] |
|                      | alkA  | alkA, [xthA or nfo], polA, [ligATP or ligNAD] |
|                      | mutY  | mutY, [xthA or nfo], polA, [ligATP or ligNAD] |
| **GO system**        | mutM, mutY, mutT, [xthA or nfo], polA, [ligATP or ligNAD] |
| **NER**              | GGR   | uvrA, uvrB, uvrC, uvrD.pcrA, polA, [ligATP or ligNAD] |
|                      | TCR   | mfd, uvrA, uvrB, uvrC, uvrD.pcrA, polA, [ligATP or ligNAD] |
| **MMR**              | methyl-directed | mutS, mutL, mutH, uvrD.pcrA, dam |
|                      | nick-directed | mutS, mutL, uvrD.pcrA (*) |
|                      | VSR-patch | vsr, dam |
| **RR**               | RecFOR | recJ, ssb, recO, recR, recA, [ruvABC or recG] |
|                      | RecBC  | recB, recC, recD, recA, recomb, priA, priB, priC, dnaT |

**NOTES.** Data are derived from KEGG pathway maps. (*): only mutH-less bacteria are included.

BER: base excision repair, NER: nucleotide excision repair, MMR: mismatch repair, RR: recombination repair.

**Explanatory Comments on SI Table 5**

A species is assumed to have a specific molecular phenotype if all corresponding loci listed in the table are present in the species genome, according to KEGG Orthology\(^\text{13}\). Each species of a given molecular phenotype is considered to be proficient in the respective repair pathway; see\(^\text{14,15}\) and references therein. For each repair pathway, we divide our dataset into a repair-proficient and a repair-deficient group. The repair-proficient group includes all DNA sequences of species of the molecular phenotype under consideration, while all the remaining DNA sequences make up the repair-deficient group.
SI Table 6. Tabulation of bacteria according to specific molecular phenotypes.

| molecular phenotypes | Actinobacteria | Bacteroidetes | Chlamydiae | Cyanobacteria | Firmicutes | α-Proteobacteria | β-Proteobacteria | γ-Proteobacteria | δ-Proteobacteria | ε-Proteobacteria | Spirochaetes | Tenericutes |
|----------------------|----------------|---------------|------------|---------------|------------|-----------------|-----------------|-----------------|-----------------|-----------------|--------------|-------------|
| phrB                 |                |               |            |               |            |                 |                 |                 |                 |                 |              |             |
| proficient           | 6.2            | 1.1           | 0.6        | 8.5           | 18.6       | 11.9            | 8.5             | 37.3            | 5.1             | 1.1             | 1.1          | 0           |
| deficient            | 6.7            | 3             | 7.5        | 1.5           | 22.4       | 18.7            | 7.5             | 11.2            | 1.5             | 5.2             | 3            | 11.9        |
| ogt                  |                |               |            |               |            |                 |                 |                 |                 |                 |              |             |
| proficient           | 8.2            | 2.2           | 4.7        | 1.3           | 20.7       | 13.4            | 10.3            | 29.3            | 4.3             | 3.9             | 0.9          | 0.9         |
| deficient            | 1.3            | 1.3           | 0          | 17.7          | 19         | 19              | 1.3             | 16.5            | 1.3             | 0               | 5.1          | 17.7        |
| alkB                 |                |               |            |               |            |                 |                 |                 |                 |                 |              |             |
| proficient           | 28             | 0             | 0          | 0             | 0          | 12              | 28              | 28              | 4               | 0               | 0            | 0           |
| deficient            | 4.5            | 2.1           | 3.8        | 5.9           | 22         | 15              | 6.3             | 25.9            | 3.5             | 3.1             | 2.1          | 5.6         |
| ada                  |                |               |            |               |            |                 |                 |                 |                 |                 |              |             |
| proficient           | 0              | 1.1           | 0          | 1.1           | 3.3        | 28.6            | 14.3            | 46.2            | 1.1             | 2.2             | 2.2          | 0           |
| deficient            | 9.1            | 2.3           | 5          | 7.3           | 27.3       | 9.1             | 5.5             | 17.7            | 4.5             | 3.2             | 1.8          | 7.3         |
| ung                  |                |               |            |               |            |                 |                 |                 |                 |                 |              |             |
| proficient           | 7.7            | 2.1           | 4.7        | 0             | 24.5       | 3.4             | 8.6             | 33.9            | 2.6             | 3.9             | 1.7          | 6.9         |
| deficient            | 2.6            | 1.3           | 0          | 21.8          | 7.7        | 48.7            | 6.4             | 2.6             | 6.4             | 0               | 2.6          | 0           |
| mug                  |                |               |            |               |            |                 |                 |                 |                 |                 |              |             |
| proficient           | 18.2           | 0             | 0          | 0             | 6.1        | 6.1             | 15.2            | 51.5            | 3               | 0               | 0            | 0           |
| deficient            | 5              | 2.2           | 4          | 6.1           | 21.9       | 15.8            | 7.2             | 23              | 3.6             | 3.2             | 2.2          | 5.8         |
| nth                  |                |               |            |               |            |                 |                 |                 |                 |                 |              |             |
| proficient           | 5.9            | 2.1           | 3.8        | 5.9           | 21.7       | 15.9            | 8.6             | 27.6            | 3.4             | 3.1             | 2.1          | 0           |
| deficient            | 14.3           | 0             | 0          | 0             | 0          | 0               | 0               | 4.8             | 4.8             | 0               | 0            | 76.2        |
| mutM                 |                |               |            |               |            |                 |                 |                 |                 |                 |              |             |
| proficient           | 7.1            | 0.8           | 0.4        | 6.4           | 22.2       | 15.8            | 9.4             | 27.8            | 4.1             | 0               | 0            | 6           |
| deficient            | 2.2            | 8.9           | 22.2       | 0             | 8.9        | 8.9             | 0               | 15.6            | 0               | 20              | 13.3         | 0           |
| nei                  |                |               |            |               |            |                 |                 |                 |                 |                 |              |             |
| proficient           | 34.2           | 0             | 0          | 2.6           | 0          | 0               | 0               | 60.5            | 2.6             | 0               | 0            | 0           |
| deficient            | 2.6            | 2.2           | 4          | 5.9           | 23.1       | 16.8            | 9.2             | 21.2            | 3.7             | 3.3             | 2.2          | 5.9         |
| tag                  |                |               |            |               |            |                 |                 |                 |                 |                 |              |             |
| proficient           | 8.5            | 1.6           | 0.5        | 0             | 24.9       | 13.8            | 11.1            | 33.9            | 3.2             | 0.5             | 1.1          | 1.1         |
| deficient            | 3.3            | 2.5           | 8.2        | 13.9          | 13.1       | 16.4            | 3.3             | 13.9            | 4.1             | 6.6             | 3.3          | 11.5        |
| alkA                 |                |               |            |               |            |                 |                 |                 |                 |                 |              |             |
| proficient           | 1.7            | 0             | 0.8        | 1.7           | 12.6       | 23.5            | 18.5            | 37.8            | 0.8             | 0               | 1.7          | 0.8         |
| deficient            | 9.4            | 3.1           | 5.2        | 7.8           | 25         | 9.4             | 1.6             | 18.8            | 5.2             | 4.7             | 2.1          | 7.8         |
| mutY                 |                |               |            |               |            |                 |                 |                 |                 |                 |              |             |
| proficient           | 7              | 2.3           | 4.3        | 4.7           | 21         | 12.5            | 9.7             | 30              | 4.3             | 3.1             | 1.2          | 0           |
|                | deficient | 0   | 0   | 9.3  | 16.7 | 25.9 | 0   | 7.4  | 0   | 1.9  | 5.6  | 29.6 |
|----------------|-----------|-----|-----|------|------|------|-----|------|-----|------|------|------|
| **GO system**  |           |     |     |      |      |      |     |      |     |      |      |      |
| proficient     | 7.8       | 0   | 0.5 | 2.4  | 24.3 | 13.6 | 12.1| 34.5 | 4.9 | 0    | 0    | 0    |
| deficient      | 3.8       | 5.7 | 9.5 | 11.4 | 12.4 | 17.1 | 0   | 9.5  | 1   | 8.6  | 5.7  | 15.2 |
| **NER**        |           |     |     |      |      |      |     |      |     |      |      |      |
| proficient     | 6.8       | 2   | 3.7 | 5.8  | 21.4 | 13.6 | 8.5 | 24.5 | 3.7 | 3.1  | 1.7  | 5.1  |
| deficient      | 0         | 0   | 0   | 0    | 0    | 35.3 | 0   | 52.9 | 0   | 0    | 5.9  | 5.9  |
| **TCR**        |           |     |     |      |      |      |     |      |     |      |      |      |
| proficient     | 7.2       | 2.2 | 4.2 | 6.2  | 22.8 | 13.4 | 9.1 | 26.1 | 4   | 3.3  | 1.8  | 0    |
| deficient      | 0         | 0   | 0   | 0    | 25.7 | 0    | 25.7| 0    | 0   | 2.9  | 45.7 |
| methyl-directed|           |     |     |      |      |      |     |      |     |      |      |      |
| proficient     | 0         | 0   | 0   | 0    | 0    | 0    | 0   | 100  | 0   | 0    | 0    | 0    |
| deficient      | 7.6       | 2.3 | 4.2 | 6.4  | 23.9 | 17.4 | 9.5 | 12.9 | 4.2 | 3.4  | 2.3  | 6.1  |
| nick-directed  |           |     |     |      |      |      |     |      |     |      |      |      |
| proficient     | 0.5       | 3   | 5.6 | 4.6  | 30.5 | 22.8 | 12.7| 13.2 | 5.6 | 0    | 1.5  | 0    |
| deficient      | 16.7      | 0   | 0   | 7    | 2.6  | 0.9  | 0   | 48.2 | 0   | 7.9  | 2.6  | 14   |
| **VSR-patch**  |           |     |     |      |      |      |     |      |     |      |      |      |
| proficient     | 0         | 0   | 0   | 0    | 0    | 3.4  | 13.8| 79.3 | 3.4 | 0    | 0    | 0    |
| deficient      | 7.1       | 2.1 | 3.9 | 6    | 22.3 | 16   | 7.4 | 20.6 | 3.5 | 3.2  | 2.1  | 5.7  |
| RecFOR         |           |     |     |      |      |      |     |      |     |      |      |      |
| proficient     | 0.8       | 2.4 | 4.4 | 6    | 24.6 | 17.9 | 9.5 | 29.4 | 4   | 0    | 1.2  | 0    |
| deficient      | 30.5      | 0   | 0   | 3.4  | 1.7  | 1.7  | 1.7 | 11.9 | 1.7 | 15.3 | 5.1  | 27.1 |
| **RecBC**      |           |     |     |      |      |      |     |      |     |      |      |      |
| proficient     | 0         | 0   | 0   | 0    | 0    | 0    | 0   | 100  | 0   | 0    | 0    | 0    |
| deficient      | 6.7       | 2   | 3.7 | 5.7  | 21.2 | 15.5 | 8.4 | 22.6 | 3.7 | 3    | 2    | 5.4  |

NOTES.- The percent participation of each phylum / class within groups of bacteria with a specific molecular phenotype. We consider the taxonomic levels at which we conducted our cladistic analysis. Molecular phenotypes indicate proficiency or deficiency in the DNA repair mechanisms under consideration (see SI Table 5).

**Explanatory Comments on SI Table 6**

Each row represents a specific molecular phenotype. Thus, the rows are indicative of the phylogenetic dispersion of the corresponding repair pathways. Most of the DNA repair systems under consideration are present in distantly diverged species, belonging to different phyla. Methyl-directed MMR and RecBC homologous recombination pathways are the only exceptions, since all bacteria in our collection that are proficient in these pathways belong to γ-Proteobacteria.
Each column corresponds to a phylum or class. Thus, the columns give an overview of the within phylum / class variation of the presence or absence of different DNA repair systems. There are no phyla / classes whose members do not differ in their molecular phenotypes regarding at least four of the examined repair pathways, with \(\gamma\)-Proteobacteria being comprised of bacteria representing all different molecular phenotypes.
SI Table 7. Analysis of $S_{Trs}^{G-C}$ distribution in relation to the presence (proficient) or absence (deficient) of specific molecular pathways involved in DNA repair and modification.

### 3rd fourfold degenerate sites

| Molecular phenotypes | $S_{Trs}^{G-C}$ | Median | P-value |
|----------------------|-----------------|--------|---------|
|                      | deficient  | proficient | initial  | adjusted |
| Direct repair        |             |             |         |          |
| phrB                 | -0.0635    | -0.0206    | 0.00189 | ** 0.0378 | *          |
| ogt                  | -0.0638    | -0.028     | 0.0289  | * 0.578   | -          |
| alkB                 | -0.0439    | 0.0103     | 0.00163 | ** 0.0326 | *          |
| ada                  | -0.058     | -0.008     | 9.77e-07| *** 1.954e-05 | ***       |
| BER                  |             |             |         |          |
| ung                  | -0.0838    | -0.0212    | 0.0361  | * 0.722   | -          |
| mug                  | -0.0458    | 0.0773     | 1.89e-08| *** 3.78e-07 | ***       |
| nth                  | -0.113     | -0.0331    | 0.049   | * 0.98    | -          |
| mutM                 | -0.0515    | -0.0307    | 0.433   | - 1       | -          |
| nei                  | -0.0444    | 0.00169    | 0.000488| *** 0.00976 | **         |
| tag                  | -0.0654    | -0.0229    | 0.251   | - 1       | -          |
| alkA                 | -0.095     | -0.00831   | 4.06e-11| *** 8.12e-10 | ***       |
| mutY                 | -0.0969    | -0.0323    | 0.0963  | - 1       | -          |
| GO system            | -0.0617    | -0.0229    | 0.0194  | * 0.388   | -          |
| NER                  |             |             |         |          |
| GGR                  | 0.0345     | -0.0412    | 0.0299  | * 0.598   | -          |
| TCR                  | -0.00305   | -0.0412    | 0.239   | - 1       | -          |
| MMR                  |             |             |         |          |
| methyl-directed      | -0.0531    | 0.0221     | 5.29e-07| *** 1.058e-05 | ***       |
| nick-directed        | -0.0242    | -0.0448    | 0.439   | - 1       | -          |
| VSR-patch            | -0.0451    | 0.0906     | 1.75e-07| *** 3.5e-06 | ***       |
| RR                   |             |             |         |          |
| RecFOR               | -0.0762    | -0.0298    | 0.0578  | - 1       | -          |
| RecBC                | -0.0437    | 0.106      | 4.94e-09| *** 9.88e-08 | ***       |

**NOTES.** - For each molecular phenotype, the median $S_{Trs}^{G-C}$ is computed for deficient and proficient bacteria, along with the $p$-value of a two-tailed Wilcoxon test, denoting whether the difference of the skew distribution in deficient versus proficient species is statistically significant. To control the familywise error in our comparisons, we use the Bonferroni correction. Both the initial (unadjusted) and the adjusted $p$-values are given. “-”: $p$-value ≥ 0.05, “*”: 0.05 > $p$-value ≥ 0.01, “**”: 0.01 > $p$-value ≥ 0.001, “***”: $p$-value < 0.001.

BER: base excision repair, NER: nucleotide excision repair, MMR: mismatch repair, RR: recombination repair.
Explanatory Comments on SI Table 7

1st.- Each molecular phenotype denotes the corresponding DNA repair system. When considering a specific repair system, a statistically significant difference between the skew distributions of repair-proficient and repair-deficient bacteria indicates that this system is linked to mutational strand-biases. Consequently, either DNA repair activity is strand-asymmetric or the repair system acts against lesions that preferentially occur in one DNA strand, in particular the coding strand.

2nd.- Based on the adjusted p-values, a highly significant correlation (p-value < 0.01) is detected between $S_{\text{Trs}}^{G\text{C}}$ and repair systems that act against alkylation adducts ($\text{ada}$, $\text{alkA}$), excise oxidized pyrimidine bases ($\text{nei}$) or 3,N4-ethenocytosine ($\varepsilon\text{C}$) and U from $\varepsilon\text{C}:\text{G}$ and U:G mispairs, respectively, (mug), or perform mismatch-repair (methyl-directed or VSR-patch MMR) and recombination-repair (RecBC). In five of the aforementioned cases (mug, nei, methyl-directed MMR, VSR-patch MMR, RecBC), there is a switch from negative to positive $S_{\text{Trs}}^{G\text{C}}$ values in repair-proficient bacteria, indicating an inversion of mutational bias when the corresponding repair systems are active.

3rd.- In all studied cases which are statistically significant (adjusted p-value < 0.05) there is a shift of the $S_{\text{Trs}}^{G\text{C}}$ distribution from lower to higher (signed) values, when a repair mechanism is absent or present, respectively (SI Tables 5 and 6). In the case of global-genome repair (GGR), the inverse holds true; yet, the difference of the $S_{\text{Trs}}^{G\text{C}}$ distribution between GGR-proficient and GGR-deficient species is non-significant after applying the Bonferroni correction (initial p-value < 0.05; adjusted p-value > 0.05).

4th.- Methyl-directed and VSR-patch MMR process mainly errors lying in the newly synthesized strand. Thus, MMR systems are not expected to be related to transcription-associated skews, contrary to what we observe (adjusted p-value < 0.001). The contradiction is lifted if we take into account that in bacteria replication and transcription may proceed simultaneously on the same DNA strand. The coverage of leading strand by coding strands is > 50% in more than 81% of our collection.

5th.- It has been reported that bacteria lacking $\text{recA}$ and $\text{priA}$, both involved in RecBC homologous recombination pathway, display strong compositional biases $^{16}$. Our analysis indicates that, though $S_{\text{Trs}}^{G\text{C}}$ correlates highly significantly with $\text{recA}$ and $\text{priA}$ (RecBC: adjusted p-value < 0.001), bacteria lacking both $\text{recA}$ and $\text{priA}$ (RecBC$^{-}$) exhibit only a mild bias compared to bacteria bearing those genes (RecBC$^{+}$) (median $S_{\text{Trs}}^{G\text{C}}$: -0.0437 & 0.106, respectively). Thus, according to our results, and in line with other previous studies $^{17}$, recombination may enhance strand asymmetries.
REFERENCES

1. Zeileis, A., Kleiber, C., Krämer, W., and Hornik, K. 2003, Testing and dating of structural changes in practice. *Comput. Stat. Data Anal.*, 44, 109–23.

2. Zeileis, A., Shah, A., and Patnaik, I. 2010, Testing, monitoring, and dating structural changes in exchange rate regimes. *Comput. Stat. Data Anal.*, 54, 1696–706.

3. Zeileis, A., Leisch, F., Hornik, K., and Kleiber, C. 2002, strucchange : An R Package for Testing for Structural Change in Linear Regression Models. *J. Stat. Softw.*, 7, 1–38.

4. Paradis, E., Claude, J., and Strimmer, K. 2004, APE: Analyses of Phylogenetics and Evolution in R language. *Bioinformatics*, 20, 289–90.

5. Kullback, S., and Leibler, R. A. 1951, On Information and Sufficiency. *Ann. Math. Stat.*, 22, 79–86.

6. Nye, T. M. W., Liò, P., and Gilks, W. R. 2006, A novel algorithm and web-based tool for comparing two alternative phylogenetic trees. *Bioinformatics*, 22, 117–9.

7. Palidwor, G. A., Perkins, T. J., and Xia, X. 2010, A general model of codon bias due to GC mutational bias. *PLoS One*, 5, e13431.

8. Yu, J. 2007, A content-centric organization of the genetic code. *Genomics. Proteomics Bioinformatics*, 5, 1–6.

9. Xiao, J.-F., and Yu, J. 2007, A scenario on the stepwise evolution of the genetic code. *Genomics. Proteomics Bioinformatics*, 5, 143–51.

10. Copley, S. D., Smith, E., and Morowitz, H. J. 2005, A mechanism for the association of amino acids with their codons and the origin of the genetic code. *Proc. Natl. Acad. Sci. U. S. A.*, 102, 4442–7.

11. Lory, J. R., and Lory, C. 1999, Evolution of DNA base composition under no-strand-bias conditions when the substitution rates are not constant. *Mol. Biol. Evol.*, 16, 719–23.

12. Charneski, C. A., Honti, F., Bryant, J. M., Hurst, L. D., and Feil, E. J. 2011, Atypical AT skew in Firmicute genomes results from selection and not from mutation. *PLoS Genet.*, 7, e1002283.

13. Du, J., Yuan, Z., Ma, Z., Song, J., Xie, X., and Chen, Y. 2014, KEGG-PATH: Kyoto encyclopedia of genes and genomes-based pathway analysis using a path analysis model. *Mol. Biosyst.*, 10, 2441–7.

14. Morita, R., Nakane, S., Shimada, A., et al. 2010, Molecular mechanisms of the whole DNA repair system: a comparison of bacterial and eukaryotic systems. *J. Nucleic Acids*, 2010, 179594.

15. Resende, B. C., Rebelato, A. B., D’Afonseca, V., et al. 2011, DNA repair in Corynebacterium model. *Gene*, 482, 1–7.

16. Klasson, L., and Andersson, S. G. E. 2006, Strong asymmetric mutation bias in endosymbiont genomes coincide with loss of genes for replication restart pathways. *Mol. Biol. Evol.*, 23, 1031–9.

17. Rocha, E. P. C., Cornet, E., and Michel, B. 2005, Comparative and Evolutionary Analysis of the Bacterial Homologous Recombination Systems. *PLoS Genet.*, 1, e15.