DNA Research, 2016, 23(4), 353–363
doi: 10.1093/dnares/dsw021
Advance Access Publication Date: 26 June 2016
Full Paper

Full Paper

A novel skew analysis reveals substitution asymmetries linked to genetic code GC-biases and PolIII a-subunit isoforms

Konstantinos Apostolou-Karampelis[1],*, Christoforos Nikolaou[2], and Yannis Almirantis[1],*

[1]Institute of Biosciences and Applications, National Center for Scientific Research "Demokritos", 15310 Athens, Greece, and [2]Computational Genomics Group, Department of Biology, University of Crete, 71409 Heraklion, Greece

*To whom correspondence should be addressed. Tel: +302106503601. Email: apostolou@bio.demokritos.gr (K.A.K.); Email: yalmir@bio.demokritos.gr (Y.A.)

Edited by Dr Yuji Kohara

Received 11 February 2016; Accepted 9 May 2016

Abstract

Strand biases reflect deviations from a null expectation of DNA evolution that assumes strand-symmetric substitution rates. Here, we present strong evidence that nearest-neighbour preferences are a strand-biased feature of bacterial genomes, indicating neighbour-dependent substitution asymmetries. To detect such asymmetries we introduce an alignment free index (relative abundance skews). The profiles of relative abundance skews along coding sequences can trace the phylogenetic relations of bacteria, suggesting that the patterns of neighbour-dependent substitution strand-biases are not common among different lineages, but are rather species-specific. Analysis of neighbour-dependent and codon-site skews sheds light on the origins of substitution asymmetries. Via a simple model we argue that the structure of the genetic code imposes position-dependent substitution strand-biases along coding sequences, as a response to GC mutation pressure. Thus, the organization of the genetic code *per se* can lead to an uneven distribution of nucleotides among different codon sites, even when requirements for specific codons and amino-acids are not accounted for. Moreover, our results suggest that strand-biases in replication fidelity of PolIII α-subunit induce substitution asymmetries, both neighbour-dependent and independent, on a genome scale. The role of DNA repair systems, such as transcription-coupled repair, is also considered.

Key words: Chargaff's second parity rule (PR2), dinucleotide relative abundances (odds ratios), GC mutational pressure, substitution strand-biases, PolIII a-subunit isoforms

1. Introduction

The evolution of genomic base composition depends, in a complex fashion, on mutation and selection.[1] When there is no strand-specific bias in mutation rates and selection pressure, the interstrand base-pairing rule reduces the maximum number of substitution rates to six: $r_{A\rightarrow T}=r_{T\rightarrow A}$, $r_{G\rightarrow C}=r_{C\rightarrow G}$, $r_{A\rightarrow G}=r_{T\rightarrow C}$, $r_{G\rightarrow A}=r_{C\rightarrow T}$, $r_{A\rightarrow C}=r_{T\rightarrow G}$, and $r_{C\rightarrow A}=r_{G\rightarrow T}$, as stated by parity rule 1 (PR1).[2] Under PR1, the average intrastrand base composition of a chromosome is expected to reach an equilibrium state, at which [A] = [T] and [G] = [C], a relationship known as PR2,[2,3] also referred to as Chargaff's second parity rule.

Most of bacteria abide by PR2 on a chromosome-wide scale.[4–6] Nonetheless, they display distinctive local deviations from this parity, as shown by previous studies.[7–16] The departure from PR2 provides a window into fundamental processes, such as replication, transcription, and repair mechanisms, which act differentially along each DNA strand and induce strand biased substitution rates, shaping DNA sequence contrary to the null expectation of [A] = [T] and [G] = [C] equilibrium.[8,16–22]

Lobry suggested that the observed compositional strand-asymmetries are due to differences in replication-associated mutagenesis and repair between leading and lagging strands,[7] and following studies argued along the same lines.[9,23] These asymmetries are also attributed to the uneven gene partition between replication strands[11,16,22,24,25] coupled with transcription-induced mutagenesis and repair,[26] as well as codon and amino acid usage biases.[27–30]

From another perspective, nucleotide composition of genes, mainly reflected on GC content, may affect amino acid biases of proteins, as suggested by compositional correlation between DNA and protein sequences,[31] and hydropathy profile studies.[32] Accordingly, mutational pressure, identified on the basis of GC content and DNA compositional biases, can have a major effect on the long-term molecular evolution of proteins.

Though there are numerous studies on mononucleotide strand asymmetries, the distribution of oligonucleotides among different strands seems to have attracted less attention.[33,34] Interestingly, no study has extensively assessed strand asymmetries at the level of dinucleotides, though these oligomers are the primary ordering units of DNA bases.[35]

Dinucleotide occurrence 'preferences' have been reported since very early[36,37] as a result of nearest-neighbour constraints. The relative abundances (odds-ratios) of dinucleotides quantify such preferences. It has been noted that within the same DNA strand the relative abundance of a given dinucleotide is approximately equal to the one of each reverse complement,[38–40] suggesting a base order intra-strand parity between relative abundances of reverse complementary dinucleotides. Mrázek and Karlin[10] considered whether deviations from this base order intra-strand parity exist. However, based on the limited number of complete genomes then available, they concluded that 'dinucleotide relative abundances tend to be symmetric and effectively constant relative to the leading and lagging strand for all dinucleotides, despite the strand compositional asymmetry'.[10]

In this work, we provide statistically solid evidence for the existence of local deviations from intra-strand parity at the level of dinucleotide relative abundances. We show that such skews along coding sequences (CDS) can trace the standard phylogenetic relationships more efficiently than other compositional measures. The skew analysis that we present supports that the GC-biased structure of genetic code shapes compositional asymmetries along coding regions. Moreover, we assess whole-genome substitution strand-biases induced by the molecular machinery of replication, transcription and DNA-repair.

2. Materials and methods

2.1 Dataset

We perform our analysis on the dataset used by Necşulea and Lobry to study DNA base composition asymmetries in prokaryotes.[22] For each bacterial sequence, we retrieved the latest version deposited in NCBI database until 20 April 2015, restricting our collection to species which belong to phyla with more than three members present in the initial dataset. For each chromosome, we retrieve the coordinates of the origin of replication (*ori*) from DoriC 5.0 database.[41] We exclude linear chromosomes whose origin of replication is located less than one-fifth of their total length away from their ends. In total, 340 DNA sequences of the initial dataset meet the above prerequisites and constitute our genomic dataset.

2.2 Definition of skews

Mononucleotide skews are computed as the following ratios: $S^{\mathrm{A-T}} = \frac{f^{\mathrm{A}} - f^{\mathrm{T}}}{f^{\mathrm{A}} + f^{\mathrm{T}}}$ and $S^{\mathrm{G-C}} = \frac{f^{\mathrm{G}} - f^{\mathrm{C}}}{f^{\mathrm{G}} + f^{\mathrm{C}}}$, where f^{X} is the observed frequency of $\mathrm{X} \in \{\mathrm{A, T, G, C}\}$ in a given DNA sequence.

For each pair of reverse complementary dinucleotides, the observed frequency skew is computed as the following ratio: $S^{\mathrm{XY-Y'X'}} = \frac{f^{\mathrm{XY}} - f^{\mathrm{Y'X'}}}{f^{\mathrm{XY}} + f^{\mathrm{Y'X'}}}$, where Y'X' is the reverse complement of XY and f^{XY}, $f^{\mathrm{Y'X'}}$ are the observed frequencies of the corresponding dinucleotides, as in reference.[39]

We introduce the relative abundance skews of reverse complementary dinucleotides, which are given by the formula: $P^{\mathrm{XY-Y'X'}} = \rho^{\mathrm{XY}} - \rho^{\mathrm{Y'X'}}$ where Y'X' is the reverse complement of XY and ρ^{XY}, $\rho^{\mathrm{Y'X'}}$ are the relative abundances of the corresponding dinucleotides. The relative abundance of a given dinucleotide is the ratio of its observed frequency to expected frequency, the later being the product of the observed frequencies of its mononucleotide components. Hence $\rho^{\mathrm{XY}} = f^{\mathrm{XY}} / f^{\mathrm{X}} f^{\mathrm{Y}}$.

Note that, from the 16 distinct dinucleotides, 4 of them are identical to their reverse complementary ones. The remaining 12 form 6 pairs of reverse complementary dinucleotides, namely AG-CT, GA-TC, GG-CC, AA-TT, AC-GT and CA-TG, for which we study the dinucleotide frequency and relative abundance skews.

2.3 Graphical representation of skews

For every chromosome in our collection, we calculate the skews along the plus strand in reference sequence in adjacent, non-overlapping windows of 10^4 bp length and plot their non-cumulative and cumulative diagrams. The coordinates of all circular chromosomes are shifted so that the *ori* is placed at the middle of their published sequence. Then, the *ter* is set as the location of *ori* plus half of the chromosomes length, as in previous genomic studies.[42,43] In this way, the *ter* region corresponds to the end (or, which is the same, to the start), of the shifted sequences, with the first half of each diagram corresponding to the lagging strand, and the second half to the leading strand.

2.4 CDS concatenates

From each DNA sequence, we retrieve the coding strand of all its protein-coding genes, in their sense direction, and concatenate them to form an artificial sequence, the 'CDS concatenate'. The ordering of these coding strands maintains their relative position along the original DNA sequence. As the *ori* usually does not overlap with protein-coding genes, we detect the region of CDS concatenates which is closer to the *ori*. For all circular sequences we shift the coordinates of CDS concatenates so that this region will be placed at the middle of these artificial sequences.

2.5 Detecting breakpoints in skew patterns

In many bacterial chromosomes, mononucleotide skews are close to constant along the leading or lagging strand, with their sign changing near the *ori* and *ter*. Thus, it is reasonable to assume that mononucleotide skews fit a linear regression model along each replication strand. In this context, *ori* (or *ter*) represents a breakpoint, where the model coefficients shift from one stable regression relationship to a different one. To test the statistical significance of these structural changes in mononucleotide skew patterns and assess whether such changes also appear in dinucleotide frequency and relative abundance skews, we use a dynamic programming algorithm developed in references,[44,45] as implemented in the R package strucchange.[46]

The algorithm allows simultaneous estimation of multiple breakpoints by assessing deviations from stability in the fitted linear regression model. It computes the number and position of the optimal breakpoints from data alone, with no a priori knowledge. We apply this algorithm in skews along both plus strand and CDS concatenates, and plot the resulting fitted models.

2.6 Correlation of strand compositional asymmetries with species phylogeny

We group the skews along the CDS concatenates into three classes; 'mononucleotide skews' (V^{MONO}): ${S_{CDS}}^{A\text{-}T}$ and ${S_{CDS}}^{G\text{-}C}$, 'dinucleotide skews' (V^{DI}): ${S_{CDS}}^{XY\text{-}Y'X'}$, and 'relative abundance skews' (V^{RA}): ${P_{CDS}}^{XY-Y'X'}$, where $XY-Y'X' \in$ {AG-CT, GA-TC, GG-CC, AA-TT, AC-GT, CA-TG}. Each of these classes can be seen as a multidimensional random variable, corresponding to a list of vectors of skew values. We compute the V^{MONO}, V^{DI}, and V^{RA} along the CDS concatenates in adjacent, non-overlapping windows of 10^4 bp length. To assess the similarity between the skew patterns of any two CDS concatenates, we pairwise compare the multivariate distribution of each of their variables (V^{MONO}, V^{DI}, V^{RA}) using the symmetric Kullback-Leibler (KL) divergence.[47] Thus, the symmetric KL-divergence quantifies the dissimilarity between each pair of bacteria present in our collection, in terms of skew profiles along the corresponding CDS concatenates.

We group the bacteria of our collection according to their phylum, and, in the case of Proteobacteria (which is by far the largest phylum in our dataset), according to their class. For each of these groups, we perform complete-linkage hierarchical clustering of the CDS concatenates, using the symmetric KL-divergence of their V^{MONO} or V^{DI} or V^{RA}, to obtain the corresponding cladograms (skew-based trees). For purposes of this analysis, if a species has more than one chromosomes, we keep the largest chromosome excluding any other (a total of 29 DNA sequences). We employ the web-based tool Compare2Trees[48] to compare the topology of the skew-based trees and the species-trees, corresponding to the phylogenetic relations of bacteria. We obtained the species-tree from NCBI Taxonomy,[49] which provides a sequence-based phylogenetic classification, manually curated to reflect the current consensus in the systematic literature. The scores resulting from Compare2Trees express the percent topological similarity of the trees and indicate whether V^{MONO}, V^{DI}, V^{RA} can trace bacterial species phylogeny. We use as benchmark the topological scores corresponding to cladograms we construct based on the δ-distance of the genomic signatures.[50,51]

In order to detect correlations of skews with bacterial phylogeny, we use the multidimensional variables V^{MONO}, V^{DI}, and V^{RA} instead of each skew separately, because if i.e. the rates of A→G transitions differer among species, this will simultaneously affect both ${S_{CDS}}^{A\text{-}T}$ and ${S_{CDS}}^{G\text{-}C}$. Moreover, in the same example, if the rates of A→G are context-dependent, this may also affect all relative abundance skews, since the six reverse complementary dinucleotide pairs we examine have all at least one dinucleotide which contains an A or a G.

2.7 Evolutionary relations of CDSs

We retrieve the non-supervised orthologous groups from EggNOG v4.0 database.[52] We consider orthologous genes at the taxonomic levels for which our cladistic diagrams are constructed. For all species within a given taxonomic level, we pick out of each chromosome the coding strands of the genes which have orthologs in no >10% of these species (or in just one of them, for taxonomic groups with few members present in our collection). These CDSs form the 'low-coverage orthologs' group of the chromosome. All other orthologs detected in the respective chromosome form the 'remaining orthologs, (complementary) group.

For species with more than one chromosomes we keet the largest one and exclude all the others. We also exclude from our analysis the chromosomes of the species which are not represented in the eggNOG v4.0 database. Thereafter, a total of 266 DNA sequences are taken into account.

2.8 Transcription- and replication-associated skews

To detect mutation-derived skews we use third 4-fold-degenerate sites (third-ff), which are not subject to selection on the encoded amino-acid. We consider nucleotide composition in the sense direction and distinguish between codon-sites of genes lying on leading and lagging strand, where replication and transcription progress in the same or opposite direction, respectively. For each skew, let $SK_{sense,leading}$ and $SK_{sense,lagging}$ be its values in the first and second set of codon-sites, respectively. For explanatory purposes, let us denote by α the transcription-induced component of this skew and by β the replication-induced component. Transcription-induced skews have opposite signs in sense versus anti-sense strands. We consider α with respect to the sense strand. Likewise, replication-induced skews have opposite signs in leading versus lagging strand. We consider β with respect to the leading strand. Then, $SK_{\mathbf{sense,leading}} \approx \alpha + \beta$ and $SK_{\mathbf{sense,lagging}} \approx \alpha - \beta$. We define the transcription-associated (Trs) skew as $\frac{SK_{sense,leading}+SK_{sense,lagging}}{2}$, which is an estimate of α. Likewise, we define the replication-associated (Rep) skew as $\frac{SK_{sense,leading}-SK_{sense,lagging}}{2}$, which is an estimate of β. The above formulation provides a heuristic decoupling of the effect of replication and transcription on stand asymmetries, represented by Rep and Trs, respectively.

Because certain base substitutions depend on the identity of the adjacent nucleotides,[53] we also have to take into account flanking neighbours of third-ff sites. Since there are no 4-fold degenerate codons with A at their second position, we only consider dinucleotides, X_3Y_1, occupying a third-ff site and the first site of the adjacent codon, as described in.[54] In this context, the relative abundance of X_3Y_1 is: $\rho^{X}{}_{3}{}^{Y}{}_{1} = f_3^{XY}{}_1/f_3^X f_1^Y$, with f_3^X and f_1^Y being the frequencies of X and Y at third-ff and their neighbouring first sites, respectively, with $X, Y \in \{A, T, G, C\}$.

2.9 Transcription-coupled repair and PolIII α-subunit isoforms

To identify which bacteria possess the *mfd* gene [transcription-coupled repair (TCR)-proficient] and what type of polymerase III (PolIII) α-subunit isoform they bear, we retrieve the corresponding data from the KEGG Orthology database.[55]

3. Results and discussion

3.1 Relative abundance skews constitute a general property of bacterial genomes

A convenient way to analyse strand biases of single base distribution is by means of S^{A-T} and S^{G-C} cumulative diagrams.[56] Likewise, cumulative diagrams of dinucleotide and relative abundance skews can be used to assess compositional asymmetries at the level of dinucleotide observed frequencies (*f*) and relative abundances (ρ), respectively. In Figure 1 we indicatively present the cumulative skew diagrams of CA–TG and AC–GT pairs along the plus strand of the chromosome of *Carboxydothermus hydrogenoformans* $Z-2901$.

The diagrams, in terms of both *f* and *ρ*, present a characteristic V/inverted-V shape, with their extreme located at the *ori* site. Relative abundance skew patterns are reported here for the first time. These highly structured patterns point to intra-strand asymmetries that should be rather significant, both statistically and biologically.

Strand asymmetries in substitution rates of a nucleotide can be drastically affected by its flanking neighbours, in a genome-wide scale. Dinucleotide frequency skews can emerge from strand-biased substitution processes which are either neighbour-dependent or-independent. On the contrary, persistent patterns of relative abundance skews can only result from strand biases of neighbour-dependent substitutions.

The S_{plus}^{CA-TG} and P_{plus}^{CA-TG} curves point towards the same direction, both displaying a clear inverted-V shape. On the other hand, the S_{plus}^{AC-GT} and P_{plus}^{AC-GT} curves point in the opposite direction, the former bearing an inverted-V and the latter a V shape. This is an indication of the different modalities acting on the formation of compositional skews (i.e. S_{plus}^{AC-GT}) and nearest-neighbour preference skews (i.e. P_{plus}^{AC-GT}).

From Figure 1b it follows that GT is more frequently occurring in leading strand than AC, in accordance with an excess of G and T versus C and A. However, the relative abundance of AC in leading strand is higher than the one of GT. Unlike observed frequencies, the relative abundances of dinucleotides factor out the effect of single-base composition and their skews cannot be inferred from the corresponding mono- and di-nucleotide ones. Instead, relative abundance skews constitute an alignment-free index which effectively detects chromosomal regions where context-dependent substitutions are strand-biased.

The patterns in Figure 1 indicate that the depicted skews are quasi-constant along the leading or lagging strand and have opposite sign upstream and downstream the *ori*. To assess the statistical significance of the observed changes in skew polarity we use a linear regression setup.[44,45] The fitted models of S_{plus}^{AC-GT} and P_{plus}^{AC-GT} display sharp, step-like breakpoints at the *ori* site (Fig. 2a and c), denoting that dinucleotide and relative abundance skews significantly correlate with the mode of replication. These breakpoints illustrate asymmetric substitution patterns in the two replication strands.[8]

Compositional skews may also emerge due to substitution strand-biases associated with CDSs.[26] To address this question, we use the same linear regression setup to model S_{CDS}^{AC-GT} and P_{CDS}^{AC-GT} along the corresponding CDS concatenates, where all genes are considered in the direction of transcription. In this way, consecutive parts of the leading and lagging strand are alternating within both halves of the genome as delimited by *ori*. Thus, along CDS concatenates the effect of replication on the observed skews is cancelled out within a window of a few thousands bps. As seen in Figure 2b and d though skew diagrams display fluctuations, there is no structural change near the position of the *ori*. The fitted models are constant along CDS concatenates, highlighting that the transcription orientation of genes is a major determinant of strand biases, at the level of both base composition (S_{CDS}^{AC-GT}) and base order (P_{CDS}^{AC-GT}).[26,57]

Analysis of 340 bacterial chromosomes with cumulative diagrams and linear regression models shows that both dinucleotide and relative abundance skews are prevalent features of bacterial genomes in general. The corresponding plots are provided in a web-page (see

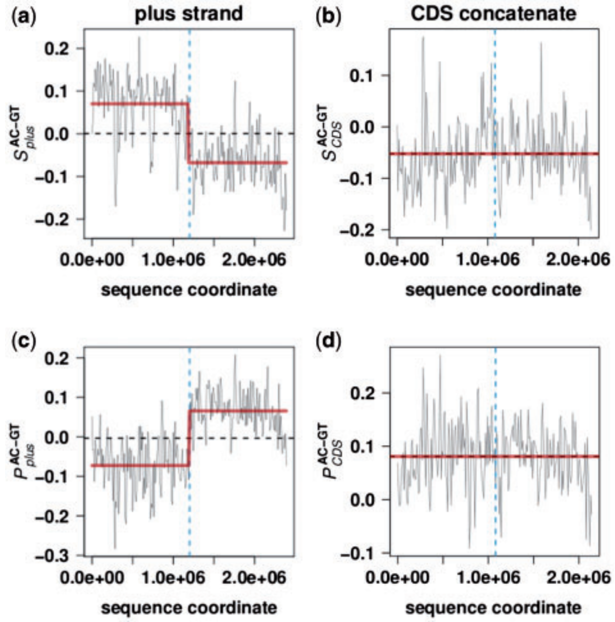

Figure 2. Diagrams (thin curve) and fitted linear models (thick line) for S^{AC-GT} and P^{AC-GT} in *Carboxydothermus hydrogenoformans* Z − 2,901, along the plus (published) strand (**a** and **c**) and the CDS concatenate (**b** and **d**). The vertical dashed straight line indicates the *ori*. The horizontal dashed straight line is drawn at the mean value of the skew. Along the plus strand skews display a step-like breakpoint at the *ori* and their mean values approximately equal zero, since skews in leading and lagging strand are almost cancelled out on a genome scale (a and c). On the contrary, skews along CDS concatenates display no structural change near the *ori* and their mean considerably departs from zero (b and d).

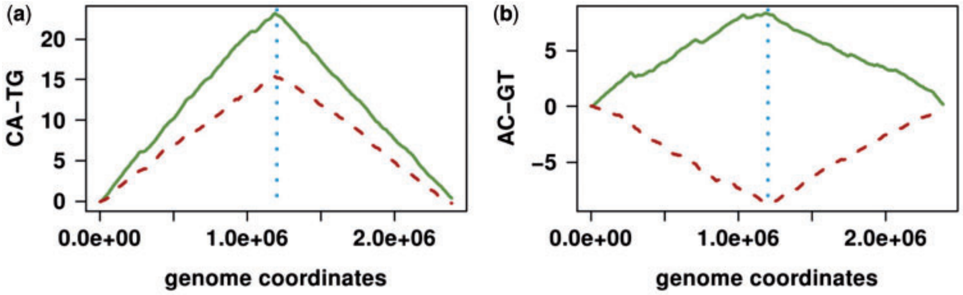

Figure 1. Cumulative skew diagrams along the plus strand of the chromosome of *Carboxydothermus hydrogenoformans* Z − 2,901. In each diagram, both dinucleotide skews (continuous line) and relative abundance skews (dashed line) are plotted. For the full set of cumulative skew diagrams, corresponding to all genomes in our collection, see link to Additional Figures and Tables.

link to Additional Figures and Tables), along with tabulated descriptive statistics of the skew distributions within our collection. The form of our cumulative skew diagrams ranges from the characteristic V or inverted-V shapes, to heavily distorted curves with irregular peaks and troughs, in line with previously known skew patterns of bacterial genomes.[15,16] Overall inspection of dinucleotide and relative abundance skew diagrams shows that their patterns cannot be directly inferred from one another. When focusing on the linear regression models of relative abundance skews, strand-asymmetries of context-dependent substitutions seem to significantly correlate with replication- and CDS-associated biases in up to >50 and 80% of our collection, respectively (Supplementary Table S1).

3.2 Strand biases of CDSs are species-specific

The set of strand-symmetrised relative abundances (ρ^*_{XY}), namely 'genomic signatures', have been used for phylogenetic reconstruction.[58] For a dinucleotide, XY, $\rho^*_{XY} = f^*_{XY}/f^*_X f^*_Y$, where f^*_{XY}, f^*_X and f^*_Y are the frequencies of dinucleotide XY and mononucleotides X, Y, respectively, computed along the concatenate of the sequence with its reverse complement.[35,58] We assess whether strand biases of relative abundances, quantified by relative abundance skews, can also track the evolutionary path of divergent species. Based on pairwise comparisons of relative abundance skew profiles along CDSs, we construct the corresponding cladograms (see Methods section and Supplementary Figure S1). Then, we compare the skew-based trees to the species trees and obtain their percent topological similarity (Fig. 3a).

As seen in the corresponding boxplots, both strand-specific ($\rho_{XY} - \rho_{Y'X'}$) and symmetrised (ρ^*_{XY}) versions of nearest-neighbour correlations are species-specific. For all taxonomic subdivisions the inferred topology largely agrees with the species phylogeny, with the resulting scores being higher than 72% in all cases except Actinobacteria and δ-Proteobacteria (Supplementary Table S2).

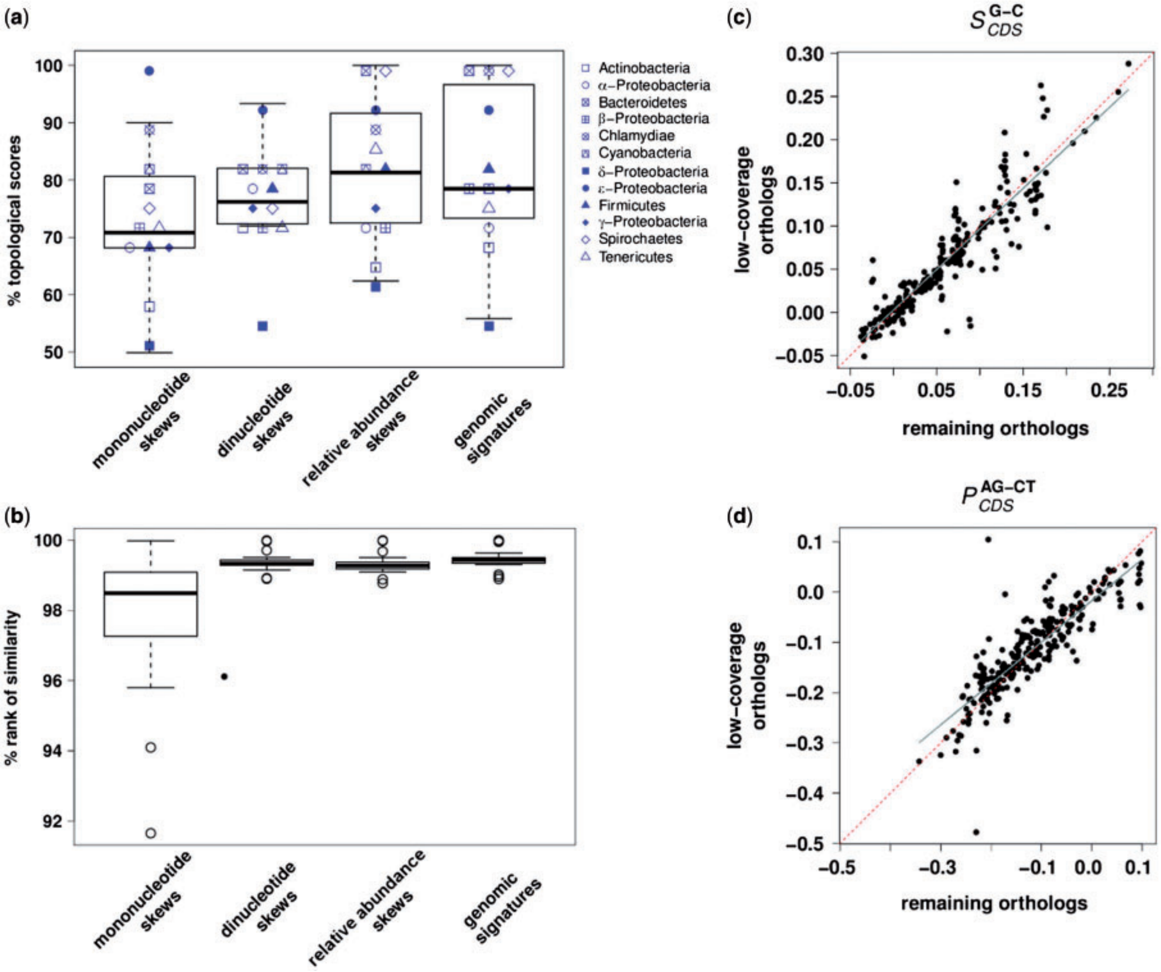

Figure 3. a. Boxplots of topological scores, depicting the percent topological similarity of the skew-based trees to the species-trees. Comparisons are made between species that belong to the same taxonomic rank (phylum or class). **b.** Boxplots of percent rank of similarity between chromosomes of the same organism, in terms of their skew distribution along their CDSs. **a and b.** Skews are grouped into three classes: mononucleotide skews (V^{MONO}), dinucleotide skews (V^{DI}), and relative abundance skews (V^{RA}). The topological scores and the percent rank of similarity based on δ-distances between genomic signatures are presented as a benchmark. **c and d.** Regression of the skews (${S_{CDS}}^{G\text{-}C}$ and ${P_{CDS}}^{AG\text{-}CT}$) 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; dashed line) indicate that skews in 'low-coverage orthologs' are identical to the skews in 'remaining orthologs'. ${S_{CDS}}^{G\text{-}C}$ regression line roughly coincides with the diagonal (slope = 0.939, y-intercept = 2.75×10^{-3}), indicating that ${S_{CDS}}^{G\text{-}C}$ of the 'low-coverage orthologs' comply with ${S_{CDS}}^{G\text{-}C}$ of the *remaining orthologs*, in almost all the chromosomes we examined. ${P_{CDS}}^{AG\text{-}CT}$ regression line intercepts the diagonal, with its slope close to unity (0.823) and y-intercept close to zero ($-1.74 \times {10^{-2}}$), indicating that $P^{AG\text{-}CT}$ is also approximately equal in 'low-coverage' and 'remaining orthologs'.

Moreover, skew-based trees tend to be more accurate compared with genomic signature trees, as evidenced by their median scores. Thus, it can be argued that the strand bias of first-neighbour correlations constitute an idiosyncratic genomic feature that is linked to the evolutionary dynamics of the DNA strands.

We also present the topological scores of mono- and di-nucleotide skew-based trees (Fig. 3a). Base frequencies and base order are fused at the level of base doublets, and hence dinucleotide skews reflect the combined effect of mononucleotide and relative abundance strand-biases. As shown by the median scores, skews contain information that enables increasingly accurate phylogenetic reconstruction as we shift from base singlets to doublets and eventually to doublet relative abundances.

To rule out the possibility that the observed skew similarities between different genomes are the result of CDS homology, we compare the skews of non-homologous parts of the same genome.

We sort in ascending order all pairs of chromosomes according to the similarity of their skews, in terms of symmetric KL-divergence, as obtained by pairwise comparisons of the distribution of each of their variables, V^{MONO}, V^{DI}, V^{RA} along the CDS concatenates. We focused on the bacteria which carry more than one chromosomes (a total of 26 in our collection; plasmids are excluded). In Figure 3b, we present the ranks of similarity between chromosomes that belong to the same organism (a total of 55 pairs). When comparing dinucleotide or relative abundance skews, the distributions of ranks are quite compact, while ranks corresponding to mononucleotide skews are more scattered. Yet in all cases the median values are higher than 98%. Since different chromosomes of the same organism share limited, if any, homologous regions, their high ranks of skew similarity cannot be attributed to sequence similarity.

Moreover, in each chromosome we distinguish between two separate groups of genes, on the basis of the abundance of their orthologs within a given phylum or class (‘low-coverage’ vs. ‘remaining’ orthologs). For each skew, a single value is computed along the coding strand of genes which belong to the same group. In Figure 3c and d we regress $S_{CDS}{}^{G\text{-}C}$ and $P_{CDS}{}^{AG\text{-}CT}$ of the ‘low-coverage’ orthologs on $S_{CDS}{}^{G\text{-}C}$ and $P_{CDS}{}^{AG\text{-}CT}$ of the ‘remaining’ ones. The slope and y-intercept are close to unity and zero, respectively, indicating that ‘low-coverage’ and ‘remaining’ orthologs exhibit highly similar patterns of $S_{CDS}{}^{G\text{-}C}$ and $P_{CDS}{}^{AG\text{-}CT}$. Similar correlation patterns are detected for all skews (Supplementary Figure S2 and Supplementary Table 3). We conclude that within a given chromosome CDSs have quite similar contribution on the overall CDS skews, irrespective of their evolutionary descent.

Taken together, our results strongly suggest that the correlation of skews with the evolutionary course of bacteria cannot be attributed to sequence divergence of CDSs originating from a shared ancestry. Regarding in particular relative abundance skews, we obtained clear indications that they reflect species-specific substitution strand-biases, which are context-dependent.

3.3 The GC biased structure of genetic code imposes site-specific strand-asymmetries

Skews can be effectively implemented in the analysis of evolutionary processes that distinguish between coding and transcribed strands, resulting in strand-biased substitutions. Here, we inquire site-specific biases in nucleotide distribution along CDSs. We compute the $S_{CDS}{}^{A\text{-}T}$ and $S_{CDS}{}^{G\text{-}C}$ at each codon site. Regarding third codon position, 4-fold (third-ff) and 2-fold (third-tf) sites are considered separately. For all chromosomes in our collection, we compute the Pearson’s product-moment correlation coefficient (Pearson’s *r*)[59] between $S_{CDS}{}^{G\text{-}C}$ at a given codon site and the frequency of each codon, and plot the corresponding values in heatmaps (Fig. 4a–d).

If the observed compositional asymmetries resulted primarily from synonymous codon usage, clear correlation patterns should appear when considering third-ff sites, where all substitutions are synonymous.[2] However, $S_{CDS}{}^{G\text{-}C}$ at third-ff sites displays only weak correlation with codon frequencies. At the first, second, and the third-tf sites $S_{CDS}{}^{G\text{-}C}$ is associated with codon frequencies in a way that distinguishes between A/T- and G/C-ending codons. Namely, at third-tf and first sites *r* is positive for A/T-ending codons and negative for G/C-ending codons, while the inverse holds true at second sites. These rather structured patterns, which are also evident in the case of $S_{CDS}{}^{A\text{-}T}$ (Supplementary Figure S4), are, *prima facie*, counter-intuitive. Third-tf sites are susceptible only to GC-changing substitutions (A↔G or C↔T) without altering the encoding amino-acid. Only codons for Arg and Leu allow synonymous, GC-changing substitutions in both first and third position, while substitutions in the remaining first and all second sites are non-synonymous. If skews at the first, second, and the third-tf sites resulted primarily from selective constraints, they should be correlated with the frequencies of codons specifying particular amino-acids.[2] Instead, their values respond to the GC% of the third codon sites. In the evolutionary timescale, GC% changes much faster at third than first and second codon sites, and thus more accurately reflects genome-wide mutational biases towards a particular GC content.[60]

Previous studies documented that species with similar GC content share similar patterns of codon and amino-acid usage.[31,61–63] These empirical relations imply a direction of causality from GC-biased substitutions to codon and amino-acid usage, and not vice versa, since codon frequencies can be arranged in many different ways resulting in the same GC content.[60,64] Motivated by this, we put forward a simple model in which the codon usage of an artificial sequence is determined solely by its GC content, taking into account the GC variability within groups of synonymous codons. We formulate the response of each codon to the GC content of the sequence as a function with three parameters: n_i, $R_i{}^{GC}$, and $R_i{}^{AT}$. n_i is the number of G and C of the *i*th codon.

$$R_i^{\mathrm{GC}} = \frac{\mathrm{GC}_i}{\frac{1}{a_i}\sum_{j=1}^{a_i} \mathrm{GC}_j}$$

where GC_i is the GC content of the *i*th codon and a_i the number of codons in its synonymous group. Similarly

$$R_i^{\mathrm{AT}} = \frac{\mathrm{AT}_i}{\frac{1}{a_i}\sum_{j=1}^{a_i} \mathrm{AT}_j}$$

where $\mathrm{AT}_i = 1 - \mathrm{GC}_i$. We introduce $R_i{}^{GC}$ and $R_i{}^{AT}$ to assess the response of the *i*th codon to the overall GC content of the sequence, given the mean GC content of its synonymous group. Thus, $R_i{}^{GC}$ and $R_i{}^{AT}$ introduce synonymous-codon GC biases in our model. For a given GC content (GC) we generate an artificial sequence comprised of 10^6 codons. Codon probabilities (P_i) are estimated as:

$$P_i = (\mathrm{GC})^{n_i R_i^{\mathrm{GC}}} (1 - \mathrm{GC})^{(3-n_i) R_i^{\mathrm{AT}}}$$

with their sum normalized to unity.

We produce a set of artificial sequences with their GC content ranging from 0.2 to 0.8, with a step of 0.005 (Supplementary Figure S3). Then, we compute Pearson’s *r* between site-specific skews and codon frequencies, as derived from the artificial sequences. Note

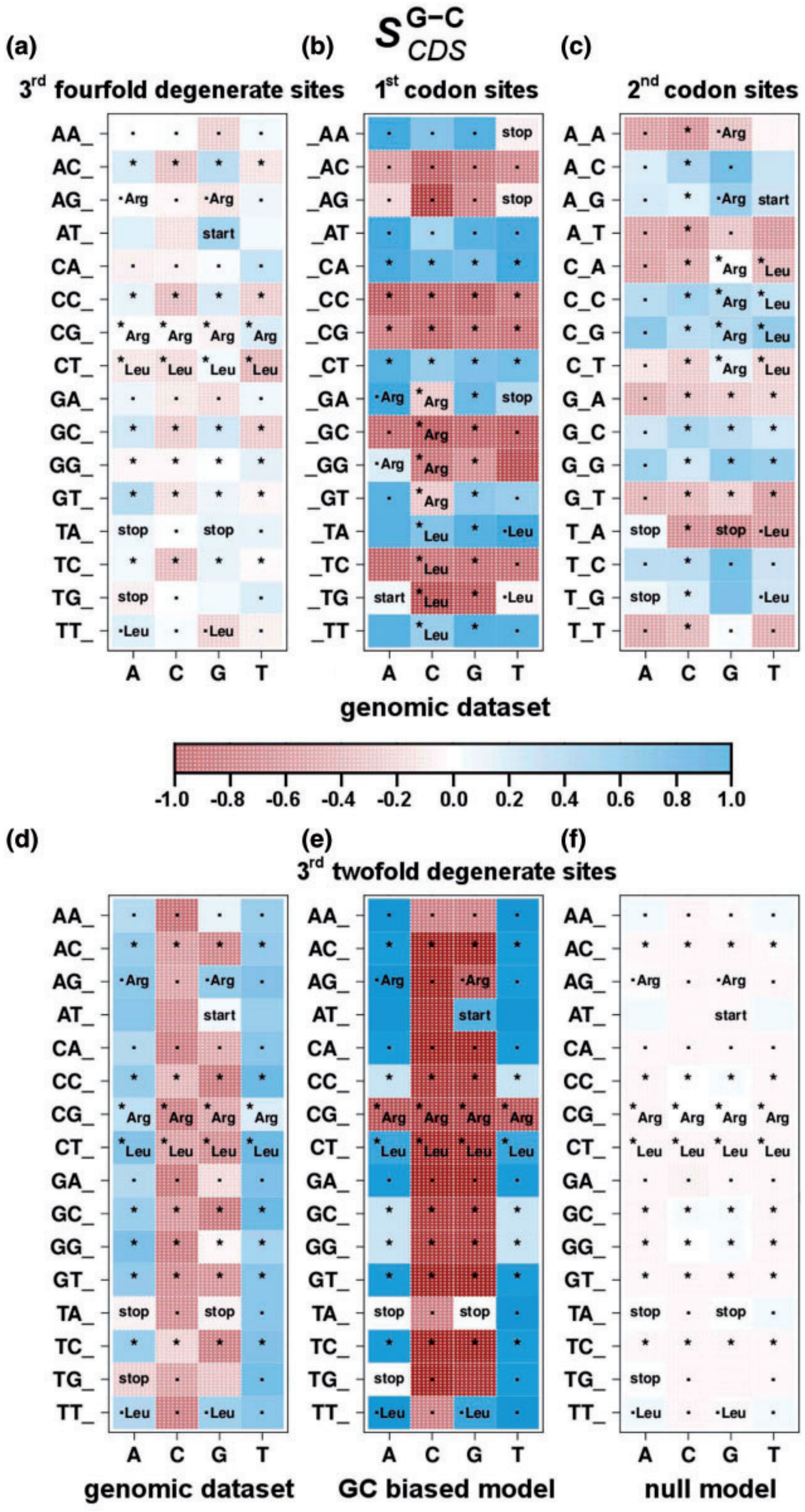

Figure 4. Heatmaps of Pearson's product-moment correlation coefficient (*r*) for codon usage versus $S_{CDS}^{G\text{-}C}$. (**a–d**) $S_{CDS}^{G\text{-}C}$ is calculated at the third-ff, the first, second, and the third-tf codon sites, in genomic data. (**e and f**) $S_{CDS}^{G\text{-}C}$ is calculated at third-tf of the artificial sequences produced by our synonymous-codon GC biased model and the null-hypothesis model. Columns indicate the base composition of the codon site at which skews are computed. Stippled cells depict negative *r*, non-stippled cells depict positive *r*. The intensity of the color indicates the degree of correlation, as recited in the color key. +1, perfect positive correlation; −1, perfect negative correlation; 0, no correlation. Start and stop codons are excluded from our calculations. ATG (start codon) and, in some cases, TGA (opal stop codon) are also found inside CDS, the first encoding for Methionine and the second for Tryptophane or Selenocysteine, and thus correlations of $S_{CDS}^{A\text{-}T}$ and $S_{CDS}^{G\text{-}C}$ with [ATG] and [TGA] appear in the heatmaps. The six-fold degenerate codons of Leucine and Arginine are labeled in the heatmaps. For further remarks see Supplementary Figure S4. Codons used in calculations of skews at third 4- and 2-fold-degenerate sites are marked with '*' and '•', respectively. 'start', initiation codon; 'stop', termination codons.

that our model is not built to predict the codon usage actually observed in bacterial genomes. Instead, it is aimed to assess whether there is a causal link between strand asymmetries in CDSs and the structure of genetic code, driven by GC content variability, as implied by the heatmaps in Figure 4b–d. The parameters n_i, ${R_i}^{\mathrm{GC}}$ and ${R_i}^{\mathrm{AT}}$ are symmetric with respect to the complementary nucleotide bases and do not account for base order within codons (Supplementary Table S4). Thus, any compositional asymmetries in the produced artificial sequences and their correlation with codon usage, can only be attributed to the partition of codons into synonymous groups and the GC varibility within these groups.

The corresponding heatmaps show no significant correlation between skews at third-ff sites and codon usage, but largely reproduce the correlation patterns detected in our genomic dataset when considering skews at third-tf sites (cf. Figure 4d,e), first and second sites (cf. Supplementary Figure S4e–p). Namely, at third-tf and first sites of the artificial sequences ${S_{\mathrm{CDS}}}^{\text{A-T}}$ and ${S_{\mathrm{CDS}}}^{\text{G-C}}$ increase with A/T-ending codons and counter to G/C-ending ones. The correlation is reversed at second codon sites of these sequences (Supplementary Figure S4n and p). The only marked discrepancy between genomic data and our model regards ${S_{\mathrm{CDS}}}^{\text{A-T}}$ at second codon sites, where the inverse pattern of correlation is detected (cf. Supplementary Figure S4m and n). Among other factors, this discrepancy can be attributed to selective pressures, which have been previously associated with the emergence of atypical ${S_{\mathrm{CDS}}}^{\text{A-T}}$ patterns.[16]

Our model suggests that given the GC bias of synonymous codons (${R_i}^{\mathbf{GC}}$ and ${R_i}^{\mathbf{AT}}$) the interspecies variation of GC content suffices to yield the observed correlations between codon usage and site-specific skews along CDSs. To further test our argument, we consider a null-hypothesis model, in which $P_i = (\mathrm{GC})^{n_i}(1-\mathrm{GC})^{(3-n_i)}$, and thus synonymous-codon GC biases (${R_i}^{\mathbf{GC}}$ and ${R_i}^{\mathbf{AT}}$) are not accounted for. Under this null-model the previously observed patterns collapse (Fig. 4f), indicating that synonymous-codon GC biases are indeed a decisive factor linking skews to codon usage.

GC mutational biases tailored with the structure of the genetic code can drastically modulate the nucleotide and codon composition of CDSs;[65,66] and references given therein. Here we presented evidence that the synonymous-codon GC biases along with GC content variability among genomes, attributed to directional mutation pressure, though not *per se* strand-biased, can lead to asymmetries between coding and transcribed strand. This holds true even when selection of specific codons and amino-acids is not accounted for. This is more prominent at the first, second, and the third-tf sites, where such asymmetries are inherently imposed by the genetic code. Our findings imply that, given the overall GC content of the genome, the partition of codons into synonymous groups provides the ground for an *a priori* estimation of codon site-specific skews. Such skews correspond to an uneven baseline distribution of nucleotides among different codon sites, which should be considered when estimating the expected number of substitutions per site in comparative genomic analysis.

As discussed earlier and shown in Figure 4a, in third-ff sites strand asymmetries exhibit weak correlations with codon usage. Subsequently, we address the role of specific molecular systems, which may induce replication- and transcription-associated biases, in shaping third-ff asymmetries.

3.4 Transcription-coupled repair and PolIII α-subunit isoforms induce whole-genome mutational biases

Intra-strand asymmetries provide valuable insight into molecular processes with distinct modes of action on each DNA strand. We analyse the distribution of skews in our genomic dataset to infer substitution asymmetries related to two such processes, namely transcription-coupled repair (TCR) and DNA replication. In the following analysis, all comparisons of skew distributions are performed using the two-tailed Wilcoxon's rank sum test.

Bacteria with active TCR possess homologs of the *mfd* gene, which codes for the transcription-repair coupling factor (TRCF).[67] To assess the effect of TCR on mutational biases, we compare the distribution of transcription-associated (Trs) skews in relation to the existence of the *mfd* gene (TCR$^+$: *mfd* present, TCR$^-$: *mfd* absent). As seen in Table 1A, the TCR$^+$ group shows a significantly different ${S_{\mathrm{Trs}}}^{\text{A-T}}$ distribution compared with the TCR$^-$ one (P-value $< 10^{-5}$). Moreover, three out of six relative abundance skews (${P_{\mathrm{Trs}}}^{\text{GG-CC}}$, ${P_{\mathrm{Trs}}}^{\text{AA-TT}}$, and ${P_{\mathrm{Trs}}}^{\text{AC-GT}}$) strongly correlate with the presence of *mfd* (P-value $< 10^{-7}$), while there is also a moderate differentiation of ${P_{\mathrm{Trs}}}^{\text{GA-TC}}$ distribution in TCR$^+$ and TCR$^-$ bacteria (P-value < 0.05).

TCR specifically targets lesions on the transcribed strand,[68] enhancing the bias of TAM, such as C→T transitions, towards the coding strand.[26,69,70] Skew analysis captures this strand-bias, with the distribution of ${S_{\mathrm{Trs}}}^{\text{A-T}}(\mathrm{TCR}^+)$ being shifted to the left of ${S_{\mathrm{Trs}}}^{\text{A-T}}(\mathrm{TCR}^-)$ distribution, in accordance with an excess of C→T over G→A substitutions on the coding strand. However, there is no correlation of ${S_{\mathrm{Trs}}}^{\text{G-C}}$ with TCR (P-value $= 0.239$), indicating that this strand-bias does not originate solely from C→T asymmetries, in line with [21]. The rates of TAM are known to be context-dependent, with C→T being greatly accelerated when C occurs in pyrimidine dimers.[71] This may lead to the observed correlation of *mfd* with ${P_{\mathrm{Trs}}}^{\text{GG-CC}}$ and possibly with ${P_{\mathrm{Trs}}}^{\text{AA-TT}}$. However, as seen in the case of ${P_{\mathrm{Trs}}}^{\text{AC-GT}}$, *mfd* also strongly correlates with relative abundance skews of dinucleotides that cannot form pyrimidine dimers. Thus, the efficiency of TCR may itself depend on sequence context.

We extend our analysis to replication-induced substitution asymmetries.[9] In bacteria, both DNA strands are synthesized by the α-subunit dimer of DNA PolIII.[72] We partition our genomic dataset into three groups, according to the dimeric combination of their α-subunit isoforms: dnaE (homodimer of DnaE1), dnaE2 (heterodimer of DnaE1 with DnaE2) and polC (heterodimer of PolC with either DnaE1 or DnaE3).[62] Note that the polC group is comprised of all Firmicutes and Tenericutes of our collection.

In a previous study[73] no correlation was found between skews and α-subunit isoforms. There, however, bacteria were grouped in relation to the existence of *pol*C only, while *dna*E2 was not accounted for. Table 1B shows that replication-associated (Rep) skews are strongly correlated to the grouping scheme we employ. Both $\mathrm{S}_{\mathrm{Rep}}^{\mathrm{A-T}}$ and $\mathrm{S}_{\mathrm{Rep}}^{\mathrm{G-C}}$ are distinctively distributed among the three groups (P-value $\ll 10^{-4}$). Relative abundance skews also significantly correlate with the type of α-subunit dimers (i.e. ${P_{\mathrm{Trs}}}^{\text{AA-TT}}$, ${P_{\mathrm{Trs}}}^{\text{GG-CC}}$, ${P_{\mathrm{Trs}}}^{\text{AC-GT}}$). Moreover, PolC largely determines the sign of $\mathrm{S}_{\mathrm{Rep}}^{\mathrm{A-T}}$ (positive; see also[23]), ${P_{\mathrm{Rep}}}^{\text{GG-CC}}$ and ${P_{\mathrm{Rep}}}^{\text{AA-TT}}$ (both negative). The other skews are of the same sign in all three groups.

When present, PolC replicates the leading strand,[74] while DnaE2 performs SOS-induced translesion synthesis.[75,76] Both PolC and DnaE2 form heterodimers whose components possess distinct nucleotide incorporation and proofreading activities. Our results reveal that such asymmetries of α-subunit induce specific strand-biased substitutions on a genome-wide scale. In many cases these biases are neighbour-dependent, as indicated by relative abundance skews ($\mathrm{P}_{\mathrm{Rep}}^{\mathrm{AA-TT}}$, $\mathrm{P}_{\mathrm{Rep}}^{\mathrm{GG-CC}}$, $\mathrm{P}_{\mathrm{Rep}}^{\mathrm{AC-GT}}$). In the dnaE group the replication of both strands is catalyzed by the same α-subunit isoform. Thus, the observed skews should be attributed to the intrinsic asymmetries of the replication fork.[7]

Table 1. Analysis of skew distributions in relation to TCR and PolIII α-subunit

A. Transcription-associated (Trs) skews	$S_{\mathrm{Trs}}^{\text{A-T}}$	$S_{\mathrm{Trs}}^{\text{G-C}}$	$P_{\mathrm{Trs}}^{\text{AG-CT}}$	$P_{\mathrm{Trs}}^{\text{GA-TC}}$	$P_{\mathrm{Trs}}^{\text{GG-CC}}$	$P_{\mathrm{Trs}}^{\text{AA-TT}}$	$P_{\mathrm{Trs}}^{\text{AC-GT}}$	$P_{\mathrm{Trs}}^{\text{CA-TG}}$
TCR^{-}	−0.0642	−0.00305	−0.178	−0.0193	−0.205	0.0333	−0.0806	0.0604
TCR^{+}	−0.156	−0.0412	−0.191	0.0358	0.173	−0.186	0.313	0.0912
***P*-value**								
TCR^{-} TCR^{+}	***	—	—	*	***	***	***	—
B. Replication-associated (Rep) skews	$S_{\mathrm{Rep}}^{\text{A-T}}$	$S_{\mathrm{Rep}}^{\text{G-C}}$	$P_{\mathrm{Rep}}^{\text{AG-CT}}$	$P_{\mathrm{Rep}}^{\text{GA-TC}}$	$P_{\mathrm{Rep}}^{\text{GG-CC}}$	$P_{\mathrm{Rep}}^{\text{AA-TT}}$	$P_{\mathrm{Rep}}^{\text{AC-GT}}$	$P_{\mathrm{Rep}}^{\text{CA-TG}}$
dnaE	−0.0357	0.0806	0.00521	0.0137	0.000299	0.0161	−0.0309	0.00614
dnaE2	−0.07	0.0386	0.017	0.0152	0.0168	0.0371	−0.0532	0.0116
polC	0.0241	0.132	0.000189	0.012	−0.0222	−0.029	−0.0376	0.0203
***P*-value**								
dnaE dnaE2	***	***	*	—	***	***	***	—
dnaE polC	***	***	–	—	**	***	—	**
polC dnaE2	***	***	***	—	***	***	**	**

For each group of genomes the median skew values are provided. We pairwise compare the skew distributions among the considered genome groups. The *p*-values for the two-tailed Wilcoxon's rank sum tests conducted denote the statistical significance of their difference. —, *P*-value ≥ 0.05; *,0.05 > *P*-value ≥ 0.01; **, 0.01 > *P*-value ≥ 0.001; ***, *P*-value < 0.001.

Analysis of skew distribution can also provide useful information about molecular processes, besides replication and transcription, which are not yet determined to act in a strand-specific manner. Since we detect no significant correlation between $S_{\mathrm{Trs}}^{\text{G-C}}$ and TCR, we compare the distribution of $S_{\mathrm{Trs}}^{\text{G-C}}$ in relation to other DNA repair systems (Supplementary Tables S5 and S6). We consider direct reversal of DNA damage, base excision repair, nucleotide excision repair, mismatch repair and recombination repair pathways. In 7 out of 20 cases studied, we detect a statistically significant correlation (adjusted P-value < 0.01, see Supplementary Table S7) between $S_{\mathrm{Trs}}^{\text{G-C}}$ and specific DNA repair systems. The presence of these systems is linked to a shift of the $S_{\mathrm{Trs}}^{\text{G-C}}$ distribution from lower to higher (signed) values, in a remarkably uniform way, indicating an inversion of mutational bias when the corresponding repair systems are active.

4. Conclusions

Strand-asymmetries are deeply rooted into DNA evolution. To our knowledge, this is the first work to provide solid evidence that nearest-neighbour preferences are a strand-biased feature of bacterial genomes. The relative abundance skews, which we herein introduce, quantify these biases to assess strand-asymmetries in context-dependent substitutions.

We show that skews along CDSs are distinct among different species but similar between genes of the same organism. Moreover, relative abundance skews are more strongly correlated with species phylogeny, even when compared with genomic signatures. Hence, CDS-associated substitution asymmetries do not emanate from evolutionary trends which are universal across bacteria, but are species-specific.

Site-specific and context-dependent skew analysis facilitates the inquiry about the origin of certain substitution asymmetries. When CDSs are subjected to mutational bias towards a particular GC content, the structure of the genetic code imposes site-specific strand-biases on nucleotide substitutions. This leads to baseline position-dependent strand-asymmetries that should be taken into account when considering the effect of directional mutation and purifying selection on biases in CDSs' composition.

Furthermore, the molecular machinery of transcription and replication are implicated in strand-asymmetries. TCR results in substitution biases towards coding strands, and our skew analysis efficiently captures these asymmetries. The isoforms of the PolIII α-subunit introduce replication errors at different rates on each DNA strand. Our results indicate that these biases are prevalent enough to induce whole-genome substitution asymmetries, both context-dependent and -independent.

Additional Figures and Tables are available in the following link:

http://bio.demokritos.gr/index.php?option=com_content&view=article&id=165&Itemid=167&lang=el

Acknowledgements

This work was supported by a scholarship from National Center for Scientific Research 'Demokritos'. We thank Prof Nikolaos Glykos, Dr George Giannakopoulos, Dr Panagiota Tsitoura, and Dr Dimitris Polychronopoulos for helpful discussion.

Conflict of interest

None declared.

Supplementary data

Supplementary data are available at www.dnaresearch.oxfordjournals.org.

Funding

K.A.K. acknowledges support through a scholarship awarded by NCSR "Demokritos". Funding to pay the Open Access publication charges for this article was provided by NCSR "Demokritos".

References

1. Lobry, J.R. and Lobry, 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.
2. Sueoka, N. 1995, Intrastrand parity rules of DNA base composition and usage biases of synonymous codons. *J. Mol. Evol.*, **40**, 318–25.
3. Lobry, J.R. 1995, Properties of a general model of DNA evolution under no-strand-bias conditions. *J. Mol. Evol.*, **40**, 326–30.
4. Prabhu, V.V. 1993, Symmetry observations in long nucleotide sequences. *Nucleic Acids Res.*, **21**, 2797–800.
5. Bell, S.J. and Forsdyke, D.R. 1999, Accounting units in DNA. *J. Theor. Biol.*, **197**, 51–61.
6. Mitchell, D. and Bridge, R. 2006, A test of Chargaff's second rule. *Biochem. Biophys. Res. Commun.*, **340**, 90–4.
7. Lobry, J.R. 1996, Asymmetric substitution patterns in the two DNA strands of bacteria. *Mol. Biol. Evol.*, **13**, 660–5.
8. Lobry, J.R. and Sueoka, N. 2002, Asymmetric directional mutation pressures in bacteria. *Genome Biol.*, **3**, RESEARCH0058.
9. Rocha, E.P., Danchin, A., and Viari, A. 1999, Universal replication biases in bacteria. *Mol. Microbiol.*, **32**, 11–6.
10. Mrázek, J. and Karlin, S. 1998, Strand compositional asymmetry in bacterial and large viral genomes. *Proc. Natl. Acad. Sci. USA*, **95**, 3720–5.
11. Nikolaou, C. and Almirantis, Y. 2005, A study on the correlation of nucleotide skews and the positioning of the origin of replication: different modes of replication in bacterial species. *Nucleic Acids Res.*, **33**, 6816–22.
12. Tillier, E.R. and Collins, R. A. 2000, The contributions of replication orientation, gene direction, and signal sequences to base-composition asymmetries in bacterial genomes. *J. Mol. Evol.*, **50**, 249–57.
13. Rocha, E.P.C., Touchon, M., and Feil, E.J. 2006, Similar compositional biases are caused by very different mutational effects. *Genome Res.*, **16**, 1537–47.
14. Morton, R.A., and Morton, B.R. 2007, Separating the effects of mutation and selection in producing DNA skew in bacterial chromosomes. *BMC Genomics*, **8**, 369.
15. Frank, A.C. and Lobry, J.R. 1999, Asymmetric substitution patterns: a review of possible underlying mutational or selective mechanisms. *Gene*, **238**, 65–77.
16. 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.
17. Nikolaou, C. and Almirantis, Y. 2006, Deviations from Chargaff's second parity rule in organellar DNA Insights into the evolution of organellar genomes. *Gene*, **381**, 34–41.
18. Rocha, E.P.C. 2008, The organization of the bacterial genome. *Annu. Rev. Genet.*, **42**, 211–33.
19. 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.
20. Danchin, A. 2003, Genomes and evolution. *Curr. Issues Mol. Biol.*, **5**, 37–42.
21. Rocha, E.P. and Danchin, A. 2001, Ongoing evolution of strand composition in bacterial genomes. *Mol. Biol. Evol.*, **18**, 1789–99.
22. Necşulea, A. and Lobry, J.R. 2007, A new method for assessing the effect of replication on DNA base composition asymmetry. *Mol. Biol. Evol.*, **24**, 2169–79.
23. Worning, P., Jensen, L.J., Hallin, P.F., Staerfeldt, H.-H., and Ussery, D.W. 2006, Origin of replication in circular prokaryotic chromosomes. *Environ. Microbiol.*, **8**, 353–61.
24. Bell, S.J. and Forsdyke, D.R. 1999, Deviations from Chargaff's second parity rule correlate with direction of transcription. *J. Theor. Biol.*, **197**, 63–76.
25. Lopez, P. and Philippe, H. 2001, Composition strand asymmetries in prokaryotic genomes: mutational bias and biased gene orientation. *C. R. Acad. Sci. III.*, **324**, 201–8.
26. Francino, M.P., Chao, L., Riley, M.A., and Ochman, H. 1996, Asymmetries generated by transcription-coupled repair in enterobacterial genes. *Science*, **272**, 107–9.
27. 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.
28. Gouy, M. and Gautier, C. 1982, Codon usage in bacteria: correlation with gene expressivity. *Nucleic Acids Res.*, **10**, 7055–74.
29. Bulmer, M. 1991, The selection-mutation-drift theory of synonymous codon usage. *Genetics*, **129**, 897–907.
30. Xia, X. 1998, How optimized is the translational machinery in Escherichia coli, Salmonella typhimurium and Saccharomyces cerevisiae? *Genetics*, **149**, 37–44.
31. Sueoka, N. 1961, Compositional correlation between deoxyribonucleic acid and protein. *Cold Spring Harb. Symp. Quant. Biol.*, **26**, 35–43.
32. D'Onofrio, G., Jabbari, K., Musto, H., and Bernardi, G. 1999, The correlation of protein hydropathy with the base composition of coding sequences. *Gene*, **238**, 3–14.
33. Salzberg, S.L., Salzberg, A.J., Kerlavage, A.R., and Tomb, J.F. 1998, Skewed oligomers and origins of replication. *Gene*, **217**, 57–67.
34. Mascher, M., Schubert, I., Scholz, U., and Friedel, S. 2013, Patterns of nucleotide asymmetries in plant and animal genomes. *Biosystems*, **111**, 181–9.
35. Karlin, S. 1998, Global dinucleotide signatures and analysis of genomic heterogeneity. *Curr. Opin. Microbiol.*, **1**, 598–610.
36. Josse, J., Kaiser, A.D., and Kornberg, A. 1961, Enzymatic synthesis of deoxyribonucleic acid. VIII. Frequencies of nearest neighbor base sequences in deoxyribonucleic acid. *J. Biol. Chem.*, **236**, 864–75.
37. Nussinov, R. 1981, Nearest neighbor nucleotide patterns. Structural and biological implications. *J. Biol. Chem.*, **256**, 8458–62.
38. Nussinov, R. 1984, Strong doublet preferences in nucleotide sequences and DNA geometry. *J. Mol. Evol.*, **20**, 111–9.
39. Shioiri, C. and Takahata, N. 2001, Skew of mononucleotide frequencies, relative abundance of dinucleotides, and DNA strand asymmetry. *J. Mol. Evol.*, **53**, 364–76.
40. Baisnée, P.-F., Hampson, S., and Baldi, P. 2002, Why are complementary DNA strands symmetric? *Bioinformatics*, **18**, 1021–33.
41. Gao, F., Luo, H., and Zhang, C.-T. 2013, DoriC 5.0: an updated database of oriC regions in both bacterial and archaeal genomes. *Nucleic Acids Res.*, **41**, D90–3.
42. Mao, X., Zhang, H., Yin, Y., and Xu, Y. 2012, The percentage of bacterial genes on leading versus lagging strands is influenced by multiple balancing forces. *Nucleic Acids Res.*, **40**, 8210–8.
43. Saha, S.K., Goswami, A., and Dutta, C. 2014, Association of purine asymmetry, strand-biased gene distribution and PolC within Firmicutes and beyond: a new appraisal. *BMC Genomics*, **15**, 430.
44. 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.
45. 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.
46. Zeileis, A., Leisch, F., Hornik, K., and Kleiber, C. 2002, strucchange t an R package for testing for structural change in linear regression models. *J. Stat. Softw.*, **7**, 1–38.
47. Kullback, S., and Leibler, R.A. 1951, On information and sufficiency. *Ann. Math. Stat.*, **22**, 79–86.
48. 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.
49. Federhen, S. 2012, The NCBI Taxonomy database. *Nucleic Acids Res.*, **40**, D136–43.
50. Karlin, S. and Burge, C. 1995, Dinucleotide relative abundance extremes: a genomic signature. *Trends Genet.*, **11**, 283–90.
51. Campbell, A., Mrázek, J., and Karlin, S. 1999, Genome signature comparisons among prokaryote, plasmid, and mitochondrial DNA. *Proc. Natl. Acad. Sci. U. S. A.*, **96**, 9184–9.
52. Powell, S., Forslund, K., Szklarczyk, D., et al. 2014, eggNOG v4.0: nested orthology inference across 3686 organisms. *Nucleic Acids Res.*, **42**, D231–9.

53. Arndt, P.F.k, and Hwa, T. 2005, Identification and measurement of neighbor-dependent nucleotide substitution processes. *Bioinformatics*, **21**, 2322–8.
54. Chamary, J.-V. and Hurst, L. D. 2004, Similar rates but different modes of sequence evolution in introns and at exonic silent sites in rodents: evidence for selectively driven codon usage. *Mol. Biol. Evol.*, **21**, 1014–23.
55. 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.
56. Grigoriev, A. 1998, Analyzing genomes with cumulative skew diagrams. *Nucleic Acids Res.*, **26**, 2286–90.
57. Francino, M.P. and Ochman, H. 1997, Strand asymmetries in DNA evolution. *Trends Genet.*, **13**, 240–5.
58. Karlin, S. and Mrázek, J. 1997, Compositional differences within and between eukaryotic genomes. *Proc. Natl. Acad. Sci. USA*, **94**, 10227–32.
59. Becker, R.A., Chambers, J.M., and Wilks, A. R. 1988, The new S language: a programming environment for data analysis and graphics. Pacific Grove, CA, USA: Wadsworth & Brooks/Cole Advanced Books & Software. ISBN 0-534-09192-X.
60. Knight, R.D., Freeland, S.J., and Landweber, L.F. 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.
61. Muto, A. and Osawa, S. 1987, The guanine and cytosine content of genomic DNA and bacterial evolution. *Proc. Natl. Acad. Sci. USA*, **84**, 166–9.
62. Hu, J., Zhao, X., Zhang, Z., and Yu, J. 2007, Compositional dynamics of guanine and cytosine content in prokaryotic genomes. *Res. Microbiol.*, **158**, 363–70.
63. Sorimachi, K. and Okayasu, T. 2008, Codon evolution is governed by linear formulas. *Amino Acids*, **34**, 661–8.
64. 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.
65. Yu, J. 2007, A content-centric organization of the genetic code. *Genomics Proteomics Bioinformatics*, **5**, 1–6.
66. Zhang, Z. and Yu, J. 2010, Modeling compositional dynamics based on GC and purine contents of protein-coding sequences. *Biol. Direct*, **5**, 63.
67. Selby, C.P., Witkin, E.M. and Sancar, A. 1991, Escherichia coli mfd mutant deficient in “mutation frequency decline” lacks strand-specific repair: in vitro complementation with purified coupling factor. *Proc. Natl. Acad. Sci. USA*, **88**, 11574–8.
68. Bockrath, R.C., and Palmer, J. E. 1977, Differential repair of premutational UV-lesions at tRNA genes in E. coli. *Mol. Gen. Genet.*, **156**, 133–40.
69. Oller, A.R., Fijalkowska, I.J., Dunn, R.L., and Schaaper, R.M. 1992, Transcription-repair coupling determines the strandedness of ultraviolet mutagenesis in Escherichia coli. *Proc. Natl. Acad. Sci. USA*, **89**, 11036–40.
70. Deaconescu, A.M. 2013, RNA polymerase between lesion bypass and DNA repair. *Cell. Mol. Life Sci.*, **70**, 4495–509.
71. Peng, W. and Shaw, B.R. 1996, Accelerated deamination of cytosine residues in UV-induced cyclobutane pyrimidine dimers leads to CC–>TT transitions. *Biochemistry*, **35**, 10172–81.
72. Zhao, X.-Q., Hu, J.-F., and Yu, J. 2006, Comparative analysis of eubacterial DNA polymerase III alpha subunits. *Genomics Proteomics Bioinformatics*, **4**, 203–11.
73. Rocha, E.P.C. 2002, Is there a role for replication fork asymmetry in the distribution of genes in bacterial genomes? *Trends Microbiol.*, **10**, 393–5.
74. Dervyn, E., Suski, C., Daniel, R., et al. 2001, Two essential DNA polymerases at the bacterial replication fork. *Science*, **294**, 1716–9.
75. Boshoff, H.I.M., Reed, M.B., Barry, C.E., and Mizrahi, V. 2003, DnaE2 polymerase contributes to in vivo survival and the emergence of drug resistance in Mycobacterium tuberculosis. *Cell*, **113**, 183–93.
76. Galhardo, R.S., Rocha, R.P., Marques, M.V. and Menck, C.F.M. 2005, An SOS-regulated operon involved in damage-inducible mutagenesis in Caulobacter crescentus. *Nucleic Acids Res.*, **33**, 2603–14.