A global perspective of codon usage

Bohdan B. Khomtchouk,[1] Claes Wahlestedt,[1] and Wolfgang Nonner[2,*]

[1] *Department of Psychiatry and Behavioral Sciences, University of Miami Miller School of Medicine,Miami, Florida 33101, USA*

[2] *Department of Physiology and Biophysics, University of Miami Miller School of Medicine, Miami, Florida 33101, USA*

Diversity of codon usage across 2730 genomes from different taxa is analyzed with non-parametric methods. Usages range from near-uniform recruitment of different codons (limited by amino acid stoichiometry of encoded proteomes) to two distinct, non-uniform recruitments of nearly exclusive subsets of different codons. Intermediate usages are described as a variation from one distinct non-uniform usage to the other that involves joint, proportionate changes in the frequencies of most different codons. The two distinct non-uniform usages define two distinct genetic codes of which one employs A/U, and the other C/G, as third bases.

INTRODUCTION

The standard genetic code specifies each but two of the 20 different amino acids by two to six synonymous codons. Usage of different synonyms thus can in principle 'encrypt' information beyond that of protein primary structure [1]. Indeed, different codons specifying the same amino acid occur with diverse frequencies within genes, within genomes, and across genomes. 'Cryptic' information contained in the DNA of genes (and the transcribed mRNA) could in a variety of ways direct mRNA processing and/or translation [1], possibilities that have motivated investigations of the evolution and consequences of 'codon usage bias' (reviewed by [2–4]).

Molecular "restoration studies" of extinct forms of life, as originally proposed by Pauling and Zuckerkandl [5] and now pursued in the reconstruction of phylogenetic trees, are based on the primary sequence information encoded in DNA [6, 7]. Forms of life close to the base of the tree have been inferred from homologous proteins found in free-living bacteria and archaea [8, 9]. On the other hand, the origin of the genetic code that links genome and proteome is still elusive, much as it was when the code was discovered [10, 11].

In this study we apply Principal Component Analysis to the genomic codon frequencies of the taxa represented in the CUTG database to create a global view of codon usage. In this view, the variation of codon usage across different genomes appears as the manifestation of an evolutionary process that transforms genomes from two antipodal usages of the genetic code toward a balanced usage of the code. Genomes thought to be close to those at the base of the phylogenetic tree are among the most diverse in their codon usages.

* Correspondence: wnonner@med.miami.edu

METHODS

We analyze genomes that are represented in the CUTG database with codon counts $\sum_{i=1}^{64} c_i \geq 10^4$ (database files 'gbxxx.spsum.txt' where xxx = bct, inv, mam, phg, pln, pri, rod, vrl, vrt). Entries for mitochondrial or other organelle genomes are excluded. The count c_i of occurrences of a different codon i ($1 \leq i \leq 64$) in an analyzed genome is converted to the frequency

$$y_i = \frac{c_i}{\sum_{i=1}^{64} c_i} \tag{1}$$

On the basis of frequencies we compute non-parametric measures to quantify variations of codon usage within and across genomes. To describe codon usage within a particular genome, we take the frequencies of the 64 different codons as the coordinates of that genome in a 64-dimensional Euclidean space. The Euclidean distance between that point and a reference point representing a genome using the 64 different codons with equal frequencies then defines a measure of frequency variation within the genome, which we call 'bias' B:

$$B = \sqrt{\sum_{i=1}^{64} \left(y_i - \frac{1}{64} \right)^2} \tag{2}$$

In a hypothetical genome in which all different codons occur with equal frequencies, bias $B = 0$. Real genomes will have bias $B > 0$: even though synonymous codons encoding for the same amino acid might be used equally, the amino acid stoichiometry of an encoded proteome requires non-uniformity in the usage of different codons.

To describe the variation of usage of a different codon i across a set of genomes $k = 1, ..., N$, we define the variance

$$V_i = \frac{1}{N-1} \sum_{k=1}^{N} \left(y_{ik} - \bar{y}_i \right)^2 \tag{3}$$

where $\bar{y}_i = \frac{1}{N} \sum_{k=1}^{N} y_{ik}$ is the average frequency of the different codon i in the set of N genomes. As a global

measure of how usage of all different codons varies across the set of genomes, we define 'diversity' of codon usage as the total variance

$$D = \sum_{i=1}^{64} V_i \tag{4}$$

Diversity across a set of genomes will be zero if all genomes have zero bias. Diversity will also be zero if bias is not zero but all genomes k of the set use particular different codons i with the same frequencies, $y_{ik} = y_i$.

In order to detect dependencies among different codons in the variation in codon usage, we describe diversity more specifically by the 64×64 covariance matrix with the elements

$$c_{ij} = \frac{1}{N-1} \sum_{k=1}^{N} (y_{ik} - \bar{y}_i)(y_{jk} - \bar{y}_j) \tag{5}$$

The covariance matrix is subjected to Singular Value Decomposition (SVD) to determine its eigenvalues and eigenvectors for Principal Component Analysis (PCA). The decomposition produces a set of orthonormal base vectors (the eigenvectors) that define a new Cartesian axis system that is rotated with respect to the original frequency axis system. The variations occurring in the directions of these new axes contribute to the total diversity of codon usage in amounts given by the eigenvalue associated with each axis. The rotation of the axis system maximizes the fraction of total diversity associated with each axis.

As will be described in Results, the eigenvector associated with the largest eigenvalue accounts for much of the observed total diversity of codon usage. We therefore focus further analysis on that first principal component. We compute the projection of the observed codon frequencies y_i of genome k onto the axis defined by the first eigenvector, (v_i^1):

$$z_k = \sum_{i=1}^{64} (y_{ik} - \bar{y}_i)\, v_i^1 \tag{6}$$

To the extent that the first principal component accounts for the diversity of the genomes, the projection z_k describes the codon usage of an entire genome k by a single coordinate, which we will use to generate a map of codon usage. Codon frequencies of the genome are reconstructed using only the first principal component:

$$y_{ik}^1 = \bar{y}_i + z_k v_i^1 \tag{7}$$

These frequencies are compared with the observed original codon frequencies y_{ik} to assess the completeness of the reconstruction.

In order to distinguish between variation due to use of alternate synonymous codons and variation due to use of alternate amino acids we perform PCA also on the amino acid frequencies derived from the genomes. Those 20 frequencies and that of the stop signal are obtained as partial sums of codon frequencies taken over the groups of codons encoding the different amino acids and the stop signal. In this translation, we use the standard genetic code, ignoring variants known for some organisms.

RESULTS

Frequencies of codon occurrence in gDNA of *Homo sapiens* and *Arabidopsis thaliana* extracted from the CUTG database (see Methods) are shown in Fig. 1A. The human frequencies have been arranged into a descending series (*red color*). The frequencies of the different codons of the plant when arranged to follow the order of the different codons in the human series do not form a descending series (*blue color*). When the codon frequencies of the plant are by themselves arranged into a descending series, the resulting curve is similar to that for the ranked human series (*line*). Particular different codons, however, are used with variable preferences in the human and plant genomes, with balance tipping more or less stochastically one way or the other across the 64 different codons.

Fig. 1B presents two further sample genomes, of the bacteria *Streptomyces griseus* (*red color*) and *Clostridium tetani E88* (*blue color*), analyzed like the genomes of Fig. 1A. Each bacterial genome uses different codons to much different degrees, but when codon frequencies are individually arranged in descending order, similar frequency/rank curves emerge. On the other hand, the two bacterial genomes encode proteins with nearly exclusive subsets of the 64 different codons.

The four genomes in Fig. 1A,B have been chosen in hindsight of an analysis that is described in the following. They are examples for the variety of codon usages that we found in 2730 genomes from all sections of the CUTG database (see Methods). We will return to the genomes of Fig. 1A,B to illustrate results of that analysis.

We assess non-uniformity of codon usage observed within an individual genome using as a non-parametric measure the 'bias' B defined by eq. 2 of Methods. Bias measures how much the observed frequencies of a genome diverge from those of a hypothetical genome in which all different codons are used with the same frequency. The histogram in Fig. 1C, summarizing the values of B for 2730 genomes of the CUTG database (*gray shade*), peaks near the lower end of the bias range. (The genomes of Fig. 1A,B locate to the bins marked by *asterisks*.) The leftmost four bins comprise 1460 genomes – more than half of the genomes locate to the lower third of the binned range of bias. The majority of genomes use different codons much the ways of the human and *Arabidopsis* genomes (Fig. 1A), rather than the ways of the bacte-

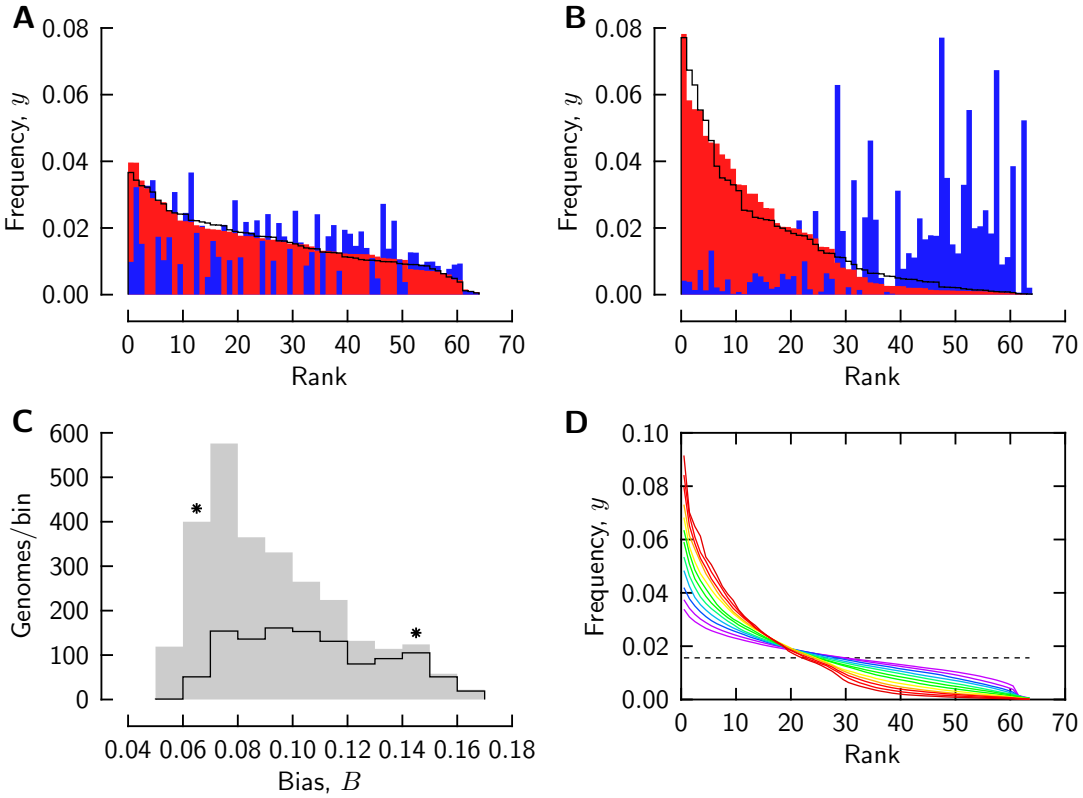


FIG. 1. *Biased usage of different codons in genomes.* Codon frequencies of *Homo sapiens* (A, red color), *Arabidopsis thaliana* (A, blue color), *Streptomyces griseus* (B, red color), and *Clostridium tetani E88* (B, blue color), arranged as described in the text. C: histograms of bias B for 2730 genomes from the CUTG database (gray shaded) and of the bacterial genomes in that set (line). Asterisks mark the bins to which the genomes in A,B locate. D: frequency/rank curves averaged over the genomes of each bin in C (with bias increasing from purple to red).

rial genomes of Fig. 1B. Overall, bacterial genomes use different codons with a wide range of bias, as is apparent when their codon frequencies are binned separately (*line* in Fig. 1C). Although bacteria as the largest CUTG database section contribute with 1134 genomes to the analyzed 2730 genomes, the peak of the overall distribution of bias is mainly due to non-bacterial genomes.

Since frequency/rank curves of genomes have been described or modeled in the past for small sets of genomes [12–14], it is of interest to inspect frequency/rank curves from a large genome set as analyzed here. Each curve in Fig. 1D is the average of the frequency/rank curves for the genomes of one bin in the histogram of Fig. 1C. The frequency corresponding to uniform usage of the different codons (1/64) is marked by the *dashed line*. There is a continuous variation of frequency/rank relation from genomes with weak (*purple lines*) to genomes with strong bias (*red lines*). A new feature, not described so far in our knowledge, is a cascading decay of frequency with rank that appears in genomes with strong bias.

Frequency/rank characteristics give an incomplete account of codon usage: genomes with similar frequency/rank characteristics actually differ in how much they use particular different codons (Fig. 1A,B). Here, we focus on the variation of codon usage *across* different genomes for further information.

For assessing codon usage across genomes, we define a non-parametric measure of 'diversity', D (eq. 4 in Methods). Diversity quantifies the frequency variation of each one of the different codons of a set of genomes with respect to the codon's average frequency in the set. We compute diversity D for the genomes in each bin of the bias histogram (Fig. 1C) and plot it versus bias B (Fig. 2A, *gray shade*). Diversity increases by about an order of magnitude with increasing bias: usage of individual different codons becomes more diverse across genomes as bias within genomes increases. That observation, already evident in the examples of Fig. 1A,B, is general for the 2730 analyzed genomes. Note that diversity among genomes is not a *necessary* consequence of genomes using different codons with bias (see eq. 3, 4).

Diversity of codon usage across different organisms is naturally restricted because their genes encode homologous proteins. That restriction is expected to generate dependencies among the frequencies of different codons that emerge when genomes are compared. For instance, a codon encoding for one amino acid might be swapped for a codon encoding for a neutral substitute of that amino acid, or the codon might be swapped for another synonym of that amino acid. To assess dependencies in the

diversity of codon usage, we apply Principal Component Analysis (PCA) as decribed in Methods. Here, PCA transforms the coordinate system of codon frequencies into a new set of Cartesian axes that are successively constructed to receive projections of as much diversity as possible. The axes are given by the eigenvectors of a covariance matrix (eq. 5), and the associated diversity projections are given by the eigenvalues. The linear algebraic operations computing PCA yield unique results.

The lines in Fig. 2A represent the first and second eigenvalues obtained by PCA. Each genome set locating to one of the bins of the bias histogram (Fig. 1C) is analyzed separately, so that there are different eigenvalues for the successive bins. The contribution of the first principal component (*black line*) follows the total diversity (*gray shaded*) as the bias within genomes increases, such that the first principal component accounts for an increasing fraction of total diversity. In contrast, the second principal component (*gray line*) contributes a smaller part of the diversity and remains small as bias increases. PCA yields a total of 64 eigenvalues, of which 62 are not shown in Fig. 2A. The combined contributions of the second to 64th principal components are given by the difference between the total diversity (*gray shaded*) and the contribution of the first eigenvalue (*black line*). The first principal component reports here a variation of codon usage that accounts for most of the diversity observed across the genomes of bins 5–12. These PCA results are universal because they are derived from a set comprising genomes from all sections of the CUTG database.

Fig. 2B summarizes first eigenvectors from PCA analyses of codon frequencies. An eigenvector is represented by the (unit) vector component associated with each different codon (the codon axis is divided into groups of codons that code for the indicated amino acid; codon order within the synonymous groups is spelled out in Table 1). A global analysis, comprising all 2730 genomes binned in Fig. 1C, yields the vector represented by the *solid line*. Bin for bin analyses of the genomes in the second to last bin yield eigenvectors whose components fall within the range marked by the *gray bars*. The axis associated with the first principal component defines a joint variation of the frequencies of most codons. Dependency among codon frequencies is spread much wider than that expected from the swapping of synonymous codons for one amino acid, which is not necessarily tied to codon swapping in the synomymous groups of other amino acids. The few different codons for which variation is small include those encoding single-codon amino acids (methionine, tryptophan) and the stop signal. Codons belonging to the same synonymous group undergo extensive and oppositely directed changes of frequency.

The described eigenvector is independently detected in the genomes of the second to last bin although the diversity of the genomes varies substantially among the bins (Fig. 2A). Diversity of codon usage across genomes follows a universal pattern of variation in the frequencies of the majority of different codons.

The covariances studied in PCA imply that the frequency of a different codon is shifted by the average frequency of that codon across the set of genomes. Fig. 2C shows the average codon frequencies computed over all 2730 genomes (*gray shaded*). In many cases the frequencies of synonymous codons have similar values (the hypothetical situation of exactly uniform usage of the synonyms is represented by the *line* in Fig. 2C to assist the eye). The variations described by the first PCA eigenvector then are relative to the codon usage expected from the average frequencies of encoded amino acids and quite uniform recruitment of synonymous codons.

In order to assess dependencies in the variations of amino acid usage, we 'translate' codon frequencies of genomes into amino acid frequencies (using the standard genetic code) and submit the amino acid frequencies to PCA. The results are shown in Fig. 2D-F, side-by-side with the codon-level PCA results. Again, the first principal component describes much of the total diversity, and the first eigenvector is universal across the genomes of the database. We find oppositely directed changes in the frequencies of the individual amino acids within the groups defined by side-chain nature. Fig. 2F shows the average amino acid frequencies about which the variations described by the first eigenvector occur. We will below assess to what extent the first principal components determined by PCA of either amino acid or individual codon frequencies describe the same variations of amino acid frequencies. An analogous analysis, with codons grouped together by side-chain nature of the encoded amino acids, reveals a substantial shift among amino acids with nonpolar and polar side chains (not shown).

A first principal component describing a large fraction of codon diversity provides an opportunity to characterize a genome k by a simple number: the projection of the codon frequency vector onto the first principal axis (eq. 6). That number, z_k, is plotted versus genome bias B_k in the panels of Fig. 3. Each panel summarizes the genomes of the indicated CUTG database section(s). The genomes used as examples in Fig. 1A,B are marked by *filled red circles*. The graphs provide a perspective of genomic codon usage in the form of a map. The individual genomes map to a V-shaped formation, and that formation is consistently observed across taxa.

The external outline of the V-shaped formation is an expression of the fact that diversity across genomes requires a certain amount of bias within the genomes. The absence of genomes from the interior of the formation is an expression of the fact that, for instance, a genome locating to a projection $z \approx 0$ would have codon frequencies close to the mean frequencies plotted in Fig. 2C, which imply a small value of bias in codon usage and would tend to map such a genome toward the junction of the V-shaped formation.

The first principal component of our PCA results describes a pattern of variation of codon usage across a genome set and allows codon usage by a genome of the set to be characterized by a simple number. A sim-

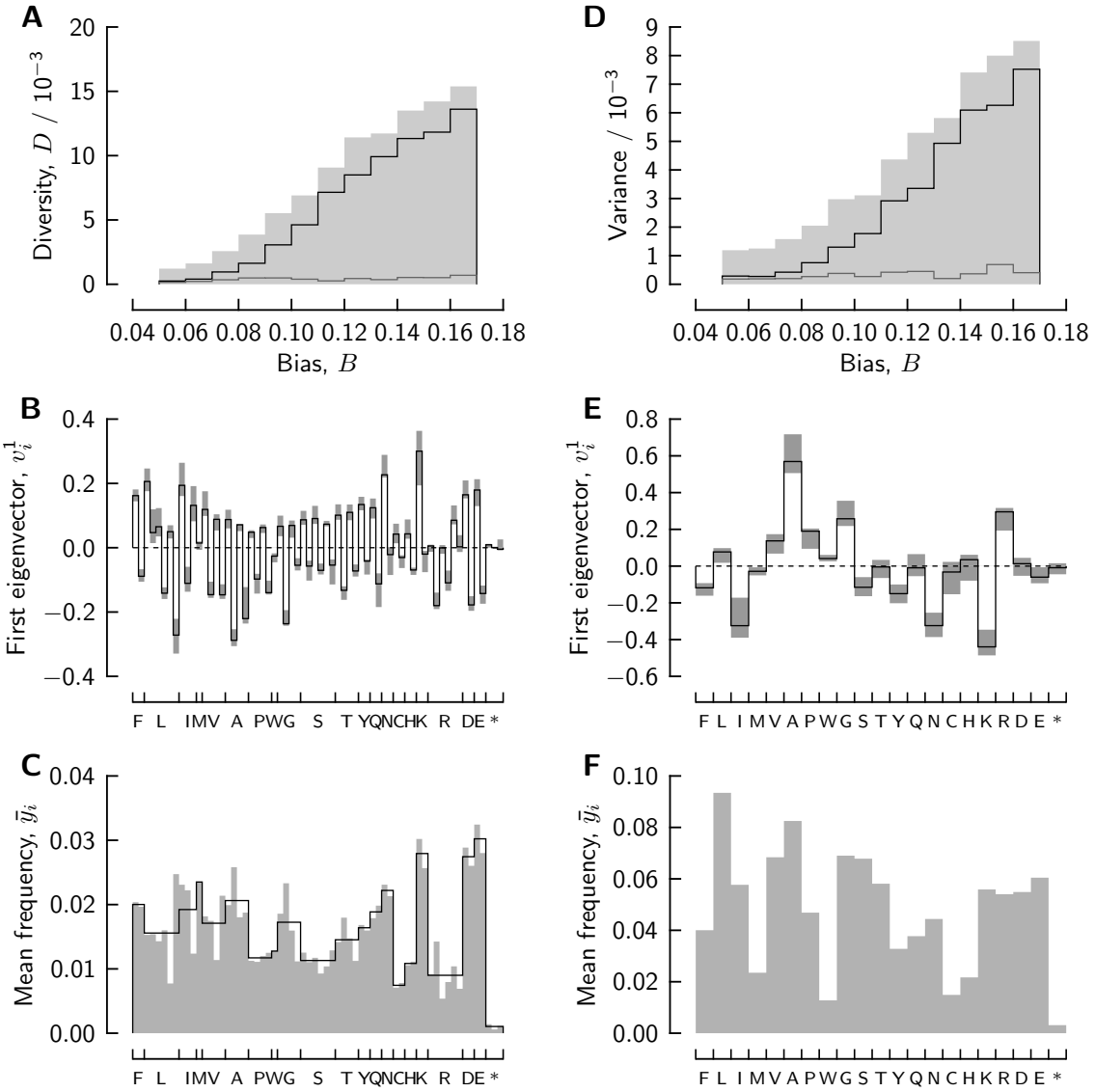


FIG. 2. *Diversity in the usage of different codons across genomes examined by PCA.* A-C: analysis of codon frequencies; D-F: analysis of amino acid frequencies. A,D: total diversity D computed for the genomes of each bin in Fig. 1C; the first and second PCA eigenvalues are shown as black and gray lines. B,E: the first PCA eigenvector components computed over the whole set of 2730 genomes (line), or over the genomes of the second to last individual bins of A,D (the gray bars indicate the range of the eigenvectors of the different bins). C,D mean frequencies computed over the full set of 2730 genomes (gray shaded); the line in C shows hypothetical mean frequencies for uniform usage of synonymous codons.

ple test to verify the accuracy of this characterization is reconstruction of a genome's codon frequencies on the basis of that information (eq. 7 of Methods). Fig. 4A-D show the reconstructed codon frequencies (*lines*) together with the observed frequencies (*gray shaded*) of the four genomes introduced as examples in Fig. 1A,B. The comparison suggests that the partial PCA information used for the reconstructions captures the diverse codon usages of these genomes quite well.

An analogous reconstruction is made for the amino acid frequencies using the PCA results shown in Fig. 2C-E. Fig. 4E-G show the amino acid frequencies derived from the observed codon frequencies (*gray shaded*), the frequencies reconstructed from the PCA of the amino acid frequencies (*black lines*), and the amino acid frequencies derived from the reconstructed codon frequencies as seen in the left-column panels of Fig. 4 (*red lines*). The amino acid frequencies reconstructed from the first principal components of the two sets of PCA results are closely similar – the variation in amino acid frequencies is also captured in the PCA results of the individual codon frequencies. Hence, the variations among genomes that affect the usage of synonymous codons and the variations that affect the usage of particular amino acids essentially occur together.

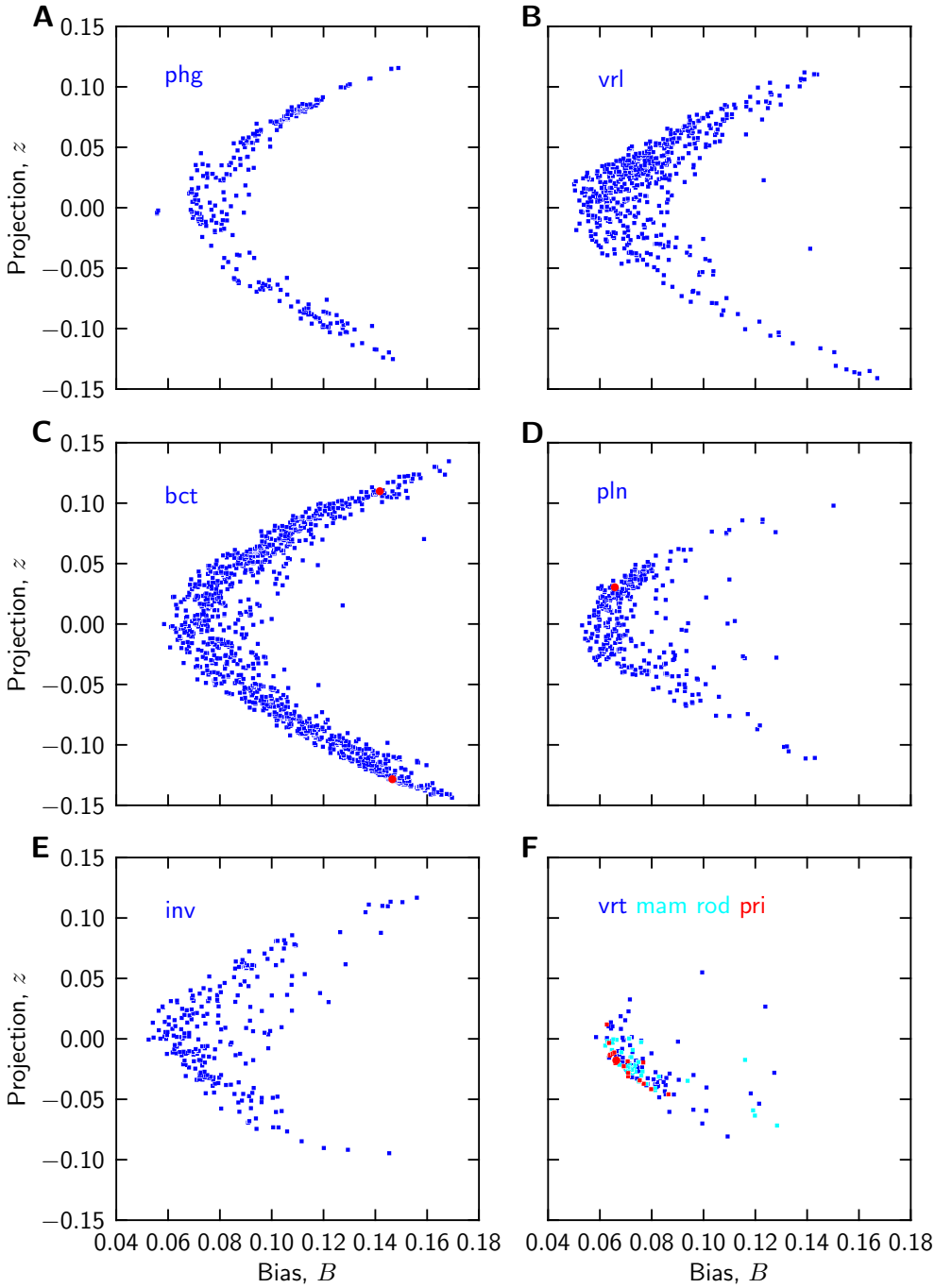


FIG. 3. *Codon usage maps of genomes.* Genomes are grouped by CUTG database sections as indicated (see Methods); the section labeled 'bct' includes both bacterial and archaeal genomes. Red filled circles mark the genomes of *Homo sapiens* (F), *Arabidopsis thaliana* (D), *Streptomyces griseus* (C, lower circle), and *Clostridium tetani* (C, upper circle). The position of a genome relates measures of codon usage bias within the genome (horizontal axis) and of codon usage diversity against average usage across all genomes (vertical axis). The Supplementary Materials include a separate file for each CUTG database, 'xxx_map.txt', tabulating species id, species name and map coordinates B and z.

DISCUSSION

We have differentiated the codon usages of 2730 genomes sampled across the kingdoms of life in a two-dimensional map (Fig. 3). The horizontal map dimension represents non-uniformity of different-codon frequencies across the individual genome. The vertical dimension represents the diversity of codon frequency in the individual genome with respect to the mean of all genomes. The one-dimensional measure describing diversity in all

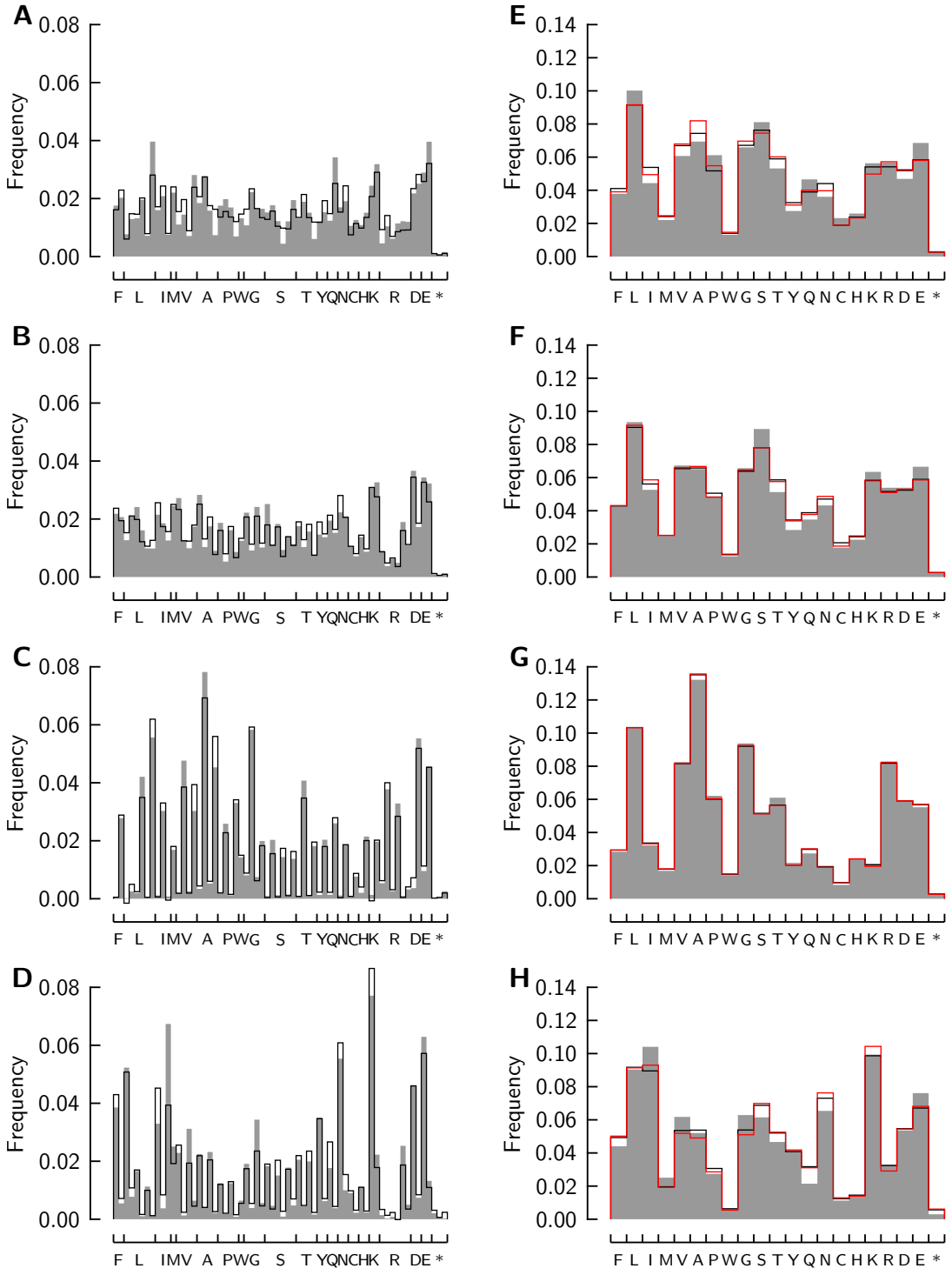


FIG. 4. *Codon (A-D) and amino acid frequencies (E-H) reconstructed from PCA results.* Gray shaded: observed codon frequencies; black lines: frequencies reconstructed from the first principal component. Example genomes: *Homo sapiens* (A, E), *Arabidopsis thaliana* (B, F), *Streptomyces griseus* (C, G), and *Clostridium tetani* (D, H). The red lines in E-H are amino acid frequencies computed from the reconstructed codon frequencies of A-D. All reconstructions are made with the PCA results obtained from the genomes in the bias bin to which the example genome locates (see Figs. 1C and 3).

64 different codons is derived from PCA of codon frequencies. A dominant first principal component indicates that the diversity arises from a proportionate variation of the frequencies of most different codons (Fig. 2A, B).

Most organisms group to map regions corresponding to weak bias and weak diversity in codon usage (see also Fig. 1C). Organisms with larger bias distribute in two distinct branches dependent on the direction in which their different codon frequencies diverge from the mean of all genomes. Genomes with weak diversity approach the average codon usage of all genomes (Fig. 2C). Strong diversity among genomes occurs between two distinct patterns of codon usage. To identify those patterns consider two sample genomes (Table 1).

Table 1 juxtaposes the observed codon frequencies of *Streptomyces griseus* and *Clostridium tetani E88*. Those bacteria are identified in the map as being among the organisms with the most biased and diverse codon usages (Fig. 3C). Indeed, *Streptomyces* and *Clostridium* use nearly exclusive subarrays of the 64 different codons (Fig. 1B). Inspection of Table 1 reveals that the two bacteria have strong preferences for different synonymous codons. *Clostridium* prefers synonyms with U or A as third base, *Streptomyces* prefers synonyms with G or C as third base. Each bacterium restricts its codon usage virtually to one half of the genetic code table (with minor irregularities involving the codons for methionine and tryptophan).

The codon usages of genomes thus map to regions interconnecting minimal bias and diversity with either maximal bias, use of codons with U/A as third base (positive branches in Fig. 3), or maximal bias, use of codons with G/C as third base (negative branches). The two extreme forms of codon usage are, in essence, restricted to using either one half of the genetic code table, whereas the unbiased form of usage recruits the full code table. The vertical dimension of the map approximately represents the fraction of G/C occurrences as third base ('GC3 content') because G/C occurrences in different codons vary in proportion, as is indicated by the PCA first eigenvector (Fig. 1B).

A second kind of bias in codon usage is observed in the encoding of the amino acids leucine (L) and arginine (R). Each of these amino acids is represented by six synonymous codons, using two different patterns in the first two bases. *Streptomyces* uses codons with one pattern in the first two bases, whereas *Clostridium* uses codons with the other pattern (Table 1). These two kinds of variation in synonymous codon usage are accompanied by a variation in amino acid usage (Figs. 4G,H).

Together, these three different aspects that render the codon usages in *Streptomyces* and *Clostridium* distinct from each other represent a universal pattern of diversity. The pattern dominates the variation of codon (and amino acid) usage across the entire set of genomes that we have sampled. The pattern is found from the genomes with strong bias in their codon usage to the genomes that approach uniform codon usage (Fig. 2B). It involves proportionate, concerted variations of the frequencies of most different codons. The simplest mechanism that one might imagine to produce this kind of variation in codon usage is lateral DNA transfer across the U/A- and G/C-preferring groups of organisms; this mechanism implies that the variation progresses from strong to weak bias of codon usage.

Table 1. Codon frequencies ($\times 10^3$) of *Streptomyces griseus* and *Clostridium tetani*

F	UUU	1	39	W	UGG	14	6	H	CAU	2	11
	UUC	28	5	G	GGU	8	19		CAC	21	2
L	UUA	0	52		GGC	58	4	K	AAA	1	77
	UUG	3	8		GGA	7	34		AAG	20	22
	CUU	2	17		GGG	20	6	R	CGU	5	1
	CUC	42	1	S	UCU	1	18		CGC	38	0
	CUA	1	11		UCC	20	5		CGA	3	1
	CUG	56	1		UCA	1	15		CGG	33	0
I	AUU	1	33		UCG	14	1		AGA	1	25
	AUC	30	4		AGU	1	18		AGG	3	5
	AUA	1	67		AGC	14	5	D	GAU	4	46
M	AUG	17	25	T	ACU	1	21		GAC	55	7
V	GUU	2	23		ACC	41	5	E	GAA	10	63
	GUC	48	1		ACA	1	20		GAG	46	13
	GUA	3	31		ACG	18	2	*	UAA	0	2
	GUG	30	6	Y	UAU	1	35		UAG	0	1
A	GCU	3	22		UAC	20	6		UGA	3	0
	GCC	78	4	Q	CAA	1	18				
	GCA	5	23		CAG	26	4				
	GCG	45	2	N	AAU	1	55				
P	CCU	2	12		AAC	19	10				
	CCC	26	2	C	UGU	1	9				
	CCA	1	12		UGC	8	2				
	CCG	33	1								

Codon usage bias decreases from present prokaryotes toward single and multi-cellular karyotic organisms (Fig. 3). Genomes with the strongest bias are found among prokaryotes that have been proposed to be metabolically most similar to the hypothetical 'last universal common ancestor' [8, 9]. Such genomes populate the most diverse regions of our codon-usage map (Fig. 3C, representing both bacteria and archaea; see file 'bct_map.txt' in Supplementary Materials). Archaea (e.g., *Methylobacterium extorquens*, *Halobacterium salinarum*) share a bias for using G/C as third base with bacteria (e.g. *Anaeromyxobacter dehalogenans* (Protobacteria), *Actinoplanes fruilensis* (Actinobacteria), *Desulfovibrio vulgaris* (Delta-protobacteria)). On the other hand, the bacteria *Mycoplasma mycoides* (Tenericutes) and, as described, *Clostridium tetani* (Clostridia) prefer U/A as third bases. While these organisms all possess proteins involved in a form of anaerobic metabolism [9], they form two genetic groups that encode proteins with different subsets of the genetic code table. Our results suggest that the pattern of codon usage is highly persis-

tent in the evolutionary record and thus might be a useful criterion for phylogenetic inquiries into early evolution.

CONCLUSIONS

Systematic, graded variations of synonymous codon usage are observed across a broad set of taxa. At one end of the range of variation, either synonyms with U or A as third base, or synonyms with G or C as third base, are used; at the other end, balance in the usage of synonyms is approached. That variation comprises all groups of synonymous codons in the genetic code. It separates genomes into two lineages that diverge as codon bias within genomes increases. In the context of evolution, these observations provide constraints regarding the time base of the evolution of codon usage, the nature of that evolution, and a hypothesized last universal common ancestor. Analysis of codon usage might be an informative complement to the analysis of amino acid sequences in molecular studies of evolution.

ACKNOWLEDGEMENTS

BBK dedicates this work to the memory of his uncle, Taras Khomchuk. BBK wishes to acknowledge the financial support of the United States Department of Defense (DoD) through the National Defense Science and Engineering Graduate Fellowship (NDSEG) Program: this research was conducted with Government support under and awarded by DoD, Army Research Office (ARO), National Defense Science and Engineering Graduate (NDSEG) Fellowship, 32 CFR 168a.

[1] Emile Zuckerkandl and Linus Pauling. Molecules as documents of evolutionary history. *Journal of Theoretical Biology*, 8:357–366, 1965.

[2] Siv G.E. Andersson and C.G. Kurland. Codon preferences in free-living microorganisms. *Microbiological Reviews*, 54(2):198–210, 1990.

[3] Laurent Duret. Evolution of synonymous codon usage in metazoans. *Current Opinion in Genetics & Development*, 12:640–649, 2002.

[4] Joshua B. Plotkin and Grzegorz Kudla. Synonymous but not the same: the causes and consequences of codon bias. *Nature Reviews Genetics*, 12:32–42, 2011.

[5] Linus Pauling and Emile Zuckerkandl. Chemical paleogenetics. molecular "restoration studies" of extinct forms of life. *Acta Chemica Scandinavica*, 17:S9–S16, 1963.

[6] Francesca D. Ciccarelli, Tobias Doerks, Christian von Mering, Christopher J. Creevey, Berend Snel, and Peer Bork. Toward automatic reconstruction of a highly resolved tree of life. *Science*, 311:1283–1287, 2006.

[7] Laura A. Hug, Brett J. Baker, Karthik Anantharaman, Christopher T. Brown, Alexander J. Probst, Cindy J. Castelle, Christina N. Butterfield, Alex W. Hernsdorf, Yuki Armano, Kotaro Ise, Yohey Suzuki, Natasha Dudek, David A. Relman, Karl M. Finstad, Ronals Amundson, Brian C. Thomas, and Jillian F. Banfield. A new view of the tree of life. *Nature Microbiology*, 1:1–6, 2016.

[8] Christos A. Ouzounis, Victor Kunin, Nikos Darzentas, and Leon Goldovsky. A minimal estimate for the gene content of the last universal common ancestor – exobiology from a terrestrial perspective. *Research in Microbiology*, 157:57–68, 2006.

[9] Madeline C. Weiss, Filipa L. Sousa, Natalia Mrnjavac, Sinje Neukirchen, Mayo Roettger, Shijalal Nelson-Sahti, and William F. Martin. The physiology and habitat of the last universal common ancestor. *Nature Microbiology*, page 16116, 2016.

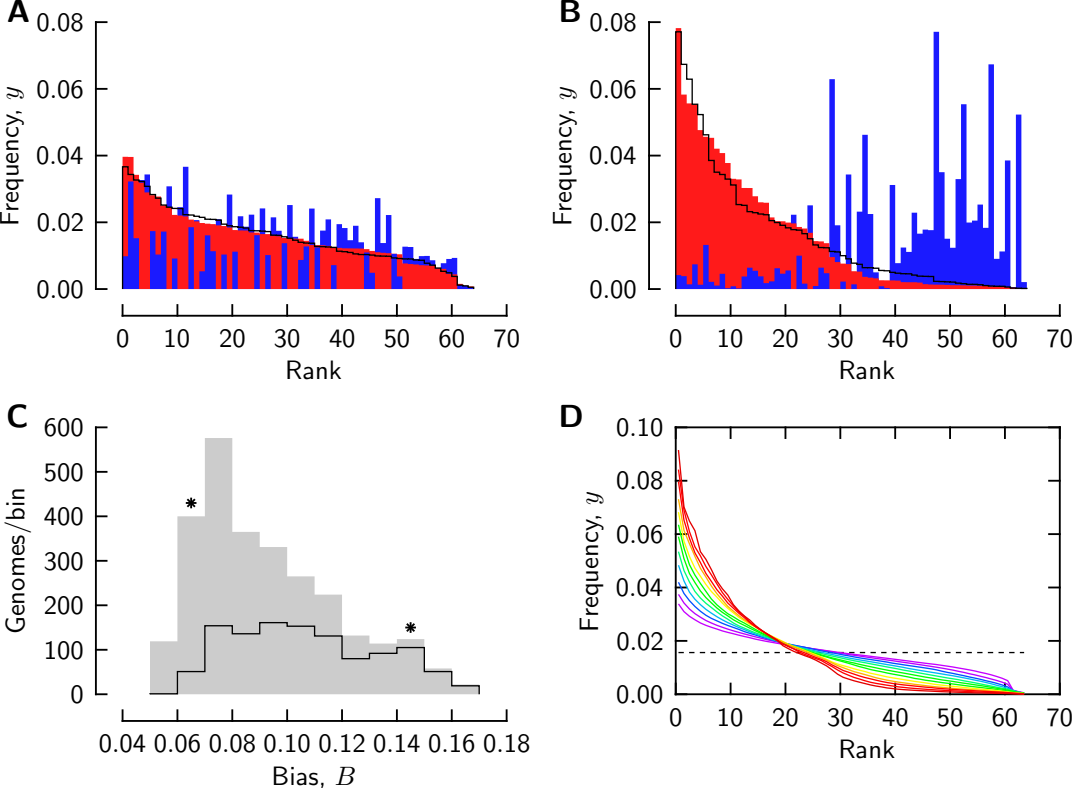
[10] M.W. Nirenberg, W. Jones, P. Leder, B.F.C. Clark, W.S. Sly, and S. Peska. On the coding of genetic information. *Cold Spring Harbour Symposia on Qnatitative Biology*, 28:549–557, 1963.

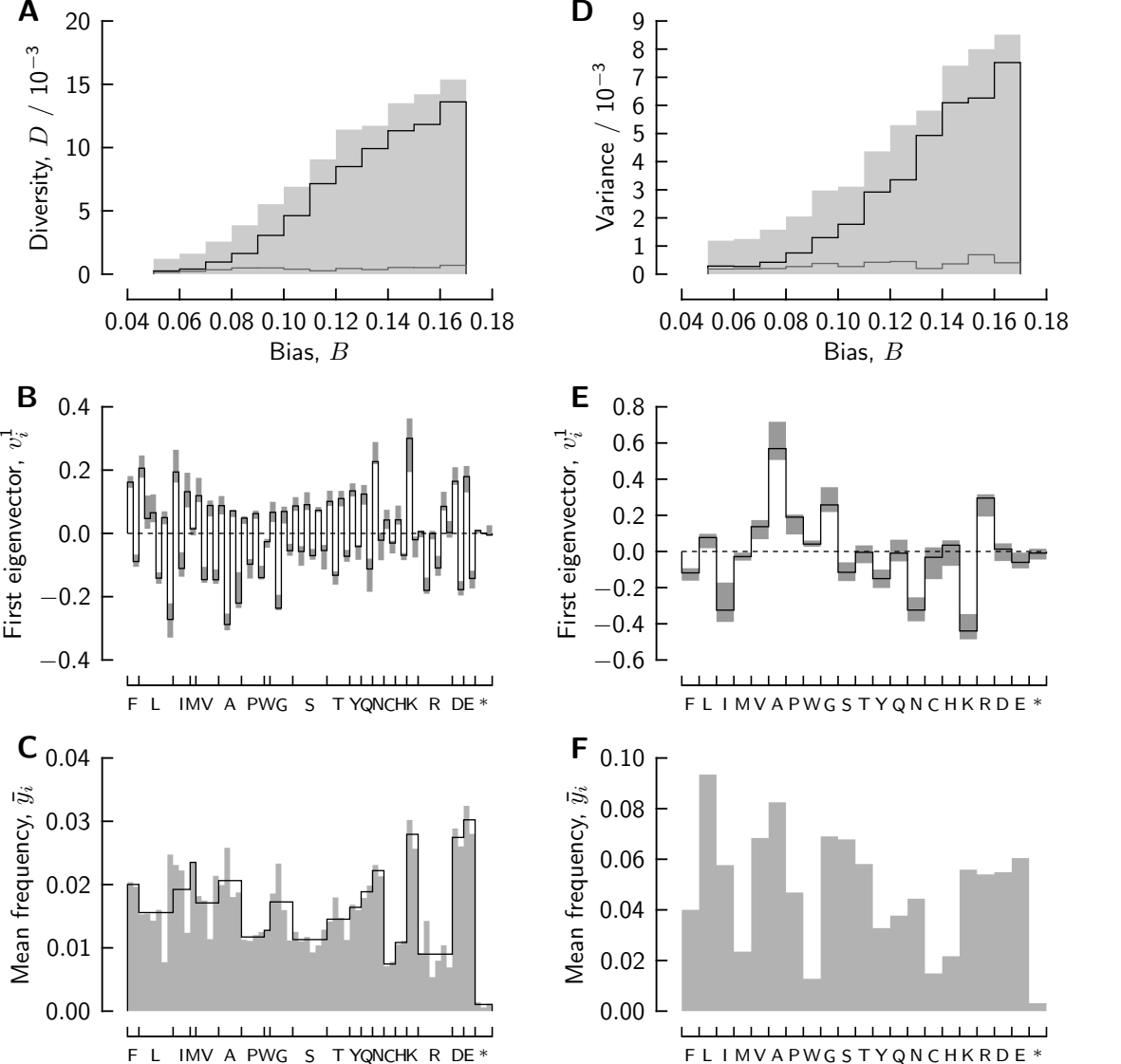
[11] F.H.C. Crick. The origin of the genetic code. *Journal of Molecular Biology*, 38:367–379, 1968.

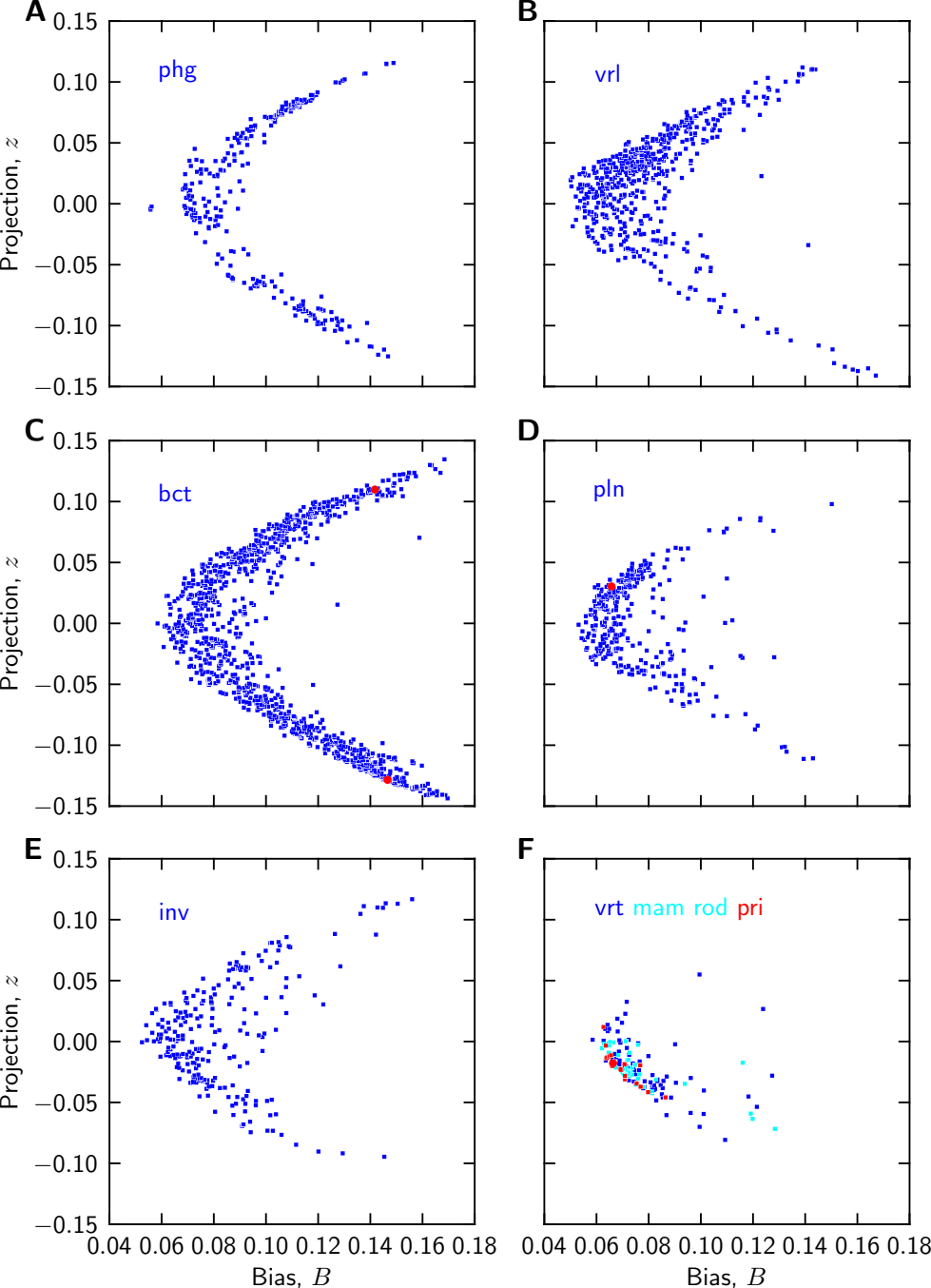
[12] M. Yu Borodovsky and S.M. Gusein-Zade. A general rule for ranged series of codon frequencies in different genomes. *Journal of Miomolecular Structure & Dynamics*, 6(5):1001–1012, 1998.

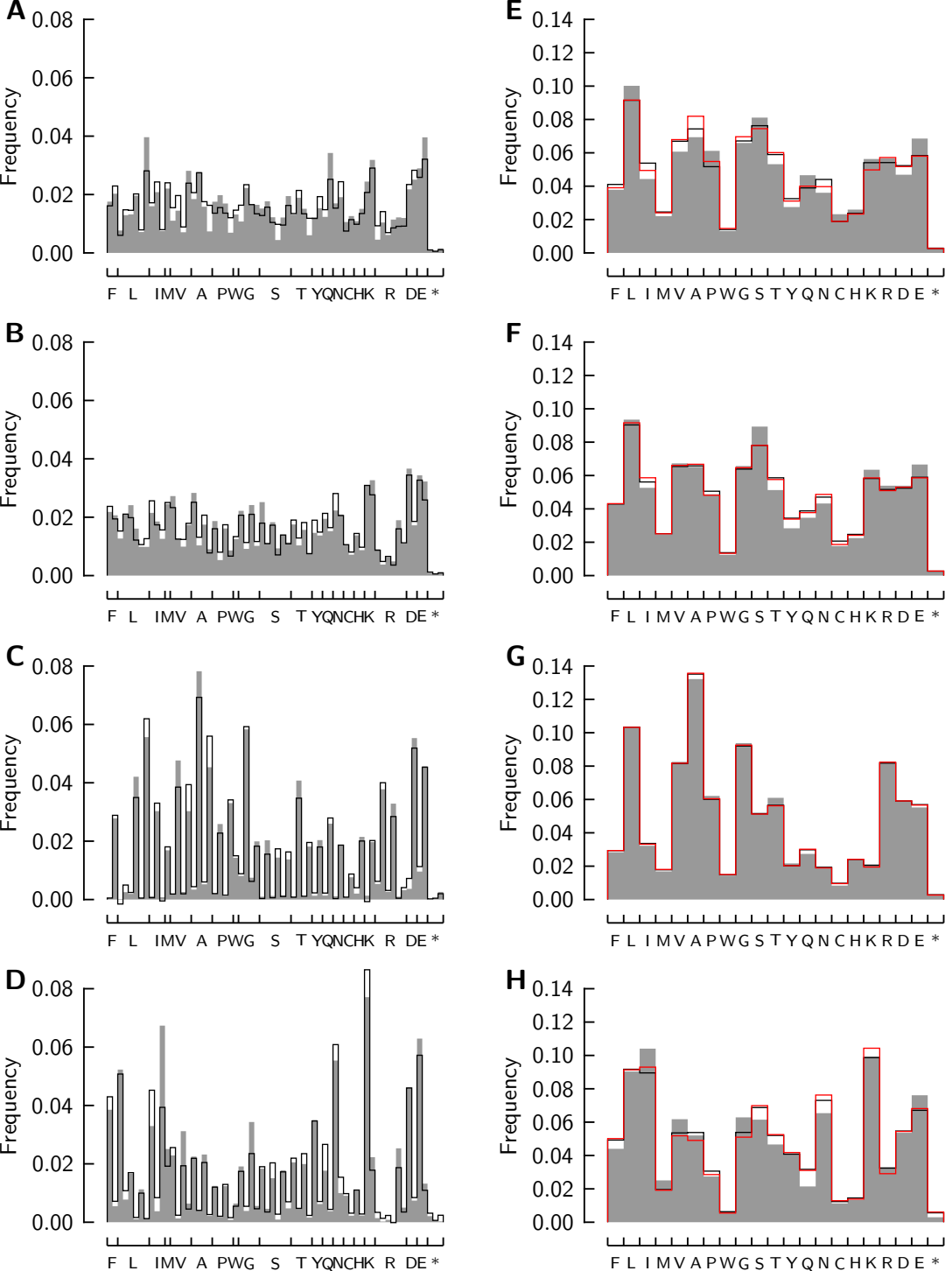
[13] Sabir M. Gusein-Zade and Mark Yu. Borodovsky. An improved distribution of codon frequencies allowing for inhomogeneity of dna's primary-structure evolution. *Journal of Biomolecular Structure & Dynamics*, 7(5):1185–1197, 1990.

[14] L. Frappat, C. Minichini, A. Sciarrino, and P. Sorba. Universality and shannon entropy of codon usage. *Physical Review E*, 68:061910, 2003.


A
B
C
D
Frequency, y
Rank
Genomes/bin
Bias, B
0.00
0.02
0.04
0.06
0.08
0.10
0
10
20
30
40
50
60
70
100
200
300
400
500
600
0.04
0.06
0.08
0.10
0.12
0.14
0.16
0.18
*
*


A
Diversity, D / 10^{-3}
0
5
10
15
20
0.04 0.06 0.08 0.10 0.12 0.14 0.16 0.18
Bias, B
D
Variance / 10^{-3}
0
1
2
3
4
5
6
7
8
9
0.04 0.06 0.08 0.10 0.12 0.14 0.16 0.18
Bias, B
B
First eigenvector, v_i^1
0.4
0.2
0.0
−0.2
−0.4
F L IMV A PWG S T YQNCHK R DE *
E
First eigenvector, v_i^1
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4
−0.6
F L I M V A P W G S T Y Q N C H K R D E *
C
Mean frequency, $\bar{y}_i$
0.04
0.03
0.02
0.01
0.00
F L IMV A PWG S T YQNCHK R DE *
F
Mean frequency, $\bar{y}_i$
0.10
0.08
0.06
0.04
0.02
0.00
F L I M V A P W G S T Y Q N C H K R D E *


A
phg
B
vrl
C
bct
D
pln
E
inv
F
vrt mam rod pri
Projection, z
Bias, B
0.15
0.10
0.05
0.00
−0.05
−0.10
−0.15
0.04 0.06 0.08 0.10 0.12 0.14 0.16 0.18


A
B
C
D
E
F
G
H
Frequency
0.00
0.02
0.04
0.06
0.08
0.10
0.12
0.14
F L IMV A PWG S T YQNCHK R DE *
F L I M V A P W G S T Y Q N C H K R D E *