



Consistent Clustering Pattern of Prokaryotic Genes Based on Base Frequency at the Second Codon Position and its Association with Functional Category Preference

Yan-Ting Jin^{1,2} · Cong Ma¹ · Xin Wang¹ · Shu-Xuan Wang¹ · Kai-Yue Zhang¹ · Wen-Xin Zheng³ · Zixin Deng² · Ju Wang⁴ · Feng-Biao Guo² 

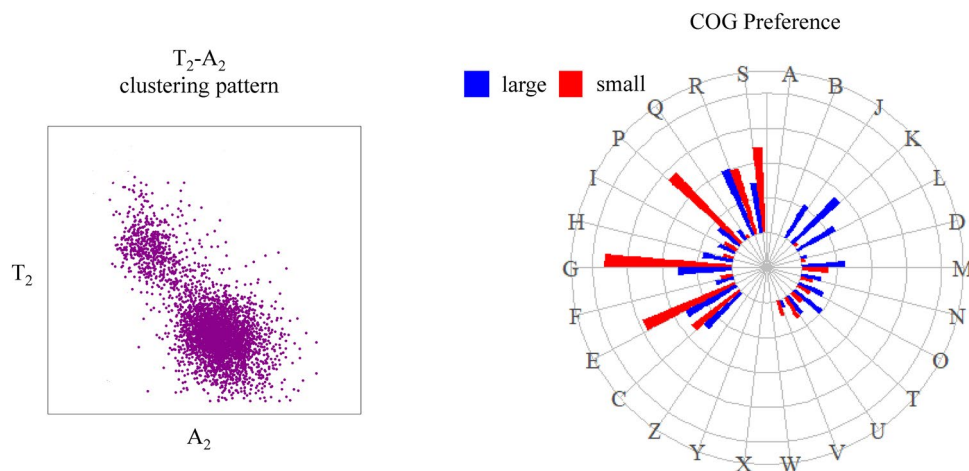
Received: 9 June 2021 / Revised: 2 November 2021 / Accepted: 7 November 2021 / Published online: 24 November 2021

© The Author(s) 2021

Abstract

In 2002, our research group observed a gene clustering pattern based on the base frequency of A versus T at the second codon position in the genome of *Vibrio cholera* and found that the functional category distribution of genes in the two clusters was different. With the availability of a large number of sequenced genomes, we performed a systematic investigation of A_2 – T_2 distribution and found that 2694 out of 2764 prokaryotic genomes have an optimal clustering number of two, indicating a consistent pattern. Analysis of the functional categories of the coding genes in each cluster in 1483 prokaryotic genomes indicated, that 99.33% of the genomes exhibited a significant difference ($p < 0.01$) in function distribution between the two clusters. Specifically, functional category P was overrepresented in the small cluster of 98.65% of genomes, whereas categories J, K, and L were overrepresented in the larger cluster of over 98.52% of genomes. Lineage analysis uncovered that these preferences appear consistently across all phyla. Overall, our work revealed an almost universal clustering pattern based on the relative frequency of A_2 versus T_2 and its role in functional category preference. These findings will promote the understanding of the rationality of theoretical prediction of functional classes of genes from their nucleotide sequences and how protein function is determined by DNA sequence.

Graphical abstract



Keywords Base frequency · A_2 versus T_2 · The second codon position · Two unequal clusters · Protein function preference

1 Introduction

The genetic code is a set of rules that defines how the four-letter code of DNA is translated into the 20-letter code of protein [1, 2]. Being employed in all organisms, the genetic code has many conserved and universal features, including successive triplets without overlapping [3], degeneracy for mutation tolerance [4] and codon usage bias that is meaningful for the regulation of gene expression [5–10]. Studies of codon and base usage can facilitate our understanding of the origin and evolution of genetic code. The choice of nucleotide at a specific codon position in coding genes has received much attention. For example, researchers have found strong C base preference at the second position of the second codons in the cell envelope-related genes [11], and clarified that the physical origins of codon positions strongly influence cotranslational protein folding [12]. For the broadest studies are of the third codon position. Codon degeneracy mainly manifests at this position and among the synonymous codons, the one matching the most abundant tRNA usually has the highest frequency [13]. This coupling pattern has been thought to benefit translation efficiency [14]. Recent genomic-scale expression data analyses have demonstrated the global effect of synonymous codon usage bias on transcription and translation efficiency [15–19]. Synonymous codon usage can also regulate protein folding type [20–23] and mutation at the synonymous site may cause intolerance and disease in human [24]. In addition, codon position specific nucleotide bias has been employed by some computational tools as an important feature for the identification of functional genes [25–28].

Although the nucleotide bias of the three codon positions has been widely used in gene identification, few studies have connected the codon position-specific nucleotide pattern with gene function distribution, particularly in the second codon position. In 2002, we observed an interesting pattern of base usage of coding genes in *Vibrio cholerae*. All coding genes could be divided into two unequal clusters according to the relative base frequencies of A and T at the second codon position, and the coding genes in the two clusters exhibited significant difference in protein functions [29]. We hypothesized that this pattern might appear widely in the prokaryotic domain and that it could be connected with gene function. Hence, a larger scale analysis was needed to validate our speculation.

In this paper, a systematic analysis of thousands of genomes across three domains revealed that the clustering phenomenon according to the base frequencies of A and T at the second codon position is almost universal and is especially remarkable for prokaryotes. Further study on 1483 prokaryotes with COG (clusters of orthologous

groups) annotation [30, 31] shows that 99.33% of genomes have a significant difference ($p < 0.01$, by Chi-squared test) in the functional distributions of genes between the two unequal clusters. Furthermore, the overrepresent categories in the two clusters were consistent among species and prokaryotic phyla. Here, we revealed a consistent of A_2-T_2 associated clustering pattern and consistent functional influence in prokaryotes.

2 Materials and Detailed Methods

2.1 Genome Data Collection

The prokaryotic genomic data used in this study were downloaded from NCBI (ftp://ftp.ncbi.nlm.nih.gov/genomes/archive/old_refseq/Bacteria/). Three files, i.e., ‘all.ffn.tar.gz’, ‘all.faa.tar.gz’ and ‘all.ptt.tar.gz’ were retrieved (on March 26, 2017), and contained FASTA files for nucleotide coding regions, FASTA for amino acids, and protein tables of all the prokaryotes available. After removing the sequences for plasmids and fragments from every genome, we obtained 2764 prokaryotic genomes with the sequences of protein coding genes (164 archaea and 2600 bacteria) for further analysis. The 1035 eukaryotic genomic data were downloaded from Ensembl (<https://asia.ensembl.org/downloads.html>), the details are shown in Table 1. Altogether, the gene sequences and annotation information of 3799 genomes from three domains of life were collected (Table 1).

2.2 K-Means Algorithm

K-means is a statistical method for partitioning observations in a data set into a given number of clusters (K). In this study, the K-means was used to divide the coding genes in a genome into a specific number of clusters (e.g., $K = 2$) based on the relative base frequencies of A and T at the second codon position. We wanted to determine whether there was a consistent clustering pattern based on A_2-T_2 in the currently available genomes.

There are many basic clustering techniques, which can be classified into five categories: partitioning methods, density-based methods, grid-based methods, hierarchical methods

Table 1 The protein coding genes of 3799 genomes collected for analysis

Domains	Genomes	Detail
Prokaryotes	2764	164 archaea and 2600 bacteria
Eukaryotes	1035	68 metazoa, 186 protists, 735 fungi, 44 plants, 1 <i>M. musculus</i> and 1 <i>H. Sapiens</i>
All genomes	3799	

and model-based methods [32]. The *K*-means algorithm is a partitioning method. It can give a definite number of final clusters and furthermore is highly efficient to implement. Here we need to cluster the genes of thousands of prokaryotic genomes hence we choose this method.

2.3 Silhouette Coefficient Analysis

As an unsupervised machine learning algorithm, *K*-means can be used to group the protein coding genes in a genome into *K* clusters ($K=2-9$ in our case). We further adopted silhouette coefficient analysis to determine the optimal number of clusters into which the genes could be divided. Briefly, the silhouette coefficient was used to quantify the separation between the resulting clusters by measuring how close each gene in one cluster is to genes in the neighboring clusters. In this study, we used the silhouette coefficient to evaluate the performance of *K* in the range from 2 to 9 for all genomes.

2.4 Clusters of Orthologous Groups (COGs) Function Annotation

Comparison of proteins encoded in numbers of complete genomes from many major phylogenetic lineages and elucidation of consistent patterns of sequence similarities allows the delineation of many clusters of orthologous groups (COGs) [31]. There are 26 function categories (Table 2) in this framework and each category is denoted with a specific letter. Among all the prokaryotic genomes analyzed, 1483 had complete COG annotation.

2.5 Chi-Squared Test

We further used the Chi-squared (χ^2) test to evaluate whether there was a significant difference in the COG function categories in the two unequal clusters obtained by *K*-means clustering. For each genome, the χ^2 test involved a 2×26 Chi-square table. The first row contained the number of genes in each COG function category in the large cluster, and the second row contained the number of genes in each COG function category in the small cluster. The differences

were significant for 1475 out of the 1483 genomes ($p < 0.05$) and highly significant for 1473 genomes ($p < 0.01$).

2.6 Measurement of the Difference in Specific Functional Category in the Two Clusters

$$F(P_{small}) = \frac{N(P_{small})}{N(small)}$$

$$F(P_{large}) = \frac{N(P_{large})}{N(large)}, \tag{1}$$

	P category	smaller	larger	
$F(P_{small}) > F(P_{large})$		1	0	, \tag{2}
$F(P_{small}) < F(P_{large})$		0	1	
$F(P_{small}) = F(P_{large})$		0	0	

Equations (1), (2): (1) $N(P_{small})$: the number of coding genes belonging to the P functional category in the small cluster. $N(small)$: the total number of coding genes in the small cluster. $F(P_{small})$: the ratio of $N(P_{small})$ to $N(small)$ and so for $F(P_{large})$. (2) Significance is determined by whether the difference is beyond 5% of the lower proportion value.

2.7 Bacteria Taxonomy

TaxonKit (<https://github.com/shenwei356/taxonkit>) [33] was used to rapidly assign the prokaryotic genomes into different phyla, and we retained only those phyla containing more than 10 genomes for further research.

3 Experimental Results

3.1 Grouping Protein Coding Genes by Base Frequency at the Second Codon Position

To analyze the base distribution at the three codon positions, $f(X_n)$ was defined to represent the frequency of a base at a certain codon position for a coding gene, with ‘X’ denoting bases A, T, C or G, and ‘n’ denoting the 1st, 2nd or 3rd codon position. For example, $f(A_2)$ represents the frequency of base A at the second codon position of a coding gene. We directly employed base T in coding DNA instead of base U in the counterpart mRNA for convenience. When the distribution of $f(X_n)$ was checked separately, no universal pattern was observed, although several frequencies, e.g., $f(X_1)$, $f(X_3)$, $f(C_2)$ and $f(G_2)$, showed skewed distributions for a few genomes. However, when the combination of $f(A_2)$ and $f(T_2)$ was applied, we found that, the protein coding genes gathered into two unequal clusters in the scatter plot in most genomes, similar to what we observed previously in *Vibrio cholerae* [29].

Table 2 The 26 function categories could be classed into four super-categories

Super-category	Number	Code letter
Information storage and processing	5	J, K, L, A, B
Cellular processes and signaling	11	D, Y, V, T, M, N, Z, W, U, O, X
Metabolism has eight categories	8	C, G, E, F, H, I, P, Q
Poorly characterized	2	R, S

Then, a systematic survey of $f(A_2)$ and $f(T_2)$ was conducted across the three domains of life (164 archaea, 2600 bacteria and 1035 eukaryotes), and it was shown that the protein coding genes in a genome could be divided into two unequal clusters. The smaller cluster has a larger $f(T_2)$ and smaller $f(A_2)$, and the larger cluster has a relatively smaller $f(T_2)$ and larger $f(A_2)$. Such a pattern was evident and almost universal in archaea and bacteria. This pattern could also be observed in eukaryotes but was usually nonsignificant (Fig. 1).

3.2 Silhouette Coefficient to Measure the Optimal Cluster Number

To check whether the genes in the genomes could be significantly divided into two clusters based on the $f(A_2)$ and $f(T_2)$, we adopted the quantitative method of K -means clustering to divide the genes in each genome into different groups. Then, the silhouette coefficient was used to measure how many clusters the genes could be divided into with the maximum intercluster distance and minimum intracluster distance. We surveyed 2764 prokaryotic genomes and 1035 eukaryotic genomes. We found that 98.17% (161/164) of the archaeal genomes and 97.42% (2533/2600) of bacterial genomes could be optimally divided into two clusters (Fig. 2; Table S1).

Although the tendency of gene clustering could still be observed for eukaryotes, particularly higher eukaryotes (multicellular organisms), the best clustering number varied widely. Of all the eukaryotes examined, only 28.31% had an optimal clustering number of two, while optimal

cluster numbers of three and four were found for 23.29% and 46.28% of the genomes, respectively (Fig. 2B). This phenomenon might be associated with the complexity of multicellular eukaryotes, which have many more genes in their genomes, with gene functions closely related to factors such as cell type and transcription regulation and also requiring more elaborate cell structures and intricate metabolic networks [34]. For simplicity, we focused on the prokaryotic genomes in the following analyses.

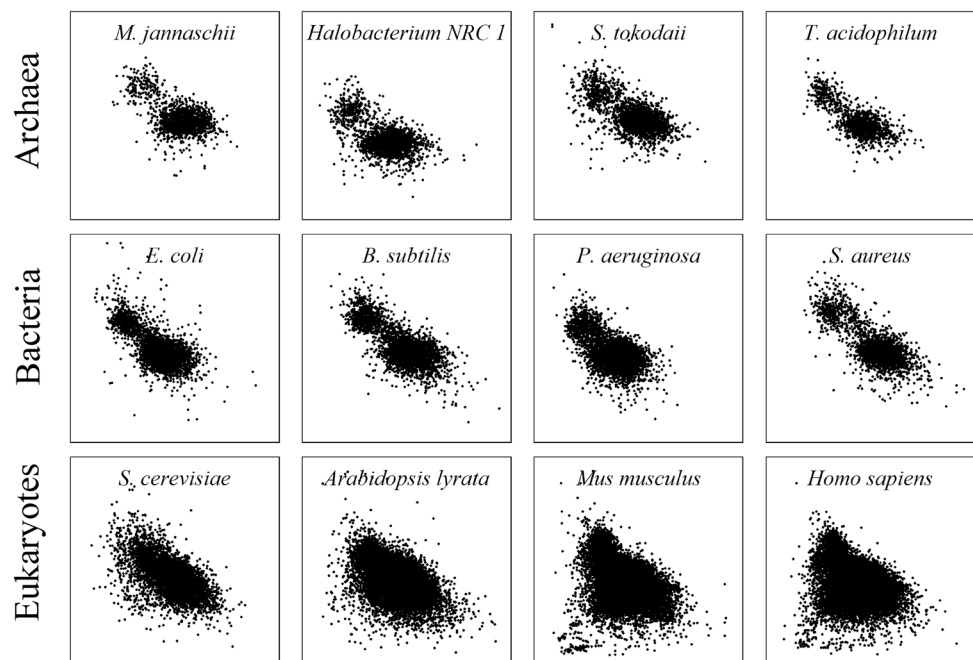
3.3 Biased Functional Distribution of Genes in the Two Clusters

We further checked the function of the genes in the two clusters divided by $f(A_2)$ and $f(T_2)$. Of the 2764 prokaryotic genomes, COG (cluster of orthologous groups) annotations could be retrieved for 1483. When the Chi-squared test was applied to the 26 COG functional categories, a significant difference ($p < 0.01$) was detected for the functions of genes included in the two clusters for 99.33% of the genomes examined (Table 3; Table S2).

3.4 Differences in the Distribution of Protein Function Categories in the Two Unequal Clusters

To see a clearer pattern of the functional difference in genes in the two clusters, *E. coli* and *M. jannaschii* were chosen as a representative of bacteria and archaea, respectively (Table S3). Since the larger cluster contained much more genes than the smaller cluster, we calculated the proportion

Fig. 1 Coding genes are divided into two unequal clusters by the base frequencies of A and T at the second position of codons. The scatter plots of 12 representative genomes from three domains with $f(A_2)$ as the x axis and $f(T_2)$ as the y axis both ranging from 0 to 0.7. The clustering phenomenon in archaea and bacteria is significant: a small cluster with much higher $f(T_2)$ and a large cluster with similar $f(T_2)$ and $f(A_2)$. This phenomenon was not significant in eukaryotes



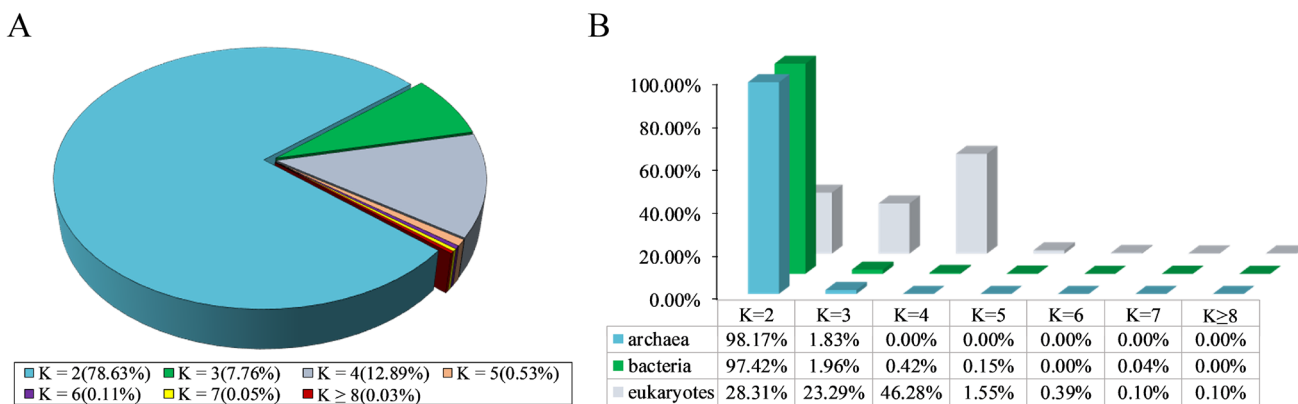


Fig. 2 The best choice of cluster number is 2. **A** Taking all three domains as a whole, 78.63% of genomes had an optimal *K* of 2. **B** General distribution of quantitative optimal *K* values indicates that two clusters are the best choice for prokaryotes and some of eukaryotes (Table S1)

Table 3 The Chi-squared test results of 1483 genomes on the protein function difference in the two unequal clusters

	$p \geq 0.05$	$p < 0.05$	$p \geq 0.01$	$p < 0.01$
Genome number	8	1475	10	1473
Frequency	0.54%	99.46%	0.67%	99.33%

of genes belonging to every functional category in each cluster through Eq. (1). By comparing the gene function distributions in the two clusters, we found that gene related to inorganic ion transport and metabolism (COG category P) were much more prevalent in the smaller cluster than in the larger cluster. In contrast, genes related to translation, ribosomal structure and biogenesis (J), transcription (K) and replication, recombination and repair (L) were more likely to be included in the larger cluster (Fig. 3A). This pattern seemed to be universal as it was observed in all the genomes checked.

Then we used Eq. (2) to examine whether a given category is overrepresented in the larger (A_2 preference) or smaller (T_2 preference) cluster. Using the P category as an example, for a certain genome, if $F(P_{small}) > F(P_{large})$ and the difference was greater than 5%, for P, we assigned a value of 1 to the smaller cluster and 0 to the larger cluster. In contrast, if $F(P_{small}) < F(P_{large})$ and the difference was greater than 5%, the smaller cluster was assigned a value of 0, while the larger cluster was assigned a value of 1. If the difference between $F(P_{small})$ and $F(P_{large})$ was less than the 5% threshold, we defined both clusters as 0, denoting that no significant difference existed in the P category between the smaller and the larger clusters. Using this method, we were able to determine how many genomes exhibited overrepresentation of each functional category in the two clusters. Among the 26 COG categories, P, U, V, S, and G were overrepresented in the smaller cluster for 98.65% (1463/1483), 95.75%

(1420/1483), 92.92% (1378/1483), 90.42% (1341/1483), and 84.49% (1253/1483) genomes, respectively. Comparatively, J, K, L, F, H, T, and Q were overrepresented in the larger cluster for 98.92% (1467/1483), 98.52% (1461/1483), 98.79% (1465/1483), 95.95% (1423/1483), 87.93% (1304/1483), 87.59% (1299/1483) and 81.66% (1211/1483) genomes, respectively (Fig. 3B; Table S4).

Since prokaryotes are classified into different phyla, we checked whether all the phyla had consistent preference of functional category. The 13 prokaryotic phyla, each containing more than 10 genomes (Table S4), were extracted for further lineage analysis. Each functional category was calculated for each genome through Eq. (2). Then we calculated the cumulative number of the overrepresented categories in each phylum. *Proteobacteria*, for example, contained 627 genomes, 528 of which had smaller cluster as 1 and 97 of which had larger cluster as 1 for C category [2 genomes had similar $F(C_{small})$ and $F(C_{large})$]. Hence the overrepresented ratios in the small and large clusters in *Proteobacteria* were 528/625 (84.48%) and 97/625 (15.52%), respectively. If the number of genomes was zero in some GOG functional categories, the ratio of both the smaller and the larger clusters was defined as zero. In such cases, the sum of both the smaller and the larger cluster ratios would be 0. However, for most phyla, the sum of the ratios was 1.

In Fig. 3C, we focus on those categories with different colors in the left panel and right panel because two clusters have distinct ratios. For the left panel of the larger cluster, if one functional category has a consistent color that means all phyla have consistent preference and similar for the right panel of the smaller cluster. Following this rule, we found that J, K, L, F, H, and Q were consistently overrepresented in the larger cluster of all phyla, while P, U, and S were in the smaller cluster of all phyla. Therefore, at the species level (Fig. 3A), prokaryote level (Fig. 3B) and phylum level (Fig. 3C), a consistent preference of functional category was

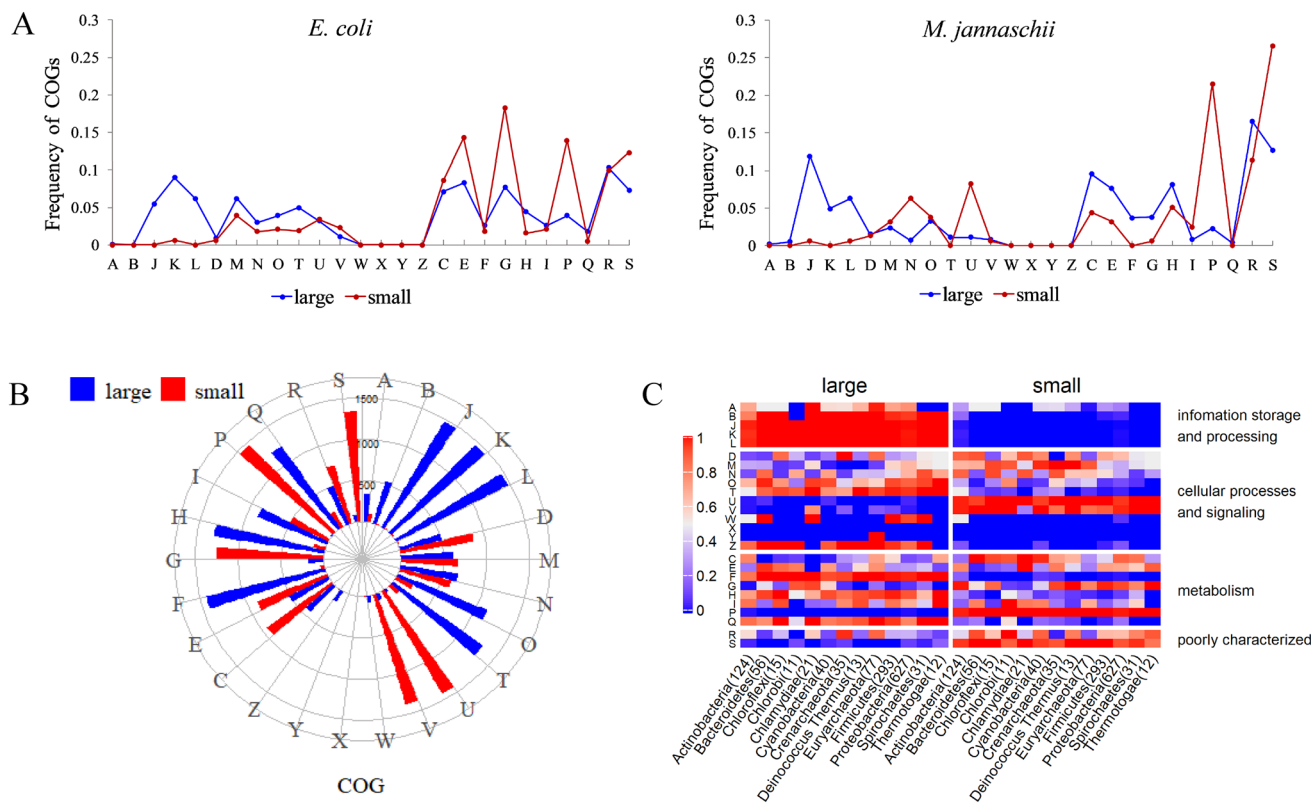


Fig. 3 The distribution and difference in COG functional categories in the two unequal clusters of 1483 genomes. **A** In two representative genomes in prokaryotes, P-related genes prevailed in the small cluster, while J-, K- and L-related genes were observed at a higher proportion in the large cluster. **B** Cumulative overrepresented genome numbers of 26 functional categories. The 26 functional categories are

observed, i.e., J, K and L were significantly overrepresented in the larger cluster, whereas P showed the opposite pattern. Several other categories exhibited bias to a lesser degree.

4 Discussions

Researchers have observed several clustering patterns in sequenced genomes. Médigue et al., analyzed codon usage in 780 *E. coli* genes [35]. Using factorial correspondence analysis, they illustrated that these genes could be divided into three classes. The first two classes are associated with expression level and the third is associated with mobility characteristics [35]. A similar pattern was found in the protein coding genes of *Bacillus subtilis* [36]. Since then, numerous studies have confirmed codon usage associated patterns in various prokaryotic genomes [37–39]. Ma and Chen defined the most deviated codon position (MDCP) and found that basing on MDCP, the CDSs of a genome can be classified into two clusters: typical and atypical [40]. Genes can also be divided into two separate clusters based

on strand associated nucleotide bias [41–43]. All these clustering patterns are based on the distribution bias of codon usage, that is, nucleotide frequencies at three codon positions. Comparatively, here our clustering pattern is associated only with the second codon position and appears similarly in almost all prokaryotes.

On the other hand, codon usage has been used to cluster coding sequences of *Arabidopsis thaliana* genes in order to improve gene prediction [44]. Amino acid composition has been combined with machine-learning method to predict protein functional families and achieved accuracy of 69.1–96.1% [45]. Although DNA sequence could be also extracted as features to prediction protein function [46], in most cases features of amino acid frequency are adopted [47]. Here, we illustrated one example of direct link between nucleotide frequency and protein function categories. Therefore, our work would help to understand why protein function could be predicted from gene sequence. We hope future researches could get highly reliable prediction of protein function from DNA sequence and we think their used features would mainly associate with the second codon

position, particularly A_2 versus T_2 frequency and in that sense our result would be well validated.

5 Conclusion

By a systematic analysis of the base frequencies at the second codon position across the three domains of life, we found that the protein coding genes of prokaryotes can be divided into two unequal clusters based on $f(A_2)$ and $f(T_2)$. Further analysis showed significant difference in the proportions of genes belonging to certain COG categories in the two clusters. P-related genes were more prevalent in the smaller cluster, while J-, K- and L-related genes were more likely to be included in the larger cluster. Lineage analysis revealed that the bias was basically consistent among different phyla. Hence, this work demonstrates an almost universal clustering pattern by the frequency of T_2 versus A_2 and its basically consistent influence on functional category distribution among prokaryotic phyla. These findings can help us understand why coding potentiality and functional category assignment could be theoretically predicted from gene sequences.

Supplementary Information The online version contains supplementary material available at <https://doi.org/10.1007/s12539-021-00493-w>.

Acknowledgements This work was supported by the National Natural Science Foundation of China [31871335] and Beijing Natural Science Foundation (BJNSF 5182006).

Author Contributions F-BG and JW designed and coordinated this project. Y-TJ did the main analysis, and S-XW drew diagrams. Y-TJ and K-YZ drafted the manuscript. CM and XW double checked the results. JW and F-BG revised the manuscript with comments from other authors. All the authors read and approved this manuscript.

Declarations

Conflict of interest On behalf of all authors, the corresponding author states that there is no conflict of interest.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

1. Crick FH (1968) The origin of the genetic code. *J Mol Biol* 38:367–379. [https://doi.org/10.1016/0022-2836\(68\)90392-6](https://doi.org/10.1016/0022-2836(68)90392-6)
2. Fischer EC, Hashimoto K, Zhang Y, Feldman AW, Dien VT, Karadeema RJ et al (2020) New codons for efficient production of unnatural proteins in a semisynthetic organism. *Nat Chem Biol* 16:570–576. <https://doi.org/10.1038/s41589-020-0507-z>
3. Nirenberg MW, Matthaei JH (1961) The dependence of cell-free protein synthesis in *E. coli* upon naturally occurring or synthetic polyribonucleotides. *Proc Natl Acad Sci USA* 47:1588–1602. <https://doi.org/10.1073/pnas.47.10.1588>
4. Caskey CT, Tompkins R, Scolnick E, Caryk T, Nirenberg M (1968) Sequential translation of trinucleotide codons for the initiation and termination of protein synthesis. *Science* 162:135–138. <https://doi.org/10.1126/science.162.3849.135>
5. Hockenberry AJ, Sireer MI, Amaral LA, Jewett MC (2014) Quantifying position-dependent codon usage bias. *Mol Biol Evol* 31:1880–1893. <https://doi.org/10.1093/molbev/msu126>
6. Tischer A, Graur D (1989) Nucleic acid composition, codon usage, and the rate of synonymous substitution in protein-coding genes. *J Mol Evol* 28:286–298. <https://doi.org/10.1007/BF02103424>
7. Zhou Z, Dang Y, Zhou M, Li L, Yu CH, Fu J et al (2016) Codon usage is an important determinant of gene expression levels largely through its effects on transcription. *Proc Natl Acad Sci USA* 113:E6117–E6125. <https://doi.org/10.1073/pnas.1606724113>
8. Boël G, Letso R, Neely H, Price WN, Wong KH, Su M et al (2016) Codon influence on protein expression in *E. coli* correlates with mRNA levels. *Nature* 529:358–363. <https://doi.org/10.1038/nature16509>
9. Liu Y, Yang Q, Zhao F (2021) Synonymous but not silent: the codon usage code for gene expression and protein folding. *Annu Rev Biochem*. <https://doi.org/10.1146/annurev-biochem-071320-112701>
10. Hanson G, Collier J (2018) Codon optimality, bias and usage in translation and mRNA decay. *Nat Rev Mol Cell Biol* 19:20–30. <https://doi.org/10.1038/nrm.2017.91>
11. Tang SL, Chang BC, Halgamuge SK (2010) Gene functionality's influence on the second codon: a large-scale survey of second codon composition in three domains. *Genomics* 96:92–101. <https://doi.org/10.1016/j.ygeno.2010.04.001>
12. Sharma AK, Bukau B, O'Brien EP (2016) Physical origins of codon positions that strongly influence cotranslational folding: a framework for controlling nascent-protein folding. *J Am Chem Soc* 138:1180–1195. <https://doi.org/10.1021/jacs.5b08145>
13. Ikemura T (1981) Correlation between the abundance of *Escherichia coli* transfer RNAs and the occurrence of the respective codons in its protein genes: a proposal for a synonymous codon choice that is optimal for the *E. coli* translational system. *J Mol Biol* 151:389–409. [https://doi.org/10.1016/0022-2836\(81\)90003-6](https://doi.org/10.1016/0022-2836(81)90003-6)
14. Novoa EM, Ribas de Pouplana L (2012) Speeding with control: codon usage, tRNAs, and ribosomes. *Trends Genet* 28:574–581. <https://doi.org/10.1016/j.tig.2012.07.006>
15. Frumkin I, Lajoie MJ, Gregg CJ, Hornung G, Church GM, Pilpel Y (2018) Codon usage of highly expressed genes affects proteome-wide translation efficiency. *Proc Natl Acad Sci USA* 115:E4940–E4949. <https://doi.org/10.1073/pnas.1719375115>
16. Jeacock L, Faria J, Horn D (2018) Codon usage bias controls mRNA and protein abundance in trypanosomatids. *Elife*. <https://doi.org/10.7554/eLife.32496.001>

17. Xu Y, Liu K, Han Y, Xing Y, Zhang Y, Yang Q et al (2021) Codon usage bias regulates gene expression and protein conformation in yeast expression system *P. pastoris*. *Microb Cell Factories* 20:91. <https://doi.org/10.1186/s12934-021-01580-9>
18. Yang Q, Lyu X, Zhao F, Liu Y (2021) Effects of codon usage on gene expression are promoter context dependent. *Nucleic Acids Res* 49:818–831. <https://doi.org/10.1093/nar/gkaa1253>
19. Zhao F, Zhou Z, Dang Y, Na H, Adam C, Lipzen A et al (2021) Genome-wide role of codon usage on transcription and identification of potential regulators. *Proc Natl Acad Sci USA*. <https://doi.org/10.1073/pnas.2022590118>
20. Yu C-H, Dang Y, Zhou Z, Wu C, Zhao F, Sachs MS et al (2015) Codon usage influences the local rate of translation elongation to regulate co-translational protein folding. *Mol Cell* 59:744–754. <https://doi.org/10.1016/j.molcel.2015.07.018>
21. Liu Y (2020) A code within the genetic code: codon usage regulates co-translational protein folding. *Cell Commun Signal* 18:145. <https://doi.org/10.1186/s12964-020-00642-6>
22. Liu Y, Yang Q, Zhao F (2021) Synonymous but not silent: the codon usage code for gene expression and protein folding. *Annu Rev Biochem* 90:375–401. <https://doi.org/10.1146/annurev-biochem-071320-112701>
23. Santoni D (2021) The impact of codon choice on translation process in *Saccharomyces cerevisiae*: folding class, protein function and secondary structure. *J Theor Biol* 526:110806. <https://doi.org/10.1016/j.jtbi.2021.110806>
24. Cheng N, Li M, Zhao L, Zhang B, Yang Y, Zheng C-H et al (2020) Comparison and integration of computational methods for deleterious synonymous mutation prediction. *Brief Bioinform* 21:970–981. <https://doi.org/10.1093/bib/bbz047>
25. Guo FB, Dong C, Hua HL, Liu S, Luo H, Zhang HW et al (2017) Accurate prediction of human essential genes using only nucleotide composition and association information. *Bioinformatics* 33:1758–1764. <https://doi.org/10.1093/bioinformatics/btx055>
26. Zhang K-Y, Gao Y-Z, Du M-Z, Liu S, Dong C, Guo F-B (2019) Vgas: a viral genome annotation system. *Front Microbiol* 10:184. <https://doi.org/10.3389/fmicb.2019.00184>
27. Zhang X, Pan W (2019) Exon prediction based on multiscale products of a genomic-inspired multiscale bilateral filtering. *PLoS One* 14:e0205050. <https://doi.org/10.1371/journal.pone.0205050>
28. Chen J, Liu Y, Liao Q, Liu B (2019) iEsGene-ZCPseKNC: identify essential genes based on Z curve pseudo k-tuple nucleotide composition. *IEEE Access* 165241–165247. <https://ieeexplore.ieee.org/document/8894693>
29. Wang J, Guo FB (2002) Base frequencies at the second codon position of *Vibrio cholerae* genes connect with protein function. *Biochem Biophys Res Commun* 290:81–84. <https://doi.org/10.1006/bbrc.2001.6174>
30. Tatusov RL, Koonin EV, Lipman DJ (1997) A genomic perspective on protein families. *Science* 278:631–637. <https://doi.org/10.1126/science.278.5338.631>
31. Tatusov RL, Fedorova ND, Jackson JD, Jacobs AR, Kiryutin B, Koonin EV et al (2003) The COG database: an updated version includes eukaryotes. *BMC Bioinform* 4:41. <https://doi.org/10.1186/1471-2105-4-41>
32. Han J, Pei J, Kamber M (2011) *Data mining: concepts and techniques*. Elsevier, Amsterdam
33. Shen W, Xiong J (2019) TaxonKit: a cross-platform and efficient NCBI taxonomy toolkit. *bioRxiv*. <https://doi.org/10.1101/513523>
34. Lane N, Martin W (2010) The energetics of genome complexity. *Nature* 467:929–934. <https://doi.org/10.1038/nature09486>
35. Médigue C, Rouxel T, Vigier P, Hénaut A, Danchin A (1991) Evidence for horizontal gene transfer in *Escherichia coli* speciation. *J Mol Biol* 222:851–856. [https://doi.org/10.1016/0022-2836\(91\)90575-Q](https://doi.org/10.1016/0022-2836(91)90575-Q)
36. Kunst F, Ogasawara N, Moszer I, Albertini AM, Alloni G, Azevedo V et al (1997) The complete genome sequence of the gram-positive bacterium *Bacillus subtilis*. *Nature* 390:249–256. <https://doi.org/10.1038/36786>
37. Das S, Ghosh S, Pan A, Dutta C (2005) Compositional variation in bacterial genes and proteins with potential expression level. *FEBS Lett* 579:5205–5210. <https://doi.org/10.1016/j.febslet.2005.08.042>
38. Angellotti MC, Bhuiyan SB, Chen G, Wan XF (2007) CodonO: codon usage bias analysis within and across genomes. *Nucleic Acids Res* 35:W132–W136. <https://doi.org/10.1093/nar/gkm392>
39. Atopkin DM, Semenchenko AA, Solodovnik DA, Ivashko YI, Vinnikov KA (2021) First next-generation sequencing data for Haploporidae (Digenea: Haploporata): characterization of complete mitochondrial genome and ribosomal operon for *Parasaccocoeilium mugili* Zhukov, 1971. *Parasitol Res*. <https://doi.org/10.1007/s00436-021-07159-y>
40. Ma BG, Chen LL (2005) The most deviated codon position in AT-rich bacterial genomes: a function related analysis. *J Biomol Struct Dyn* 23:143–149. <https://doi.org/10.1080/07391102.2005.10507055>
41. Guo FB, Yuan JB (2009) Codon usages of genes on chromosome, and surprisingly, genes in plasmid are primarily affected by strand-specific mutational biases in *Lawsonia intracellularis*. *DNA Res* 16:91–104. <https://doi.org/10.1093/dnares/dsp001>
42. Das S, Paul S, Chatterjee S, Dutta C (2005) Codon and amino acid usage in two major human pathogens of genus *Bartonella*—optimization between replicational-transcriptional selection, translational control and cost minimization. *DNA Res* 12:91–102. <https://doi.org/10.1093/dnares/12.2.91>
43. Chen WH, Lu G, Bork P, Hu S, Lercher MJ (2016) Energy efficiency trade-offs drive nucleotide usage in transcribed regions. *Nat Commun* 7:11334. <https://doi.org/10.1038/ncomms11334>
44. Mathé C, Peresetsky A, Déhais P, Van Montagu M, Rouzé P (1999) Classification of *Arabidopsis thaliana* gene sequences: clustering of coding sequences into two groups according to codon usage improves gene prediction. *J Mol Biol* 285:1977–1991. <https://doi.org/10.1006/jmbi.1998.2451>
45. Cai CZ, Han LY, Ji ZL, Chen X, Chen YZ (2003) SVM-Prot: web-based support vector machine software for functional classification of a protein from its primary sequence. *Nucleic Acids Res* 31:3692–3697. <https://doi.org/10.1093/nar/gkg600>
46. Han L, Cui J, Lin H, Ji Z, Cao Z, Li Y et al (2006) Recent progresses in the application of machine learning approach for predicting protein functional class independent of sequence similarity. *Proteomics* 6:4023–4037. <https://doi.org/10.1002/pmic.20050938>
47. Zhao B, Katuwawala A, Oldfield CJ, Dunker AK, Faraggi E, Gsponer J et al (2021) DescribePROT: database of amino acid-level protein structure and function predictions. *Nucleic Acids Res* 49:D298–d308. <https://doi.org/10.1093/nar/gkaa931>

Authors and Affiliations

Yan-Ting Jin^{1,2} · Cong Ma¹ · Xin Wang¹ · Shu-Xuan Wang¹ · Kai-Yue Zhang¹ · Wen-Xin Zheng³ · Zixin Deng² · Ju Wang⁴ · Feng-Biao Guo² 

✉ Ju Wang
wangju@tmu.edu.cn

✉ Feng-Biao Guo
fbguoy@whu.edu.cn

¹ School of Life Science and Technology, University of Electronic Science and Technology of China, Chengdu 611731, China

² Key Laboratory of Combinatorial Biosynthesis and Drug Discovery, Ministry of Education and School

of Pharmaceutical Sciences, Wuhan University, Wuhan 430071, China

³ School of Biomedical Engineering, Capital Medical University in Beijing, Beijing 100069, China

⁴ School of Biomedical Engineering, Tianjin Medical University, Tianjin 300070, China