BMC Bioinformatics

SOFTWARE

Open Access

GenRCA: a user-friendly rare codon analysis tool for comprehensive evaluation of codon usage preferences based on coding sequences in genomes

Kunjie Fan[1†], Yuanyuan Li[2†], Zhiwei Chen[2†] and Long Fan[1*]

†Kunjie Fan, Yuanyuan Li and Zhiwei Chen have contributed equally to this study.

*Correspondence: leo.fan@genscript.com

[1] Production and R&D Center I of LSS, GenScript (Shanghai) Biotech Co., Ltd., Shanghai, China
[2] Production and R&D Center I of LSS, GenScript Biotech Corporation, Nanjing, China

Abstract

Background: The study of codon usage bias is important for understanding gene expression, evolution and gene design, providing critical insights into the molecular processes that govern the function and regulation of genes. Codon Usage Bias (CUB) indices are valuable metrics for understanding codon usage patterns across different organisms without extensive experiments. Considering that there is no one-fits-all index for all species, a comprehensive platform supporting the calculation and analysis of multiple CUB indices for codon optimization is greatly needed.

Results: Here, we release GenRCA, an updated version of our previous Rare Codon Analysis Tool, as a free and user-friendly website for all-inclusive evaluation of codon usage preferences of coding sequences. In this study, we manually reviewed and implemented up to 31 codon preference indices, with 65 expression host organisms covered and batch processing of multiple gene sequences supported, aiming to improve the user experience and provide more comprehensive and efficient analysis.

Conclusions: Our website fills a gap in the availability of comprehensive tools for species-specific CUB calculations, enabling researchers to thoroughly assess the protein expression level based on a comprehensive list of 31 indices and further guide the codon optimization.

Keywords: Codon usage, Rare codon analysis, Protein expression, Gene design

Introduction

The Codon Usage Bias (CUB) refers to the non-random usage of synonymous codons encoding the same amino acid within a genome or set of genes [1]. It reflects the preference or bias in the selection of specific codons over others during translation. Various factors, including expression level, GC content, recombination rates, RNA stability, codon position, gene length, environmental stress, and population size, can influence CUB within and among species [2, 3]. Understanding CUB is important

as it provides insights into evolutionary processes, gene expression regulation, protein folding, and adaptation to different environments [4–6]. CUB indices are effective tools to study the pattern of codon usage bias, allowing for straightforward and computationally efficient evaluations of species-specific codon usage, eliminating the requirement for extra experiments, providing valuable insights into the genetic characteristics of organisms.

Over the past few years, a number of measures have been proposed to quantify CUB. There exist some free websites that support the calculation of codon usage preferences, as outlined in Table 1. Among them, the first version of our publicly available website, GenScript Rare Codon Analysis, initially launched in 2008, has gained significant popularity with a high daily user visit count and over 300 citations by supporting 3 types of CUB indices and 17 species along with informative visualizations. As for other websites, they either only provide basic codon usage frequency distribution or support simply one index, which is not useful. Given that there is no universal index that fits all species, there is a significant need for a comprehensive platform that supports the calculation of multiple CUB indices for various species, enabling thorough evaluation of codon usage preferences for coding sequences.

Hence, in this study, we release GenRCA, a user-friendly website for all-inclusive rare codon analysis, which is freely available at https://www.genscript.com/tools/rare-codon-analysis. Compared to our previous version, the available number of codon preference indices is extended from 3 to 31 and the number of supported expression host species is increased from 17 to 65, enabling users to explore and compare codon usage biases across a wider range of species in a more comprehensive manner (Fig. 1). Furthermore, we introduced a new feature that enables batch processing of multiple gene sequences. With a number of supported indices and species, informative visualizations, and user-friendly interface, our website holds great potential to improve the overall convenience and accuracy of protein expression optimization (Table 2).

Table 1 Description of existing rare codon analysis websites

URL	Content	Species	Batch	Download
GenRCA rare Codon Analysis Tool (genscript.com)	Codon usage frequency distribution 31 CUB indices	65	Yes	Yes
https://www.biologicscorp.com/tools/RareCodonAnalyzer	Codon usage frequency distribution 1 CUB index (CAI)	14	No	No
http://www.detaibio.com/tools/rare-codon-analyzer.html	Codon usage frequency distribution	14	No	Yes
E. coli codon usage analyzer (ucr.edu)	Codon usage frequency distribution	1	No	No
Rare Codon Caltor, Programmed by Edmund Ng (ucla.edu)	Frequency of codon occurrence	1	No	No
Rare Codon Search (bioline.com)	Searching for rare codons	6	No	No
http://www.bitgene.net/dna/rare_codon	Codon usage frequency distribution	1	No	No

"Batch" refers to whether the website supports batch processing of multiple sequences. "Download" denotes whether the website supports the download of analysis report

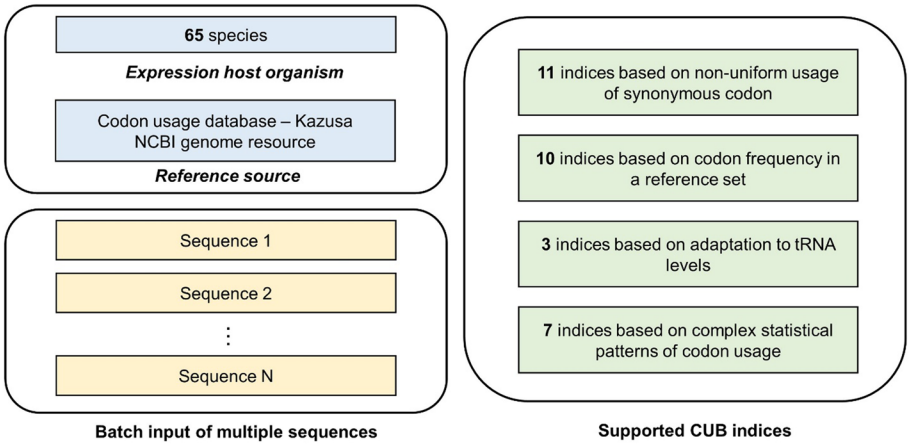

Fig. 1 Overview of supported functionalities in GenRCA website. GenRCA includes the calculation of 31 CUB indices for 65 species and two reference sources, as well as the batch processing of multiple input sequences

Table 2 List of 31 supported indices on our website

Category	Indices
Indices based on non-uniform usage of synonymous codon	RSCU (Relative Synonymous Codon Usage) [7]
	ENC (Effective Number of Codons) [8, 9]
	RCBS (Relative Codon Bias Strength) [10]
	DCBS (Directional Codon Bias Score) [11]
	CDC(Codon Deviation Coefficient) [12]
	MILC (Measure Independent of Length and Composition) [13]
	ICDI (Intrinsic Codon Deviation Index) [14]
	SCUO (Synonymous Codon Usage Order) [15, 16]
	Ew (Weighted Sum of Relative Entropy) [17]
	P (Codon Preference) [18]
	MCB (Maximum-likelihood Codon Bias) [19]
Indices based on codon frequency in a reference set of genes	CAI (Codon Adaptation Index) [20]
	FOP (Frequency of Optimal Codons) [21, 22]
	COUSIN (Codon Usage Similarity Index) [23]
	CBI (Codon Bias Index) [24]
	Dmean (Mean Dissimilarity-based Index) [25]
	RCA (Relative Codon Adaptation) [26]
	CUFS (Codon Usage Frequency Similarity) [27]
	B (Codon Usage Bias) [28]
Indices based on adaptation to the tRNA levels and their supply	tAI (tRNA Adaptation Index) [29]
	gtAI (Genetic tRNA Adaptation Index) [30]
	P2 index [31]
Indices based on complex patterns of codon usage	GC content [32]
	ENcp (Effective Number of Codon Pairs) [33]
	CPS (Codon Pair Score) [34, 35]
	Codon volatility [36]

COUSIN includes COUSIN18 and COUSIN59

GC content can be subdivided into GC, GC1, GC2, GC3

Implementation

Our implemented rare codon analysis tool is provided as a free and user-friendly website. We manually collected almost all available CUB-related papers and re-implemented 31 proposed methods in Python programming language. We also aggregated reference genome for up to 65 species into our website. We detail the supported functionalities and workflow of our website in the following section.

Supported indices and species

To achieve a more comprehensive evaluation of rare codon usage, 31 commonly utilized indices and 2 motif-based metrics were implemented and integrated them into our website. These indices can be categorized into four groups according to the way they process the gene expression, as outlined in Table 1. Detailed description of all these included indices is in the Supplementary material.

The first category calculates the deviation of codon usage frequency from a "uniform" distribution, which provides an informative measurement of codon usage bias without requiring prior knowledge, indicating selection's influence on gene expression levels, such as RSCU (Relative Synonymous Codon Usage) [7] and ENC (Effective Number of Codons) [8, 9]. Codon bias indices in the second category compare codon frequency between a reference set of genes and the host organism, using different methodologies to calculate similarity scores and identify coding sequences with higher gene expression based on closely resembling codons in the reference set, such as CAI (Codon Adaptation Index) [20]and FOP (Frequency of Optimal Codons) [21, 22]. Considering that codons decoded by more frequent tRNAs are used more frequently, the third type of methods are proposed based on this positive correlation between tRNA levels and codon usage, including tAI (tRNA Adaptation Index) [29] and P2 index [31]. To model the influence of longer sequences and regulatory codes on gene expression and intracellular processes, the fourth category of methods employ advanced statistical methods to analyze complex patterns of codon usage, such as GC content [32] and ENcp (Effective Number of Codon Pairs) [33].

Our website offers an extensive selection of 65 expression host species, with each reference species providing users with access to two codon tables. One of codon tables is directly cited from commonly used Codon Usage Database (https://www.kazusa.or.jp/codon/), while the other is calculated on the basis of genomic coding sequences downloaded from well-annotated CDS of NCBI Genomes FTP (https://www.ncbi.nlm.nih.gov/home/genomes/). Detailed information about supported species can be found in the Supplementary materials.

Web server

The website interface of our Rare Codon Analysis Tool allows users to submit one or multiple sequences and choose from a comprehensive list of 65 expression host organisms. Users can submit sequences and receive analysis results in just three simple steps:

Step 1: Select your desired expression host organism along with their reference sources.

Step 2: Enter your DNA or RNA sequence(s) into the table.

Step 3: Click "Analysis" button to view detailed analysis results and attached references.

As indicated in Fig. 2A, when multiple sequences are provided, unique gene names should be specified. You can directly copy multiple sequences from your file and paste them into the spreadsheet in our website. We implemented a "Load examples" button that, when clicked, loads two selected sequences into the spreadsheet to serve as a simple illustrative example. We restrict our analysis to sequences with lengths between 60 and 12,000 bp. Once sequences are provided, several preprocessing steps will be conducted. Firstly, we remove all symbols apart from alphabetic characters, including special symbols such as \t and \n. Additionally, any stop codons located at the end of the sequence will be eliminated. If a stop codon or concatenated bases appears in the middle, which refer to sequences containing repeated or continuous bases that do not form valid codons, it will be excluded from further analysis. Additionally, sequences whose lengths are not multiples of 3 will be removed, as they are not valid coding sequences.

Once the "Analysis" button is clicked, the comprehensive analysis results would be displayed in just a few seconds. The results on our website are presented in an interactive manner, allowing users to choose metrics of most interest to display on the top from 31 indices (Fig. 2B), serving as a quick view of the analysis. Indices shown on the top is totally customized and determined by the user. By placing the mouse over the index in the navigation bar, users can view the corresponding description and mathematical definition. In the main results panel, a graphical representation of CAI is first presented, followed by two motif-based values that have been proved to be related to gene expression: number of negative CIS elements and number of negative repeat elements. Then, detailed results and reference ranges for all 31 indices are provided below, organized into

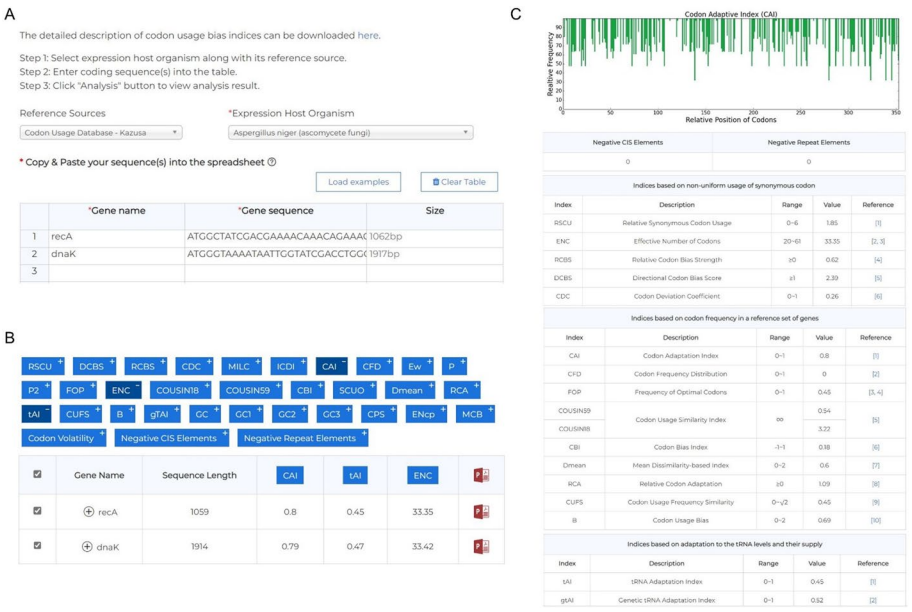

Fig. 2 GenRCA website interface. **A** The input interface of GenRCA. Users need to choose the reference source and expression host organism, and input the queried sequences. **B** Highlighted indices of interest shown on the top. Users can customize which metrics to show on the top by clicking the corresponding button, and learn the detailed description of the index by placing the mouse over the button. **C** Main results panel. The graphical representation of CAI, two motif-based metrics and 31 CUB indices are displayed

four sections based on their categories (Fig. 2C). Additionally, users can download an analysis report in PDF format for further reference.

Results and discussion

The main contribution of our study is integrating a large number of CUB indices for a comprehensive rare codon analysis, compared to available tools (Table 1). As an example, we used recA gene as a query to search on two existing tools: rare codon analyzer on Biologics International Corp and Detaibio. Both tools presented only codon usage frequency distribution and codon adaptive index (CAI) score (Fig. 3), while our tool provides thirty more CUB indices for enriched information (Fig. 2C).

To validate the usefulness and necessity of a comprehensive rare codon analysis tool, we conducted correlation analysis between protein expression and a set of CUB indices for four species: *Saccharomyces cerevisiae* [37], *Cricetulus griseus* [38], Escherichia coli [39], and Mus musculus [40], based on Spearman correlation coefficients. We extracted protein IDs and protein expression values from the supplementary files of these four papers, and obtained corresponding DNA coding sequences by searching on the Ensembl database. Then, all indices were calculated by inputting coding sequences to our website (Supplementary data). Surprisingly, it is found that indices showing the strongest correlation with protein expression differ significantly among species, suggesting that there is no universally applicable index for all species (Fig. 4A). Notably, commonly used indices for assessing protein expression, such as CAI, did not perform as expected across these species, which is consistent with previous studies [41–43], indicating that there is no one-fits-all index for all species. Hence, our website offers up to 31 indices for researchers to comprehensively assess the protein expression levels rather than only relying on one or two commonly used indices. We also conducted a principal component analysis (PCA) on the Spearman correlation results used above, and showed a scatter plot of these four species. As shown in Fig. 4B, Saccharomyces cerevisiae and Escherichia coli, both unicellular organisms, are clustered closely, while being far away from other two multicellular organisms, indicating that the use of multiple CUB indices can reveal evolutionary processes. In the future, these comprehensive list of 31 indices

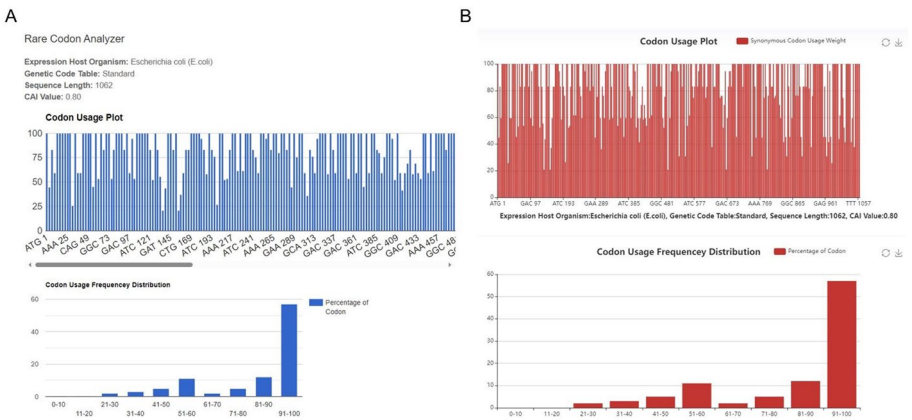

Fig. 3 Results shown on the website of available tools. recA gene is used as an example, which is the same as in Fig. 2. **A** https://www.biologicscorp.com/tools/RareCodonAnalyzer **B** http://www.detaibio.com/tools/rare-codon-analyzer.html

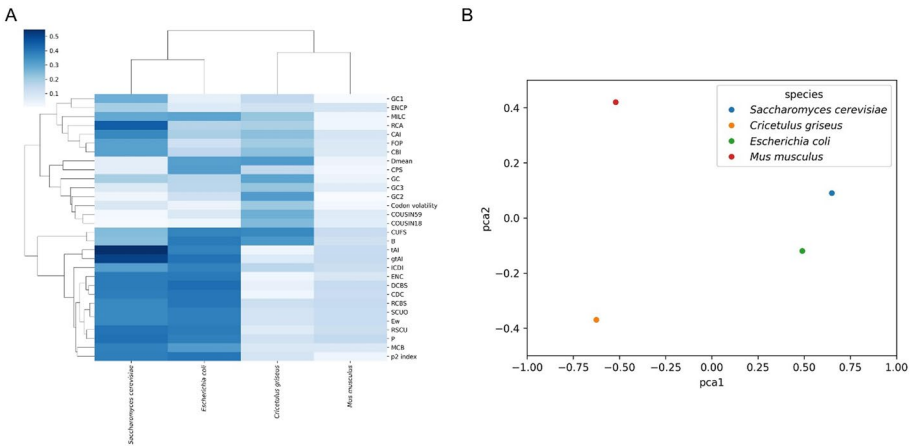

Fig. 4 Correlation analysis between protein expression and CUB indices. **A** Heatmap of Spearman correlation coefficients between protein expression and CUB indices for four species with dendrogram. **B** PCA plot of Spearman correlation coefficients for four species

can serve as input features for a machine learning model to better predict the protein expression. For example, a simple linear regression model could be trained to associate the relationship between the expression level and a set of 31 indices.

Though being the most comprehensive rare codon analysis website to date, our proposed website still has some limitations, in terms of the incompleteness of supported species and reference sources. In the future, we will integrate more types of species into our website, aiming to include as many species as possible from the Codon Usage Database and the NCBI genome database. Besides, we will regularly update reference sources information, such as synchronizing periodically with the Codon Usage Database. Moreover, we may allow users to customize species and reference sources by uploading a codon table.

Conclusions

In this work, we made significant updates to our highly cited rare codon analysis website, which allows users to evaluate and determine whether codon optimization is necessary to enhance gene expression in the target host organism. We expanded the number of supported CUB indices to 31 and included codon tables for 65 expression host species. Additionally, we incorporated a batch processing feature, allowing users to conveniently analyze multiple sequences simultaneously, thereby improving the overall user-friendliness of the website. Considering that there is no single index suitable for every species, our website offers opportunities for researchers to comprehensively evaluate the protein expression level of coding sequences by considering all 31 supported indices together for their species of interest.

Abbreviations

CUB	Codon usage bias
RSCU	Relative synonymous codon usage
ENC	Effective number of codons
CAI	Codon adaptation index
FOP	Frequency of optimal codons
tAI	TRNA adaptation index
ENcp	Effective number of codon pairs

PCA Principal component analysis

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12859-024-05934-z.

Additional file1 (DOCX 121 KB)

Additional file2 (XLSX 4102 KB)

Author contributions

F.K. and L.Y wrote the main manuscript text, F.K. prepared all figures, L.Y and C.Z implemented the software, F.L designed the study. All authors reviewed the manuscript.

Funding

This study was supported by Pearl Plan (Pearl Elite Talent Award) of Pudong New Area of Shanghai Municipality.

Availability of data and materials

Project name: GenRCA rare codon analysis tool. Project home page: https://www.genscript.com/tools/rare-codon-analysis. Operating system: Platform independent. Programming language: Python 3. Other requirements: Not applicable. License: Apache License 2.0. Any restrictions to use by non-academics: None. The Datasets used in this study are available in the Supplementary data file.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they are all employees of GenScript Biotech Corporation, or GenScript (Shanghai) Biotech Co., Ltd, which is a wholly-owned subsidiary of GenScript Biotech Corporation. The company has no financial or personal interests that could potentially influence the research findings or the objectivity of the authors.

Received: 20 December 2023 Accepted: 17 September 2024
Published online: 27 September 2024

References

1. Bahiri-Elitzur S, Tuller T. Codon-based indices for modeling gene expression and transcript evolution. Comput Struct Biotechnol J. 2021;19:2646–63.
2. Hershberg R, Petrov DA. Selection on codon bias. Annu Rev Genet. 2008;42:287–99.
3. Parvathy ST, Udayasuriyan V, Bhadana V. Codon usage bias. Mol Biol Rep. 2022;49:539–65.
4. Liu Y. A code within the genetic code: codon usage regulates co-translational protein folding. Cell Commun Signal. 2020;18:145.
5. Athey J, et al. A new and updated resource for codon usage tables. BMC Bioinf. 2017;18:391.
6. Quax TEF, Claassens NJ, Söll D, van der Oost J. Codon bias as a means to fine-tune gene expression. Mol Cell. 2015;59:149–61.
7. Sharp PM, Li WH. An evolutionary perspective on synonymous codon usage in unicellular organisms. J Mol Evol. 1986;24:28–38.
8. Wright F. The 'effective number of codons' used in a gene. Gene. 1990;87:23–9.
9. Satapathy SS, Sahoo AK, Ray SK, Ghosh TC. Codon degeneracy and amino acid abundance influence the measures of codon usage bias: improved Nc (N_c) and ENCprime (N'_c) measures. Genes Cells. 2017;22:277–83.
10. Roymondal U, Das S, Sahoo S. Predicting gene expression level from relative codon usage bias: an application to *Escherichia coli* genome. DNA Res. 2009;16:13–30.
11. Sabi R, Tuller T. Modelling the efficiency of codon–tRNA interactions based on codon usage bias. DNA Res. 2014;21:511–26.
12. Zhang Z, et al. Codon Deviation Coefficient: a novel measure for estimating codon usage bias and its statistical significance. BMC Bioinf. 2012;13:43.
13. Supek F, Vlahoviček K. Comparison of codon usage measures and their applicability in prediction of microbial gene expressivity. BMC Bioinf. 2005;6:182.
14. Freire-Picos MA, et al. Codon usage in Kluyveromyces lactis and in yeast cytochrome c-encoding genes. Gene. 1994;139:43–9.
15. Wan X-F, Xu D, Kleinhofs A, Zhou J. Quantitative relationship between synonymous codon usage bias and GC composition across unicellular genomes. BMC Evol Biol. 2004;4:19.
16. Wan X-F, Zhou J, Xu D. CodonO: a new informatics method for measuring synonymous codon usage bias within and across genomes. Int J Gen Syst. 2006;35:109–25.

17. Suzuki H, Saito R, Tomita M. The 'weighted sum of relative entropy': a new index for synonymous codon usage bias. Gene. 2004;335:19–23.
18. Gribskov M, Devereux J, Burgess RR. The codon preference plot: graphic analysis of protein coding sequences and prediction of gene expression. Nucleic Acids Res. 1984;12:539–49.
19. Urrutia AO, Hurst LD. Codon usage bias covaries with expression breadth and the rate of synonymous evolution in humans, but this is not evidence for selection. Genetics. 2001;159:1191–9.
20. Sharp PM, Li WH. The codon adaptation index-a measure of directional synonymous codon usage bias, and its potential applications. Nucleic Acids Res. 1987;15:1281–95.
21. Ikemura T. 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. 1981;151, 389–409.
22. Ikemura T. Correlation between the abundance of yeast transfer RNAs and the occurrence of the respective codons in protein genes. J Mol Biol. 1982;158:573–97.
23. Bourret J, Alizon S, Bravo IG. COUSIN (COdon usage similarity INdex): a normalized measure of codon usage preferences. Genome Biol Evol. 2019;11:3523–8.
24. Bennetzen JL, Hall BD. Codon selection in yeast. J Biol Chem. 1982;257:3026–31.
25. Suzuki H, Saito R, Tomita M. Measure of synonymous codon usage diversity among genes in bacteria. BMC Bioinf. 2009;10:167.
26. Fox JM, Erill I. Relative codon adaptation: a generic codon bias index for prediction of gene expression. DNA Res. 2010;17:185–96.
27. Diament A, Pinter RY, Tuller T. Three-dimensional eukaryotic genomic organization is strongly correlated with codon usage expression and function. Nat Commun. 2014;5:5876.
28. Karlin S, Mrázek J, Campbell AM. Codon usages in different gene classes of the *Escherichia coli* genome. Mol Microbiol. 1998;29:1341–55.
29. Reis Md. Solving the riddle of codon usage preferences: a test for translational selection. Nucleic Acids Res. 2004;32, 5036–5044.
30. Anwar AM, et al. gtAI: an improved species-specific tRNA adaptation index using the genetic algorithm. Front Mol Biosci. 2023;10:1218518.
31. Gouy M, Gautier C. Codon usage in bacteria: correlation with gene expressivity. Nucleic Acids Res. 1982;10:7055–74.
32. Stenico M, Lloyd AT, Sharp PM. Codon usage in *Caenorhabditis elegans*: delineation of translational selection and mutational biases. Nucleic Acids Res. 1994;22:2437–46.
33. Alexaki A, et al. Codon and codon-pair usage tables (CoCoPUTs): facilitating genetic variation analyses and recombinant gene design. J Mol Biol. 2019;431:2434–41.
34. Kunec D, Osterrieder N. Codon pair bias is a direct consequence of dinucleotide bias. Cell Rep. 2016;14:55–67.
35. Coleman JR, et al. Virus attenuation by genome-scale changes in codon pair bias. Science. 2008;320:1784–7.
36. Plotkin JB, Dushoff J, Fraser HB. Detecting selection using a single genome sequence of *M. tuberculosis* and *P. falciparum*. Nature 428, 942–945 (2004).
37. Ghaemmaghami S, et al. Global analysis of protein expression in yeast. Nature. 2003;425:737–41.
38. Baycin-Hizal D, et al. Proteomic analysis of Chinese hamster ovary cells. J Proteome Res. 2012;11:5265–76.
39. Lu P, Vogel C, Wang R, Yao X, Marcotte EM. Absolute protein expression profiling estimates the relative contributions of transcriptional and translational regulation. Nat Biotechnol. 2007;25:117–24.
40. Schwanhäusser B, et al. Global quantification of mammalian gene expression control. Nature. 2011;473:337–42.
41. Welch M, et al. Design parameters to control synthetic gene expression in *Escherichia coli*. PLoS ONE. 2009;4: e7002.
42. Kudla G, Murray AW, Tollervey D, Plotkin JB. Coding-sequence determinants of gene expression in *Escherichia coli*. Science. 2009;324:255–8.
43. Friberg M, von Rohr P, Gonnet G. Limitations of codon adaptation index and other coding DNA-based features for prediction of protein expression inSaccharomyces cerevisiae. Yeast. 2004;21:1083–93.

Publisher's Note