frontiers | Frontiers in Neuroscience

ORIGINAL RESEARCH
published: 04 July 2022
doi: 10.3389/fnins.2022.895607

Codon Usage Bias Correlates With Gene Length in Neurodegeneration Associated Genes

Rekha Khandia[1], Mohd. Saeed[2], Ahmed M. Alharbi[3], Ghulam Md. Ashraf[4,5], Nigel H. Greig[6] and Mohammad Amjad Kamal[7,8,9,10]*

[1] Department of Biochemistry and Genetics, Barkatullah University, Bhopal, India, [2] Department of Biology, College of Sciences, University of Hail, Hail, Saudi Arabia, [3] Department of Clinical Laboratory Sciences, College of Applied Medical Sciences, University of Hail, Hail, Saudi Arabia, [4] Pre-clinical Research Unit, King Fahd Medical Research Center, King Abdulaziz University, Jeddah, Saudi Arabia, [5] Department of Medical Laboratory Sciences, Faculty of Applied Medical Sciences, King Abdulaziz University, Jeddah, Saudi Arabia, [6] Drug Design and Development Section, Translational Gerontology Branch, Intramural Research Program National Institute on Aging, NIH, Baltimore, MD, United States, [7] Institutes for Systems Genetics, Frontiers Science Center for Disease-Related Molecular Network, West China Hospital, Sichuan University, Chengdu, China, [8] King Fahd Medical Research Center, King Abdulaziz University, Jeddah, Saudi Arabia, [9] Department of Pharmacy, Faculty of Allied Health Sciences, Daffodil International University, Dhaka, Bangladesh, [10] Enzymoics, Novel Global Community Educational Foundation, Hebersham, NSW, Australia

OPEN ACCESS

Edited by:
Khurshid Ahmad, Yeungnam University, South Korea

Reviewed by:
Simona Cavalu, University of Oradea, Romania
Ratna Prabha, Indian Council of Agricultural Research, India

***Correspondence:**
Rekha Khandia
rekha.khandia@bubhopal.ac.in; bu.rekha.khandia@gmail.com

Specialty section:
This article was submitted to Neurogenomics, a section of the journal Frontiers in Neuroscience

Received: *14 March 2022*
Accepted: *08 June 2022*
Published: *04 July 2022*

Citation:
Khandia R, Saeed M, Alharbi AM, Ashraf GM, Greig NH and Kamal MA (2022) Codon Usage Bias Correlates With Gene Length in Neurodegeneration Associated Genes. Front. Neurosci. 16:895607. doi: 10.3389/fnins.2022.895607

Codon usage analysis is a crucial part of molecular characterization and is used to determine the factors affecting the evolution of a gene. The length of a gene is an important parameter that affects the characteristics of the gene, such as codon usage, compositional parameters, and sometimes, its functions. In the present study, we investigated the association of various parameters related to codon usage with the length of genes. Gene expression is affected by nucleotide disproportion. In sixty genes related to neurodegenerative disorders, the G nucleotide was the most abundant and the T nucleotide was the least. The nucleotide T exhibited a significant association with the length of the gene at both the overall compositional level and the first and second codon positions. Codon usage bias (CUB) of these genes was affected by pyrimidine and keto skews. Gene length was found to be significantly correlated with codon bias in neurodegeneration associated genes. In gene segments with lengths below 1,200 bp and above 2,400 bp, CUB was positively associated with length. Relative synonymous CUB, which is another measure of CUB, showed that codons TTA, GTT, GTC, TCA, GGT, and GGA exhibited a positive association with length, whereas codons GTA, AGC, CGT, CGA, and GGG showed a negative association. GC-ending codons were preferred over AT-ending codons. Overall analysis indicated that the association between CUB and length varies depending on the segment size; however, CUB of 1,200–2,000 bp gene segments appeared not affected by gene length. In synopsis, analysis suggests that length of the genes correlates with various imperative molecular signatures including A/T nucleotide disproportion and codon choices. In the present study we additionally evaluated various molecular features and their correlation with different indices of codon

usage, like the Codon Adaptation Index (CAI) and Relative Dynonymous Codon Usage (RSCU) of codons. We also considered the impact of gene fragment size on different molecular features in genes related to neurodegeneration. This analysis will aid our understanding of and in potentially modulating gene expression in cases of defective gene functioning in clinical settings.

Keywords: gene length, compositional bias, codon usage bias, neurodegeneration, codon preference

INTRODUCTION

Several disorders, including neurodegenerative disorders, have been linked to age-related illnesses and dementias. Neurodegenerative disorders, of which there are multiple, relate to diseases instigated by dysfunction of neurons and pose a significant hazard to mental health (Gitler et al., 2017). They can severely impact a patient's ability to move, speak, think, and even breathe (Przedborski et al., 2003). Ataxias (problems with mobility) and dementias are also common to neurodegeneration. Neurodegenerative disorders occur due to either inherited genetic changes, wherein defective genes are passed from one generation to the next (e.g., Huntington's disease and familial Alzheimer's disease) or a combination of hereditary and environmental factors (e.g., sporadic Parkinson's disease which may potentially be caused by long-term exposure to toxic chemicals and/or pesticides, or an initiating factor such as head injury). However, aging is the most well-known factor contributing to neurodegenerative disorders (Research et al., 2017). Attempts are being conducted to halt or at least reduce the progression of neurodegeneration, and numerous natural and synthesized compounds as well as life-style changes are being evaluated in this context (Katsuno et al., 2012; Martier and Konstantinova, 2020; Sharifi-Rad et al., 2020; Bhattacharya et al., 2022; Onikanni et al., 2022). Dementia is described as cognitive and behavioral impairment involving "functional" impairments that significantly impact one's daily activities (Josephs et al., 2009). Dysfunction and degeneration of brain cells and their interconnections can lead to irreversible dementia. Alzheimer's disease (AD) dementia, vascular dementia (VaD), Lewy body dementia (LBD), and frontotemporal lobar dementia (FTD) are the four leading types of neurodegenerative dementias. The most well-known form of dementia, Alzheimer's disease (AD), accounts for approximately 60% of all dementias, with the remaining 40% deriving from VaD, LBD, and FTD (Ripich and Horner, 2004). Major risk factors for dementia include hypertension, smoking, obesity, depression, physical inactivity, diabetes, minimal social interaction, hearing impairment, excessive alcohol consumption, traumatic brain injury, and air pollution (including high nitrogen oxide and carbon monoxide concentration) (Livingston et al., 2020). Interestingly, a genetic connection also has been found for dementia, and, in this context, next-generation sequencing has made large-scale molecular studies feasible. The genes associated with neurodegeneration need to be evaluated to decipher the molecular features of the subset of genes involved in dementias.

The role of various factors, such as selection, mutation, and composition in the evolution of a gene can be determined via studies on codon usage bias (CUB). CUB also provides insights into the differential architecture of a gene based on the gene's function. Several synonymous mutations are involved in various ailments and underscore the fact that silent mutations can be detrimental. Furthermore, several studies have focused on various factors that affect CUB within and between species, such as the expression level, RNA stability, recombination rates, GC content, codon position, and gene length (Grishkevich and Yanai, 2014). Gene length is a particularly crucial parameter in the study of CUB that increases with time (partly owing to the insertion of transposable elements) and decreases with partial gene duplication. However, long gene length and high gene expression contribute equally to gene duplication and alternative splicing. In contrast, short gene length and low gene expression result in large gene families; thus pointing toward the origin of new trends in the genome (Behura and Severson, 2012). According to Urrutia and Hurst (2003), smaller gene length results in higher expression, smaller proteins, high amino acid bias, high codon bias, and less intronic material, which can evoke selective pressure to increase the efficiency of protein synthesis. The relevance of gene length to the early transcriptional responses of human fibroblasts toward serum stimulation has been studied, and the genome-wide transcriptional response was reported to be affected by gene length. Length indirectly regulates gene expression, i.e., shorter genes lead to faster protein production and subsequently participate in the control of longer proteins, which are produced later in the reaction (Kirkconnell et al., 2017). Many studies have suggested that gene length is involved in various biological processes that can ultimately lead to disorders such as cancer, neuronal dysfunction, and cardiomyopathies. Studies on different multigenic diseases show that gene length and splicing complexity are partially separated in defining cancer-linked pathways (Sahakyan and Balasubramanian, 2016).

Human genome studies on the functions of gene length, especially protein-coding genes, suggest that longer genes are expressed more in the brain, as well as in heart conditions and cancer. In contrast, smaller genes are involved mainly in the immune system and skin development; hence, genes with longer transcripts are predominantly associated with functions in the initial phase of development, whereas genes with shorter transcripts are crucial in everyday functions (Lopes et al., 2021). Accordingly, the compositional analysis of transcripts of AD

Abbreviations: CUB, Codon usage bias; A, adenine; T, Thymine; G, Guanine; C, Cytosine; RSCU, Relative synonymous codon usage; ENC, Effective number of codons; CAI, Codon Adaptation Index; SCS, Scaled chi-square; PCA, Principal component analysis.

shows that GC-ending codons are heavily skewed in the protein sequences (Yang et al., 2010), which suggests that the gene expression level plays a major role in the alteration of genes (Kalisz and Purugganan, 2004).

Gene length has been associated with gene expression in bacteria (Chiaromonte et al., 2003). In animals, short genes show strong expression, whereas, in plants, long genes show strong expression (Ren et al., 2007). Similarly, composition and gene length are correlated; short genes (<2,000 bp) show an increased percentage of GC3 than do long genes (Yang et al., 2021). A significant positive correlation has been reported between gene length and synonymous CUB in *Escherichia coli*, whereas a negative correlation has been reported in *Drosophila melanogaster* and *Saccharomyces cerevisiae*, which depicts the strong bias toward longer genes to enhance translational efficiency because of selection pressure (Moriyama and Powell, 1998). A significant correlation between the effective number of codons (ENC) and different skews (GC, AT, purine, pyrimidine, amino, and keto skews) has also been reported in *Wuchereria bancrofti* and *Schistosoma haematobium*, wherein all the skews were negative in both the organisms, except for that of pyrimidine in *S. haematobium*. This suggests a significant role of CUB on nucleotide disproportion (Mazumder et al., 2017); however, the different skews represent a marker for a specific genus and species.

The major purpose of this study was to analyze the characteristics that affect codon usage in neurodegenerative genes using various parameters such as relative synonymous codon usage (RSCU), ENC, nucleotide composition, and nucleotide skew calculations. An analysis of codon usage helps us better appreciate the role of gene length and evolutionary processes in determining codon usage of genes that may cause neurodegenerative diseases, thereby expanding our knowledgebase to improve current treatments and develop new therapeutic strategies.

MATERIALS AND METHODS

Sequence Retrieval and Analysis

The genes responsible for neurodegeneration were obtained from the NCBI Genetic Testing Registry NGS Neurodegenerative disorders Multi-Gene Panel. A total of 183 transcripts belonging to 60 genes were downloaded and analyzed for the study. Both the start and stop codons were in the reading frame of the qualified genes and there were no ambiguous codons.

Analysis of Composition and GC Content

The nucleotide components of the transcripts were analyzed. The overall percentage composition and percentage at the first, second, and third codon positions were also determined. Furthermore, the overall GC content and that at the three codon positions were determined. Compositional analysis was performed using the CAIcal software, as developed by Puigbò et al. (2008).

Determination of Relative Synonymous Codon Usage

Not all 59 codons (excluding methionine and tryptophan encoded by a single codon and stop codons) are used equally to encode a protein. A bias in usage of the codon is depicted as an RSCU value, which denotes the relative frequency for which a codon is used, as compared to other codons of the family.

Codon Adaptation Index

The Codon Adaptation Index (CAI) measures the similarity between codon usage and a reference set. It is a directional measure of codon bias and a predicting tool for the level of gene expression in an organism (Sharp and Li, 1987). The CAI values were calculated using the COUSIN tool, as developed by Bourret et al. (2019). The value of the CAI ranges between zero and one.

Effective Number of Codons Determination

The ENC is a non-directional measure of codon bias (Song et al., 2017) with values ranging from 20 to 61. An ENC of 20 shows the highest bias, which means that only one codon of the codon family will be used for coding a single amino acid. An ENC of 61 shows the most negligible bias, indicating that all codons that are coding for a single amino acid will be used (Wang et al., 2018). The ENC values were calculated using the COUSIN tool developed by Bourret et al. (2019).

Data Segregation Based on Length

The lengths of the transcripts ranged from 168 to 4,536 bp. Eight groups were prepared for a detailed investigation. The groups encompassed segments 1 (1–400 bp), 2 (400–800 bp), 3 (800–1,200 bp), 4 (1,200–1,600 bp), 5 (1,600–2,000 bp), 6 (2,000–2,400 bp), and 7 (2,400–4,500 bp). The eighth group contained full-length transcripts.

Skew Calculations

Nucleotide skew indicates the bias in nucleotide usage. In a single strand of DNA, the frequencies of bases are not necessarily equal to those of its complementary strand; this difference often arises at the time of replication. For example, in leading strands, G and T are more abundant than C and A, which is attributed to mutational forces (Charneski et al., 2011). The skew is the normalized excess of one nucleotide over another in any given sequence (nucleotide A – nucleotide B/nucleotide A + nucleotide B). If the GC bias is zero, nucleotide A = nucleotide B.

Scaled Chi-Square

Scaled chi-square (SCS) is another index for determining the codon bias, and it is a directional measure of CUB. Chi-square values indicate the deviation from the expected value; as the chi-square value highly depends on the length of the gene, the value is scaled by dividing the chi value by the number of codons present in a gene, excluding methionine and tryptophan, which do not contribute in chi. Therefore, using the SCS, genes of varying lengths can be compared (Shields and Sharp, 1987).

Correlation Between Length and Codon Usage Bias

To investigate the effects of CUB on gene segments of various lengths, we divided our genes into different-sized segments. Eight groups were prepared: one group had complete gene lengths, whereas the other groups contained segments of 1–400, 400–800, 800–1,200, 1,200–1,600, 1,600–2,000, 2,000–2,400, and 2,400–4,500 bp.

Statistical Analysis

Routine calculations such as addition and subtraction were performed in Microsoft Excel 2010. Statistical analysis, such as correlation analysis and principal component analysis (PCA), was performed using the Minitab 17 statistical software.

RESULTS

Compositional Analysis

The compositional analysis showed that the overall mean GC content was higher than the AT content (52.41 and 47.58%, respectively). The same trend was observed in the composition at the first and third codon positions; however, at the second codon position, the AT2 content was greater than the GC2 content (58.94 and 41.05%, respectively). At the third codon position, a variance was observed for the A and G nucleotides (36.3 and 52.71%, respectively). The lowest variance was observed for T and G in the overall composition (13.82 and 9.74%, respectively) and at codon positions one (8.90 and 11.13%, respectively) and two (9.39 and 5.21%, respectively). Nucleotide composition at different codon positions is presented in **Figure 1**.

The GC content at the third codon position was widely distributed (from 27.73 to 89.60%). The distribution of the GC content at the second codon position was the lowest, ranging from 30.13 to 52.33% with one outlier at 76.82% (**Figure 2**).

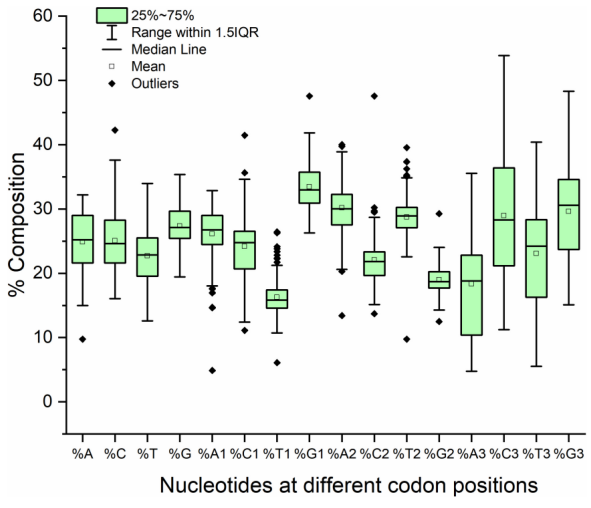

FIGURE 1 | Box plot with outliers presenting nucleotide composition in neurodegeneration-related genes.

Correlation of Length With Composition

Correlation analysis is a statistical method used to determine the strength of a relationship between two variables. Such a relation could be either positive or negative. A negative correlation indicates that upon increasing one variable the second variable decreases and vice versa. In a positive correlation with an increase in one variable the other variable also increases. A higher correlation coefficient indicates a strong relationship, whereas a lower correlation indicates a weaker relationship. Correlation analysis was performed between the 20 compositional constraints (%A, %C, %T, %G, %GC, %A1, %C1, %T1, %G1, %GC1, %A2, %C2, %T2, %G2, %GC2, %A3, %C3, %T3, %G3, and %GC3) and the length of transcripts. The length was positively correlated with the overall nucleotide T component (T, $r = 0.162$, $p < 0.05$) and the T component at the first (T1, $r = 0.147$, $p < 0.05$) and second (T2, $r = 0.258$, $p < 0.001$) codon positions. A negative correlation was observed between C1 ($r = -0.200$, $p < 0.01$) and G2 ($r = -0.198$, $p < 0.05$).

Correlation Between Length and Nucleotide Disproportion

All six nucleotide skews were subjected to correlation analysis with the length of genes. The length was negatively correlated with the AT skew ($r = -0.204$, $p < 0.01$). The nucleotide skews had no correlation with length.

Correlation of Codon Usage Bias and Gene Expression With Skews

The correlation between various skews, gene expression level, and codon bias was evaluated, and it shows that SCS had a significant positive correlation only with the pyrimidine and keto skews. The CAI had a significant negative association with all skews except AT. This indicates that in the case of increased skew, protein expression is decreased; this association is independent of the AT skew.

Correlation of Length With Codon Usage Bias

The SCS values were statistically positively associated with the complete length ($r = 0.724$, $p < 0.0001$) of genes associated with neurodegeneration. We investigated segments of various lengths for their association with CUB. In segments that had lengths below 1,200 bp and above 2,400 bp, a positive association was estimated ($r = 0.940$, $p < 0.01$; $r = 0.638$, $p < 0.001$; $r = 0.332$, $p < 0.05$, and $r = 0.689$, $p < 0.001$ for lengths 1–400, 400–800, 800–1,200, and 2,400–4,500 bp, respectively), whereas for segments ranging from 1,200 to 2,400 bp, CUB was not affected by length. The CUBs of the 400–800 and 2,000–2,400 bp segments were negatively correlated with each other ($r = -0.448$, $p < 0.05$).

Correlation of Effective Number of Codons With GC3 Component

Both ENC and GC3 component are measures of bias. The ENC is a non-directional measure of CUB, with higher ENC values indicating lower bias (Hambuch and Parsch, 2005). GC3 is the

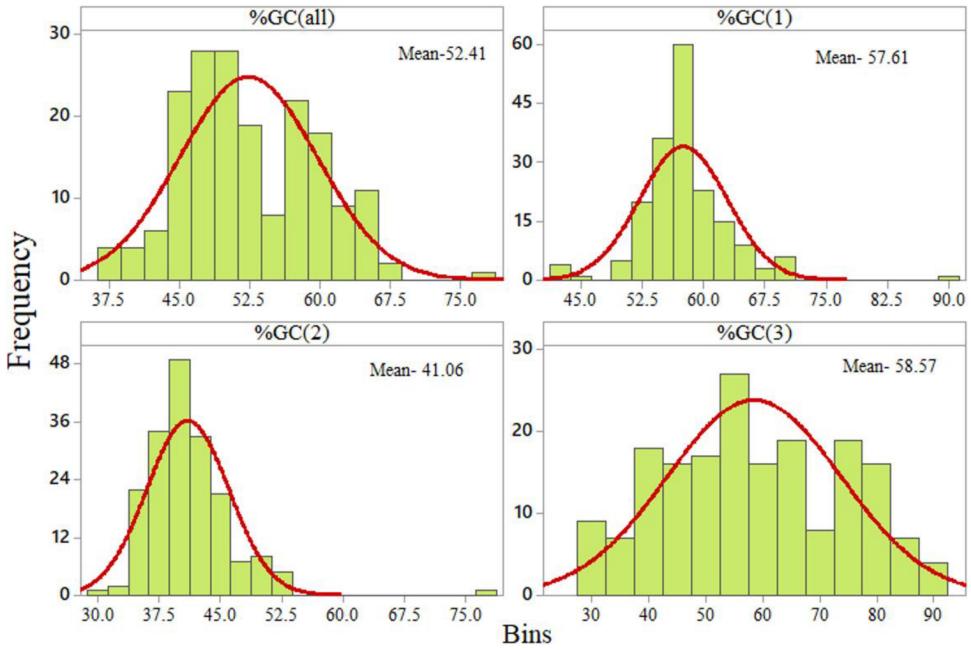

FIGURE 2 | Distribution of GC content at various codon positions.

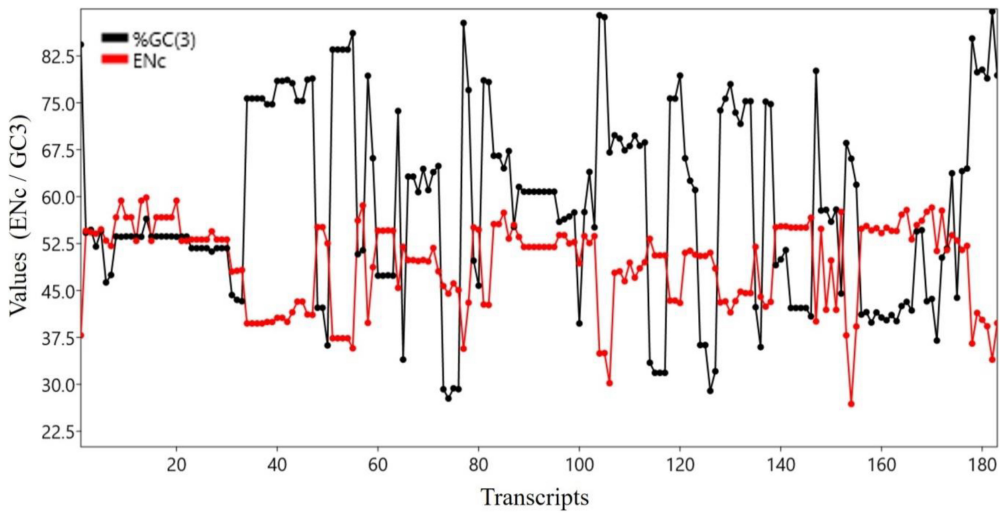

FIGURE 3 | Trend analysis of ENC and GC3 components.

frequency of G or C nucleotides at the third codon position and can be used to measure the CUB (Sahoo et al., 2019). There was a significant negative correlation between the ENC and GC3 ($r = 0.683$, $p < 0.0001$). **Figure 3** shows the trends of GC3 and ENC for various transcripts.

Effect of Length on Relative Synonymous Codon Usage Values

We performed a correlation analysis between the lengths and RSCU values of 183 transcripts. Eleven codons exhibited an association with the length of transcripts based on the RSCU values. Codons TTA ($r = 0.181$, $p < 0.05$), GTT ($r = 0.194$, $p < 0.01$), GTC ($r = 0.18$, $p < 0.05$), TCA ($r = 0.225$, $p < 0.01$), GGT ($r = 0.182$, $p < 0.05$), and GGA ($r = 0.161$, $p < 0.05$) showed a positive correlation with the length of the transcript (**Figure 4A**), whereas GTA ($r = -0.236$, $p < 0.001$), AGC ($r = -0.252$, $p < 0.001$), CGT ($r = -0.200$, $p < 0.001$), CGA ($r = -0.153$, $p < 0.05$), and GGG ($r = -0.190$, $p < 0.01$) showed a negative association with length. To investigate the correlation of these codons with the GC3 content, regression analysis was performed. The GC3 content was not correlated with length;

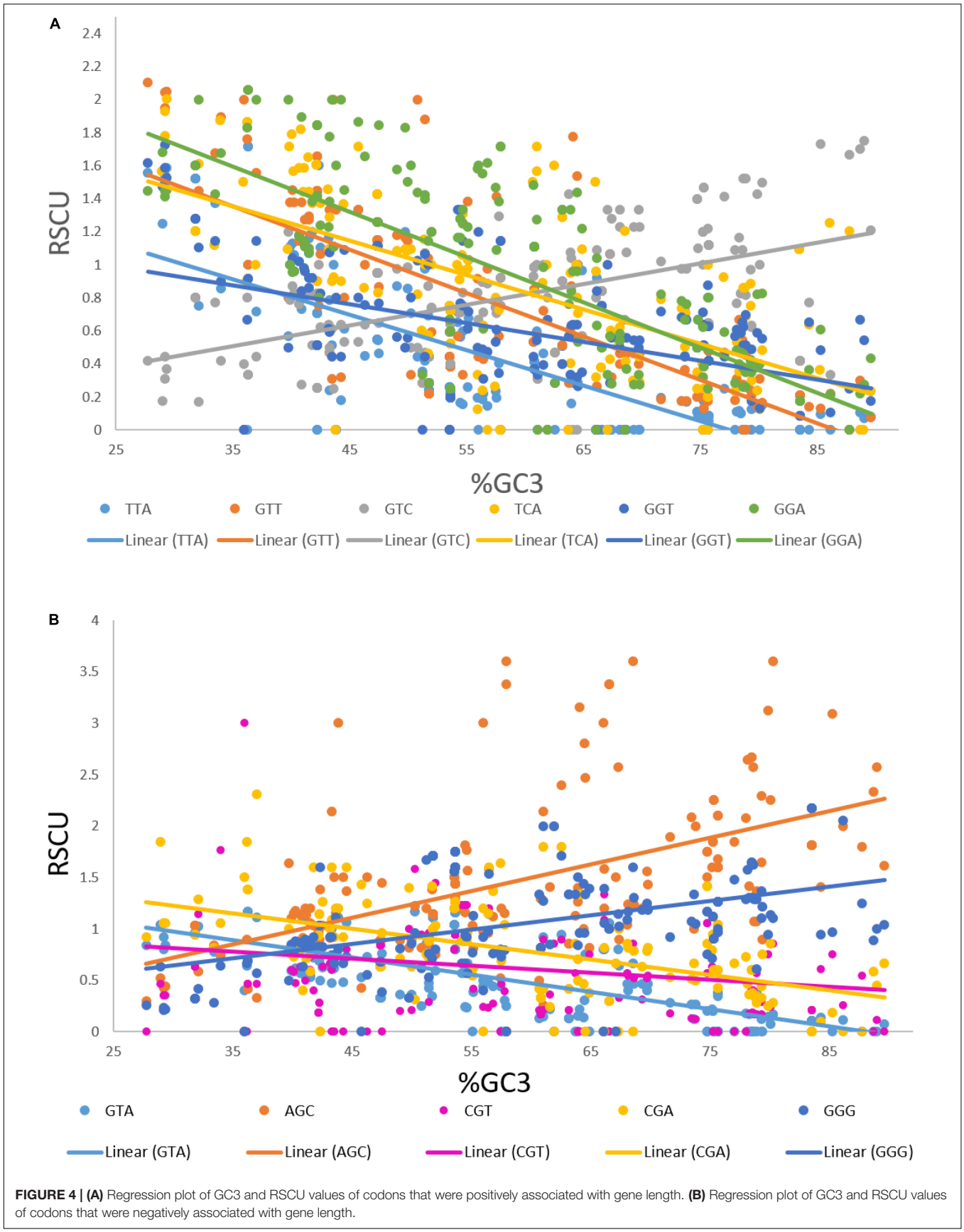

FIGURE 4 | (A) Regression plot of GC3 and RSCU values of codons that were positively associated with gene length. **(B)** Regression plot of GC3 and RSCU values of codons that were negatively associated with gene length.

however, an analysis was performed to investigate the effect of%GC3 on codons affected by length (**Figures 4A,B**). As shown in **Table 1**, among positively associated codons, GTT showed maximum variation (54.43%), whereas GGT showed minimum variation (23.73%); this was attributed to the GC3 content. GTA showed the maximum (44.38%) variation among the negatively associated codons, whereas CGT showed the minimum (2.16%) variation, which could be explained based on the GC3 content.

Effects of Different Lengths on Codon Usage Bias of Codons

Based on the RSCU values, we determined that TTA, GTT, GTC, TCA, GGT, and GGA exhibited a positive association with gene length, whereas GTA, AGC, CGT, CGA, and GGG showed a negative association. The manner via which CUB is affected over different ranges of transcript lengths was evaluated via regression analysis. Whether an association was constant along the length of the transcript or changed with it, was tested by generating eight groups based on different lengths, and then performing regression analysis. The eight groups encompassed segments of 1–400, 400–800, 800–1,200, 1,200–1,600, 100–2,000, 2,000–2,400, 2,400–4,500, and 1–4,500 bp.

Negative Correlation of Length With Codon Usage Bias of Codons

The post-regression r-squared values for each slope are presented in **Table 2**. At the whole transcript level, the RSCU values did not change considerably with the length of codons. However, in different ranges, a clear pattern was observed. Specifically, in the range 1–400 bp, the RSCU of the CGA codon was mostly affected by length (82.06%), followed by GTA (71.74%) and GGG (4.2%). In the ranges 400–800 and 1,200–2,000 bp, length did not affect the CUB of any codons. The overall analysis revealed that CGA experienced maximum variation in CUB, which was attributed to different length ranges. **Figure 5** depicts the results, with the regression slopes.

Positive Correlation of Length With Codon Usage Bias of Codons

The post-regression r-squared values for each slope are given in **Table 3**. At the whole transcript level, no considerable change in the RSCU values due to the length of codons was evident.

TABLE 1 | Regression coefficients (GC3 vs. RSCU) of codons that were significantly associated with transcript length.

Codons positively associated with length	R^2	Codons negatively associated with length	R^2
TTA	0.5293	GTA	44.38
GTT	0.5443	AGC	31.12
GTC	0.2773	CGT	02.16
TCA	0.3825	CGA	25.39
GGT	0.2373	GGG	20.55
GGA	0.4778	–	–

TABLE 2 | R-squared values (R^2) ≥ 0.25 are shown in bold.

S. no.	Length	Negatively associated codons				
		GTA	AGC	CGT	CGA	GGG
1	0–4,400	0.0344	0.0561	0.0002	0.0105	0.0553
2	1–400	**0.7174**	0.0219	0.0376	**0.8206**	**0.472**
3	400–800	0.0137	0.2286	0.0135	0.1207	0.0566
4	800–1,200	0.1843	**0.5396**	0.0643	0.041	0.0811
5	1,200–1,600	0.1093	0.0107	0.0708	0.1714	0.0041
6	1,600–2,000	0.1669	0.0905	0.1246	0.0064	0.1544
7	2,000–2,400	0.1646	**0.3854**	0.1246	**0.4199**	0.0067
8	2,400–4,400	0.1804	Nil	0.0729	**0.4536**	0.0034

The R^2-value indicates the percentage of variation in the CUB of a codon that can be explained by length.

The overall analysis of codons that were positively correlated to length revealed that GTT experienced the maximum variation in CUB (85.24%) followed by GGA (79.58%), across different length ranges. Similar to negatively linked codons, length did not affect the CUB of any positively correlated with codons in the ranges 400–800 and 1,200–2,000 bp. GGA experienced the maximum variation in CUB, which was attributed to the length of different ranges. **Figure 6** depicts the results of the regression slopes.

Principal Component Analysis

PCA is commonly used in dimensionality reduction. In **Figure 7**, axis 1 and axis 2 are plotted and the two axes account for most of the component. PCA is widely used to determine the major trend of codon usage. RSCU values of 59 synonymous codons are taken as a 59-dimensional vector. Methionine and tryptophan, which are encoded by single codons and three stop codons, are excluded from the study. PCA showed most genes to be scattered across the X-axis, and all genes (except *ARG1*, *PCBD1*, and *PTC*) were within the 90% confidence interval (**Figure 7**). There was not much variation in the codon usage of these genes. The first and second components contributed 51.42 and 6.07% of the variation, respectively. CUB was hence determined to be at a medium to low level.

Genes With Special Feature

We analyzed the genes as to whether any display a feature different than present in other genes. For the same, we evaluated RSCU values of genes. Genes showing different behavior are presented in **Table 4**.

The analysis revealed that there are 10 genes that reflect a different behavior than others. Here it is noteworthy that the codons TCG, GCG, and CGC are underrepresented in 81.67, 71.67, and 61.67% genes. This is an expected result since this codon contains the CpG dinucleotide, and CpG dinucleotide containing codons are found underrepresented in neurodegenerative disorders associated with brain iron accumulation (Alqahtani et al., 2021). CAG repeats are common in the *HTT* gene associated with Huntington's disease (Guo et al., 2018) and overrepresentation of CAG may associate with it. Contrary to our result, the TTT codon encoding

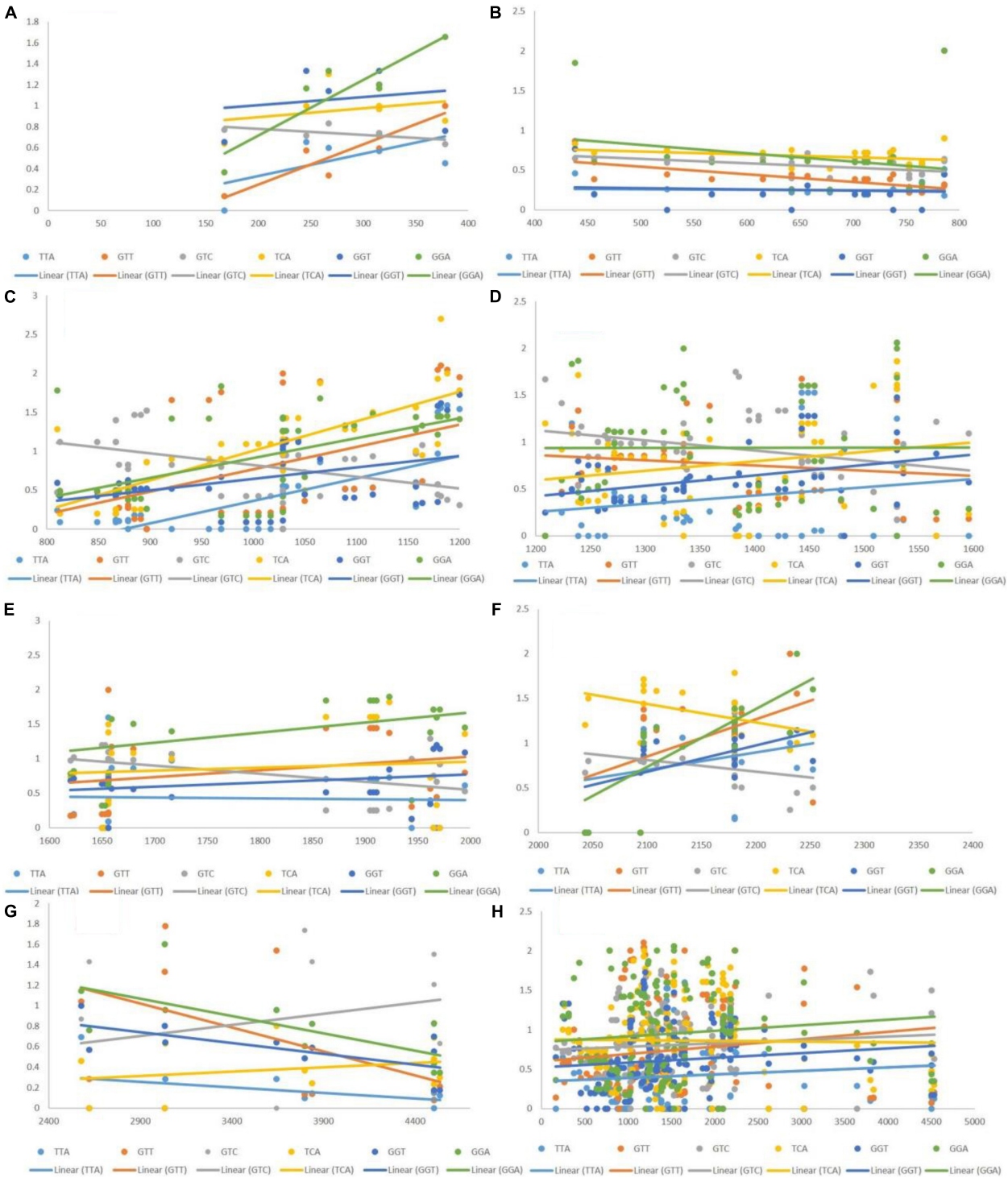

FIGURE 5 | Regression analysis of RSCU values of codons (negatively associated with gene length) with gene segments of various lengths. Each dot represents a codon, the trend line is depicted as a solid line, and color coding for each codon and trend line is given inside the figure. **(A)** 1–400 bp; **(B)** 400–800 bp; **(C)** 800–1,200 bp; **(D)** 1,200–1,600 bp; **(E)** 1,600–2,000 bp; **(F)** 2,000–2,400 bp; **(G)** 2,400–4,500 bp; **(H)** 1–4,500 bp.

for phenylalanine is found overrepresented in cyclin genes. Underrepresentation of the CAA codon can potentially be understood by the fact that mutation in CAA might culminate into the TAA stop codon, and thus such underrepresentation might be to avoid premature termination of protein translation (DeRonde et al., 2022). Some genes exhibiting a codon usage

TABLE 3 | R-squared values (R^2) ≥ 0.25 are shown in bold.

S. no.	Length	Positively associated codons					
		TTA	GTT	GTC	TCA	GGT	GGA
1	0–4,400	0.0059	0.0177	0.0083	0.0002	0.016	0.008
2	1–400	**0.338**	**0.8524**	**0.3967**	0.0778	0.0354	**0.7958**
3	400–800	0.0067	**0.513**	0.2169	0.0985	0.0092	0.0413
4	800–1,200	**0.5401**	0.2379	0.216	**0.5965**	0.1652	**0.3005**
5	1,200–1,600	0.0316	0.0166	0.1056	0.0442	0.1599	Nil
6	1,600–2,000	0.0017	0.0709	0.2279	0.0097	0.0856	0.0707
7	2,000–2,400	0.0684	0.2166	0.0985	0.2106	**0.3292**	**0.6082**
8	2,400–4,400	0.1741	**0.3166**	0.0673	0.0608	**0.4513**	**0.4911**

The R^2-value indicates the percentage of variation in the CUB of a codon that can be explained by length.

pattern contradictory to maximum of the genes display their uniqueness and also might be associated with their physical and biological properties.

DISCUSSION

The frequencies of codon usage vary, which has been termed as codon bias. This is an imperative evolutionary phenomenon that operates from lower organisms to higher eukaryotes. Various hypotheses, such as selection-mutation equilibrium, genetic drift, and GC-biased gene conversion, have been presented to explain CUB in genes. In the present study, we evaluated the effect of transcript length on various parameters, such as codon bias, gene expression, and RSCU values of codons.

The nucleotide composition of any genome is an integral part of the molecular architecture of any gene. Composition is known to affect codon usage, as well as the choice and frequency of amino acids. An analysis of codon-pair repeats in 12 *Drosophila* genomes revealed the predominant usage of hydrophilic amino acids in NNG-CNN (a codon pair with G at the 3' end of the first codon and C at the 5' end of the second codon) (Behura and Severson, 2012). The importance of nucleotide composition was further underscored by Jørgensen et al. (2007), who analyzed the GC content in relation to bias and found a high association. Furthermore, genes located in GC-poor domains show higher deviation in bias and amino acid usage in *Apis mellifera*. In our analysis, G was most abundant, followed by A and C, with T being the least abundant. The GC3 content showed maximum variation and ranged between 30.26 and 87.75%, and it was crucial in compositional constraints. Wide variation in the GC3 content has been observed in freshwater turbellarians, cestodes, and trematodes, which could be easily separated into two distinct clusters based on the GC3 content range (Lamolle et al., 2019).

Estimation of a non-linear correlation between GC3 and a CUB measure revealed GC3 as a key factor in determining codon usage, and this application is independent of species (Wan et al., 2004). GC3 components of grasses resemble warm-blooded vertebrates, whereas those of dicot species resemble cold-blooded vertebrates (Montero et al., 1990).

In our study, ENC exhibited a negative association with GC3; in the case of higher GC3 content, a higher bias was seen (ENC is a non-directional measure of CUB). Our results are concordant with those of Huang et al. (2017), who also found a negative correlation (r = –0.243, $p < 0.01$) between ENC and GC3. Few studies have focused on the effects of the GC3 component on gene length (Duret and Mouchiroud, 1999; Wang and Hickey, 2007). In the present work, the GC3 component showed a correlation with the length of genes, although it was not significant. However, a negative correlation between GC3 and the length of the gene was found, suggesting that shorter genes have a higher GC3 content. Similar to our observation, gene length has been negatively associated with the GC3 component in *C. elegans*, *D. melanogaster*, *A. thaliana* (Duret and Mouchiroud, 1999), and *O. sativa* (Wang and Hickey, 2007). Reports have determined associations between gene expression level and length, with shorter genes showing strong expression levels in bacteria (Chiaromonte et al., 2003). A contrary trend has been observed in plants and animals, where longer genes in plants and shorter genes in animals are expressed more (Ren et al., 2007).

The codon composition is mainly affected by mutational and selective forces, including transcription, translation, mRNA stability, RNA and DNA methylation, co-translational folding, mRNA splicing, transport (Sharp and Li, 1986; Gebauer and Hentze, 2004; Bergman and Tuller, 2020), and genetic drift (Shah and Gilchrist, 2011), with CUB being affected by the codon composition (Bahiri-Elitzur and Tuller, 2021). When the effects of gene length on nucleotide composition were investigated, only the composition of T was found to be affected. Furthermore, the composition of T was affected by gene length at the first and second codon positions, while no nucleotide was correlated with gene length at the third codon position. The overall result indicated that the composition of T is imperative in deciding the gene length; in other words, gene lengths affect the composition of T in genes associated with neurodegeneration. Yang et al. (2021) linked a compositional parameter of the GC3 content to gene length, wherein genes shorter than 2,000 bp had a higher GC3 content than longer genes.

Significant amounts of the GC and AT skews can be explained based on the mutational differences between leading or lagging strands (Tillier and Collins, 2000). A comparison of the AT and GC skews in 30 avian mitochondrial genomes revealed that parrots have unusually strong compositional asymmetry (AT- and GC-skew) in their coding sequences (Eberhard and Wright, 2016). In *W. bancrofti* and *S. haematobium*, the GC, AT, purine, pyrimidine, amino, and keto skews were negatively associated with CUB, except for the pyrimidine skew in *S. haematobium* (Mazumder et al., 2017). Contrarily, in amphibians *Bombina bombina*, *B. fortinuptialis*, *B. lichuanensis*, *B. maxima*, *B. microdeladigitora*, *B. orientalis*, and *B. variegate*, no skew was correlated to CUB (Barbhuiya et al., 2019). The reports from other investigators suggest variable skewness in different genera and species, which could be a signature of the corresponding organisms. In the present study, the analysis of genes associated with neurodegeneration revealed the effect of CUB on the pyrimidine and keto skews.

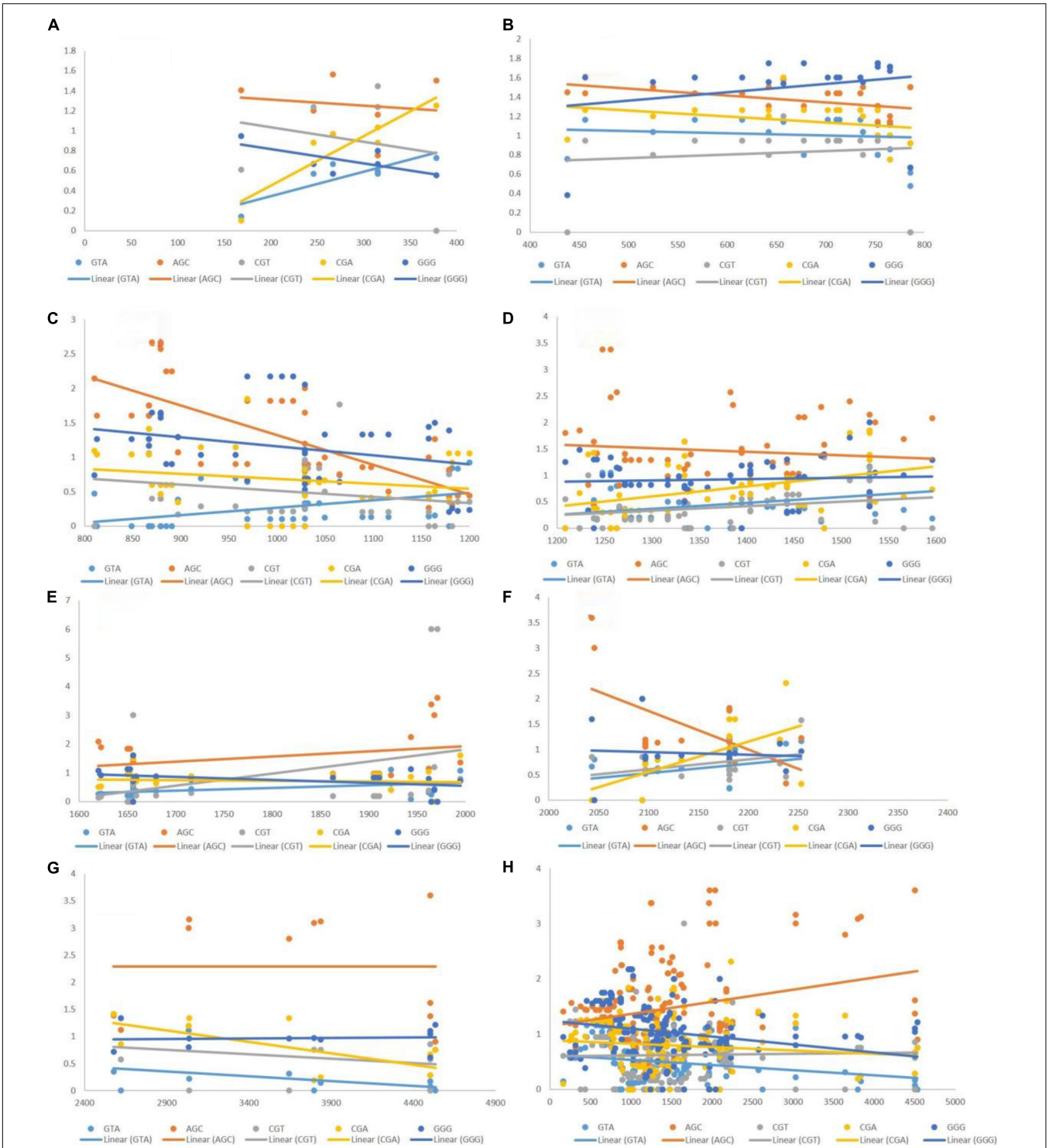

FIGURE 6 | Regression analysis of RSCU values of codons (positively associated with gene length) with gene segments of various lengths. Each dot represents a codon. The trend line is depicted as a solid line, and color coding for each codon and trend line is given inside the figure. **(A)** 1–400 bp; **(B)** 400–800 bp; **(C)** 800–1,200 bp; **(D)** 1,200–1,600 bp; **(E)** 1,600–2,000 bp; **(F)** 2,000–2,400 bp; **(G)** 2,400–4,500 bp; **(H)** 1–4,500 bp.

In *S. cerevisiae*, *D. melanogaster*, *C. elegans*, and *A. thaliana*, a negative correlation between codon bias and gene length has been observed (Moriyama and Powell, 1998), while a positive correlation has been observed for *C. elegans*. In the present study, we investigated the effects of length on CUB in 60 genes and found no correlation. However, when we investigated various

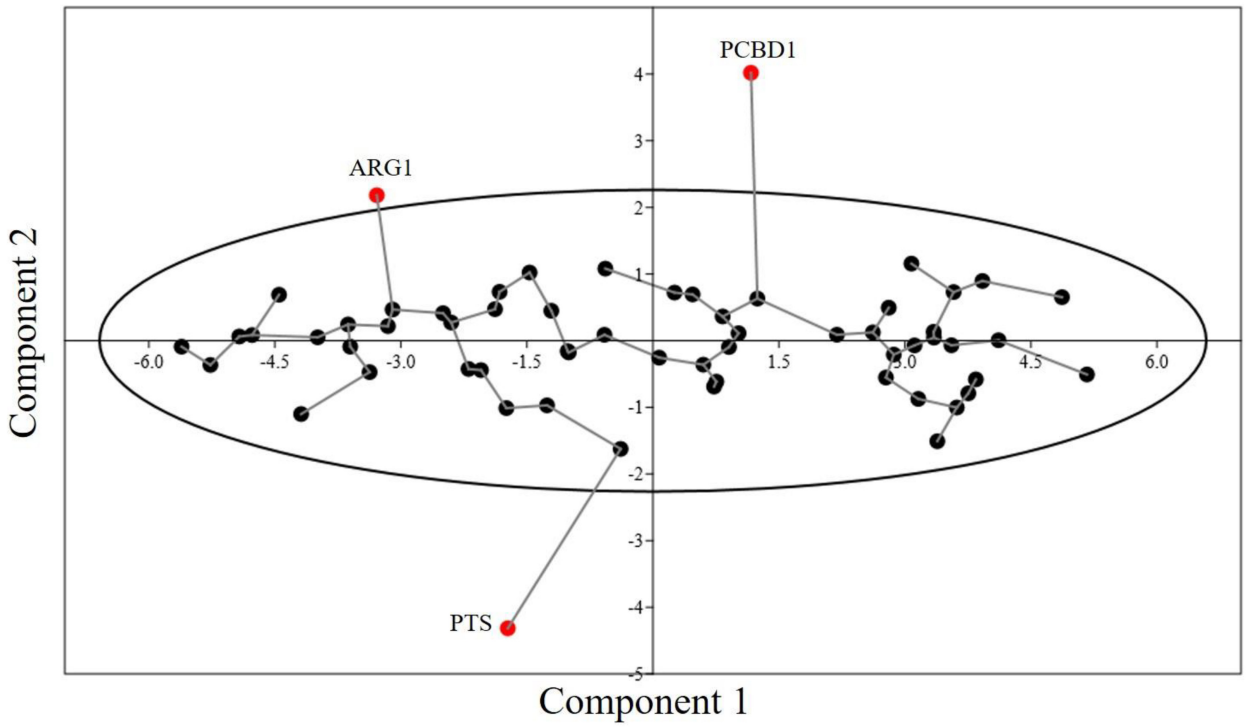

FIGURE 7 | Minimum spanning tree graph for PCA of 60 genes involved in neurodegeneration. The eclipse shows a 90% concentration of genes within the area. Three genes are present outside the concentration eclipse, showing different behavior.

TABLE 4 | The genes showing unusual behavior in terms of codon usage.

S. no.	Codons	Condition in other genes	% of genes showing common condition	Genes exhibiting contrary behavior
1	TTT	Underrepresentation	71.67	*MMADHC*
2	TCG	Underrepresentation	81.67	*ABCD1*
				GCH1
				MMACHO
3	GCG	Underrepresentation	71.67	*GAMT*
				IVD
				MMAB
4	CAA	Underrepresentation	75	*MAN2B1*
5	CGC	Underrepresentation	61.67	*DBT*
				NPC2
6	CAG	Overrepresentation	53.33	*MAN2B1*

length ranges, a clear pattern of association between CUB and length was observed. In gene segments below 1,200 bp and above 2,400 bp, CUB was significantly positively associated with length. These results suggest that CUB operates in either small or large genes that are associated with neurodegeneration. Possibly, a selection pressure is acting principally on more minor genes, which are supposed to be highly expressed, and larger genes, which are energetically expensive. However, selection pressure cannot completely explain the correlation between length and codon bias. Additionally, the translational accuracy and speed model cannot explain the negative association between CUB and length in *S. cerevisiae*, *D. melanogaster*, *C. elegans*, and *A. thaliana* (Eyre-Walker, 1996; Duret and Mouchiroud, 1999). Marais and Duret (2001) reported that a negative correlation between codon bias and gene length is found only in eukaryotes, which contradicts our results, as we found a positive association of length and CUB in all cases.

When the correlation of the RSCU values of individual codons with length was investigated, only 11 out of 59 codons were found to be affected by the length of genes. Codons TTA, GTT, GTC, TCA, GGT, and GGA exhibited positive association, whereas codons GTA, AGC, CGT, CGA, and GGG showed a negative association with length. The GC3 percentage is an indicator of both compositional bias (Deka and Chakraborty, 2016) and

codon bias (Shen et al., 2015). Therefore, we investigated the effect of the GC3 content on the CUB of these codons. Codons GTT and GTA (where G was at the first codon position and T at the second) were affected by the GC3 content, and 54.43 and 44.38% of the variation in the RSCU values of GTT and GTA, respectively, could be explained by the GC3 content.

No association between gene length and RSCU of vaccine-derived polioviruses, wild viruses, and live attenuated viruses was observed (Zhang et al., 2011). When the correlation between gene length and synonymous CUB was investigated for *D. melanogaster*, *E. coli*, and *S. cerevisiae*, a significant positive association was found in *E. coli* but negative correlations were found in *D. melanogaster* and *S. cerevisiae* genes. For *D. melanogaster* and *S. cerevisiae* genes, the ENC distribution, with respect to CUB, was different for short genes (300–500 bp) (Moriyama and Powell, 1998).

Our study found a significant positive association of SCS, a measure of CUB, with length, when the overall gene length was considered. Detailed investigation revealed that below 1,200 bp and above 2,400 bp, CUB had positive linkages with length, while in segments of 1,200–2,400 bp, there was no observed association, indicating that codon bias is present in both shorter and longer genes. However, in middle-length genes, CUB does not appear to occur. The results of the present study are concordant with those of a study by Moriyama and Powell (1998), wherein a higher bias was noted in energetically expensive longer genes, which occurred to maximize the translational efficiency, owing to selection pressure.

To further elucidate the effect of the length of various segments on RSCU, we performed regression analysis. Overall, the association of gene length with CUB was not evident. However, the RSCU values for both the GTT and GTA codons were notably affected by length (85.24 and 71.74%, respectively) in gene segments up to 400 bp. specifically, up to 400 bp, four out of six positively associated codons and three out of five negatively associated codons showed that the length was significantly correlated with RSCU values. In segments with lengths in the range 1,200–2,000 bp, CUB was not affected by length. Overall analysis indicated that the association between CUB and length varies depending on the segment size. The results were similar for both the overall CUB and CUB of specific codons.

Hia et al. (2019) demonstrated that human cells adopt a unique codon bias mechanism to modulate mRNA stability. Specifically, genes can be clustered in GC- and AT-ending codons; GC-ending codons enhance mRNA stability, while AT-ending codons destabilize mRNA. In the present study, GC-ending codons were noted to be preferred over AT-ending codons.

In contrast to that in other non-mammalian eukaryotes, CUB in humans is very high in both the highly and lowly expressed genes; selection possibly plays a role in both the enhancement and reduction of gene expression by promoting and hindering the use of optimal codons for the former and latter conditions. PCA based on the RSCU values revealed that the genes were not very scattered, and they were near the first axis. Overall, the genes exhibited a medium to low CUB.

CONCLUSION

Codon usage analysis is an integral part of the molecular characterization of a gene. We studied how various compositional factors and other features, such as codon usage, might be affected by gene length. Compositional analysis revealed that G was the most abundant nucleotide, followed by A and C, with T being the least abundant. The distribution of the GC1 and GC3 contents was similar, with GC2 being the minimum. ENC was negatively associated with the GC3 content, which indicated that in genes with high GC3 content, the bias will be higher (ENC is a non-directional measure of CUB). Additionally, a negative correlation between the GC3 content and length indicated that longer genes had a lower GC3 content and lesser bias. In the present study, no effect of length on gene expression was observed. Of all four nucleotides, an association was found only between T and gene length; this effect was observed only at the first and second codon positions. Length affects the composition of gene nucleotides when T is considered. CUB in different organisms has been reported to be associated with different nucleotide skews; in this study, the pyrimidine and keto skews were found to be associated with CUB. To investigate the effect of gene length on CUB, we performed a correlation analysis, which showed no significant association between CUB and length at an overall level; however, when a segment-wise study was undertaken, a clear pattern was observed. CUB was statistically positively associated with length in gene segments below 1,200 bp and above 2,400 bp. This suggests that selection pressure is possibly acting on smaller genes, which are supposed to be highly expressed, and on larger genes, which are energetically expensive.

The RSCU values of eleven codons were significantly correlated with length. Codons TTA, GTT, GTC, TCA, GGT, and GGA exhibited a positive association with length, whereas codons GTA, AGC, CGT, CGA, and GGG showed a negative association. Analysis was performed to determine how the RSCU values of these codons are affected by the GC3 composition, and showed that codons GTT and GTA (containing G at the first codon position and T at the second) were affected the most by GC3. This revealed the positional significance of selective nucleotides and ultimately the role of selective forces balancing the RSCU values of codons and the composition and length of genes. The distribution of CUB is a documented variable for longer and shorter genes. In our study, CUB was affected by the length of genes when the gene length was shorter than 1,200 bp and longer than 2,400 bp. Among the codons that showed a significant association with the length of genes, GTT and GTA were maximally affected (85.24 and 71.74%, respectively). GC-ending codons were preferred over AT-ending codons, and PCA indicated that there was not much variation in the codon usage of genes.

Overall analysis indicated that gene length affects compositional bias and CUB (in the form of GC3 content). Gene length is correlated with the T content at the first and second codon positions, which are affected by selective forces, as well as the CUB of a few codons, with the maximal effect on codons that have G at the first position and T at the second. Gene length also affects the CUB of genes smaller than 1,200 bp and

larger than 2,000 bp; the length of the gene significantly affects various parameters associated with codon usage.

DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found in the article/supplementary material.

AUTHOR CONTRIBUTIONS

RK and MS: conceptualization. RK, MS, AMA, and GMA: methodology, software, analysis, writing, and editing. RK, NHG, and MAK: validation, writing-review and editing, and formal analysis. RK, MS, AMA, GMA, NHG, and MAK: funding acquisition. NHG: analysis/interpretation, writing/editing, approved, and manuscript accountability. All authors have read and agreed to the published version of the manuscript and read and agreed on the final version of the manuscript.

FUNDING

We supported in part by their respective institutions: RK by Barkatullah University, India; MS and AMA by University of Hail, Saudi Arabia; GMA by King Abdulaziz University, Saudi Arabia; NHG by the Intramural Research Program, NIA, NIH, United States; and MAK by (i) Sichuan University, China; (ii) King Abdulaziz University, Saudi Arabia; (iii) Daffodil University, Bangladesh; and (iv) the Novel Global Community Educational Foundation, Australia.

ACKNOWLEDGMENTS

We are thankful to their respective universities/institutions for providing an environment and support to conduct the study.

REFERENCES

Alqahtani, T., Khandia, R., Puranik, N., Alqahtani, A. M., Almikhlafi, M. A., and Algahtany, M. A. (2021). Leucine encoding codon TTG shows an inverse relationship with GC content in genes involved in neurodegeneration with iron accumulation. *J. Integr. Neurosci.* 20, 905–918. doi: 10.31083/j.jin2004092

Bahiri-Elitzur, S., and Tuller, T. (2021). Codon-based indices for modeling gene expression and transcript evolution. *Comput. Struct. Biotechnol. J.* 19:2646.

Barbhuiya, P. A., Uddin, A., and Chakraborty, S. (2019). Genome-wide comparison of codon usage dynamics in mitochondrial genes across different species of amphibian genus *Bombina*. *J. Exp. Zool. B Mol. Dev. Evol.* 332, 99–112. doi: 10.1002/jez.b.22852

Behura, S. K., and Severson, D. W. (2012). Comparative analysis of codon usage bias and codon context patterns between dipteran and hymenopteran sequenced genomes. *PLoS One* 7:e43111. doi: 10.1371/journal.pone.0043111

Bergman, S., and Tuller, T. (2020). Widespread non-modular overlapping codes in the coding regions. *Phys. Biol.* 17:031002. doi: 10.1088/1478-3975/ab7083

Bhattacharya, T., Soares, G. A. B. E., Chopra, H., Rahman, M. M., Hasan, Z., Swain, S. S., et al. (2022). Applications of phyto-nanotechnology for the treatment of neurodegenerative disorders. *Materials* 15:804. doi: 10.3390/ma15030804

Bourret, J., Alizon, S., and Bravo, I. G. C. O. U. S. I. N. (2019). (COdon Usage Similarity INdex): a normalized measure of codon usage preferences. *Genome Biol. Evol.* 11, 3523–3528. doi: 10.1093/gbe/evz262

Charneski, C. A., Honti, F., Bryant, J. M., Hurst, L. D., and Feil, E. J. (2011). Atypical AT skew in firmicute genomes results from selection and not from mutation. *PLoS Genet.* 7:e1002283. doi: 10.1371/journal.pgen.1002283

Chiaromonte, F., Weber, R. J., Roskin, K. M., Diekhans, M., Kent, W. J., and Haussler, D. (2003). The share of human genomic DNA under selection estimated from human-mouse genomic aligmnents. *Cold Spring Harb. Symp. Quant. Biol.* 68, 245–254. doi: 10.1101/sqb.2003.68.245

Deka, H., and Chakraborty, S. (2016). Insights into the usage of nucleobase triplets and codon context pattern in five influenza a virus subtypes. *J. Microbiol. Biotechnol.* 26, 1982–1992. doi: 10.4014/jmb.1605.05016

DeRonde, S., Deuling, H., Parker, J., and Chen, J. (2022). Identification of a novel SARS-CoV-2 variant with a truncated protein in ORF8 gene by next generation sequencing. *Sci. Rep.* 12:4631. doi: 10.1038/s41598-022-08780-2

Duret, L., and Mouchiroud, D. (1999). Expression pattern and, surprisingly, gene length shape codon usage in *Caenorhabditis*, *Drosophila*, and *Arabidopsis*. *Proc. Natl. Acad. Sci. U.S.A.* 96, 4482–4487. doi: 10.1073/pnas.96.8.4482

Eberhard, J. R., and Wright, T. F. (2016). Rearrangement and evolution of mitochondrial genomes in parrots. *Mol. Phylogenet. Evol.* 94:34. doi: 10.1016/j.ympev.2015.08.011

Eyre-Walker, A. (1996). Synonymous codon bias is related to gene length in *Escherichia coli*: selection for translational accuracy? *Mol. Biol. Evol.* 13, 864–872. doi: 10.1093/oxfordjournals.molbev.a025646

Gebauer, F., and Hentze, M. W. (2004). Molecular mechanisms of translational control. *Nat. Rev. Mol. Cell Biol.* 5, 827–835.

Gitler, A. D., Dhillon, P., and Shorter, J. (2017). Neurodegenerative disease: models, mechanisms, and a new hope. *Dis. Model Mech.* 10, 499–502. doi: 10.1242/dmm.030205

Grishkevich, V., and Yanai, I. (2014). Gene length and expression level shape genomic novelties. *Genome Res.* 24:1497. doi: 10.1101/gr.169722.113

Guo, Q., Cheng, J., Seefelder, M., Engler T., Pfeifer G., et al (2018). Bin huang null. the cryo-electron microscopy structure of huntingtin. *Nature.* 555, 117–120. doi: 10.1038/nature25502

Hambuch, T. M., and Parsch, J. (2005). Patterns of synonymous codon usage in drosophila melanogaster genes with sex-biased expression. *Genetics.* 170:1691. doi: 10.1534/genetics.104.038109

Hia, F., Yang, S. F., Shichino, Y., Yoshinaga, M., Murakawa, Y., Vandenbon, A., et al. (2019). Codon bias confers stability to human mRNAs. *EMBO Rep.* 20:e48220. doi: 10.15252/embr.201948220

Huang, X., Xu, J., Chen, L., Wang, Y., Gu, X., Peng, X., et al. (2017). Analysis of transcriptome data reveals multifactor constraint on codon usage in taenia multiceps. *BMC Genomics* 18:308. doi: 10.1186/s12864-017-3704-8

Jørgensen, F. G., Schierup, M. H., and Clark, A. G. (2007). Heterogeneity in regional GC content and differential usage of codons and amino acids in GC-poor and GC-rich regions of the genome of *Apis mellifera*. *Mol. Biol. Evol.* 24, 611–619. doi: 10.1093/molbev/msl190

Josephs, K. A., Ahlskog, J. E., Parisi, J. E., Boeve, B. F., Crum, B. A., Giannini, C., et al. (2009). Rapidly progressive neurodegenerative dementias. *Arch. Neurol.* 66:201.

Kalisz, S., and Purugganan, M. D. (2004). Epialleles via DNA methylation: consequences for plant evolution. *Trends Ecol. Evol.* 19, 309–314. doi: 10.1016/j.tree.2004.03.034

Katsuno, M., Tanaka, F., and Sobue, G. (2012). Perspectives on molecular targeted therapies and clinical trials for neurodegenerative diseases. *J. Neurol. Neurosurg. Psychiatry* 83, 329–335. doi: 10.1136/jnnp-2011-301307

Kirkconnell, K. S., Magnuson, B., Paulsen, M. T., Lu, B., Bedi, K., and Ljungman, M. (2017). Gene length as a biological timer to establish temporal transcriptional regulation. *Cell Cycle* 16:259. doi: 10.1080/15384101.2016.1234550

Lamolle, G., Fontenla, S., Rijo, G., Tort, J. F., and Smircich, P. (2019). Compositional analysis of flatworm genomes shows strong codon usage biases across all classes. *Front. Genet.* 10:771. doi: 10.3389/fgene.2019.00771

Livingston, G., Huntley, J., Sommerlad, A., Ames, D., Ballard, C., Banerjee, S., et al. (2020). Dementia prevention, intervention, and care: 2020 report of the lancet commission. *Lancet* 396, 413–446. doi: 10.1016/S0140-6736(20)30367-6

Lopes, I., Altab, G., Raina, P., and Magalhães, JP de (2021). Gene size matters: an analysis of gene length in the human genome. *Front. Genet.* 12:559998 doi: 10.3389/fgene.2021.559998

Marais, G., and Duret, L. (2001). synonymous codon usage, accuracy of translation, and gene length in *Caenorhabditis elegans*. *J. Mol. Evol.* 52, 275–280. doi: 10.1007/s002390010155

Martier, R., and Konstantinova, P. (2020). Gene therapy for neurodegenerative diseases: slowing down the ticking clock. *Front. Neurosci.* 14:580179. doi: 10.3389/fnins.2020.580179

Mazumder, G. A., Uddin, A., and Chakraborty, S. (2017). Expression levels and codon usage patterns in nuclear genes of the filarial nematode Wucheraria bancrofti and the blood fluke *Schistosoma haematobium*. *J. Helminthol.* 91, 72–79. doi: 10.1017/S0022149X16000092

Montero, L. M., Salinas, J., Matassi, G., and Bernardi, G. (1990). Gene distribution and isochore organization in the nuclear genome of plants. *Nucleic Acids Res.* 18:1859. doi: 10.1093/nar/18.7.1859

Moriyama, E. N., and Powell, J. R. (1998). Gene length and codon usage bias in *Drosophila melanogaster*, *Saccharomyces cerevisiae* and *Escherichia coli*. Nucleic Acids Res. 26, 3188–3193.

Onikanni, A. S., Lawal, B., Oyinloye, B. E., Mostafa-Hedeab, G., Alorabi, M., Cavalu, S., et al. (2022). Therapeutic efficacy of Clompanus pubescens leaves fractions via downregulation of neuronal cholinesterases/Na+-K+ATPase/IL-1β, and improving the neurocognitive and antioxidants status of streptozotocin-induced diabetic rats. *Biomed. Pharmacother.* 148:112730. doi: 10.1016/j.biopha.2022.112730

Przedborski, S., Vila, M., and Jackson-Lewis, V. (2003). Neurodegeneration: what is it and where are we? *J Clin Invest.* 111, 3–10. doi: 10.1172/JCI17522

Puigbò, P., Bravo, I. G., and Garcia-Vallve, S. (2008). CAIcal: a combined set of tools to assess codon usage adaptation. *Biol. Direct.* 6:3. doi: 10.1186/1745-6150-3-38

Ren, L., Gao, G., Zhao, D., Ding, M., Luo, J., and Deng, H. (2007). Developmental stage related patterns of codon usage and genomic GC content: searching for evolutionary fingerprints with models of stem cell differentiation. *Genome Biol.* 8:R35. doi: 10.1186/gb-2007-8-3-r35

Research, T. W., and Ehi of M. Neurodegenerative disorders (2017). *WEHI. The Walter and Eliza Hall Institute of Medical Research*. Available online at : https://www.wehi.edu.au/research-diseases/development-and-ageing/neurodegenerative-disorders(accessed on May 5, 2022)

Ripich, D. N., and Horner, J. (2004). The neurodegenerative dementias: diagnoses and interventions. *ASHA Lead.* 9, 4–15.

Sahakyan, A. B., and Balasubramanian, S. (2016). Long genes and genes with multiple splice variants are enriched in pathways linked to cancer and other multigenic diseases. *BMC Genomics* 17:225. doi: 10.1186/s12864-016-2582-9

Sahoo, S., Das, S. S., and Rakshit, R. (2019). Codon usage pattern and predicted gene expression in *Arabidopsis thaliana*. *Gene: X* 1:2.

Shah, P., and Gilchrist, M. A. (2011). Explaining complex codon usage patterns with selection for translational efficiency, mutation bias, and genetic drift. *PNAS* 108, 10231–10236. doi: 10.1073/pnas.1016719108

Sharifi-Rad, M., Lankatillake, C., Dias, D. A., Docea, A. O., Mahomoodally, M. F., Lobine, D., et al. (2020). Impact of natural compounds on neurodegenerative disorders: from preclinical to pharmacotherapeutics. *J. Clin. Med.* 9:E1061. doi: 10.3390/jcm9041061

Sharp, P. M., and Li, W. H. (1986). An evolutionary perspective on synonymous codon usage in unicellular organisms. *J. Mol. Evol.* 24, 28–38. doi: 10.1007/BF02099948

Sharp, P. M., and Li, W. H. (1987). The codon adaptation index–a measure of directional synonymous codon usage bias, and its potential applications. *Nucleic Acids Res.* 15, 1281–1295. doi: 10.1093/nar/15.3.1281

Shen, W., Wang, D., Ye, B., Shi, M., Ma, L., Zhang, Y., et al. (2015). GC3-biased gene domains in mammalian genomes. *Bioinformatics* 31, 3081–3084. doi: 10.1093/bioinformatics/btv329

Shields, D. C., and Sharp, P. M. (1987). Synonymous codon usage in *Bacillus subtilis* reflects both translational selection and mutational biases. *Nucleic Acids Res.* 15:8023. doi: 10.1093/nar/15.19.8023

Song, H., Liu, J., Song, Q., Zhang, Q., Tian, P., and Nan, Z. (2017). Comprehensive analysis of codon usage bias in seven *Epichloë* species and their peramine-coding genes. *Front. Microbiol.* 8:1419. doi: 10.3389/fmicb.2017.01419

Tillier, E. R. M., and Collins, R. A. (2000). The contributions of replication orientation, gene direction, and signal sequences to base-composition asymmetries in bacterial genomes. *J. Mol. Evol.* 50, 249–257. doi: 10.1007/s002399910029

Urrutia, A. O., and Hurst, L. D. (2003). The Signature of selection mediated by expression on human genes. *Genome Res.* 13:2260. doi: 10.1101/gr.641103

Wan, X. F., Xu, D., and Kleinhofs, A. (2004). Quantitative relationship between synonymous codon usage bias and GC composition across unicellular genomes. *BMC Evol. Biol.* 4:19. doi: 10.1186/1471-2148-4-19

Wang, H. C., and Hickey, D. A. (2007). Rapid divergence of codon usage patterns within the rice genome. *BMC Evol. Biol.* 7(Suppl. 1):S6. doi: 10.1186/1471-2148-7-S1-S6

Wang, L., Xing, H., Yuan, Y., Wang, X., Saeed, M., Tao, J., et al. (2018). Genome-wide analysis of codon usage bias in four sequenced cotton species. *PLoS One* 13:e0194372. 10.1371/journal.pone.0194372 doi: doi:

Yang, C., Zhao, Q., Wang, Y., Zhao, J., Qiao, L., Wu, B., et al. (2021). Comparative analysis of genomic and transcriptome sequences reveals divergent patterns of codon bias in wheat and its ancestor species. *Front. Genet.* 12:732432. doi: 10.3389/fgene.2021.732432

Yang, J., Zhu, T. Y., Jiang, Z. X., Chen, C., Wang, Y. L., Zhang, S., et al. (2010). Codon usage biases in Alzheimer's disease and other neurodegenerative diseases. *Protein Pept. Lett.* 17, 630–645. doi: 10.2174/092986610791112666

Zhang, J., Wang, M., Liu, W. Q., Zhou, J. H., Chen, H. T., Ma, L. N., et al. (2011). Analysis of codon usage and nucleotide composition bias in polioviruses. *Virol. J.* 8, 1–8. doi: 10.1186/1743-422X-8-146

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.