Original Research

Leucine encoding codon TTG shows an inverse relationship with GC content in genes involved in neurodegeneration with iron accumulation

Taha Alqahtani[1], Rekha Khandia[2,*], Nidhi Puranik[2], Ali M Alqahtani[1], Mohannad A. Almikhlafi[3], Mubarak Ali Algahtany[4]

[1] *Department of Pharmacology, College of Pharmacy, King Khalid University, 62529 Abha, Saudi Arabia*

[2] *Department of Biochemistry and Genetics, Barkatullah University, 462026 MP Bhopal, India*

[3] *Department of Pharmacology and Toxicology, Taibah University, 42311 Madinah, Saudi Arabia*

[4] *Division of Neurosurgery, Department of Surgery, College of Medicine, King Khalid University, 62529 Abha, Saudi Arabia*

*Correspondence: rekha.khandia@bubhopal.ac.in (Rekha Khandia)

DOI:10.31083/j.jin2004092

Submitted: 19 August 2021 Revised: 29 September 2021 Accepted: 1 November 2021 Published: 30 December 2021

We determined various forces involved in shaping codon usage of the genes linked to brain iron accumulation and infantile neuroaxonal dystrophy. The analysis paved the way for determining the forces responsible for composition, expression level, physical properties and codon bias of a gene. An interesting observation related to composition was that, on all the three codon positions, any two of the four nucleotides had similar compositions. CpG, TpA, and GpT dinucleotides were underrepresented with the overrepresentation of TpG dinucleotide. CpG and TpA containing codons ATA, CTA, TCG, and GCG were underrepresented, while TpG dinucleotide containing codon CTG was overrepresented, indicative of compositional constraints importance. GC ending codons were favored when the genome is GC rich, except leucine encoding codon TTG, which exhibits an inverse relationship with GC content. Nucleotide disproportions are found associated with the physical properties of proteins. The values of CAI and ENc are suggestive of low codon bias in genes. Considering the results of neutrality analysis, parity analysis, underrepresentation of TpA and CpG codons, and over-representation of TpG codons, the correlation between the compositional constraints and skew relationships with protein properties suggested the role of all the three selectional, mutational and compositional forces in shaping codon usage with the dominance of selectional pressure.

Keywords

Iron accumulation; Neurodegeneration; CTG codon; Selectional force; Mutational force; Bioinformatics; Neurogenetics

1. Introduction

Iron in the brain plays a crucial role in sustaining homeostatic activity by participating in different physiological processes such as mitochondrial respiration, myelin formation, neurotransmitter synthesis, DNA synthesis, and metabolism [1]. A high iron level can harm brain cells by generating radicals, leading to oxidative damage. Abnormal iron homeostasis seems critical in potentiating neurodegeneration in disease conditions like traumatic brain injury, Alzheimer's, Parkinson's, Huntington's diseases, and multiple sclerosis [2]. Neurological manifestations of disturbances in iron homeostasis are described in increased prevalence of headache [3], movement disorders [4], secondary dementia [5], epilepsy [6], multiple sclerosis [7], human immunodeficiency virus dementia [8], Freidrich ataxia [9], and Alzheimer and Parkinson diseases [10]. Iron accumulation can be visualized in MRI, and iron deposition in deep brain nuclei has been described as differing between patients with multiple system atrophy and progressive supranuclear palsy [11]. In a few rare neurodegenerative disorders, iron accumulation has been observed in various parts of the brain. The examples are Woodhouse–Sakati syndrome (iron deposition in the small pituitary gland), Kufor–Rakeb disease (deposition in the basal ganglia), Coenzyme A synthase protein-associated neurodegeneration (increased pallidal iron content), SCPx syndrome (accumulation in the thalamus, the pons, and the occipital region), and GM1 type3 gangliosidosis (accumulation in lobus pallidus with progression to the substantia nigra, the red nucleus and the subthalamic nucleus). However, in a few instances, hypo-intensities for iron content in MRI have been reported in diseases including restless legs syndrome and periodic limb movements in sleep [5], adaptor protein complex-4 deficiency, DDHD Domain Containing 1 gene (*DDHD1*) pathogenic variants, and GTP Binding Protein 2 (GTPBP2) pathogenic variants [12].

Neurodegeneration with brain iron accumulation (NBIA) condition was first named Hallervorden-Spatz syndrome, and it is associated with high quantities of basal ganglia iron and axonal spheroids. Pantothenate kinase-associated neurodegeneration (PKAN), caused by mutations in the *PANK2* gene, is the utmost reason for NBIA [13, 14]. Presently, identifying various genes associated with NBIA and recognizing the growing phenotypes within this group has made significant progress. Several NBIA disease genes are involved in processes critical for brain health, including membrane lipid homeostasis regulation. The membrane potential required

Journal of Integrative Neuroscience

for normal neurotransmission and the synaptic connection is created and maintained by the membranes of neurons, which play a critical role in their function. Many of these genes are also important for mitochondrial health and appropriate cellular functions [15]. Discovering how the common networks among these genes and proteins intersect, considering lipid, iron, and energy metabolism, will help us better understand disease pathophysiology and point us in the direction of potential therapeutic targets [14]. As a result, several NBIA illnesses are caused by proteins involved in iron metabolism malfunctioning, interrupting normal cellular iron trafficking and storage. Others are caused by aberrant cellular membrane metabolism and myelin maintenance, which leads to an iron buildup through defective phospholipid and ceramide metabolism. Hence, NBIA is a physiologically and genetically heterogeneous entity with several pathophysiological pathways, varying patterns of iron accumulation, and controversy over whether iron is the major cause of degeneration or only a minor contributing element in each illness [13]. Although a reliable diagnostic criterion for NBIA disorders is unknown, they are rare. NBIA affects roughly 1 in 500,000 persons globally, according to incidences for pantothenate kinase-associated neurodegeneration (PKAN) that contributes to almost 50% of cases [16].

Several studies have been done for mutant genes associated with iron accumulation [17]. Codon preferences are one of the essential factors of a gene associated with NBIA. Codons are triplicates of nitrogenous base and are responsible for synthesizing specific amino acid chains, i.e., peptide or protein. Codons are universal, but due to the degeneracy of codons, sometimes more than one codon code for the same amino acid during protein synthesis is known as synonymous codons. These synonymous codons are not used uniformly in an organism. Thus few codons are preferred over others, and this phenomenon is known as codon usage bias (CUB). CUB study is significant in heterologous gene expression, gene expression level prediction, primer designing, and gene function analysis. Protein synthesis (translation process) is a biological process prone to mistakes. CUB-induced translation mistakes are determined by their impact on protein structure, function, and cellular level. CUB studies were done in various animal models for many diseases. Still, the first work on the nucleotide makeup and CUB of genes related to the central nervous system (CNS) was done by Uddin and Chakraborty [18]. The CUB pattern is unique to each organism and varies from one to another. CUB is thought to be influenced by several parameters, including GC content, expression levels, genetic factors, hydropathicity, and protein aromaticity. It could be used to understand evolutionary phenomena such as adaptive radiation, lateral gene transfer, gene expression level, and codon optimization [19].

We examined the nucleotide composition and codon usage of neurodegeneration with brain iron accumulation to recognize the resemblances and alterations in codon usage preferences between the genes. The analysis also identifies the codons that are over-represented and under-represented in NBIA and housekeeping genes.

2. Materials and methods

2.1 Data retrieval

Genetic Testing Registry (GTR) from National Centre for Biotechnology Information (NCBI) was accessed for neurodegeneration with brain iron accumulation and infantile neuroaxonal dystrophy. A list of genes was available from Prevention Genetics, a United States-based company. Fifteen genes were taken from the list, and genetic sequences were retrieved from NCBI nucleotide. The codon usage data of Homo sapiens was obtained from the codon usage database (https://www.kazusa.or.jp/codon/). The information regarding the genes envisaged is given in the Appendix Table 3.

2.2 Compositional analysis

Compositional analysis was done for genes using CAICal software available online at link http://genomes.urv.es/CAIcal/ [1]. The average percent composition of all nucleotides and their composition at all the three codon positions were determined; composition at the third codon position was used to determine the effects of mutational force. Average nucleotide composition was utilized in skew determination.

2.3 GC content analysis

GC content at all three codon positions was determined. An average of percent composition at GC1 and GC2 was taken as %GC12 and regressed with %GC3 to obtain neutrality. Also, the correlation analysis between GC compositional constraints was carried out.

2.4 Odds ratio analysis

After individual nucleotide analysis, dinucleotide frequency was determined. Since 16 dinucleotide combinations are possible out of four nucleotides, and dinucleotides are not present in the expected frequency. Hence odds ratio for dinucleotide frequency was obtained, that is, the ratio of expected to observed frequency. A ratio value below 0.78 indicates under-representation, and above 1.23 indicates over-representation [20, 21].

2.5 Relative synonymous codon usage (RSCU) analysis

RSCU value of any codon is the proportion of codons' observed frequency compared to the expected frequency of codons when all the synonymous codons are used equally [22]. RSCU values below 0.6 and above 1.6 were considered underrepresented and overrepresented, respectively.

2.6 Codon adaptation index (CAI) determination

The CAI is a measure of gene expression level, and a high CAI value indicates a high expression level and vice versa. The value ranges between 0 and 1. A high CAI value indicates the usage of highly optimized codons supporting the high-level expression of the protein. The CAI value was calculated using the formula given in [23].

2.7 Effective number of codons (ENc) determinations

ENc is a non-directional measure of CUB independent of gene length and the number of amino acids [24]. The ENc values ranged between 20 and 61. The value 20 shows extreme bias where only one codon will be used from all available synonymous codons. Likely, ENc value 61 indicates no bias, and all the available synonymous codons coding for a single amino acid will be used equally. The bias is considered generally high when the ENc value is $\leq$35 [25].

2.8 Parity analysis (P2)

A parity plot is generated to know the impact of mutation and selectional forces on codon usage of a gene [26]. GC bias [G3 / (G3 + C3)] is taken at abscissa, and AT bias A3 / (A3 + U3) is taken at ordinate [27]. The parity plot indicates the bias of nucleotides at the third codon position. If the AT bias value is below 0.5, indicate the preference of pyrimidine over purine [28].

2.9 Neutrality analysis

Regression analysis between %GC3 and %GC12 was done to determine the quantity of mutational force. The slope value is indicative of the balance between mutation and selection forces with other forces. Steep slope reaching towards 1 indicates mutational pressure dominance, and slope values near zero indicate the dominance of selection pressure.

2.10 Intrinsic codon bias index (ICDI)

ICDI refers to the CUB index, and ICDI does not use the frequency of optimal codons. The ICDI values range between 0 and 1, and opposite to ENc, higher values indicate high bias while lower values (below 0.3) indicate low bias [29]. The ICDI values were obtained using COUSIN software online available at https://cousin.ird.fr/ [30].

2.11 Skews calculation

Nucleotide skew is a bias in nucleotide usage. Nucleotide skews, including AT skew, GC skew, purine skew, pyrimidine skew, amino skew, and keto skew values for genes responsible for neurodegeneration with iron accumulation, were computed [31]. The formula used was (A – B) / (A + B), where A and B were respective nucleotides for each skew.

2.12 Correspondence analysis

Correspondence analysis (CA) is used to determine the codon usage pattern among various genes. The data is plotted in a multidimensional space of 59 axes representing 59 codons, excluding the codons coding for methionine and tryptophan [32]. The analysis was done in statistical software Minitab version 17. CA analysis explains the major trends of codon usage.

2.13 Physical properties of the protein (Protein indices)

Protein's physical properties affect the biological properties of protein and CUB and are driven through selectional forces [33]. Protein indices like GRAVY (grand average of hydropathy), AROMA (aromaticity-the frequency of aromatic amino acids), isoelectric point (PI), aliphatic index (AI), hydrophobicity index (HY), instability index (INSTAB), and the numbers of acidic amino acid residues (acidic AA), basic amino acid residues (basic AA), and neutral amino acid residues (neutral AA) were determined using Protparam Expasy [34] and PEPTIDE 2 tool available at commercial website https://www.peptide2.com/.

3. Results

3.1 Nucleotide composition

The average nucleotide analysis revealed that the average percent composition was almost similar for nucleotides C and G (26.16% and 26.96%). Nucleotide A composition was 24.08%, while %T was least between all (22.82%). Here we observed a change in nucleotide percent with the position of the codon. On the one hand, where we observed % composition similar for nucleotide C and G at an overall level, at first codon position, nucleotide composition was similar for %A1 and %C1 (approx. 25%). %G1 was highest (32.54%) while %T1 was the least (16.85%). At the second position of codon %A2 and %T2 were similar (approx. 29.70%), with the lowest percentage of %G2 (18.56). At the third codon, position %C3 and %G3 were very similar (30.86% and 29.79% respectively) with the least occurrence of %A3 (17.23%). When overall %GC was evaluated along with all the three positions of codon (Fig. 1), it was observed that %GC1 was maximum (ranging between 49.43% and 65.49) and %GC2 was least (ranging between 32.95% and 49.21%). Overall %GC composition had significant positive correlation with %GC1 (r = 0.891, $p < 0.001$) and %GC3 (r = 0.958, $p < 0.001$), while no correlation with %GC2.

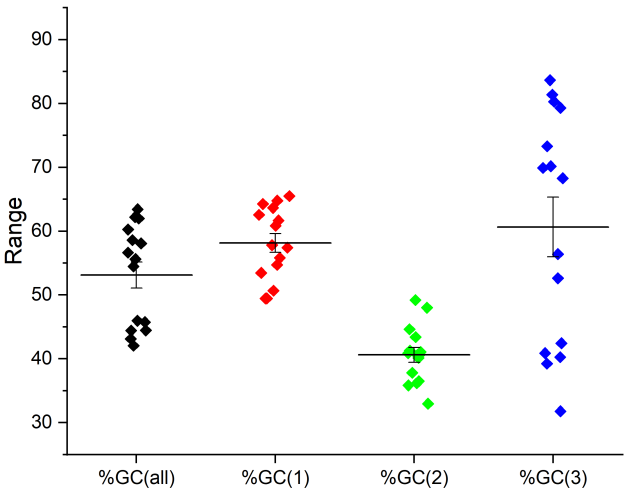

Fig. 1. The %GC compositional analysis in coding sequences. %GC(2) was minimum, while %GC(1) composition was maximum.

A correlation has been found between various nucleotide constraints, and it indicated the presence of both the mutational and selection forces (Fig. 2) affecting codon usage [27]. The figure shows all the kinds of relationships between nucleotides. A positive, negative and no correlation has been observed between the nucleotides. The results sug-

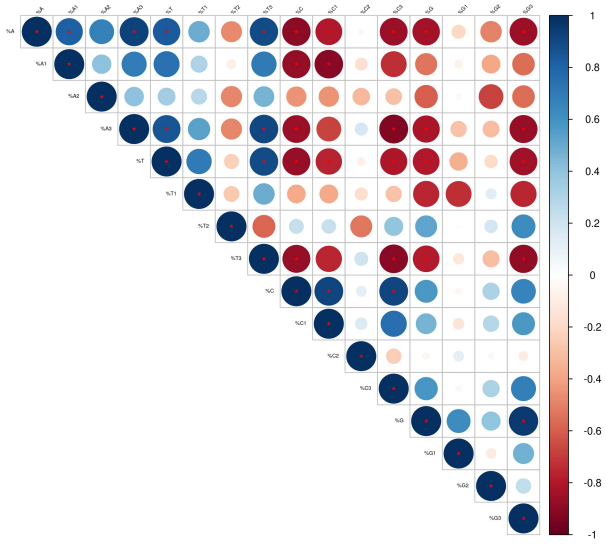

Fig. 2. Mirror plot showing the correlation between various nucleotide compositional constraints. Rectangular boxes show a significant relationship. Blue boxes show a significant positive correlation; red boxes show a significant negative correlation. Empty spaces are indicative of the insignificant correlation.

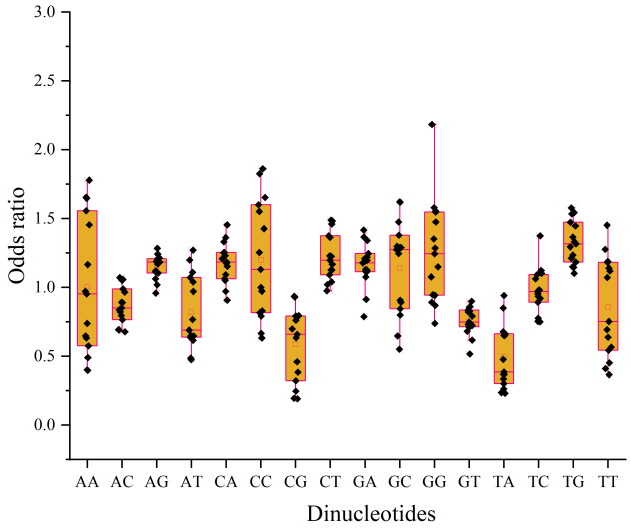

Fig. 3. Jitter and box plot showing the values of the odds ratio of dinucleotides. The figure depicted that maximum variation in odds ratio was present in the case of ApA and CpC. At the same time, dinucleotides ApC, ApG, CpA, CpT, GpA, GpT, TpC, and TpG showed least the variation in odds ratio.

gest the presence of mutation, selection, and compositional constraints in shaping codon usage in a set of genes supposed to be involved in neurodegeneration with brain iron accumulation.

3.2 Dinucleotide odds ratio analysis

Dinucleotide odds ratio analysis indicated that CpG, TpA and GpT dinucleotide were underrepresented (Fig. 3). ApG dinucleotide was randomly used in all the genes. None of the dinucleotides exhibited an odds ratio of more than 1.23 in more than 50% of the genes envisaged.

The heat map (Fig. 4A) shows that GpC, GpG, and TpG were the dinucleotides exhibiting odds ratio >1.23 in more than 50% of genes. ApT, CpG, GpT, TpA and TpT were the dinucleotides underrepresented in more than 50% of the genes.

3.3 RSCU heat map

RSCU analysis revealed that GC ending codons were preferred over AT ending codons with high RSCU values (RSCU >1.6). GTA, ATA, CTA, TTA, GTT, GGT, TCG, ACG, GCG were the codons that were underrepresented (RSCU 0.6) in more than 50% of genes. GCG, TCG, ATA, and CTA were the codons underrepresented in 80% of genes, while codon CTG was overrepresented in 80% of genes (Fig. 4B).

3.4 Gene expression pattern (CAI analysis) in genes

CAI values indicate the gene expression level in host cells. The value ranges from 0 to 1. The values towards one indicate higher expression, while closer to zero indicates low-level expression. Gene expression pattern indicates that gene expression level was moderate to high with CAI values 0.683 (gene *TBCE*) to 0.833 (gene *PLA2G6*). The CAI value had a negative association with CUB (r = –0.923, $p < 0.001$), indicating that with increasing bias, gene expression also increases.

3.5 Gene expression level and odds ratio

Gene expression level had significant positive correlations with an odds ratio of CC, CG, CT, GC, and GG. In contrast, the negative correlation with AA, AT, GA, TA, and TT (Table 1 shows Pearson r value and significance level).

3.6 ENc analysis

The ENc value is a non-directional measure of CUB. The ENc values range from 20 to 61, the value 20 shows the highest bias, and the value 61 shows the least bias. Commonly if a value is ≤35 is called highly biased genes [20], and a value ≥50 is considered near no bias. We found a range of ENc ranging between 39.36 (*ATP13A2*) to 56.06 (*TBCE*), indicating moderate to less bias in codon usage.

3.7 Relationship of nucleotide disproportion and protein indices with CUB

All the six nucleotide skews were calculated (values given in Table 2) and correlated with CUB to evaluate their effects. Upon investigation, it was observed that CUB had a negative association with AT skew (r = –0.893, $p < 0.001$) and purine skew (r = –0.854, $p < 0.001$) while the positive relationship with GC skew (r = 0.904, $p < 0.001$). In an unpublished work of Khandia *et al.* [20], it has been found that protein indices are positively associated with the CUB and PI, acidic and basic residues of proteins, and negatively associated with GRAVY, hydropathicity, and neutral amino acid. We did not find an association between CUB and any protein indices. AT skew

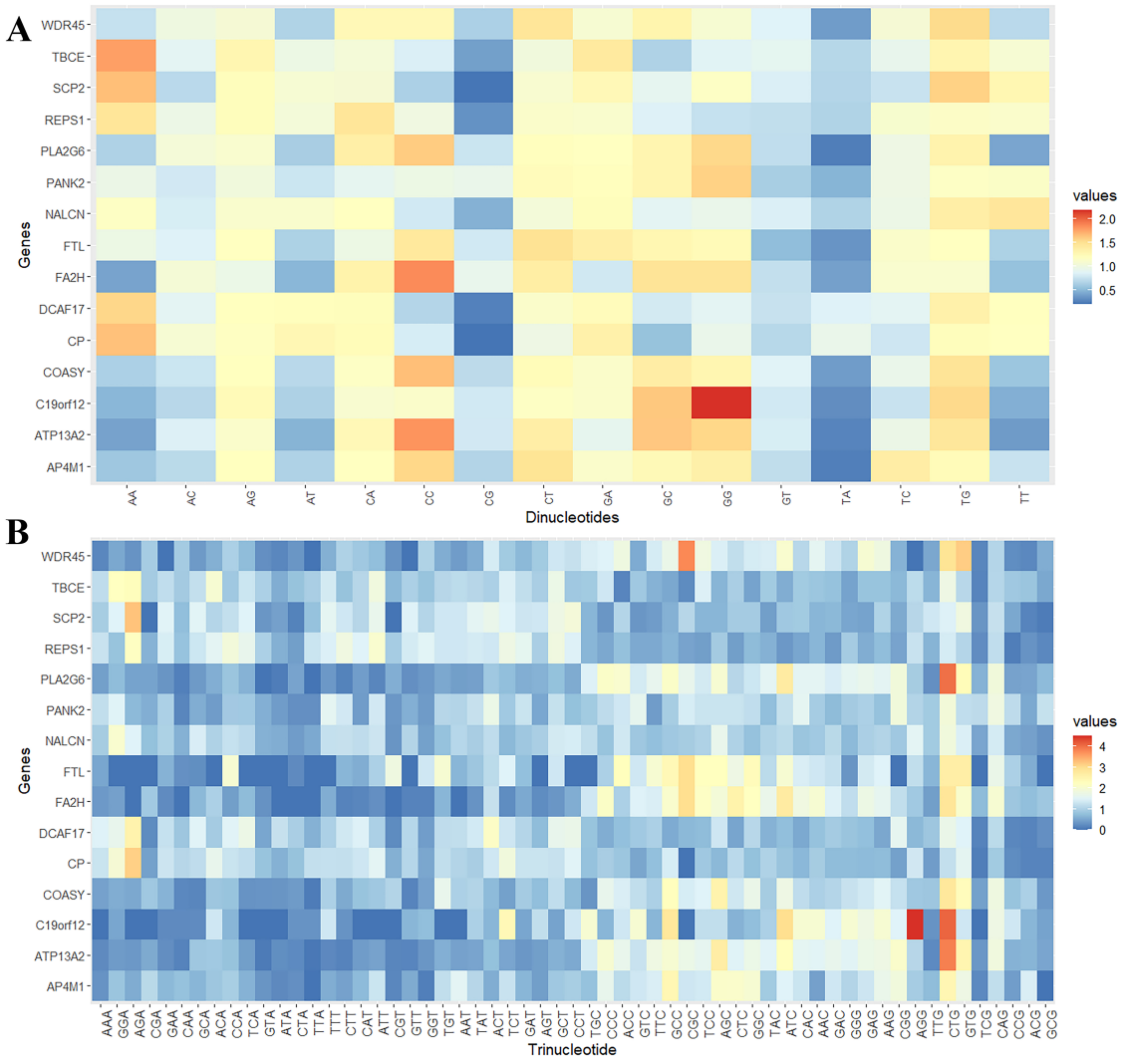

Fig. 4. Heat map for the dinucleotide odds ratio. (A) Each box represents a single gene. Odds ratios are displayed as color-coding, and the color code is given beside the figure. Dinucleotides GpC, GpG, and TpG showed higher odds ratios while ApT, CpG, GpT, TpA and TpT showed lower odds ratios. In the *C19orf12* gene, the odds ratio for GpG was reported highest. (B) Each box represents a single gene. RSCU values are displayed as color-coding, and the color code is given on the side of the figure. The left side of the heat map displays A/T ending codons, while the right-side displays C/G ending codons. CTG codon is overrepresented, while TA and CG ending codons are underrepresented.

was further found to be positively associated with aliphatic index (r = 0.544, $p < 0.05$) and hydrophobicity (r = 0.624, $p < 0.05$). Similarly, purine skew also had positive association with aliphatic index (r = 0.543, $p < 0.05$ and hydrophobicity (r = 0.680, $p < 0.01$). Contrary to AT and purine skews, GC skew had a negative association with aliphatic index and hydrophobicity aliphatic index (r = –0.530, $p < 0.05$ and hydrophobicity (r = –0.631, $p < 0.05$).

3.8 Relationship of CUB with various codons

CUB was significantly positively correlated with few A/T ending codons ($p < 0.05$) and significantly ($p < 0.05$) negatively associated with C/G ending codons (Fig. 5) except for TTG codon that showed a positive association with ENc.

3.9 Parity analysis

Parity analysis gives information regarding bias between A and T; and C and G nucleotides at the third codon position. The average value of GC bias (%G3 / %G3 + %C3) was 0.497 ± 0.06, and AT (%A3 / %A3 + %T3) bias was 0.425 ± 0.064. It indicated that T is preferred over A, and C is preferred over G (Fig. 6).

3.10 Neutrality analysis

The analysis indicated that mutational force played a minor role in shaping codon usage in a set of genes presently envisaged. Mutational force contributed 12.58%, while selectional pressure contributed 87.42% (Fig. 7). The plot also indicated that GC3 is responsible for 31.16% variation in %GC12.

Table 1. Correlation analysis between the odds ratio and CAI value.

	AA	AC	AG	AT	CA	CC	CG	CT
CAI (pearson r)	–0.829	0.176	–0.427	–0.803	0.237	0.812	0.669	0.701
p value	0.000	0.531	0.112	0.000	0.394	0.000	0.006	0.004
Significance	***			***		***	**	**
	GA	GC	GG	GT	TA	TC	TG	TT
CAI (pearson r)	–0.534	0.778	0.593	–0.173	–0.770	0.253	0.158	–0.865
p value	0.040	0.001	0.020	0.539	0.001	0.363	0.573	0.000
Significance	*	**	*		**			***

*** indicates $p < 0.001$, ** indicates $p < 0.01$ and * indicates $p < 0.05$.

Table 2. The genes show the AT skew, GC skew, purine skew, pyrimidine skew, amino skew and keto skew.

Gene	AT skew	GC skew	Purine skew	Pyrimidine skew	Amino skew	Keto skew
ATP13A2	0.596	–0.239	2.341	–1.228	3.205	–2.242
C19orf12	0.547	–0.096	1.423	–1.144	9.200	–1.284
COASY	0.444	0.121	0.572	–0.651	–15.511	–0.919
CP	0.013	0.979	–0.975	455.645	–1.004	1.004
DCAF17	0.057	0.909	–0.882	67.636	–1.026	1.031
FA2H	0.339	0.232	0.188	0.105	0.285	–0.463
FTL	0.300	0.360	–0.091	1.676	–1.115	4.971
PANK2	0.004	0.993	–0.992	1882.615	–1.001	1.001
PLA2G6	0.525	–0.123	1.612	–1.165	6.215	–1.462
TBCE	0.094	0.839	–0.798	40.661	–1.040	1.053
WDR45	0.377	0.295	0.123	0.411	–0.540	–7.409
AP4M1	0.243	0.512	–0.357	5.609	–1.136	1.508
NALCN	–0.031	1.043	–1.060	–121.532	–0.983	0.984
REPS1	0.012	0.979	–0.976	659.908	–1.003	1.003
SCP2	0.106	0.847	–0.778	23.561	–1.068	1.095

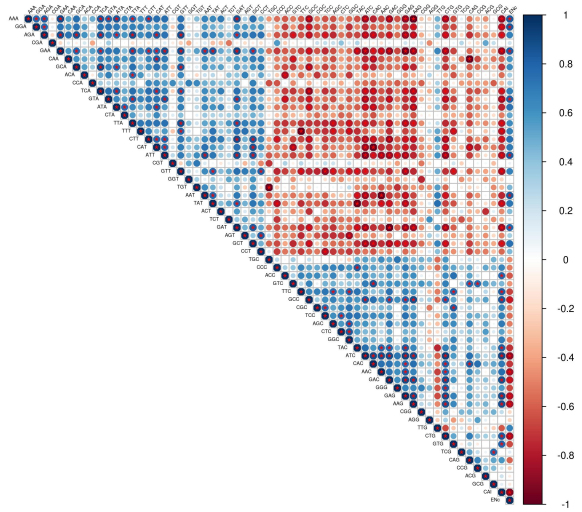

Fig. 5. A plot showing the correlation between ENc value, CAI and RSCU values of codons. Blue circles are showing positive correlation, while red circles are showing negative correlation. Siginificant correlation ($p < 0.05$) is shown by red star within the circles.

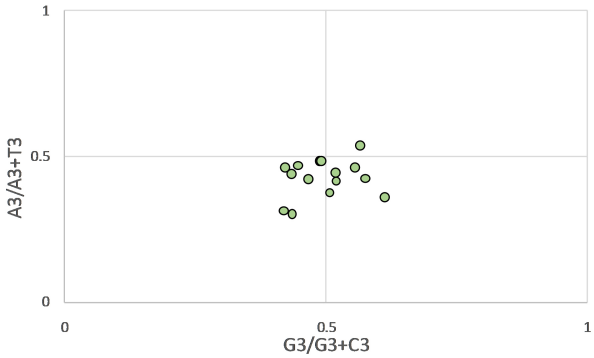

Fig. 6. Parity plot showing AT and GC bias at third codon position. The results indicate the preference of T and C over A and G, respectively. On X-axis, GC bias is displayed, while on the Y axis, AT bias is displayed.

3.11 Relationship between CUB and action of mutational pressure

Though we found a minor role of mutational force in the present analysis, we were still curious to know about the impact of mutational pressure on CUB. Regression analysis was carried out between the third codon position of each nucleotide and ENc. It revealed mutational force contributed a maximum of 67.69% variation in ENc for A nucleotide, while it was least (56.46%) for G nucleotide (Fig. 8).

3.12 Intrinsic codon bias index (ICDI)

The ICDI value is a directional measure of CUB. The values of ICDI ranged between 0.04 (*NALCN*) and 0.466

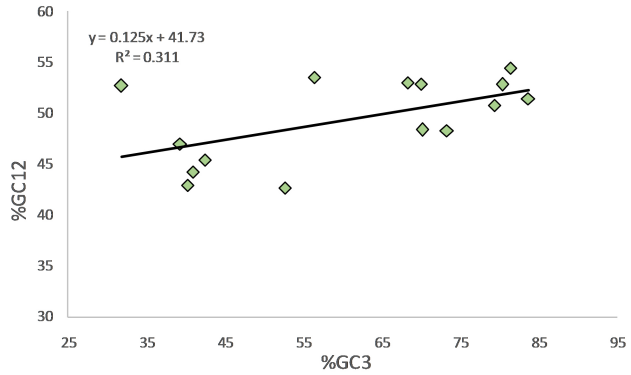

Fig. 7. Regression analysis between %GC3 and %GC12. The plot indicates the dominance of selectional force (87.42%) over mutational pressure (12.58%).

(*C19orf1*). Eighty percent of the genes (12/15) had an ICDI value below (0.3), indicating a general trend of low bias in codon usage.

3.13 Principal component analysis

Principal component analysis revealed that both the PC1 and PC2 are positively associated with the protein hydrophobicity (r = 0.535, $p < 0.05$; r = 0.572, $p < 0.05$ for PC1 and PC2, respectively) while PC1 had a negative association with the frequency of neutral amino acid (r = –0.594, $p < 0.05$) and PC2 had a negative association with frequency of basic amino acid (r = –0.6854, $p < 0.01$). The first two components accounted for 54.82% and 15.90% variation, respectively. Cumulatively first four axes accounted for 83.48% variation. Since most of the genes were found near the X-axis and were not much scattered, these genes did not exhibit much variation. With that, genes like *PANK2, NALCN, TBCE, CP, SCP2, DCAF1, and REPS1* were present in a half where most of the codons were A/T ending, showing that most of these genes are affected by A/T ending codons. On the other hand, *C19orf12, PLA2G6, FA2H, ATP13A2, WDR45, AP4M1* and *COASY* genes were present on the other half of the plot, indicating that these genes are affected by G/C ending codons (Fig. 9). Fig. 10 depicts the summary of each method used and the results obtained.

4. Discussion

Molecular evolutionary investigations suggested that codon usage bias is widespread across the genomes and profoundly contributes to genome evolution [33]. Changes in specific codons with synonymous codons have been quantified at the transcriptome level, and alterations are profoundly present in diseased conditions [34]. Codon usage is associated with specific gene functions also. Genes involved in cell growth and proliferation prefer A/T ending codons, while genes involved in differentiation and specialized cell function prefer G/C ending codons [34]. Also, cellular programs like proliferation and differentiations use a different set of preferred codons [35].

Non-optimal codons are used to regulate the rate of translation [36]. During the process of host adaptation, a differential usage of codon has been displayed in three of the HIV1 genes that are suggestive of host-specific codon usage in a given pathogen [37].

Codon usage bias directly affects mRNA stability and expression level, so synonymous mutations are un-doubtfully involved in various kinds of ailments. They have been reported to link with more than thirty different diseases [38]. The etiology behind this effect is alternative splicing. Alternative splicing produces a highly dynamic proteome responsible for modulating regulating machinery with roles in various diseases [39]. For example, one segment in the *BRCA1* gene showed silent mutations very rarely and evolved very slowly in both human and rodent counterparts. Such synonymous mutations were present in the splicing enhancer region and were so harmful that their carriers died out [40]. Altered codon usage in autophagy gene *IRGM* contributes to Crohn's disease susceptibility [41]. Rare synonymous variants in 4 genes (*CDH23, SLC9A3R1, RHBDD2*, and *ITIH2*) were studied in 67 Caribbean Hispanic families affected with Alzheimer's disease, found at a higher frequency in comparison to reference population of the same ancestry and contribute to Alzheimer etiology [42]. The codon usage might differ in population and elucidate how susceptible a population is to an ailment [43].

Codon usage analysis of the genes involved in neurodegeneration and brain iron accumulation has been envisaged for codon usage analysis and estimating various forces affecting it. Except for methionine and tryptophan, all the amino acids are encoded by two or more two codons known as synonymous codons. Synonymous single nucleotide variants (sSNVs), altered codon encoding for similar amino acids, have been implicated in several genetic disorders due to pre-mRNA splicing, mRNA structure, and miRNA regulation alterations. It also affects the protein translation rate. Effects of synonymous single nucleotide variants have been studied in the Sonic Hedgehog (*SHH*) gene, where usage of sSNVs results in holoprosencephaly, a congenital brain defect resulting in incomplete forebrain cleavage. Eight sSNVs have been observed in holoprosencephaly patients compared to healthy individuals. Out of eight, five sSNVs have been found to reduce the protein expression level ranging from 5% to 23% in cell assays [44]. It underlines the importance of synonymous codon usage studies in clinical setup and helps understand the biased codon usage-based pathologies in human diseases.

There are mainly compositional constraints, selection pressure, and mutational pressure involved in shaping codon usage bias [27]. Suppose a significant correlation is present between various compositional constraints. In that case, it is indicative of the presence of all the three compositional, mutational, and selection forces against translation efficiency, accuracy, and RNA folding [45], and the same is found here, indicating the presence of these three forces in shaping codon usage.

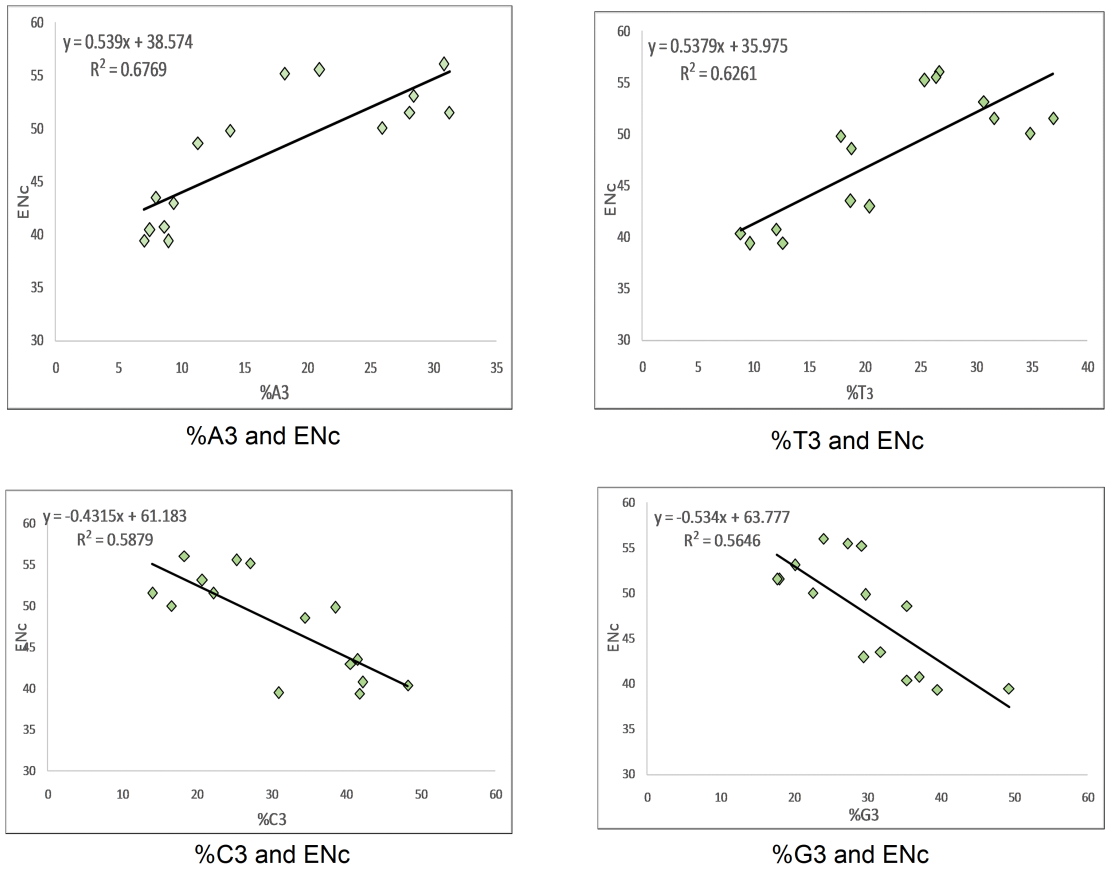

Fig. 8. Regression analysis between ENc and nucleotide composition at third codon position to determine effects of mutational force on ENc. The contribution of mutational force on the CUB was least for nucleotide C, while the mutational force on three nucleotides (A, T and G) contributed almost equally for CUB (approximately 53%).

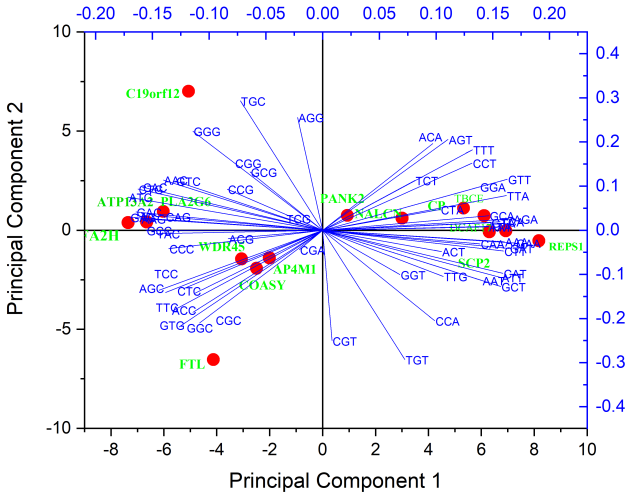

Fig. 9. PCA analysis of a set of genes involved in brain degeneration associated with iron accumulation. The genes have been shown with the red color font. Each red dot represents the position of a gene. The codons have been depicted with the blue color font. *C19orf12* and *FTL* genes showed a distinct codon usage pattern from other genes. The remaining genes scattered around X-axis indicated not much variation in their RSCU values.

The average composition of nucleotide C and G was equal with similar composition at the third codon position; also, the composition of C and G nucleotide did not follow the same trend at first and second codon positions. The average nucleotide %T composition, which was least among all nucleotides, was not least at the third codon position. We observed an interesting feature in compositional constraints related to the set of genes. Two out of four nucleotides have similar compositions at all codon positions. For the first codon position, it is %A1 and %C1 (approx. 25%), for the second position %A2 and %T2 (approx. 29.70%) and the third codon position %G3 and %C3 is similar (approx. 30%). However, we do not have any reasonable explanation for this on the current date.

In the double-stranded DNA, the GC molar content ranges between 13% and ≈ 75%, and the GC content exhibit a strong correlation with mean values for GC1, GC2, and GC3 [46]. Genome GC content is linked with amino acid and codon usage [47]. The higher the overall GC composition, the higher the GC1, GC2, and GC3. A significant positive correlation and a considerable variation in the GC content will result in greater CUB [48]. However, in our results, overall GC content had a significant positive correlation with %GC1 and %GC3 ($p < 0.001$), while no correlation

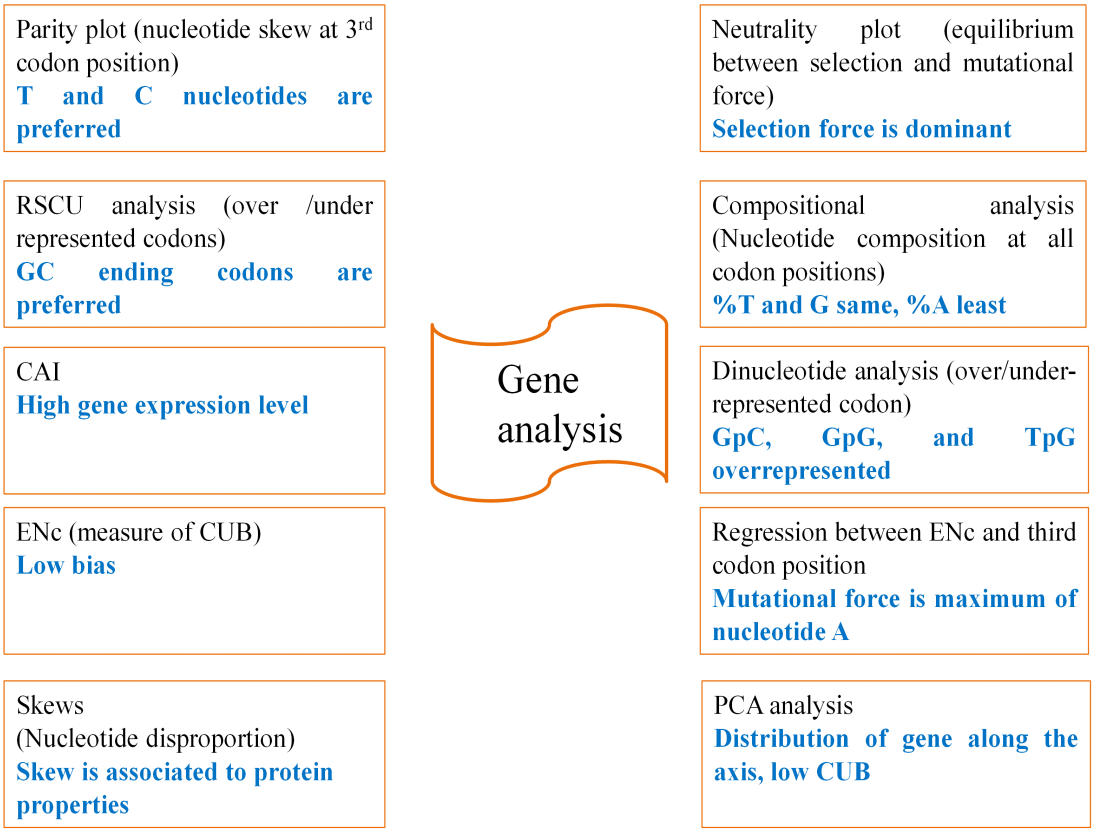

Fig. 10. Diagrammatic presentation of the different methods used and the outcomes. Overall analysis revealed that C ending codons are preferred based on the parity and RSCU analysis. The genes exhibited high gene expression levels with an association of nucleotide skew with protein properties. Overall, genes exhibit a lower bias, which is also affirmed by PCA analysis. The selectional force was dominant in shaping codon usage, and overrepresented dinucleotides were ascertained.

with %GC2, which indicated that significant enhancement or reduction in the overall GC component would not affect the %GC2 component as well as bias.

ApC, ApG, CpA, GpT and TpG dinucleotides have been over-represented in *pp1a*, and *pp1ab* genes of infectious bronchitis virus [49]. Contrary to the finding of Franzo *et al.* [11], Kunec and Osterrieder found GpT dinucleotide as the third-lowest relatively abundant dinucleotide in human viruses. Along with CpG and TpA, GpT has been found underrepresented, indicating that both in viruses and humans, the dinucleotide abundance could follow a similar pattern. We also can speculate that the genes involved in neurodegeneration with iron accumulation in the brain originated from viral genes that are incorporated and became part of the human genome and gained functionality over a long period of evolution. Underrepresentation of CpG and TpA can be considered an effect of selectional pressure [50] since TpA dinucleotide is present in two canonical stop codons (TAA and TAG). Also, TpA dinucleotides are prone to degradation through cellular RNases. Reducing TpA will help eliminate the risk of mutations and subsequent protein truncation and degradation through RNases. CpG suppression could be explained based on the phenomenon of cytosine methylation at the 5th carbon of cytosine and subsequent deamination that results in the C→ T transition. As a result, CpG may convert into TpG/CpA [51].

Contrary to the underabundance of CpG in other organisms, in *Mycobacterium lepromatosis,* CpG dinucleotide possesses a high odds ratio, possibly owing to its association with higher proinflammatory cytokine production and NF-κB activation that eventually leads to high pathogenicity of *M. lepromatosis* [40]. TpG dinucleotide had a maximum odds ratio (1.327), indicating its over-representation. TpG dinucleotide is a part of the CTG codon, and the CTG codon is over-represented in genes common to cancer and primary immunodeficiencies [52]. We found the same phenomenon, and CTG was over presented in 80% of the genes envisaged. The codons which were underrepresented in 80% of genes were ATA, CTA, TCG and GCG. Underrepresentation of these codons can be understood because these codons contain either TpA or CpG dinucleotides as a canonical part, and the same is reflected in codons.

CAI value determines gene expression and ranges between 0 and 1 [53]. Higher CAI values exhibit a higher expression level of a gene and are determined by comparing its codon usage frequency with the reference set. For any syn-

thetic construct to be expressed in a heterologous host at a higher level, several algorithms are available that maximize the CAI to obtain a higher level of expression [54]. For *E. coli* genes highest CAI value of 0.85 is for the most abundant *lpp* gene, which encodes for an outer membrane lipoprotein. The *pnp* gene also used a biased codon with moderate CAI 0.63 [55]. We observed CAI values 0.683 (gene *TBCE*) to 0.833 (gene PLA2G6), suggestive of the presence of CUB ranging from moderate to high.

Similar to CAI, ENc is also a measure of codon usage. Considering the ENc, in Drosophila, female-biased genes showed lower ENc and higher CUB while male-biased genes showed high ENc and lower codon bias. ENc also indicates host-specific distribution in the LT-Ag gene of polyomaviruses in aves, fishes and mammalian hosts [56]. ENc indicated a range of 39.36 (*ATP13A2*) to 56.06 (*TBCE*), indicating moderate to less bias in codon usage, further supporting the findings obtained from CAI analysis. Similarly, lower CUB is observed in genes involved in Krabbe disease [57] and the central nervous system [18]. The bias is increased in longer genes to use optimal codons to prevent non-sense errors in energetically costly lengthy genes [58]. Highly expressed genes show greater bias than other genes [59]. During the process of evolution, codons undergo adaptive selection to obtain a required level of protein synthesis. Housekeeping and other proteins like DNA binding proteins must use more favored codons in higher amounts.

Krabbe disease is a rare neurodegenerative disorder characterized by a mutation in the *GALC* gene, resulting in the deficiency of the enzyme galactosylceramidase. It is a lysosomal storage disorder and causes the accumulation of cytotoxic metabolites, viz. galactosylsphingosine or psychosine, in the lysosome [60]. CUB studies in four isoforms of the *GALC* gene were performed by Das *et al.* [57], and the mean ENc for the four isoforms was 56.98, which suggests a low CUB of the *GALC* gene.

Compositional bias tends to affect the CUB, and it is considered that GC bias in any genome affects CUB. Palidwor *et al.* [61] observed in species of prokaryotes, plants, and human genes that G/C ending codon usage also will be increased with an increment in genomic GC content and vice versa for A/T ending codons [62]. In the genes associated with brain iron accumulation, a preference of GC ending codons was observed over AT ending codons. Our observation differed from Das *et al.* [57]. They reported preference of AT ending codons in the *GALC* gene of humans, responsible for rare neurodegenerative Krabbe disease. In the genes associated with neurodegeneration and iron accumulation, it was observed that ENc has a significant positive association with many of the A/T ending codons, and a significant negative association with G/C ending codons signifies that G/C ending codons are associated with higher bias. In concordance to the results of Palidwor *et al.* [61], with the increase in the GC content of the gene, an increase in usage of G/C ending codons also is observed. However, we found an exception to this general phenomenon. G ending codon TTG encoding for leucine was observed to have positive association with ENc, indicating that if the genomic GC content is increased, the usage of leucine encoding is increased codon TTG will be decreased.

Unlike our result, Palidwar *et al.* [61] observed two G-ending codons, AGG (arginine) and TTG (leucine), showed a decrease in overall usage with an increment of GC component [61]. In the genes related to Alzheimer's and other neurodegenerative diseases, Yang *et al.* [62] found that C ending codons are preferred in all the highly expressed genes. In contrast, G ending codons either remained constant (AGG, coding for Arg) or showed a marked fall, even disappearing (TTG, encoding Leu). Similar to the finding of Yang *et al.* [62], in genes associated with iron accumulation in the brain, a positive association of expression level and C ending codons was observed. However, TGC (encoding Cys) and GTC (encoding for Val) were two codons, which did not influence the gene expression [62].

On the one hand, where G ending codons either remained constant (AGG, coding for Arg) or exhibited a marked fall (TTG encoding Leu), in Alzheimer's and other neurodegenerative disease genes, we observed no association between gene expression and CGG and AGG (both encoding for Arg) in our case. However, for codon TTG (encoding for Leu), a negative influence on gene expression has been observed, similar to Yang *et al.* [62]. The results suggest a typical feature adapted by genes related to neurodegenerative disorders. The driving force for codon usage bias could be mutation, genetic drift, and/or biased gene conversion, and selection forces like the section for translational efficiency and/or accuracy. Therefore, a biased codon might reflect the nucleotide disproportion in local base composition [63].

We found a negative relationship between AT skew and purine skew, while a positive relationship between CUB and GC skew indicates the impact of nucleotide disproportion of CUB. AT, GC and purines Skews had been found linked to aliphatic index and hydrophobicity, indicating the effects of nucleotide disproportion of physical properties of the protein. The results concordance with Khandia *et al.* [52], who also observed an association between nucleotide skew and protein properties [52]. Effects of nucleotide skew on codon position were observed, and it was observed that T is preferred over A and C is preferred over G at the third codon position as depicted by parity analysis. Similar results were observed in *T. saginata*, where C and T were used more frequently than G and A [50]. Differed to parity analysis results, which indicated nucleotide T and C, GC ending codons are preferred as per RSCU analysis. The difference in parity analysis and RSCU analysis results, as depicted in Kumar *et al.* [64], also found dominance of A and G in parity analysis and Dominance of G/C in RSCU analysis.

Similarly, in the equine Influenza virus coding sequences, all polymerase genes had AT bias GC content at the third codon position in fourfold degenerate codon families [65].

The bias also indicated the role of selectional force. In a regression plot between %GC3 and %GC12, it was evident that relative neutrality was less (12.58%), and therefore the role of selection pressure is dominant. Similar results have been obtained by He *et al.* [66], where they found selectional force as the dominant force in shaping codon usage in *Ginkgo biloba*, one of the most ancient tree species [66]. Contrary results had been obtained by Nasrullah *et al.* [67], who found mutational force as the dominant force over the selection on the genome of the Marburg virus [67]. The mutational force was not the major force in shaping the codon usage pattern.

Still, we were anxious to know the amplitude of mutation pressure on CUB. Mutational pressure affected A and T nucleotide positively and C and G nucleotide negatively (Regression plot between ENc and nucleotide composition at third codon position), and the mutational force on A and T nucleotides equally affected CUB, while in the case of G and C nucleotides, mutational force on C nucleotide, affected codon usage least. Since ENc is a non-directional measure of codon usage [68], the negative regression coefficient for C and G indicated that these two nucleotides would tend to increase bias if their content is increased [18]. Unlike ENc, ICDI is a directional measure of CUB [69], and ICDI trends indicated low bias.

Principal component analysis revealed that first and second components are associated with protein hydrophobicity, and basic and amino acid frequencies indicate selective forces choosing protein properties [70]. Also, few genes were grouped with A/T ending codons while others grouped with G/C ending codons. It indicates the effects of composition on codon usage. Upon PCA, almost all genes gathered near the X-axis and displayed less bias.

5. Conclusions

Codon usage of any gene is mainly driven by mutational, selectional, and compositional forces. A significant correlation between compositional constraints suggests that the above forces shape codon usage patterns in the envisaged genes. Average C and G nucleotide composition was equal in the genes. Still, the same is not observed at all codon positions, indicating the presence of forces responsible for skewness in nucleotide composition. CpG, TpA, and GpT dinucleotides were underrepresented with the overrepresentation of TpG dinucleotide. Like the dinucleotide frequency, ATA, CTA, TCG, and GCG codons encompassing TpA and CpA dinucleotides were underrepresented, and CTG codon encompassing TpG dinucleotide is over-represented, showing effects of compositional parameters on codon usage. G/C ending codons are prevalent in the GC-rich genome, while in AT-rich sequences, A/T ending codons are prevalent. However, a single exception of this phenomenon was codon TTG, which had a positive association with ENc, suggesting a decrease in TTG codon frequency with an increase in GC content. Based on moderate to high CAI values, moderate to high-level gene expression indicates translation efficacy and translational selection. The results of CAI are supported by the results of ENc, which is indicative of moderate to high CUB. Indices of compositional skews like AT, GC and purines skews are associated with some of the protein's physical properties. T and C nucleotides are preferred over A and G nucleotides at the third codon position, as indicated by parity analysis. The relative neutrality value of 12.58% refers to the minor role of mutational force over selectional force. Mutational force acted equally on G, A, and T nucleotide, while acted least on nucleotide C. PCA analysis revealed a low level of codon bias and role of selection pressure. Cumulatively looking at all analyses performed, it was observed that compositional, mutational, and selectional forces were applied on the set of genes considered to participate in neurodegeneration with iron accumulation with selection force as dominant.

Abbreviations

CAI, Codon adaptation index; RSCU, Relative synonymous codon usage; ENc, effective number of codon; SCS, scaled chi-square; ICDI, Intrinsic codon bias index.

Author contributions

Conceptualization—RK, TA, MAAlg and NP; methodology—TA, RK, NP, AMA, MAAlg and MAAlm; software—RK, TA, NP, and AMA; validation—RK, TA, and MAAlg and MAAlm; formal analysis—RK, TA, and MAAlg, MAAlm; investigation—RK, NP, MAAlg, AMA and MAAlm; resources—RK, TA, and AMA. MAAlg; experimentation, data curation—RK, NP, TA, MAAlg and MAAlm; writing—original draft preparation—RK, NP, AMA, and MAA. MAAlg and MAAlm; writing—review and editing—RK, TA, and AMA, MAAlg and MAAlm; visualization—RK, TA; supervision—RK; project administration—RK, TA, NP and AMA. MAAlg and MAAlm; funding acquisition—TA, AMA, MAAlg and MAAlm. All authors have read and agreed to the published version of the manuscript. All authors read and agreed on the final version of the manuscript.

Ethics approval and consent to participate

Not applicable.

Acknowledgment

All authors acknowledge their respective institutes and organization for providing support.

Funding

Personal funding.

Conflict of interest

The authors declare no conflict of interest.

Appendix

See Table 3.

Table 3. Gene information is associated with neurodegeneration with brain iron accumulation and infantile neuroaxonal dystrophy.

S. No	Gene	Synonyms	Chromosome location	Associated Disease	Function
1	ATP13A2 (ATPase cation transporting 13A2)	CLN12; KRPPD; PARK9; SPG78; HSA9947	1p36.13	Neurodegeneration with Brain Iron Accumulation	Transports inorganic cations as well as other substrates
2	C19orf12 (chromosome 19 open reading frame 12)	MPAN; NBIA3; NBIA4; SPG43	19q12	Neurodegeneration with Brain Iron Accumulation 4	Specific function of the encoded protein is unknown
3	COASY (Coenzyme A synthase)	NBP; DPCK; PPAT; UKR1; NBIA6; PCH12; pOV-2	17q21.2	Pontocerebellar Hypoplasia, Type 12	Biosynthesis of CoA from pantothenic acid (vitamin B5)
4	CP (Ceruloplasmin)	CP-2	3q24–q25.1	Aceruloplasminemia	Binds most of the copper in plasma and is involved in the peroxidation of Fe(II)transferrin to Fe(III) transferrin
5	DCAF17 (DDB1 and CUL4 associated factor 17)	C2orf37; C20orf37	2q31.1	Woodhouse-Sakati Syndrome	Associates with cullin 4A/damaged DNA binding protein 1 ubiquitin ligase complex
6	FA2H (fatty acid 2-hydroxylase)	FAAH; FAH1; SCS7; SPG35; FAXDC1	16q23.1	Leukodystrophy demyelinating	Many cellular processes and their structural diversity arises
7	FTL (ferritin light chain)	LFTD, NBIA3	19q13.33	Hyperferritinemia with or Without Cataract	Major intracellular iron storage protein in prokaryotes and eukaryotes
8	PANK2 (pantothenate kinase2)	C20orf48, HARP, HSS, NBIA1, PKAN	20p13	Neurodegeneration with Brain Iron Accumulation 1	Key regulatory enzyme in the biosynthesis of coenzyme A (CoA) in bacteria and mammalian cells.
9	PLA2G6 (phospholipase A2 group VI)	GVI; PLA2; INAD1; NBIA2; iPLA2; NBIA2A; NBIA2B; PARK14; PNPLA9; CaI-PLA2; IPLA2-VIA; iPLA2beta	22q13.1	Neurodegeneration with Brain Iron Accumulation 2a	Play a role in phospholipid remodeling, arachidonic acid release, leukotriene and prostaglandin synthesis, fas-mediated apoptosis, and transmembrane ion flux in glucose-stimulated B-cells
10	TBCE (tubulin folding cofactor E)	HRD; KCS; KCS1; pac2; PEAMO	1q42.3	Hypoparathyroidism-Retardation-Dysmorphism Syndrome	Play a role in capturing and stabilizing beta-tubulin intermediates in a quasi-native confirmation
11	WDR45 (WD repeat domain 45B)	NEDSBAS, WDR45L, WIPI-3, WIPI3	17q25.3	Dystonia	Mediates protein-protein interactions and a conserved motif for interaction with phospholipids.
12	AP4M1 (adaptor related protein complex 4 subunit mu 1)	MU-4; CPSQ3; SPG50; MU-ARP2	7q22.1	Spastic Paraplegia 50, Autosomal Recessive	Involved in recognizing and sorting cargo proteins with tyrosine-based motifs from the trans-Golgi network to the endosomal-lysosomal system.
13	NALCN (sodium leak channel, non-selective)	CLIFAHDD, CanIon, IHPRF, IHPRF1, INNFD, VGCNL1, bA430M15.1	13q32.3–q33.1	Congenital Contractures of the Limbs and Face, Hypotonia, and Developmental Delay	Family of voltage-gated sodium and calcium channels that regulates the resting membrane potential and excitability of neurons
14	REPS1 (RALBP1 associated Eps domain-containing protein 1)	NBIA7; RALBP1	6q24.1	Neurodegeneration with Brain Iron Accumulation 7	Proteins that participate in signaling, endocytosis and cytoskeletal changes
15	SCP2 (sterol carrier protein 2)	NLTP, NSL-TP, SCOX, SCP-2, SCP-CHI, SCP-X, SCPX	1p32.3	Leukoencephalopathy with Dystonia and Motor Neuropathy	SCPx protein is a peroxisome-associated thiolase that is involved in the oxidation of branched-chain fatty acids, while the SCP2 protein is thought to be an intracellular lipid transfer protein

References

[1] Puigbò P, Bravo IG, Garcia-Vallve S. CAIcal: A combined set of tools to assess codon usage adaptation. Biology Direct. 2008; 3: 38.

[2] Daglas M, Adlard PA. The Involvement of Iron in Traumatic Brain Injury and Neurodegenerative Disease. Frontiers in Neuroscience. 2018; 12: 981.

[3] Kruit MC, Launer LJ, Overbosch J, van Buchem MA, Ferrari MD. Iron Accumulation in Deep Brain Nuclei in Migraine: A Population-Based Magnetic Resonance Imaging Study. Cephalalgia. 2009; 29: 351–359.

[4] Hogarth P. Neurodegeneration with Brain Iron Accumulation: Diagnosis and Management. Journal of Movement Disorders. 2015; 8: 1–13.

[5] Haba-Rubio J, Staner L, Petiau C, Erb G, Schunck T, Macher JP. Restless legs syndrome and low brain iron levels in patients with hemochromatosis. Journal of Neurology, Neurosurgery and Psychiatry. 2005; 76: 1009–1010.

[6] Campbell KA, Bank B, Milgram NW. Epileptogenic effects of electrolytic lesions in the hippocampus: Role of iron deposition. Experimental Neurology. 1984; 86: 506–514.

[7] Burgetova A, Seidl Z, Krasensky J, Horakova D, Vaneckova M. Multiple sclerosis and the accumulation of iron in the basal ganglia: quantitative assessment of brain iron using MRI t(2) relaxometry. European Neurology. 2010; 63: 136–143.

[8] Miszkiel KA, Paley MN, Wilkinson ID, Hall-Craggs MA, Ordidge R, Kendall BE, *et al.* The measurement of R2, R2* and R2′ in HIV-infected patients using the prime sequence as a measure of brain iron deposition. Magnetic Resonance Imaging. 1997; 15: 1113–1119.

[9] Waldvogel D, van Gelderen P, Hallett M. Increased iron in the dentate nucleus of patients with Friedrich's ataxia. Annals of Neurology. 1999; 46: 123–125.

[10] Brar S, Henderson D, Schenck J, Zimmerman EA. Iron accumulation in the substantia nigra of patients with Alzheimer's disease and parkinsonism. Archives of Neurology. 2009; 66: 371–374.

[11] Lee JH, Lee MS. Brain Iron Accumulation in Atypical Parkinsonian Syndromes: *in vivo* MRI Evidences for Distinctive Patterns. Frontiers in Neurology. 2019; 10: 74.

[12] Lehéricy S, Roze E, Goizet C, Mochel F. MRI of neurodegeneration with brain iron accumulation. Current Opinion in Neurology. 2020; 33: 462–473.

[13] Schneider SA. Neurodegeneration with Brain Iron Accumulation. Current Neurology and Neuroscience Reports. 2016; 16: 1–9.

[14] Gregory A, Hayflick SJ. Genetics of Neurodegeneration with Brain Iron Accumulation. Current Neurology and Neuroscience Reports. 2011; 11: 254–261.

[15] Meyer E, Kurian MA, Hayflick SJ. Neurodegeneration with Brain Iron Accumulation: Genetic Diversity and Pathophysiological Mechanisms. Annual Review of Genomics and Human Genetics. 2015; 16: 257–279.

[16] Hayflick SJ, Kurian MA, Hogarth P. Neurodegeneration with brain iron accumulation. Handbook of Clinical Neurology. 2018; 147: 293–305.

[17] Mcneill A, Chinnery PF. Neurodegeneration with brain iron accumulation. Handbook of Clinical Neurology. 2011; 100: 161–172.

[18] Uddin A, Chakraborty S. Codon Usage Pattern of Genes Involved in Central Nervous System. Molecular Neurobiology. 2019; 56: 1737–1748.

[19] Chakraborty S, Barbhuiya PA, Paul S, Uddin A, Choudhury Y, Ahn Y, *et al.* Codon usage trend in genes associated with obesity. Biotechnology Letters. 2020; 42: 1865–1875.

[20] Khandia R, Singhal S, Kumar U, Ansari A, Tiwari R, Dhama K, *et al.* Analysis of Nipah Virus Codon Usage and Adaptation to Hosts. Frontiers in Microbiology. 2019; 10: 886.

[21] Kunec D, Osterrieder N. Codon Pair Bias is a Direct Consequence of Dinucleotide Bias. Cell Reports. 2016; 14: 55–67.

[22] Khattak S, Rauf MA, Zaman Q, Ali Y, Fatima S, Muhammad P, *et al.* Genome-Wide Analysis of Codon Usage Patterns of SARS-CoV-2 Virus Reveals Global Heterogeneity of COVID-19. Biomolecules. 2021; 11: 912.

[23] Brandão PE. The evolution of codon usage in structural and non-structural viral genes: the case of Avian coronavirus and its natural host Gallus gallus. Virus Research. 2013; 178: 264–271.

[24] Wright F. The 'effective number of codons' used in a gene. Gene. 1990; 87: 23–29.

[25] Butt AM, Nasrullah I, Tong Y. Genome-wide analysis of codon usage and influencing factors in chikungunya viruses. PLoS ONE. 2014; 9: e90905.

[26] Gun L, Yumiao R, Haixian P, Liang Z. Comprehensive Analysis and Comparison on the Codon Usage Pattern of whole Mycobacterium tuberculosis Coding Genome from Different Area. BioMed Research International. 2018; 2018: 3574976.

[27] Nath Choudhury M, Uddin A, Chakraborty S. Codon usage bias and its influencing factors for Y-linked genes in human. Computational Biology and Chemistry. 2017; 69: 77–86.

[28] Wang L, Xing H, Yuan Y, Wang X, Saeed M, Tao J, *et al.* Genome-wide analysis of codon usage bias in four sequenced cotton species. PLoS ONE. 2018; 13: e0194372.

[29] Freire-Picos MA, Gonzalez-Siso MI, Rodríguez-Belmonte E, Rodríguez-Torres AM, Ramil E, Cerdan ME. Codon usage in Kluyveromyces lactis and in yeast cytochrome c-encoding genes. Gene. 1994; 139: 43–49.

[30] Bourret J, Alizon S, Bravo IG. COUSIN (COdon Usage Similarity INdex): A Normalized Measure of Codon Usage Preferences. Genome Biology and Evolution. 2019; 11: 3523–3528.

[31] Deb B, Uddin A, Chakraborty S. Codon usage pattern and its influencing factors in different genomes of hepadnaviruses. Archives of Virology. 2020; 165: 557–570.

[32] Hassan S, Mahalingam V, Kumar V. Synonymous Codon Usage Analysis of Thirty Two Mycobacteriophage Genomes. Advances in Bioinformatics. 2009; 2009: 316936.

[33] Deb B, Uddin A, Chakraborty S. Composition, codon usage pattern, protein properties, and influencing factors in the genomes of members of the family Anelloviridae. Archives of Virology. 2021; 166: 461–474.

[34] Gasteiger E, Hoogland C, Gattiker A, Duvaud S, Wilkins MR, Appel RD, *et al.* Protein Identification and Analysis Tools on the ExPASy Server. In Walker JM (ed.) The Proteomics Protocols Handbook (pp. 571–607). Humana Press: Totowa. 2005.

[35] Gingold H, Tehler D, Christoffersen NR, Nielsen MM, Asmar F, Kooistra SM, *et al.* A Dual Program for Translation Regulation in Cellular Proliferation and Differentiation. Cell. 2014; 158: 1281–1292.

[36] Whittle CA, Kulkarni A, Extavour CG. Evidence of multifaceted functions of codon usage in translation within the model beetle Tribolium castaneum. DNA Research. 2019; 26: 473–484.

[37] Pandit A, Sinha S. Differential Trends in the Codon Usage Patterns in HIV-1 Genes. PLoS ONE. 2011; 6: e28889.

[38] Sauna ZE, Kimchi-Sarfaty C. Understanding the contribution of synonymous mutations to human disease. Nature Reviews Genetics. 2011; 12: 683–691.

[39] Wang GS, Cooper TA. Splicing in disease: disruption of the splicing code and the decoding machinery. Nature Reviews Genetics. 2007; 8: 749–761.

[40] Chamary JV, Hurst LD. The Price of Silent Mutations. Scientific American. 2009; 300: 46–53.

[41] Parkes M, Barrett JC, Prescott NJ, Tremelling M, Anderson CA, Fisher SA, *et al.* Sequence variants in the autophagy gene IRGM and multiple other replicating loci contribute to Crohn's disease susceptibility. Nature Genetics. 2007; 39: 830–832.

[42] Tang M, Alaniz ME, Felsky D, Vardarajan B, Reyes-Dumeyer D, Lantigua R, *et al.* Synonymous variants associated with Alzheimer's disease in multiplex families. Neurology: Genetics. 2020; 6: e450.

[43] Hodgman MW, Miller JB, Meurs TE, Kauwe JSK. CUBAP: an interactive web portal for analyzing codon usage biases across populations. Nucleic Acids Research. 2020; 48: 11030–11039.

[44] Kim A, Le Douce J, Diab F, Ferovova M, Dubourg C, Odent S, *et al.*

Synonymous variants in holoprosencephaly alter codon usage and impact the Sonic Hedgehog protein. Brain. 2020; 143: 2027–2038.
[45] Zahdeh F, Carmel L. Nucleotide composition affects codon usage toward the 3'-end. PLoS ONE. 2019; 14: e0225633.
[46] Simón D, Cristina J, Musto H. Nucleotide Composition and Codon Usage Across Viruses and Their Respective Hosts. Frontiers in Microbiology. 2021; 12: 646300.
[47] Ermolaeva MD. Synonymous Codon Usage in Bacteria 91 Synonymous Codon Usage in Bacteria. Current Issues in Molecular Biology. 2001; 3: 91–97.
[48] Knight RD, Freeland SJ, Landweber LF. A simple model based on mutation and selection explains trends in codon and amino-acid usage and GC composition within and across genomes. Genome Biology. 2001; 2: research0010.1–0010.13.
[49] Franzo G, Tucciarone CM, Legnardi M, Cecchinato M. Effect of genome composition and codon bias on infectious bronchitis virus evolution and adaptation to target tissues. BMC Genomics. 2021; 22: 244.
[50] Munjal A, Khandia R, Shende KK, Das J. Mycobacterium lepromatosis genome exhibits unusually high CpG dinucleotide content and selection is key force in shaping codon usage. Infection, Genetics and Evolution. 2020; 84: 104399.
[51] Bird AP. DNA methylation and the frequency of CpG in animal DNA. Nucleic Acids Research. 1980; 8: 1499–1504.
[52] Khandia R, Alqahtani T, Alqahtani AM. Genes Common in Primary Immunodeficiencies and Cancer Display Overrepresentation of Codon CTG and Dominant Role of Selection Pressure in Shaping Codon Usage. Biomedicines. 2021; 9: 1001.
[53] Sharp P, Li W. The codon adaptation index-a measure of directional synonymous codon usage bias, and its potential applications. Nucleic Acids Research. 1987; 15: 1281–1295.
[54] Jansen R. Revisiting the codon adaptation index from a whole-genome perspective: Analyzing the relationship between gene expression and codon occurrence in yeast using a variety of models. Nucleic Acids Research. 2003; 31: 2242–2251.
[55] DiRienzo JM, Nakamura K, Inouye M. The Outer Membrane Proteins of Gram-Negative Bacteria: Biosynthesis, Assembly, and Functions. Annual Review of Biochemistry. 1978; 47: 481–532.
[56] Cho M, Kim H, Son HS. Codon usage patterns of LT-Ag genes in polyomaviruses from different host species. Virology Journal. 2019; 16: 137.
[57] Wilkins MR, Gasteiger E, Bairoch A, Sanchez JC, Williams KL, Appel RD, *et al.* Protein identification and analysis tools in the ExPASy server. Methods in molecular biology (Clifton, N.J.). 1999; 112: 531–552.
[58] Duret L, Mouchiroud D. Expression pattern and, surprisingly, gene length shape codon usage in Caenorhabditis, Drosophila, and Arabidopsis. Proceedings of the National Academy of Sciences. 1999; 96: 4482–4487.
[59] Powell JR, Moriyama EN. Evolution of codon usage bias in Drosophila. Proceedings of the National Academy of Sciences. 1997; 94: 7784–7790.
[60] Foss AH, Duffner PK, Carter RL. Lifetime risk estimators in epidemiological studies of Krabbe Disease. Rare Diseases. 2013; 1: e25212.
[61] Palidwor GA, Perkins TJ, Xia X. A general model of codon bias due to GC mutational bias. PLoS ONE. 2010; 5: e13431.
[62] Yang J, Zhu TY, Jiang ZX, Chen C, Wang YL, Zhang S, *et al.* Codon Usage Biases in Alzheimer's Disease and other Neurodegenerative Diseases. Protein and Peptide Letters. 2010; 17: 630–645.
[63] Cutter AD, Wasmuth JD, Blaxter ML. The Evolution of Biased Codon and Amino Acid Usage in Nematode Genomes. Molecular Biology and Evolution. 2006; 23: 2303–2315.
[64] Kumar U, Khandia R, Singhal S, Puranik N, Tripathi M, Pateriya AK, *et al.* Insight into Codon Utilization Pattern of Tumor Suppressor Gene EPB41L3 from Different Mammalian Species Indicates Dominant Role of Selection Force. Cancers. 2021; 13: 2739.
[65] Bera BC, Virmani N, Kumar N, Anand T, Pavulraj S, Rash A, *et al.* Genetic and codon usage bias analyses of polymerase genes of equine influenza virus and its relation to evolution. BMC Genomics. 2017; 18: 652.
[66] He B, Dong H, Jiang C, Cao F, Tao S, Xu L. Analysis of codon usage patterns in Ginkgo biloba reveals codon usage tendency from a/U-ending to G/C-ending. Scientific Reports. 2016; 6: 35927.
[67] Nasrullah I, Butt AM, Tahir S, Idrees M, Tong Y. Genomic analysis of codon usage shows influence of mutation pressure, natural selection, and host features on Marburg virus evolution. BMC Evolutionary Biology. 2015; 15: 174.
[68] Hambuch TM, Parsch J. Patterns of Synonymous Codon Usage in Drosophila melanogaster Genes with Sex-Biased Expression. Genetics. 2005; 170: 1691–1700.
[69] Supek F, Vlahovicek K. Comparison of codon usage measures and their applicability in prediction of microbial gene expressivity. BMC Bioinformatics. 2005; 6: 182.
[70] Huang X, Xu J, Chen L, Wang Y, Gu X, Peng X, *et al.* Analysis of transcriptome data reveals multifactor constraint on codon usage in Taenia multiceps. BMC Genomics. 2017; 18: 308.