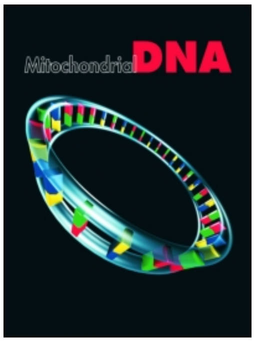
# Appendix II – List of Publications

## (a) Scientific Publications

1. **Arif Uddin**, Supriyo Chakraborty: *Synonymous codon usage pattern in mitochondrial CYB gene in pisces, aves, and mammals*. **Mitochondrial DNA 12/2015**; (**IF-1.2**).
2. **Arif Uddin,** Tarikul Huda Mazumder, Monisha Nath Choudhury, Supriyo Chakraborty**: *Codon bias and gene expression of mitochondrial ND2 gene in chordates*. Bioinformation 08/2015; (IF-0.5).**
3. Gulshana A. Mazumder**, Arif Uddin,** Supriyo Chakraborty: *Prediction of gene expression and codon usage in human parasitic helminths.* **Genes and Genomics 12-2015; (IF-0.59).**
4. Tarikul Huda Mazumder, **Arif Uddin**, Supriyo Chakraborty: *Transcription factor gene GATA2: Association of leukemia and nonsynonymous to the synonymous substitution rate across five mammals.* **Genomics 02/16; (IF-2.3).**
5. Gulshana A. Mazumder, **Arif Uddin** and Supriyo Chakraborty: *Expression levels and codon usage patterns in nuclear genes of the filarial nematode, W. bancrofti and the blood fluke, S. haematobium.* **Journal of Helminthology (Accepted IF-1.4).**
6. Arif Uddin, Supriyo Chakraborty: ***Analysis of codon usage trend, expression level and influencing factors for MT-ND1 gene among pisces, aves and mammals*** **(Revision 1 stage)**
7. Arif Uddin, Supriyo Chakraborty: ***Codon usage trend in mitochondrial CYB gene.*** **(Accepted IF-2.1)**

## (b) Conference Proceedings

1. **Arif Uddin,** Tarikul Huda Mazumder, Monisha Nath Choudhury Supriyo Chakraborty (as oral). *Expression level and codon usage trends in* MT-ATP6 *gene across fish species*, **National seminar, Pub Kamrup college,** Guwahati (September, 2015).
2. **Arif Uddin**, Supriyo Chakraborty (as poster). *Codon usage in Human Mitochondrial genes in the Context of Cancer,* **1st International Conference on Pharmacy and Pharmaceutical Sciences,** Innovare academics, Bhopal (October, 2015).
3. Supriyo Chakraborty, Himangshu Deka, **Arif Uddin**, Tarikul Huda Mazumder, Arup Kumar Malakar, Binata Halder, Monisha Nath Choudhury (as oral). *Multiple Peptide-Epitope vaccine against Leishmania major,* **IMMUNOCON**, Madurai Kamaraj University (December 2014).



Taylor & Francis
Taylor & Francis Group

Mitochondrial DNA

The Journal of DNA Mapping, Sequencing, and Analysis

ISSN: 1940-1736 (Print) 1940-1744 (Online) Journal homepage: http://www.tandfonline.com/loi/imdn20

# Synonymous codon usage pattern in mitochondrial CYB gene in pisces, aves, and mammals

Arif Uddin & Supriyo Chakraborty

**To cite this article:** Arif Uddin & Supriyo Chakraborty (2015): Synonymous codon usage pattern in mitochondrial CYB gene in pisces, aves, and mammals, Mitochondrial DNA, DOI: 10.3109/19401736.2015.1115842

**To link to this article:** http://dx.doi.org/10.3109/19401736.2015.1115842

Published online: 04 Dec 2015.

MITOCHONDRIAL DNA, 2015
http://dx.doi.org/10.3109/19401736.2015.1115842

Taylor & Francis
Taylor & Francis Group

RESEARCH ARTICLE

# Synonymous codon usage pattern in mitochondrial *CYB* gene in pisces, aves, and mammals

Arif Uddin and Supriyo Chakraborty

Department of Biotechnology, Assam University, Silchar, Assam, India

**ABSTRACT**
Cytochrome b (*CYB*) protein plays an important role in complex III of the mitochondrial oxidative phosphorylation. Codon usage is the phenomenon of non-uniform usage of synonymous codons. In the present study, we report the pattern of codon usage in *MT-CYB* gene using various codon usage parameters. Nucleotide composition such as % of C and T was higher than A and G in pisces. In aves, % of A and C was higher than T and G but in mammals, A and T was higher than C and G. Heat map shows that AT-ending codons were mostly negative and GC-ending codons were mostly positive. From the heat map based on RSCU values, it is evident that codon usage prefers A/C at the third codon position and it was less towards T/G in its third codon position. The codons absent in pisces were AGT (except *Toxotes chatareus*), TGT, and CAG (except *Elasma zonatum*). The codons such as AGT (except *Falco peregrinus*), CGT (except *Vidua chalybeata*), and ACG (except *Aythya americana*) were absent in aves whereas, in mammals, the absent codons were namely CAG (except *Canis familiaris)* and ACG (except *Rattus norvegicus*). Codon usage bias was low in pisces, aves, and mammals. The frequency of leucine was the highest in the amino acid and cysteine was the lowest. Correlation analysis further suggests that mutation pressure is mainly responsible for codon usage pattern. Natural selection might also play a vital role in codon usage pattern but it was weaker than mutation pressure.

**ARTICLE HISTORY**
Received 24 August 2015
Revised 22 October 2015
Accepted 30 October 2015
Published online
3 December 2015

**KEYWORDS**
Codon usage; *CYB* gene; mutation pressure; natural selection

## Introduction

Cytochrome b is one of the proteins involved in mitochondrial oxidative phosphorylation that makes up complex III. It is encoded by *CYB* gene of mitochondrial genome (Anderson et al., 1981) while others are encoded by nuclear genome and they together form a functional complex. The transfer of electrons from ubiquinol to cytochrome c along with translocation of protons across inner mitochondrial membrane is catalyzed by mitochondrial bc1 complex (higher eukaryotes) or complex III. It is a membrane-bound enzyme. The catalytic core of the enzyme is formed by *CYB* protein along with cytochrome c1 and it helps in assembly and function of complex III. In recent years, it has been found that mitochondrial respiratory chain disorders due to mitochondrial DNA mutation are the most common human metabolic diseases (Chinnery et al., 2000). The non-sense, missense, or frame shift mutation in *CYB* gene which causes the respiratory complex III deficiency has been observed in a number of patients (MITOMAP, 2000). The mitochondrial *CYB* gene mutations are linked with mitochondrial encephalopathy (De Coo et al., 1999; Keightley et al., 2000), myoglobinuria (Andreu et al., 1998; Andreu et al., 1999a; Andreu et al., 1999b; Bruno et al., 2003; Dumoulin et al., 1996; Lamantea et al., 2002; Mancuso et al., 2003), and cardiomyopathy (Andreu et al., 2000; Valnot et al. 1999). *CYB* gene is considered a useful gene for phylogenetic work. Based on the structure and the function of protein product, probably it is the best known mitochondrial gene (Esposti et al., 1993). Several workers have shown phylogenetic usefulness of *CYB* gene in different vertebrates (Irwin et al., 1991; Moritz et al., 1992). *CYB* gene contains more conservative as well as rapidly evolving codon position and variable region and so this gene is widely used in systematics (Izeni et al., 2001; Meyer & Wilson, 1990).

The genetic code comprises 64 codons which encode 20 standard amino acids including three non-sense or termination codons such as TAA, TAG, and TGA in standard genetic code but there are four termination codons such as TAA, TAG, AGA, and AGG in vertebrate mitochondrial genetic code (NCBI). The amino acids such as met and trp are encoded by single codon in standard genetic code but in mitochondrial genetic code, met is encoded by two codons namely ATG and ATA and trp is also encoded by two codons such as TGG and TGA. The genetic code is degenerate, i.e., some codons encode the same amino acid and are referred to as synonymous codons. The relative abundance of tRNAs and the speed of the synonymous codons in translation process might differ which are recognized by the ribosome (Butt et al., 2014). The efficiency and the accuracy of protein production depend on synonymous codons. The phenomenon of non-uniform usage of synonymous codons, i.e., some codons encoding the same amino acid are used more frequently than others, is the codon usage bias and it is species

**CONTACT** Supriyo Chakraborty ✉ supriyoch_2008@rediffmail.com Department of Biotechnology, Assam University, Dargakona, Silchar, Assam 788011, India
Supplemental data for this article can be accessed here.

specific (Gupta et al., 2004). The codon usage phenomenon has been studied in prokaryotes, eukaryotes, and viruses (Gu et al., 2004; Liu et al., 2011; Ma et al., 2013; Moratorio et al., 2013; Sharp et al., 1988; Tao et al., 2009). Various factors such as mutation pressure, natural selection, expression level, gene length, compositional bias (GC% and GC skew), replication, RNA stability, hydrophobicity, and hydrophilicity of the protein (Akashi, 1997; Moriyama & Powell, 1998; Powell & Moriyama, 1997; Powell et al., 2003) affect codon usage pattern. Among these, the compositional constraint in the presence of mutation pressure and natural selection are the major factors (Sharp & Li, 1986; Sharp & Matassi, 1986). Codon usage bias study is gaining increased attention of researchers since the advent of whole genome sequencing of many organisms. Codon usage bias varies within the same genome and also across genomes and is important to understand the evolution of genome (Jenkins & Holmes, 2003). The study of the pattern of codon usage reveals the forces that influence the evolution, so the investigation on codon usage bias and evolutionary forces might provide further clues in understanding the process of evolution at molecular level.

In the present study, we have attempted to analyze the pattern of codon usage bias of *CYB* gene across a diverse group of vertebrates, i.e., pisces, aves, and mammals to identify the commonality and the difference, if any, in base composition and other related parameters of codon usage pattern in mitochondria.

Mutations occur frequently in mtDNA and these mutations accumulate in vicious cycle with an increase in the age of an organism. Both replication errors and free radical damage were identified as the major factors involved in mtDNA mutations. It was suggested that massive mtDNA replication occurring during embryogenesis results in replication errors due to the inherent error rate of mtDNA polymerase enzyme (Larsson, 2010). These mutations are subjected to segregation and clonal expansion in postnatal life. Alternatively, it was also proposed that accumulated mtDNA damage caused by free radical (reactive oxygen species) overwhelms the mitochondrial repair machinery and results in mtDNA mutation accumulation.

Analysis of pattern of codon usage in *MT-CYB* gene is of special interest to know how the energy consumption of pisces, aves, and mammals influences the codon usage against quick environmental alterations during the course of evolution. The pisces, aves, and mammals survive in three different habitats and so their means of respiration and energy demand are also unusual. Therefore, this study helps in understanding the factors influencing gene evolution. The present investigation was undertaken with the following hypothesis.

Analysis of codon usage is a useful technique to understand the genetic and the evolutionary relationship of different species belonging to diverse habitats. Moreover, mitochondrial genes are very significant and suitable tools for these kinds of studies. In the current study, we investigated the codon usage pattern in *MT-CYB* gene among pisces, aves, and mammals thriving in different habitats to understand the pattern of codon usage. Moreover, this study would give insight into the factors influencing the codon usage pattern among the species under study.

Table 1. Family and accession number of species analyzed in the present study.

| Species | Family | Accession no |
|---|---|---|
| *Toxotes chatareus* | Toxotidae | AP006806 |
| *Elasma zonatum* | Elassomatidae | AP006813 |
| *Jordanella floridae* | Cyprinodontidae | AP006778 |
| *Platax orbicularis* | *Ephippidae* | AP006825 |
| *Latimeria menadoensis* | *Latimeriidae* | AP006858 |
| *Gallus gallus* | *Phasianidae* | X52392 |
| *Aythya americana* | *Anatidae* | AF090337 |
| *Vidua chalybeata* | Viduidae | AF090341 |
| *Falco peregrinus* | Falconidae | AF090338 |
| *Smithornis sharpei* | Calyptomenidae | AF090340 |
| *Canis familiaris* | *Canidae* | U96639 |
| *Myoxus glis* | Myoxidae | AJ001562 |
| *Rattus norvegicus* | *Muridae* | X148148 |
| *Dasypus novemcinctus* | *Dasypodidae* | Y11832 |
| *Oryctolagus cuniculus* | Leporidae | AJ001588 |

## Materials and methods

### Sequence data

The coding sequences of *MT-CYB* gene for different species of pisces, aves, and mammals were retrieved from the National Center for Biotechnology Information (http://www.ncbi.nlm.nih.gov/Genbank/). We analyzed only those cds sequences that are exact multiple of three bases. The different species along with accession number, family, and gene length are shown in Table 1.

### Compositional properties

The compositional properties of *CYB* gene were calculated for the different species of pisces, aves, and mammals namely (i) overall nucleotide composition (A, C, T, and G %) and nucleotide composition in its third codon position. (ii) The frequency of occurrence of overall GC % and GC contents at the first, second, and third position. (iii) The AT, GC, purine, pyrimidine, amino, and keto skew. All calculations were done using a perl script developed by S. C. (corresponding author).

### Measures of synonymous codon usage bias

Some of the most relevant and widely used measures of codon usage bias analyzed in this study are discussed below.

### Relative synonymous codon usage (RSCU)

Relative synonymous codon usage (RSCU) was calculated as the ratio of the observed frequency of a codon to the expected frequency if all synonymous codons of a particular amino acid are used equally. RSCU value greater than 1.0 indicates that the corresponding codons are used more frequently than the expected frequency whereas the reverse is true for RSCU values less than 1.0 indicating that the particular codons were used less frequently (Jenkins & Holme, 2003). Moreover, the RSCU value $>1.6$ was treated as over represented and RSCU value $<0.6$ was considered as under-represented (Ma et al., 2013)

$$RSCUij = \frac{Xij}{\frac{1}{ni}\sum_{j=1}^{ni} Xij}$$

where $X_{ij}$ is the frequency of occurrence of the *j*th codon for *i*th amino acid (any $X_{ij}$ with a value of zero is arbitrarily assigned a value of 0.5) and $n_i$ is the number of codons for the *i*th amino acid (*i*th codon family).

### Effective number of codons (ENC)

The effective number of codons used by a gene (ENC) is the most widely used parameter to measure the usage bias of synonymous codons (Sharma et al., 2014). ENC value of a gene is used to quantify the codon usage bias irrespective of composition of amino acid and gene length. The ENC value ranges from 20 (when only one codon is used for each amino acid) to 61 (when all codons are used randomly). If the calculated ENC is greater than 61 (because codon usage is more evenly distributed than expected), it is adjusted to 61. Higher ENC value means low codon usage bias. ENC value <35 is generally considered as the significant codon usage bias

$$ENC = 2 + \frac{9}{F_2} + \frac{1}{F_3} + \frac{5}{F_4} + \frac{3}{F_6}$$

where *Fk* ($k = 2, 3, 4, 6$) is the mean of *Fk* values for the *k*-fold degenerate amino acids, 2 stands for two amino acids, i.e., met and trp; 9, 1, 5, and 3 stand for the total number of amino acids with degeneracy class of 2, 3, 4, and 6 codons, respectively.

### Codon adaptation index (CAI)

The codon adaptation index (CAI) (Wright, 1990) is a very extensively used parameter to measure the codon bias in prokaryotes (Eyre-Walker & Bulmer, 1993; Gutierrez et al., 1994; Perriere et al., 1994) as well as eukaryotes (Akashi, 1994; Jun, 2013; Touchon & Rocha, 2008). CAI is also a quantitative predictor of gene expression. CAI values range from 0 to 1; with higher values indicating a higher proportion of the most abundant codons (Xing et al., 2014). CAI is a measure of the relative adaptiveness of the codon usage of a gene towards the codon usage of highly expressed genes. The relative adaptiveness ($\omega$) of each codon is the ratio of the usage of each codon, to that of the most abundant codon within the same synonymous family. Non-synonymous codons and termination codons (dependent on genetic code) are excluded from this analysis. The CAI is calculated as

$$CAI = \exp\left(\frac{1}{L}\sum_{k=1}^{L} \ln \omega k\right)$$

where $\omega k$ is the relative adaptiveness of the *k*th codon and *L* is the number of synonymous codons in the gene.

*Hierarchal clustering:* The RSCU values of codons from different species of pisces, aves, and mammals were clustered by the hierarchal clustering method using Netwalker 1.0 software (Qiagen, Valencia, CA).

*Phylogenetic tree:* The Phylogenetic tree was constructed by the maximum-likelihood approach based on clustal W2 alignment using MEGA 6.0 software (MEGA Inc., Englewood, NJ).

### Statistical analysis

Correlation analysis was used to identify the relationship between overall nucleotide composition and each base at the third codon position. All the statistical analyses were done using the SPSS software (SPSS Inc., Chicago, IL).

## Result and discussion

### Nucleotide composition of different species of pisces, aves, and mammals

Overall nucleotide composition may influence the codon usage bias of a genome (Jenkins & Holmes, 2003). We, therefore, analyzed nucleotide composition of coding sequences of *CYB* gene in different species of pisces, aves, and mammals. In pisces, mean ± SD of C % was the highest, followed by T % and A %, with the G % being the lowest as shown in Table 1. The mean ± SD of overall GC % was 45.34 ± 1.88 in pisces. In aves, mean ± SD of C % was the highest, followed by A % and T %, with the G % being the lowest. The mean ± SD of overall GC % was 47.12 ± 1.86 in aves. The mean ± SD of A % was the highest, followed by T % and C %, with the G % being the lowest in mammals. The mean ± SD of overall GC % in mammals was 41.66 ± 1.45. This suggests unequal distribution of A, T, G, and C % among the codons in different species of pisces, aves, and mammals with more preference for C ending codons in pisces and aves followed by T/A ending codons in pisces, but A/T-ending codons in aves. In mammals, increase preference of A-ending codons was observed which was followed by T/C-ending codons. The preference of G ending codons was less in aves followed by mammals and then pisces. The overall GC % was the lowest in mammals followed by pisces and aves. However, analysis of nucleotide composition at the third position of codons (A3%, T3%, G3%, and C3%) and GC1%, GC2%, and GC 3% provides a clear picture about the preference of codon usage in different species of pisces, aves, and mammals. The mean ± SD % of C3 in pisces and in aves was the highest, followed by A3% and T3%. The mean ± SD of % of A3 was the highest followed by C3 and T3 in mammals. The G3% was the lowest in mammals followed by aves and pisces. The mean ± SD of GC3% in aves was the highest followed by pisces and then mammals, as shown in Table 2. Therefore, from the initial nucleotide composition analysis, it was expected that nucleobase C/A might be more preferred. From the GC3 content analysis, it was evident that in pisces and mammals, AT-ending codons might be more preferred to GC-ending codons, whereas in aves, both AT- and GC-ending codons might be equally preferred at the third codon position. From the overall GC content analysis, it was found that AT-ending codons were more preferred to GC-ending codons. Nucleotide composition influenced the codon usage (Jenkins & Holmes, 2003). These results suggest that there might be compositional constraint in the presence of mutation pressure which affects the mitochondrial *CYB* gene.

We performed regression analysis between the nucleobase at the third position of codon (Wobble position) and the effective number of codons. The linear model of ENC versus A3 + T3 + G3 + C3 was fixed with the observed value of ENC. The coefficient of regression is shown in Table 3. It corresponds to difference in ENC value with a change in nucleotide composition at the third position of codons. These results show that the effective number of codon was negatively affected by T3

Table 2. Nucleotide composition among pisces, aves, and mammals.

| Species | A % | T % | G % | C % | A3% | T3% | G3% | C3% |
|---|---|---|---|---|---|---|---|---|
| *T. chatareus* | 25.63 | 26.94 | 14.34 | 33.07 | 32.8 | 17.32 | 3.93 | 45.93 |
| *E. zonatum* | 26.4 | 30.35 | 15.39 | 27.73 | 33.33 | 23.35 | 8.92 | 34.38 |
| *J. floridae* | 24.47 | 30 | 15.43 | 30.08 | 30.78 | 23.94 | 5.78 | 39.47 |
| *P.orbicularis* | 25.45 | 27.2 | 14.26 | 33.07 | 32.28 | 17.58 | 4.19 | 45.93 |
| *L. menadoensis* | 31.32 | 24.84 | 14.26 | 29.57 | 42.78 | 13.12 | 6.29 | 37.79 |
| Mean ± SD | 26.65 ± 2.69 | 27.86 ± 2.30 | 14.73 ± 0.61 | 30.70 ± 2.33 | 34.39 ± 4.74 | 19.06 ± 4.54 | 5.82 ± 2.00 | 40.70 ± 5.11 |
| *G. gallus* | 27.47 | 24.14 | 12.07 | 36.3 | 34.9 | 10.49 | 3.14 | 51.44 |
| *A. americana* | 26.85 | 23.44 | 14.17 | 35.52 | 34.64 | 7.87 | 7.34 | 50.13 |
| *V. chalybeate* | 30.18 | 23.44 | 13.38 | 32.98 | 44.88 | 8.92 | 3.14 | 43.04 |
| *F. peregrinus* | 29.74 | 24.2 | 12.42 | 33.42 | 40.41 | 12.33 | 4.19 | 43.04 |
| *S. sharpei* | 27.99 | 26.68 | 11.63 | 33.68 | 35.17 | 17.58 | 2.88 | 44.35 |
| Mean ± SD | 28.44 ± 1.44 | 24.38 ± 1.33 | 12.73 ± 1.02 | 34.38 ± 1.44 | 38.00 ± 4.52 | 11.43 ± 3.82 | 4.13 ± 1.85 | 46.40 ± 4.06 |
| *C. familiaris* | 29.03 | 29.12 | 13.94 | 27.89 | 39.21 | 20.78 | 5 | 35 |
| *M. glis* | 29.94 | 31.84 | 12.36 | 26.84 | 38.68 | 26.31 | 2.36 | 32.63 |
| *R. norvegicus* | 30.35 | 27.2 | 12.59 | 29.83 | 42.25 | 14.96 | 2.62 | 40.15 |
| *D. novemcinctus* | 31.49 | 26.57 | 13.15 | 28.77 | 43.94 | 14.73 | 3.38 | 37.63 |
| *O. cuniculus* | 28.07 | 28.94 | 12.63 | 30.35 | 35 | 22.89 | 2.36 | 39.73 |
| Mean ± SD | 29.77 ± 1.30 | 28.73 ± 2.05 | 12.93 ± 0.63 | 28.73 ± 1.42 | 39.81 ± 3.45 | 19.93 ± 5.04 | 3.14 ± 1.11 | 37.02 ± 3.19 |

Table 3. ENC, GC contents, and CAI values of *MT-CYB* gene in pisces, aves, and mammals.

| Species | ENC | Overall GC % | GC1% | GC2% | GC3% | CAI |
|---|---|---|---|---|---|---|
| *T. chatareus* | 60 | 47.04 | 53.8 | 38.6 | 49.9 | 0.7814 |
| *E. zonatum* | 59 | 43.1 | 47 | 49.9 | 43.3 | 0.8761 |
| *J. floridae* | 60 | 45.5 | 52.9 | 38.4 | 45.3 | 0.8267 |
| *P. orbicularis* | 60 | 47.3 | 52.5 | 39.4 | 50.1 | 0.8842 |
| *L. menadoensis* | 60 | 43.8 | 48.3 | 39.1 | 44.1 | 0.7621 |
| Mean ± SD | 59.80 ± 0.44 | 45.34 ± 1.88 | 50.90 ± 3.03 | 41.08 ± 4.94 | 46.54 ± 3.23 | 0.82 ± 0.05 |
| *G. gallus* | 60 | 48.4 | 50.4 | 40.2 | 54.6 | 0.7391 |
| *A. americana* | 59 | 49.7 | 51.7 | 39.9 | 57.5 | 0.7685 |
| *V. chalybeata* | 60 | 46.4 | 53.8 | 39.1 | 46.2 | 0.6923 |
| *F. peregrinus* | 60 | 45.8 | 50.4 | 39.9 | 47.2 | 0.8251 |
| *S. sharpei* | 60 | 45.3 | 49.6 | 39.1 | 47.2 | 0.8109 |
| Mean ± SD | 59.80 ± 0.44 | 47.12 ± 1.86 | 51.18 ± 1.64 | 39.64 ± 0.50 | 50.54 ± 5.14 | 0.76 ± 0.05 |
| *C. familiaris* | 58 | 41.8 | 46.8 | 38.7 | 40 | 0.8059 |
| *M. glis* | 56 | 39.2 | 44.2 | 38.4 | 35 | 0.7977 |
| *R. norvegicus* | 59 | 42.4 | 47.2 | 37.3 | 42.8 | 0.7984 |
| *D. novemcinctus* | 59 | 41.9 | 46.3 | 38.2 | 41.3 | 0.7618 |
| *O. cuniculus* | 59 | 43 | 48.4 | 38.4 | 42.1 | 0.7824 |
| Mean ± SD | 58.20 ± 1.30 | 41.66 ± 1.45 | 46.58 ± 1.54 | 38.20 ± 0.53 | 40.24 ± 3.10 | 0.78 ± 0.01 |

and G3 in pisces, G3 and C3 in aves, and T3 in mammals whereas positively affected by A3 and C3 in pisces, A3 and T3 in aves, and A3, G3, and C3 in mammals. ENC is a non-directional measure of codon usage bias and it means that the extent of codon usage bias decreases with an increase in ENC. The negative value of T3 and G3 in pisces, G3 and C3 in aves, and T3 in mammals on ENC indicates that T3 and G3 in pisces, G3 and C3 in aves, and only T3 in mammals positively influenced the codon usage.

## *Codon usage bias among* MT-CYB *gene in pisces, aves, and mammals*

The effective number of codon (ENC) was calculated to quantify the degree of codon usage bias among different species of pisces, aves, and mammals. The ENC values of different species are shown in Figure 1. The high ENC value indicates conserved genomic composition among pisces, aves, and mammals. The ENC of *MT-CYB* gene was quite higher than that of *ATP 6* and *ATP 8* in mammals. The ENC values of mitochondrial *ATP 6* and *ATP 8* in mammals range from 57 to 60 (Uddin & Chakraborty, 2014a) and 42 to 60 (Uddin & Chakraborty, 2014b), respectively, but in *MT-CYB* gene, the ENC value ranges from 59 to 60 in pisces and aves and 56 to 59 in mammals in the present study indicating low codon usage bias. The ENC values ranged from 51 to 60 in *B. mandarina* and from 52 to 60 in *O. furnacalis's* nuclear genes (Uddin & Chakraborty, 2014a). The reason for low bias might be advantageous to efficient replication in vertebrates with different cell types having different codon choices (Jenkins & Holmes, 2003).

The CAI is a directional measure of codon usage bias and its higher value indicates higher gene expression so that there is higher codon usage bias. There is an opposite relation between ENC and CAI, it means higher ENC value indicates lower bias and higher value of CAI indicates higher gene expression (Uddin & Chakraborty, 2014a). In many bacteria and small eukaryotes, highly expressed gene revealed stronger bias. CAI measures the gene expression level with respect to a reference set of genes (Sharma et al., 2014). Comparison of ENC and CAI is used to judge the nucleotide composition and codon selection (Wright, 1990, Sharp & Li, 1987). We correlated ENC and CAI to know the nucleotide composition variation and codon selection among different species of pisces, aves, and mammals for *MT-CYB* gene. We found negative correlation between ENC and CAI in pisces, aves, and in mammals. Pearson correlation coefficient was −0.511 ($p > 0.05$) in pisces, –0.014 ($p > 0.05$) in

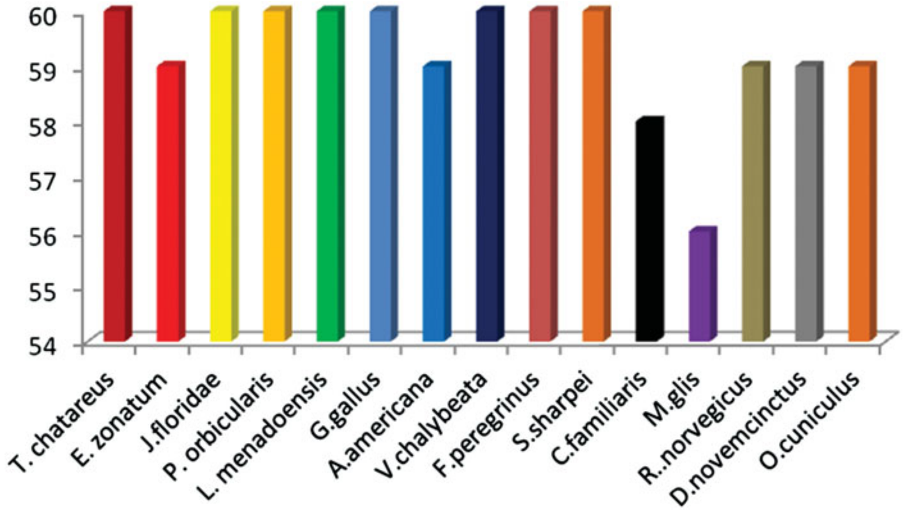


Figure 1. Distribution of ENC in different species of pisces, aves, and mammals. This figure clearly shows that *M. glis* has the lowest ENC among pisces, aves, and mammals.

aves, and −0.459 ($p > 0.05$) in mammals and these suggest codon usage bias has a very weak relationship with nucleotide composition. Dipteran and Hymenopteran species had shown negative correlation between ENC and CAI (Behura & Severson, 2012).

We performed correlation analysis between codon usage and GC3 to know the general trend in codon usage variation and GC bias. From Figure 2(a)–(c), it was found that, in pisces, aves, and mammals, most of the AT-ending codons were negative and most of the GC-ending codons were positive. In pisces, except codons like ATA, TGA, CAA, CTA, CGA, ACT, AGT, TTT, CTT, CCT, GTT, GCT, and GGT, all other AT-ending codons were negative and except ATG, TTG, TCG, TGG, CAG, CTT, CCG, CGG, GAG, GTG, GGG, ACC, and AGC, all other GC-ending codons were positive. In aves, other than codons namely TTA, CAA, CGA, GTA, GCA, ACT, TAT, CTT, and GTT, all other AT-ending codons were negative and except ATG, CCG, GCG, TAC, CGC, and GTC, all other GC ending codons were positive. In mammals, except the codons namely AAA, ACA, TCA, TGA,CTA, CCA, GTA, GCA, TCT, CTT, CGT, and GTT, all other AT-ending codons were negative and except AAG, TTG, TCG, TGG, CAG, CGG, GCG, AGC, TCC, CTC, CCC, GCC, and GGC, all other GC-ending codon were positive. From the heat map study, we found GC-ending codons were mostly positive and AT-ending codons were mostly negative. It reveals that GC-ending codons would show increasing usage with increasing GC3 and similarly AT-ending codons show decreasing usage with increasing GC3 bias.

## Relative synonymous codon usage (RSCU) analysis of MT-CYB gene in pisces, aves, and mammals

We performed RSCU analysis to determine the pattern of synonymous codon usage and the extent of G/C-ending codons. Furthermore, we divided the RSCU data into two groups: (a) RSCU value >1.6: overrepresented and (b) RSCU value <0.6: underrepresented. From the heat map, the over represented and the underrepresented codons are clearly evident as shown in Figure 3. In pisces, codons namely TCA, TCC, CTA, CCA, CAA, CGA, ACA, ACC, GCA, AAA, GGA, and TGA were overrepresented. But the underrepresented codons were TCG, AGC, AGT, TTT, TTG, CTG, CCG, CAT, CAG, CGG, CGT, ACG, ACT, AAT, GAT, GTG, GCG, AAG, GAG, GGG, and GGT. The codons TCA, TCC, TTC, CTA, TAC, CCA, CAC, CAA, CGA, ATA, ACC, AAC, GAC, GTA, GTC, GCC, AAA, GAA, GGA, and TGA were overrepresented in aves. However, the underrepresented codons were TCG, AGC, AGT, TTT, TTA, TTG, CTG, CTT, TAT, TGT, CCG, CCT, CAT, CAG, CGG, CGT, ATG, ATT, ACG, ACT, AAT, GAT, GTG, GTT, GCG, GCT, AAG, GAG, GGG, GGT, and TGG. In mammals, the overrepresented codons included TCA, CTA, TGC, CCA, CAA, CGA, ACA, GTA, GCA, AAA, GAA, GGA, and TGA, but the codons TCG, AGT, TTG, CTG, TGT, CCG, CAT, CAG, CGG, CGT, ATG, ACG, AAT, GAT, GTG, GCG, AAG, GAG, GGG, GGT, and TGG were underrepresented. Based on RSCU analysis and nucleotide composition, we deduced that the existence of preferred codons in coding sequences has been mostly influenced by compositional constraints, which account for the presence of mutation pressure.

## Effect of evolutionary forces in shaping the codon usage pattern in MT-CYB gene in pisces, aves, and mammals

### Effect of mutation pressure in shaping the codon usage pattern

Two major evolutionary forces namely mutation pressure and natural selection are considered to shape the codon usage pattern in a species. Mutation pressure affects the whole genome, which accounts for the majority of codon usage among different RNA viruses (Vicario et al., 2007). We performed correlation analysis between general nucleotide composition and nucleotide composition at the third codon position to determine whether evolutionary process is driven by mutation pressure alone or by both mutation pressure and natural selection. In pisces, aves, and mammals, highly significant positive correlation was found among A and A3%, C and C3%, and GC and GC3%, respectively, and negative correlation was

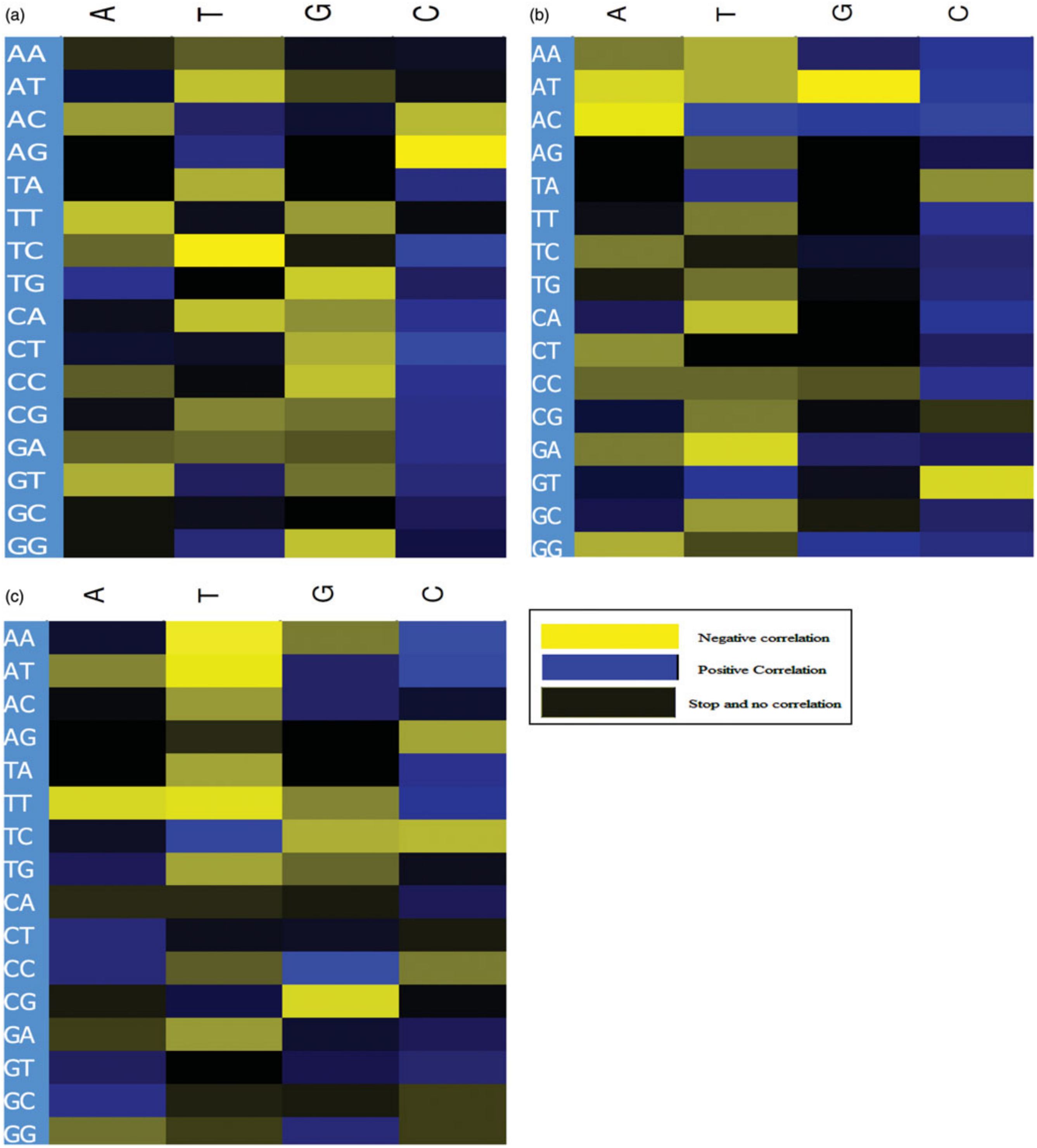


Figure 2. (a) Heat maps of correlation coefficient between codon usage and GC3 in different species in pisces. (b) Heat maps of correlation coefficient between codon usage and GC3 in different species in aves. (c) Heat maps of correlation coefficient between codon usage and GC3 in different species in mammals.

observed for most of the various nucleotide comparisons as shown in Table 4. These results suggest that the compositional constraint arising from mutation pressure determines the pattern of codon usage in *MT-CYB* gene. Furthermore, positive correlation was observed in pisces among ENC and GC %, ENC and GC3%, GC and GC3%, and GC1 and GC3%. In aves, positive correlation was observed between GC and GC3% and in mammals, positive correlation was observed among ENC and GC %, ENC and GC3%, GC and GC3%, and GC1 and GC3% as shown in Tables S1–S3. This finding suggests that the nucleotide composition resulting from mutation pressure was one of the main factors for synonymous codon usage in *MT-CYB* gene. Earlier study also found negative correlation between GC12 and GC3 in mitochondrial genome of ribbon worms (Behura & Severson, 2012). Zhicheng et al. (Zhang et al., 2013) also found similar results in TTSuV1 virus.

The neutrality plot (GC12 and GC3) as shown in Figure 5(a)–(c) reveals that pisces, aves, and mammals had a wide

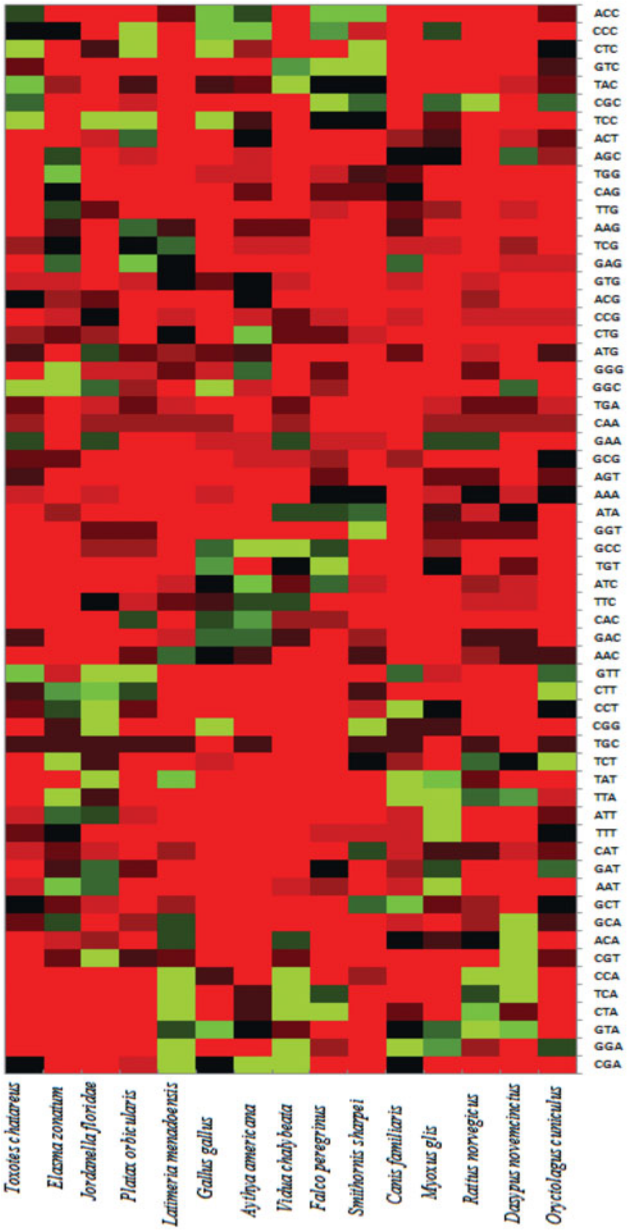


Figure 3. Heat map of RSCU value of each codon among different species of pisces, aves, and mammals. The color and the degree of intensity represent the RSCU value. The color varies from green to red with low value of RSCU to high value, respectively. Light green indicates RSCU value zero, green indicates <0.06, dark black RSCU value >1, and red indicates >1.6. (In black and white figure, grey indicates RSCU value zero, dark grey indicates RSCU < 0.06, dark black indicates RSCU > 1, and light black indicates RSCU > 1.6.)

range of GC3 distributions. Except pisces, aves and mammals had a positive correlation with GC3 although not statistically significant ($p > 0.05$) suggesting that mutation pressure might play a role in codon usage pattern.

## *Effect of natural selection in shaping the codon usage pattern*

If the pattern of synonymous codon usage is solely governed by mutation pressure, then the frequency of nucleotide A and T should be equal to that of G and C at synonymous the third codon position (Xing et al., 2014). In case of pisces, aves, and mammals, the frequencies of those nucleotides were not the same indicating that other factor such as natural selection might have played a role in codon usage pattern. However, positive correlation among T and T3%, G and G3% in pisces, aves, and mammals but except between G and C3 in aves, negative correlation between G and C3%, GC and T3 in pisces, aves, and mammals suggest that natural selection in addition to mutation pressure might have played a role in codon usage in *MT-CYB* gene.

Table 4. Linear regression between (ENC) effective numbers of codon of genes as function of base composition at the third codon position.

| ENC | A3 | T3 | G3 | C3 |
|---|---|---|---|---|
| Pisces | 0.124 | −0.527 | −0.864 | 0.691 |
| Aves | 0.415 | 0.522 | −0.963 | −0.513 |
| Mammals | 0.223 | −0.759 | 0.085 | 0.913 |

Table 5. Correlation between overall nucleotide composition and its third codon position in pisces, aves, and mammals.

| | Nucleotide | A3% | T3% | G3% | C3% | GC3% |
|---|---|---|---|---|---|---|
| Pisces | A % | 0.997** | −0.732 | 0.249 | −0.379 | −0.446 |
| | T % | −0.746 | 0.995** | −0.379 | 0.489 | −0.295 |
| | G % | −0.471 | 0.929* | 0.668 | −0.647 | −0.607 |
| | C % | −0.289 | −0.394 | −0.967** | 0.999** | 0.979** |
| | GC % | −0.499 | −0.162 | −0.934* | 0.977** | 0.965** |
| Aves | A % | 0.942* | 0.027 | −0.485 | −0.854 | −0.851 |
| | T % | −0.393 | 0.974** | −0.476 | −0.261 | −0.382 |
| | G % | 0.217 | −0.837 | 0.782 | 0.189 | 0.435 |
| | C % | −0.754 | −0.34 | 0.369 | 0.991** | 0.919* |
| | GC % | −0.463 | −0.736 | 0.717 | 0.881* | 0.959* |
| Mammals | A % | 0.943* | −0.625 | −0.012 | −0.06 | −0.035 |
| | T % | −0.613 | 0.956* | −0.173 | −0.761 | −0.869 |
| | G % | 0.178 | −0.244 | 0.985** | −0.158 | 0.197 |
| | C % | −0.117 | −0.475 | −0.3 | 0.980** | 0.902* |
| | GC % | −0.047 | −0.563 | 0.124 | 0.893* | 0.967** |

**, *significant at p=0.01 and 0.05 respectively.

## *Effect of the hydrophobicity and aromaticity of protein in shaping codon usage pattern*

A few studies have shown that hydrophobicity and aromaticity of encoded protein play a role in shaping codon usage pattern (Tillier & Collins, 2000). Only in mammalian species, GRAVY score was significantly correlated with ENC which suggests that variation in codon usage was associated with degree of hydrophobicity of encoded protein as shown in Table S4. GRAVY score was found positively correlated with ENC, GC, and GC3 in chikungunya virus (Butt et al., 2014).

## *Relation between skewness and codon usage bias*

The GC skews in different species of pisces, aves, and mammals were all negative, AT skews for most of the species were positive as shown in Table 6. These suggest that asymmetrical nucleotide composition between the two strands of DNA, one with abundance of C over G and the other with abundance of A over T (Beletskii & Bhagwat, 2001). This is similar in all mitochondrial genomes and suggests the asymmetrical compositional pattern between the two strands of DNA. As a result of mutational deamination, the nucleobases T and G give rise to C and A, respectively. The amino skew (T–G/T + G) for *CYB* gene in different species was positive indicating that T outnumbered G as a compositional constraint towards deamination. Excess T in comparison with G in this gene might have resulted from the need of maintaining optimum GC content in the coding sequence during evolutionary process. Deamination gradients exist along mitochondrial genomes during replication and transcription (Seligmann, 2011). In some taxonomic groups, the inversion of the mitochondrial control region inverts the direction of these gradients (Hassanin et al., 2005). Base composition is connected to transcription process which is exposed from skewness (Fujimori et al., 2005; Jun, 2013; Touchon & Rocha, 2008). We performed correlation analysis between ENC and each of skewness and, likewise between CAI and each skewness. Only in aves highly significant correlation between ENC and purine skew was observed and in mammals, between ENC and pyrimidine skew. Although positive correlation was found between the parameters, the estimates were not statistically significant as shown in Table S5. These results suggest that in aves, purine skew, and in mammals, pyrimidine skew might affect the codons usage (Table 6).

## Amino acid usage

The amino acid frequency of the encoded proteins in different species of pisces, aves, and mammals was estimated. Comparison of amino acid frequencies among pisces, aves, and mammals shows highly significant correlation between fishes and aves ($r = 0.981$, $p < 0.001$) than between pisces and mammals ($r = 0.964$, $p < 0.01$). The frequency of leucine was the highest in the amino acid composition of pisces, aves, and mammals, while cysteine residue was the least in the proteins as shown in Figure 4.

Table 6. GC, AT, purine, pyrimidine, keto, and amino skew values.

| Species | GC skew | AT skew | Purine | Pyrimidine | Keto | Amino skew |
|---|---|---|---|---|---|---|
| *T. chatareus* | −0.39 | −0.02 | 0.28 | −0.1 | −0.13 | 0.31 |
| *E. zonatum* | −0.29 | −0.07 | 0.27 | 0.05 | −0.02 | 0.33 |
| *J. floridae* | −0.32 | −0.1 | 0.23 | 0 | −0.1 | 0.32 |
| *P. orbicularis* | −0.4 | −0.03 | 0.28 | −0.1 | −0.13 | 0.31 |
| *L. menadoensis* | −0.35 | 0.12 | 0.37 | −0.09 | 0.03 | 0.27 |
| *G. gallus* | −0.5 | 0.06 | 0.39 | −0.2 | −0.14 | 0.33 |
| *A. americana* | −0.43 | 0.07 | 0.31 | −0.2 | −0.14 | 0.25 |
| *V. chalybeata* | −0.42 | 0.13 | 0.39 | −0.17 | −0.04 | 0.27 |
| *F. peregrinus* | −0.46 | 0.1 | 0.41 | −0.16 | −0.06 | 0.33 |
| *S. sharpei* | −0.49 | 0.02 | 0.41 | −0.12 | −0.09 | 0.39 |
| *C. familiaris* | −0.33 | 0 | 0.35 | 0.02 | 0.02 | 0.35 |
| *M. glis* | −0.37 | −0.05 | 0.4 | 0.09 | 0.04 | 0.44 |
| *R. norvegicus* | −0.41 | 0.05 | 0.41 | −0.05 | 0.01 | 0.37 |
| *D. novemcinctus* | −0.37 | 0.08 | 0.41 | −0.04 | 0.05 | 0.34 |
| *O. cuniculus* | −0.41 | −0.02 | 0.38 | −0.02 | −0.04 | 0.39 |

## Phylogenetic and cluster tree

To elucidate the variation and conservation of codon usage among different species, we performed phylogenetic analysis among 15 species belonging to pisces, aves, and mammals for maximum likelihood. The phylogenetic analysis revealed that maximum likelihood was 0.05 and three major branches corresponded to pisces, aves, and mammals. Each branch had several smaller branches as shown in Figure 5, which again supported the correlation between different classes. The cluster tree was generated with RSCU value in different species of pisces, aves, and mammals using cluster methods as shown in Figure 6. The different species of pisces, aves, and mammals were divided into three main clusters. Some species of different classes for *MT-CYB* gene were clustered in the same lineage while other species were clustered in different lineages.

# Conclusion

In our study on analysis of synonymous codon usage among pisces, aves, and mammals, the codon usage bias in *CYB* gene was low, which might be due to the interaction of mutation pressure and natural selection. The most frequent codons in *CYB* gene of pisces, aves, and mammals favoring A or C at the third codon position firmly decide the compositional constraint in the presence of mutation pressure. The present analysis of codon usage provides an insight into the codon usage variation at molecular level among pisces, aves, and mammals.

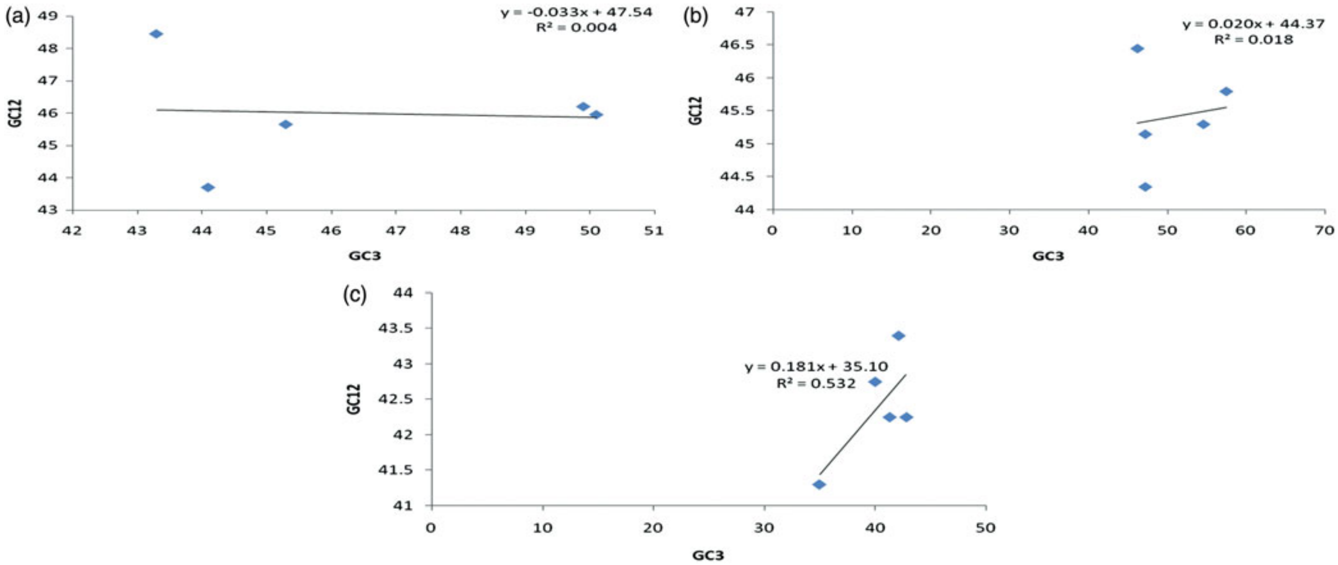


Figure 4. (a) Neutrality plot (between GC12 against GC3%) in pisces, regression line $y = 2.943X - 84.54$; $R^2 = 0.532$. (b) Neutrality plot (between GC12 against GC3%) in aves, regression line $y = 0.887X + 10.23$; $R^2 = 0.0018$. (c) Neutrality plot (between GC12 against GC3%) in mammals, regression line $y = -0.122X - 52.15$; $R^2 = 0.0004$.

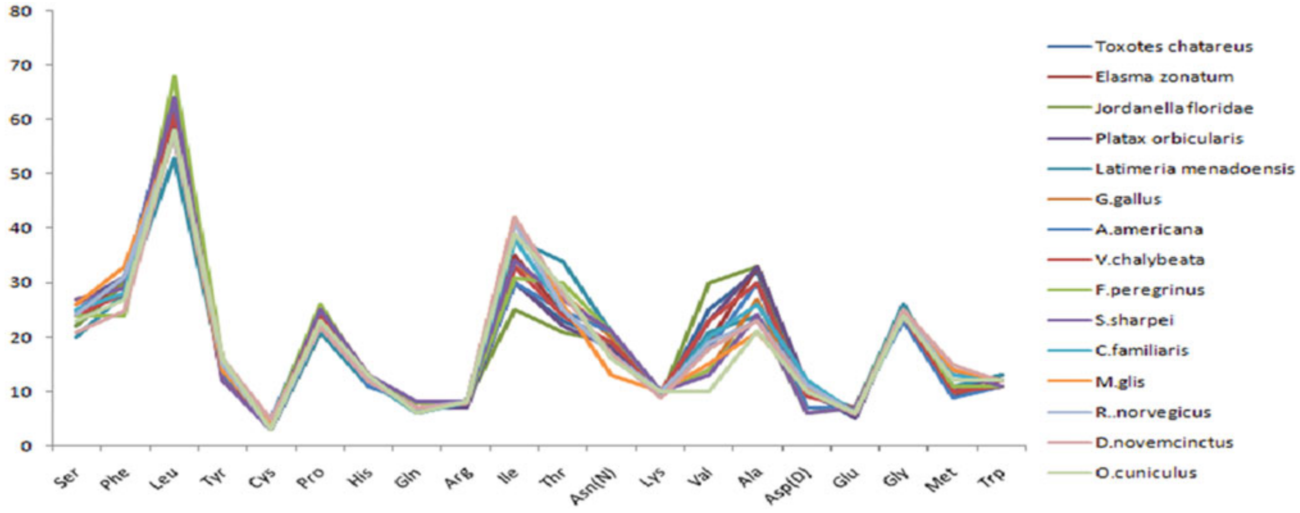


Figure 5. Comparison of amino acids in different species of pisces, aves, and mammals. This figure clearly shows leu frequency was the highest and cys frequency was the lowest.

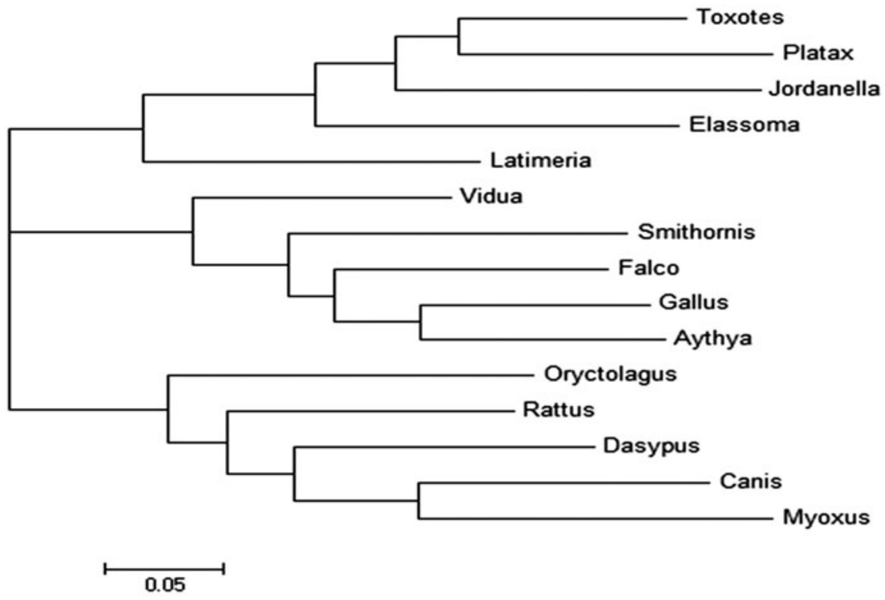


Figure 6. Phylogenetic tree of different species pisces, aves, and mammals.

## Acknowledgements

The authors are thankful to Assam University, Silchar, Assam, India, for providing necessary facilities.

## Declaration of interest

No fund was received from DST or DBT, Government of India, New Delhi, for this research work.

## References

Akashi H. (1997). Codon bias evolution in Drosophila. Population genetics of mutation-selection drift. Gene 205:269–78.

Akashi H. (1994). Synonymous codon usage in *Drosophila melanogaster*: Natural selection and translational accuracy. Genetics 136:927–35.

Anderson S, Bankier AT, Barrell BG, de Bruijn MH, Coulson AR, Drouin J, Eperon IC, et al. (1981). Sequence and organization of the human mitochondrial genome. Nature 290:457–65.

Andreu AL, Bruno C, Shanske S, Shtilbans A, Hirano M, Krishna S, Hayward L, et al. (1998). Missense mutation in the mtDNA cytochrome b gene in a patient with myopathy. Neurology 51:1444–7.

Andreu AL, Hanna MG, Reichmann H, Bruno C, Penn AS, Tanji K, Pallotti F, et al. (1999a). Exercise intolerance due to mutations in the cytochrome b gene of mitochondrial DNA. N Engl J Med 341:1037–44.

Andreu AL, Bruno C, Dunne TC, Tanji K, Shanske S, Sue CM, Krishna S, et al. (1999b). A nonsense mutation (G15059A) in the cytochrome b gene in a patient with exercise intolerance and myoglobinuria. Ann Neurol 45:127–30.

Andreu AL, Checcarelli N, Iwata S, Shanske S, DiMauro S. (2000). A missense mutation in the mitochondrial cytochrome b gene in a revisited case with histiocytoid cardiomyopathy. Pediatr Res 48:311–14.

Behura SK, Severson DW. (2012). Comparative analysis of codon usage bias and codon context patterns between Dipteran and Hymenopteran sequenced genomes. PLoS One 7:e43111.

Beletskii A, Bhagwat AS. (2001). Mutations induced by bacteriophage T7 RNA polymerase and their effects on the composition of T7 genome. J Bacteriol 183:6491–3.

Bruno C, Santorelli FM, Assereto S, Tonoli E, Tessa A, Traverso M, Scapolan S, et al. (2003). Progressive exercise intolerance associated with a new muscle-restricted nonsense mutation (G142X) in the mitochondrial *cytochrome b* gene. Muscle Nerve 28:508–11.

Butt AM, Nasrullah I, Tong Y. (2014). Genome-wide analysis of codon usage and influencing factors in Chikungunya viruses. PLoS One 9:e90905.

Chen H, Sun S, Norenburg JL, Sundberg P. (2014). Mutation and selection cause codon usage and bias in mitochondrial genomes of Ribbon Worms (Nemertea). PLoS One 9:e85631.

Chinnery PF, Johnson MA, Wardell TM, Singh-Kler R, Hayes C, Brown DT. (2000). The epidemiology of pathogenic mitochondrial DNA mutations. Ann Neurol 48:188–93.

De Coo IF, Renier WO, Ruitenbeek W, Ter Laak HJ, Bakker M, Schagger H, Oost van, et al. (1999). A 4-base pair deletion in the mitochondrial cytochrome b gene associated with parkinsonism/MELAS overlap syndrome. Ann Neurol 45:130–3.

Dumoulin R, Sagnol I, Ferlin T, Bozon D, Stepien G, Mousson B. (1996). A novel gly290asp mitochondrial cytochrome b mutation linked to a complex III deficiency in progressive exercise intolerance. Mol Cell Probes 10:389–91.

Esposti DM, De Vries S, Crimi M, Ghelli A, Patarnello T, Meyer A. (1993). Mitochondrial cytochrome *b*: Evolution and structure of the protein. Biochim Biophys Acta 1143:243–71.

Eyre-Walker A, Bulmer M. (1993). Reduced synonymous substitution rate at the start of enterobacterial genes. Nucleic Acids Res 21:4599–603.

Fujimori S, Washio T, Tomita M. (2005). GC-compositional strand bias around transcription start sites in plants and fungi. BMC Genomics. BMC Genomics 6:26.

Gu W, Zhou T, Ma J, Sun X, Lu Z. (2004). Analysis of synonymous codon usage in SARS Coronavirus and other viruses in the Nidovirales. Virus Res 101:155–61.

Gupta SK, Bhattacharyya TK, Ghosh TC. (2004). Synonymous codon usage in *Lactococcus lactis*: Mutational bias versus translational selection. J Biomol Struct Dyn 21:527–36.

Gutierrez G, Casadesús J, Oliver JL Marin A. (1994). Compositional heterogeneity of the *Escherichia coli* genome: A role for VSP repair?. J Mol Evol 39:340–6.

Hassanin A, Leger N, Deutsch J. (2005). Evidence for multiple reversals of asymmetric mutational constraints during the evolution of the mitochondrial genome of Metazoa, and consequences for phylogenetic inferences. Syst Biol 54:277–98.

Irwin DM, Kocher TD, Wilson AC. (1991). Evolution of cytochrome *b* gene in mammals. J Mol Biol Evol 2:13–34.

Izeni PF, Ort G, Sampaio I, Schneider H, Meyer A. (2001). The cytochrome *b* gene as a phylogenetic marker: The limits of resolution for analyzing relationships among Cichlid fishes. J Mol Evol 53:89–103.

Jenkins GM, Holme EC. (2003). The extent of codon usage bias in human RNA viruses and its evolutionary origin. Virus Res 92:1–7.

Ma JJ, Zhao F, Zhang J, Zhou JH, Ma Li-Na, Ding YZ, Chen HT, et al. (2013). Analysis of synonymous codon usage in Dengue viruses. J Anim Vet Adv 12:88–98.

Keightley JA, Anitori R, Burton MD, Quan F, Buist NR Kennaway NG. (2000). Mitochondrial encephalomyopathy and complex III deficiency associated with a stop-codon mutation in the cytochrome b gene. Am J Hum Genet 67:1400–10.

Lamantea E, Carrara F, Mariotti C, Morandi L, Tiranti V, Zeviani M. (2002). A novel nonsense mutation (Q352X) in the mitochondrial *cytochrome b* gene associated with a combined deficiency of complexes I and III. Neuromuscul Disord 12:49–52.

Larsson NG. (2010). Somatic mitochondrial DNA mutations in mammalian aging. Annu Rev Biochem 79:683–706.

Liu YS, Zhou JH, Chen HT, Ma LN, Pejsak Z, Dinga Y-Z, Zhanga J, et al. (2011). The characteristics of the synonymous codon usage in enterovirus 71 virus and the effects of host on the virus in codon usage pattern. Infect Genet Evol 11:1168–73.

Ma JJ, Zhao F, Zhang J, Zhou JH, Ma LN, Ding YZ, Chen HT, et al. (2013). Analysis of synonymous codon usage in Dengue viruses. J Anim Vet Adv 12:88–98.

Mancuso M, Filosto M, Stevens JC, Patterson M, Shanske S, Krishna S, Di Mauro S. (2003). Mitochondrial myopathy and complex III deficiency in a patient with a new stop-codon mutation (G339X) in the cytochrome b gene. J Neurol Sci 209:61–3.

Meyer A, Wilson C. (1990). Origin of tetrapods inferred from their mitochondrial DNA affiliation to lungfish. J Mol Evol 31:359–64.

MITOMAP. (2000). A human mitochondrial genome database. Available at: http://www.mitomap.org.

Moratorio G, Iriarte A, Moreno P, Musto H, Cristina J. (2013). A detailed comparative analysis on the overall codon usage patterns in West Nile virus. Infect Genet Evol 14:396–400.

Moritz C, Schneider CL, Wake DB. (1992). Evolutionary relationships within the *Ensatina eschscholtzii* complex confirm the ring species interpretation. Syst Biol 41:273–91.

Moriyama EN, Powell JR. (1998). Gene length and codon usage bias in *Drosophila melanogaster*, *Saccharomyces cerevisiae* and *Escherichia coli*. Nucleic Acids Res 26:3188–93.

NCBI. Translation Table 1. Available at: www.ncbi.nlm.nih.gov

Perriere G, Gouy M, Gojobori T. (1994). NRSub: A nonredundant data base for the *Bacillus subtilis* genome. Nucleic Acids Res 22:5525–9.

Powell JR, Moriyama EN. (1997). Evolution of codon usage bias in Drosophila. Proc Natl Acad Sci USA 94:7784–90.

Powell JR, Sezzi E, Moriyama EN, Gleason JM, Caccone A. (2003). Analysis of a shift in codon usage in Drosophila. J Mol Evol 57:S214–25.

Seligmann H. (2011). Mutation pressure due to converging mitochondrial replication and transcription increase lifespan, and cause growth rate-longevity tradeoffs. In: Herve S, editor, DNA replication – Current advances, doi: 10.5772/24319.

Sharma J, Chakraborty S, Uddin A. (2014). Comparative analysis of codon usage bias between two Lepidopteran insect species: *Bombyx mandarina* And *Ostrinia* furnacalis. Int J Sci Res 3:47–50.

Sharp PM, Li WH. (1986). Codon usage in regulatory genes in *Escherichia coli* does not reflect selection for 'rare' codons. Nucleic Acids Res 14:7737–49.

Sharp PM, Li WH. (1987). The codon Adaptation Index – A measure of directional synonymous codon usage bias, and its potential applications. Nucleic Acids Res 15:1281–95.

Sharp PM, Matassi G. (1986). Codon usage and genome evolution. Curr Opin Genet Dev 4:851–60.

Sharp PM, Cowe E, Higgins DG, Shields DC, Wolfe KH, Wright F, et al. (1988). Codon usage patterns in *Escherichia coli*, *Bacillus subtilis*, *Saccharomyces cerevisiae*, *Schizosaccharomyces pombe*, *Drosophila melanogaster* and *Homo sapiens*; a review of the considerable within-species diversity. Nucleic Acids Res 16:8207–11.

Tao P, Dai L, Luo M, Tang F, Tien P, Dai L, Luo M, et al. (2009). Analysis of synonymous codon usage in classical swine fever virus. Virus Genes 38:104–12.

Tillier ER, Collins RA. (2000). The contributions of replication orientation, gene direction, and signal sequences to base-composition asymmetries in bacterial genomes. J Mol Evol 50:249–57.

Touchon M, Rocha EP. (2008). From GC skews to wavelets: A gentle guide to the analysis of compositional asymmetries in genomic data. Biochimie 90:648–59.

Uddin A, Chakraborty S. (2014a). Analysis of codon usage pattern in mitochondrial *atpase6* in some mammalian species. Int J Recent Sci Res Res 5:883–8.

Uddin A, Chakraborty S. (2014b). Mutation pressure dictates codon usage pattern in mitochondrial *atpase8* in some mammalian species. Int J Sci Res (IJSR) 3:2206–12.

Valnot I, Kassis J, Chretien D, de Lonlay P, Parfait B, Munnich A, Kachaner J, et al. (1999). A mitochondrial cytochrome b mutation but no mutations of nuclearly encoded subunits in ubiquinol cytochrome c reductase (complex III) deficiency. Hum Genet 104:460–6.

Vicario S, Moriyama EN, Powell JR. (2007). Codon usage in twelve species of Drosophila. BMC Evol Biol 7:226.

Wright F. (1990). The 'effective number of codons' used in a gene. Gene 87:23–9.

Xing Y, Xuenong L, Xuepeng C. (2014). Analysis of codon usage pattern in Taenia saginata based on a transcriptome dataset. Parasites Vectors 2014. 7:527.

Zhang Z, Dai W, Wang Y, Lu C, Fan H. (2013). Analysis of synonymous codon usage patterns in torque teno sus virus 1 (TTSuV1). Arch Virol 158:145–54.

BIOINFORMATION
Discovery at the interface of physical and biological sciences

open access www.bioinformation.net Hypothesis
Volume 11(8)

# Codon bias and gene expression of mitochondrial *ND2* gene in chordates

**Arif Uddin, Tarikul Huda Mazumder, Monisha Nath Choudhury & Supriyo Chakraborty***

Department of Biotechnology, Assam University, Silchar-788011, Assam, India; Supriyo Chakraborty - Email: supriyoch_2008@rediffmail.com; *Corresponding author

Received August 05, 2015; Accepted August 24, 2015; Published August 31, 2015

**Abstract:**
**Background**: Mitochondrial *ND* gene, which encodes NADH dehydrogenase, is the first enzyme of the mitochondrial electron transport chain. Leigh syndrome, a neurodegenerative disease caused by mutation in the *ND2* gene (T4681C), is associated with bilateral symmetric lesions in basal ganglia and subcortical brain regions. Therefore, it is of interest to analyze mitochondrial DNA to glean information for evolutionary relationship. This study highlights on the analysis of compositional dynamics and selection pressure in shaping the codon usage patterns in the coding sequence of *MT-ND2* gene across pisces, aves and mammals by using bioinformatics tools like effective number of codons (ENC), codon adaptation index (CAI), relative synonymous codon usage (RSCU) etc. **Results**: We observed a low codon usage bias as reflected by high ENC values in *MT-ND2* gene among pisces, aves and mammals. The most frequently used codons were ending with A/C at the 3rd position of codon and the gene was AT rich in all the three classes. The codons TCA, CTA, CGA and TGA were over represented in all three classes. The F1 correspondence showed significant positive correlation with G, T3 and CAI while the F2 axis showed significant negative correlation with A and T but significant positive correlation with G, C, G3, C3, ENC, GC, GC1, GC2 and GC3. **Conclusions**: The codon usage bias in *MT-ND2* gene is not associated with expression level. Mutation pressure and natural selection affect the codon usage pattern in *MT-ND 2* gene.

**Key words:** Codon usage, *MT-ND2* gene, natural selection, mutation pressure

## Background

The mitochondrial genome is ideal as the molecular marker for species identification as well as systematic phylogenetic studies due to its small size. It is easily amplified and mostly conserved in gene content and characterized by lack of recombination, maternal inheritance and high evolutionary rate **[1]**. The respiratory chains of mitochondrial genome comprise of four complexes (complex, I-IV) and are encoded by 37 genes consisting of two ribosomal RNA (rRNA), twenty-two transfer RNA (tRNA) and thirteen protein coding genes. The complex-I of mitochondrial respiratory chain includes the first enzyme NADH dehydrogenase and its seven subunits (ND1-6 & ND4L) play a pivotal role in diverse pathological processes **[2].** The subunit 2 of NADH dehydrogenase is encoded by *ND2* gene and its function is not yet fully understood. However, literature suggests that a mutation in the *ND2* (T4681C) gene was found in patients with Leigh syndrome, a neurodegenerative disease characterized by bilateral symmetric lesions in basal ganglia and subcortical brain regions **[3].**

Urrutia and Hurst (2003) reported that the codon usage in human is positively related to gene expression but is inversely related to the rate of synonymous substitution **[4].** Several genomic factors such as gene expression level, protein secondary structure, and translational preferences balancing between the mutational pressure and natural selection contribute to the synonymous codon usage variation in different organisms **[5,6]**. Therefore, gaining the information on the synonymous codon usage pattern provides significant insights pertaining to the prediction, classification, and evolution of a gene at molecular level and also helps in designing highly expressed genes. In the present study we have carried out a comparative analysis of the *ND2* gene codon usage and codon context patterns among the mitochondrial genomes of three chordate classes (pisces, aves and mammals) in order to understand the molecular mechanism along with functional conservation of gene expression during the period of evolution using several bioinformatics tools.

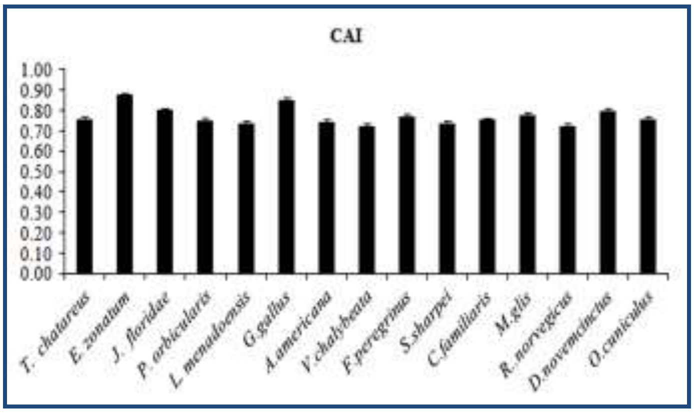


**Figure 1:** Distribution of CAI in MT-ND2 gene among different species

**Methodology:**

***Retrieval of Sequence data***

The coding sequences (cds) of *MT-ND2* gene from five species of pisces, aves and mammals each were retrieved from National Center for Biotechnology Information, USA (http://www.ncbi.nlm.nih.gov/) using the following accession numbers. The accession numbers of different species are AP006806, AP006813, AP006778, AP006825, AP006858, X52392, AF090337, AF090341, AF090338, AF090340, U96639, AJ001562, X14848, Y11832 and AJ001588. A perl programme was used to analyze the compositional features and codon usage bias parameters.

***Compositional properties***

The overall composition of A, T, G, C bases and its composition at 3rd position along with GC, GC1, GC2 and GC3 contents were calculated using the perl script.

***Codon adaptation Index (CAI)***

Codon adaptation index (CAI) is used to estimate gene expression level. The CAI is calculated as

$$CAI = \exp\left( \frac{1}{L} \sum_{k=1}^{L} \ln \omega k \right)$$

Where, ωk is the relative adaptiveness of the $k_{th}$ codon and L is the number of synonymous codons in the gene **[7]**.

***Effective Number of Codons (ENC)***

The effective number of codons (ENC) is the most extensively used parameter to measure the usage bias of the synonymous codons **[8].** The ENC value ranges from 20 (when only one codon is used for each amino acid) to 61 (when all codons are used randomly). It is calculated as:

$$ENC = 2 + \frac{9}{F_2} + \frac{1}{F_3} + \frac{5}{F_4} + \frac{3}{F_6}$$

Where $F_k$(k= 2,3,4,6) is the mean of Fk values for the k-fold degenerate amino acids.

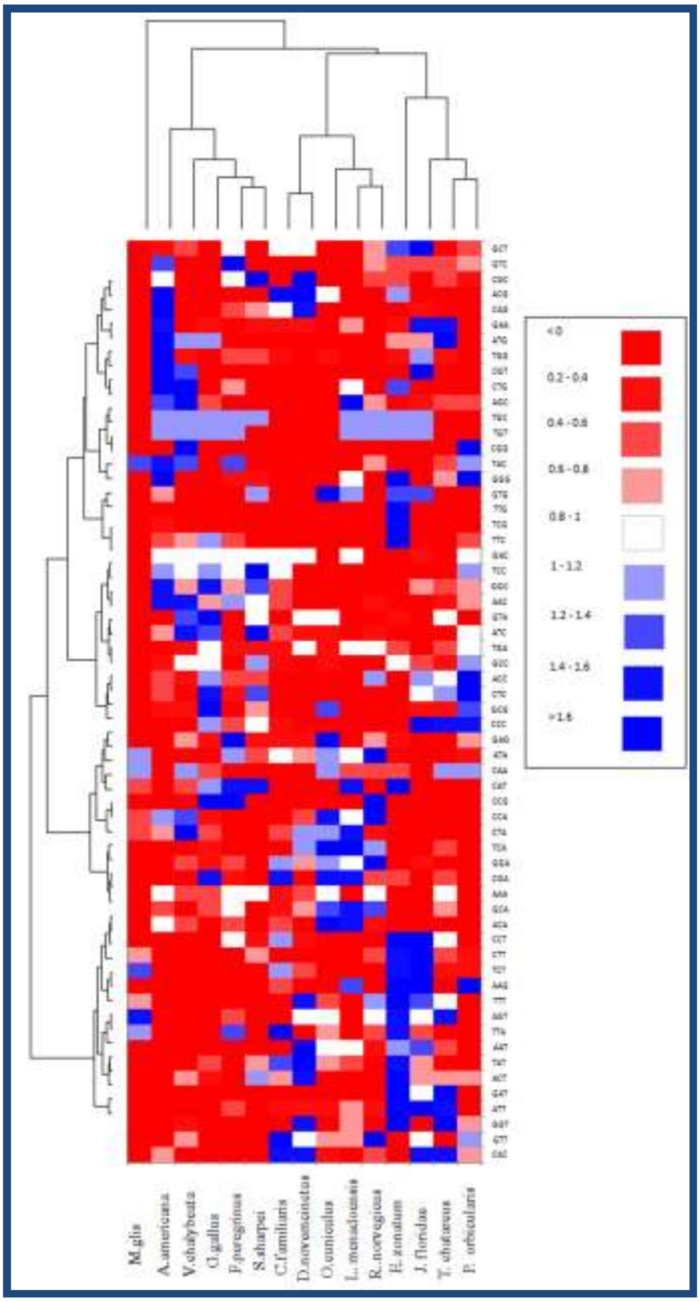


**Figure 2:** Hierarchal clustering of RSCU value in different species of MT-ND2 gene

***Relative Synonymous Codon Usage (RSCU)***

Relative synonymous codon usage was calculated as the ratio of the observed frequency of a codon to its expected frequency if all the synonymous codons of a particular amino acid are used equally **[9].** The RSCU value is calculated using the formula

$$RSCUij = \frac{Xij}{\frac{1}{ni}\sum_{j=1}^{ni} Xij}$$

where, $X_{ij}$ is the frequency of occurrence of the jth codon for ith amino acid (any $X_{ij}$ with a value of zero is arbitrarily assigned a value of 0.5) and $n_i$ is the number of codons for the $i^{th}$ amino acid ( $i^{th}$ codon family).

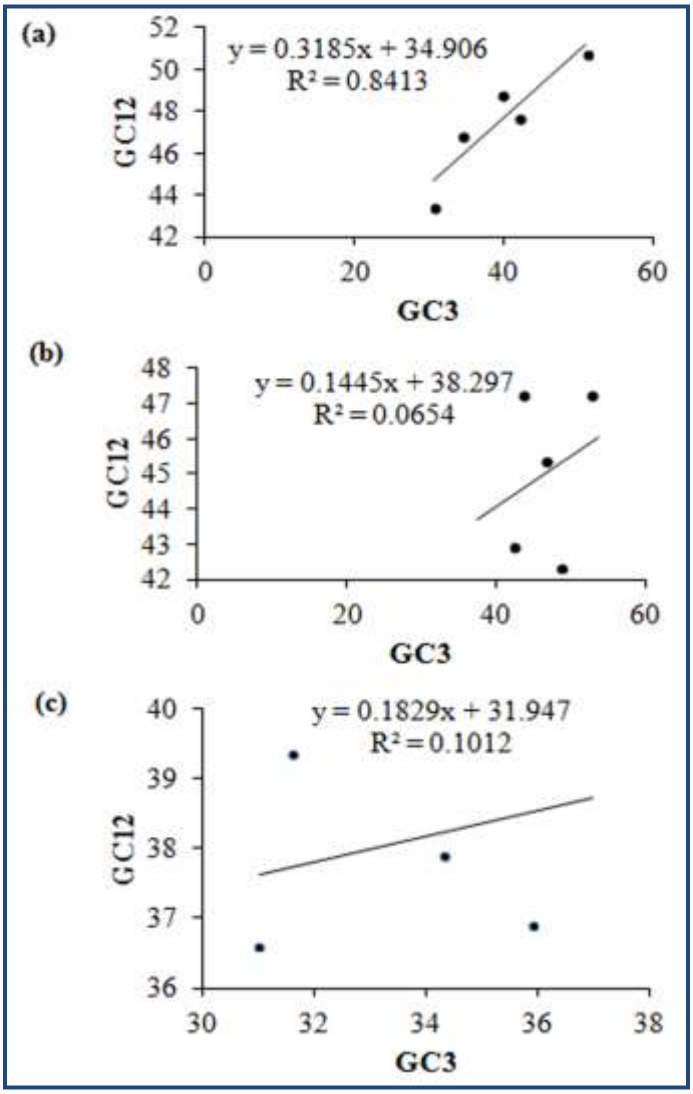


**Figure 3:** Neutrality plot of GC12 versus GC3 in (a) Pisces (b) Aves (c) mammals. GC12: average of GC1 and GC2.

*GRAVY*

GRAVY (Grand Average of Hydropathicity) values are the sum of the hydropathy values of all the amino acids in the encoded protein of the gene divided by the number of residues in the sequence **[10].**

*Aromaticity* **(Aromo)**

Aromo refers to the frequency of aromatic amino acids (Phe, Tyr, Trp) in the translated gene product **[11].**

*Correspondence analysis (COA)*

Correspondence analysis is a multivariate statistical method used to study the major trends in synonymous codon usage variation in coding sequences and distributes the codons in axis1 and axis2 with these trends **[12].**

*Software used*

Novel software developed by SC (corresponding author) using Perl script was used to calculate all the codon usage bias parameters and nucleotide composition. The genetic code of vertebrate mitochondria having 60 sense codons available in NCBI database was used for the present analysis. The RSCU values of each codon from different species were clustered by hierarchal clustering method using XLSTAT.

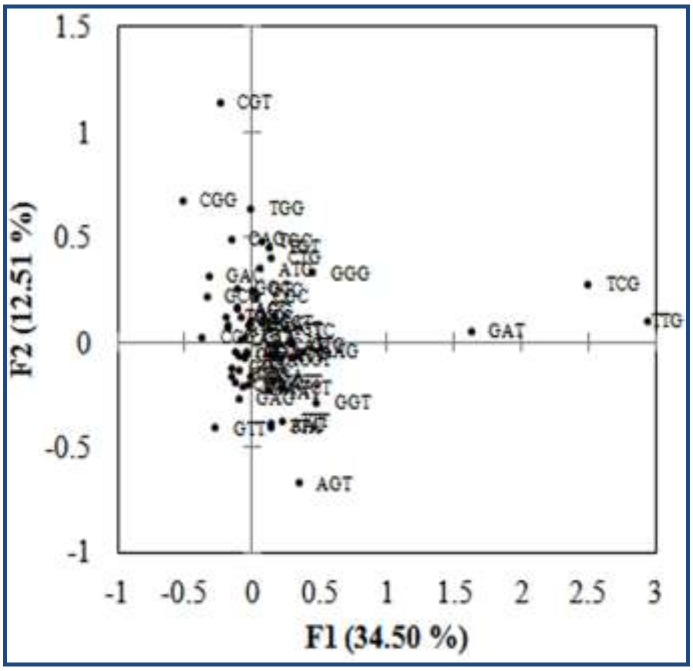


**Figure 4:** Correspondence analysis of the synonymous codon usage in MT-ND2 gene. The analysis was based on the RSCU value of the 60 synonymous codons.

***Statistical analysis***

Correlation analysis was used to identify the relationship between overall nucleotide composition and each base at 3rd codon position. All the statistical analyses were done using the SPSS software**.**

**Results & Discussion:**

The overall nucleotide compositions in the coding sequence of *MT-ND2* gene among pisces, aves and mammals were analyzed **Table 1 (see supplementary material)**. Our results showed that the nucleobase C was the highest (%) in pisces and aves but the nucleobase A was the highest in mammals whereas G was the lowest in pisces, aves and mammals. For the 3rd position of codon, A3 was the highest in pisces, aves and mammals but G3 the lowest. This clearly indicates that compositional constraint might influence the codon usage pattern of *MT-ND2* gene **[13].**

The effective number of codon (ENC) values for *MT-ND2* gene among pisces, aves and mammals were estimated **Table 2 (see supplementary material)**. The effective number of codon (ENC) is a non directional measure of codon usage bias. Its value ranges from 20-61. ENC value 20 indicates that for each amino acid only one codon is used (extreme bias) and 61 means all codons equally encode the same amino acids (no bias) [8]. We observed that the ENC value in *MT-ND2* gene was (Mean±SD) 57±2.91, 59±0.44 and 55±1.58 among pisces, aves and mammals, reflecting a weak codon bias which was similar to the findings of Jia X *et al*, 2015 in *B.mori* **[14]** . It was also found that the overall GC % was less than 50% and the gene was AT rich. This phenomenon was also reported in AT rich species such as *Plasmodium falciparum* **[15].**

We calculated the codon adaptation index (CAI) values for *MT-ND2* gene in order to find out the expression level among pisces, aves and mammals (**Figure 1**). In our analysis, the CAI values were (Mean±SD) 0.7851±0.05, 0.7667±0.05, 0.7635±0.02 in pisces, aves and mammals, respectively. We used unpaired t test between pisces and aves as well as between pisces and mammals but the difference was not statistically significant. Wei *et.al* 2014, also reported the average value of CAI in mitochondrial protein coding genes ranged from 0.5-0.7 in *B.mori* **[16]**. In addition, we performed a correlation analysis between ENC and gene expression level as measured by CAI and found no significant relationship suggesting that the codon usage bias in *MT-ND2* gene is not associated with expression level among the three classes.

Moreover, we calculated the relative synonymous codon usage (RSCU) values in the coding sequences of *MT-ND2* gene among pisces, aves and mammals **Table 3 (see supplementary material)**. In our analysis, the RSCU value > 1 means the codon is more frequently used, RSCU value <1 as less frequently used codon. But RSCU value >1.6 means over represented codon whereas the RSCU value <0.06 as under represented codon. The RSCU values of 60 codons also indicated that the codon usage bias of *MT-ND2* gene is low. Approximately half of the codons were frequently used i.e. 30 codons in pisces, 28 in aves and 21 in mammals. In all three classes the most frequent codons ended with A or C at the 3rd codon position (A/T ended: G/C ended = 16:14 in pisces, 13: 15 in aves and 14: 7 in mammals). The RSCU values were analyzed using heat map (**Figure 2**) which represented the more frequently, less frequently, over-represented and under-represented codons. The results showed that the most preferred codons were TTC, CTA and TAC in all the three classes. In addition, the four codons namely TCA, CTA, CGA and TGA encoding the amino acid serine, leucine, arginine and tryptophan respectively were over represented in *MT-ND2* gene among pisces, aves and mammals. This result suggests that compositional features played an important role in codon usage in *MT-ND2* gene **[13]**.

The overall percentage of GC contents at different codon positions were calculated (see supplementary material **Table S2**). In order to find out the role of mutation pressure and natural selection, we constructed a neutrality plot of GC12 against GC3 (**Figure 3, a-c**) **[17].** The linear regression coefficient of GC12 on GC3 indicated that natural selection plays a major role while mutation pressure plays a minor role in shaping the codon usage patterns in *MT-ND2* gene. Our result was similar to the findings of Wei *et.al* (2014) in the mitochondrial DNA codon usage analysis of *B.mori* **[16].**

We performed correspondence analysis (CoA) based on RSCU values to analyze the codon usage variation in *MT-ND2* gene among pisces, aves and mammals. In our analysis, the 1st axis (F1) accounted for 34.50% of the total variation and the 2nd axis accounted for 12.51% of the total variation (**Figure 4**). Further, correlation analysis was done to determine the interrelationships between the first two principle axes (F1 and F2), nucleotide constraints and indices of natural selection (CAI, Gravy, Aromo) on *MT-ND2* gene. The F1 axis showed significant positive correlation with G, T3 and CAI whereas the F2 showed significant positive correlation with G, C, G3, C3, ENC, GC and GC1-3 but significant negative correlation with A and T **Table 4 (see supplementary material)**. These results suggest both compositional constraint under mutation pressure and natural selection affect the codon usage pattern in *MT-ND2*. The results were similar to the findings of Butt *et.al.* **[18].**

## Conclusion:

The codon usage bias in *MT-ND2* gene is weak with high expression level. It is found that natural selection and mutation pressure affect the codon usage pattern in *MT-ND 2* gene.

## Acknowledgment:

We are thankful to Assam University, Silchar, Assam, India, for providing the necessary infrastructural facilities for this work.

## References:

**[1]** Modica-Napolitano JS & Singh KK, *Mitochondrion*. 2004 **4**: 755 [PMID: 16120430]

**[2]** Bellance N *et al. Front Biosci*. 2009 **14:** 4015 [PMID: 19273331]

**[3]** Ugalde C *et al. Mol Genet Metab*. 2007 **90:** 10 [PMID: 16996290]

**[4]** Urrutia AO & Hurst LD, *Genome Res*. 2003 **13**: 2260 [PMID: 12975314]

**[5]** Sharp PM *et al. Biochem Soc Trans*. 1993 **21**: 835 [PMID: 8132077]

**[6]** Stenico M *et al. Nucleic Acids Res*. 1994 **22:** 2437 [PMID: 8041603]

**[7]** Sharp PM *et al. Nucleic Acids Res*. 1986 **14**: 5125 [PMID: 3526280]

**[8]** Wright F, *Gene*. 1990 **87**: 23 [PMID: 2110097]

**[9]** Gupta SK & Ghosh TC, *Gene* 2001 **273**: 63 [PMID: 11483361]

**[10]** Kyte J & Doolittle RF, *J Mol Biol* . 1982 **157**: 105 [PMID: 7108955]

**[11]** Lobry JR & Gautier C, *Nucleic Acids Res*. 1994 **22**: 3174 [PMID: 8065933]

**[12]** Shields DC & Sharp PM, *Nucleic Acids Res*. 1987 **15**: 8023 [PMID: 3118331]

**[13]** Zhang Z *et al. PloS one*. 2013 **8**: e81469 [PMID: 24303050]

**[14]** Jia X *et al. BMC Genomics*. 2015 **16:** 356 [PMID: 25943559]

**[15]** Peixoto L *et al. Parasitology*. 2004 **128:** 245 [PMID: 15074874]

**[16]** Wei L *et al. BMC Evol Biol*. 2014 **14:** 262 [PMID: 25515024]

**[17]** Sueoka N, *Proc Natl Acad Sci U S A*. 1988 **85:** 2653 [PMID: 3357886]

**[18]** Butt AM *et al. PloS one*. 2014 **9**: e90905 [PMID: 24595095]

**Edited by P Kangueane**

**Citation**: **Uddin *et al.*** Bioinformation 11(8): 407-412 (2015)

## Supplementary material:

**Table 1**: Nucleotide composition in MT-ND2 gene

| Species | A % | T % | G % | C % | A3 % | T3 % | G3 % | C3 % |
|---|---|---|---|---|---|---|---|---|
| *T. chatareus* | 29.2 | 24.9 | 11.4 | 34.5 | 41 | 19.2 | 3.7 | 36.1 |
| *E. zonatum* | 28.7 | 28.7 | 13.8 | 28.8 | 37.8 | 27.8 | 8.9 | 25.5 |
| *J. floridae* | 25.2 | 28.9 | 11.6 | 34.3 | 29.8 | 28.1 | 4.9 | 37.2 |
| *P. orbicularis* | 26.3 | 22.9 | 11.5 | 39.3 | 33.8 | 15.2 | 4.3 | 46.7 |
| *L. menadoensis* | 38.2 | 22.6 | 10.8 | 28.4 | 55.9 | 13.5 | 5.4 | 25.2 |
| ***Mean±SD*** | **29.52±5.12** | **25.60±±3.05** | **11.82±1.14** | **33.06±4.53** | **39.66±10** | **20.76±6.88** | **5.44±2.03** | **34.14±9.02** |
| *G.gallus* | 32.6 | 23 | 8.6 | 35.8 | 42.4 | 8.9 | 4.3 | 44.4 |
| *A.americana* | 29.1 | 21.8 | 13.1 | 36 | 38 | 9.2 | 10.7 | 42.1 |
| *V.chalybeata* | 31 | 23 | 11.9 | 34.1 | 45.5 | 11 | 6.1 | 37.4 |
| *F.peregrinus* | 32.2 | 25.1 | 9.4 | 33.3 | 43.5 | 14.1 | 4.6 | 37.8 |
| *S.sharpei* | 29.7 | 24.5 | 8.7 | 37.1 | 37.8 | 15.6 | 3.2 | 43.4 |
| ***Mean±SD*** | **30.92±1.52** | **23.48±1.31** | **10.34±2.04** | **35.26±1.53** | **41.44±3.41** | **11.76±2.98** | **5.78±2.93** | **41.02±3.22** |
| *C.familiaris* | 35.4 | 28 | 9.2 | 27.4 | 46 | 18.1 | 5.2 | 30.7 |
| *M.glis* | 33.2 | 30.1 | 9.3 | 27.4 | 42.7 | 23.1 | 3.5 | 30.7 |
| *R..norvegicus* | 36 | 25 | 7.9 | 31.1 | 47.1 | 15.9 | 2 | 35 |
| *D.novemcinctus* | 39.1 | 26.1 | 8.3 | 26.5 | 45.7 | 22.7 | 4 | 27.6 |
| *O.cuniculus* | 35 | 28.3 | 9.1 | 27.6 | 54.6 | 14.4 | 4.6 | 26.4 |
| ***Mean±SD*** | **35.74±2.14** | **27.50±±1.99** | **8.76±0.62** | **28±1.78** | **47.22±4.43** | **18.84±3.93** | **3.86±1.21** | **30.08±3.34** |

**Table 2**: GC contents, ENC and CAI values

| Species | ENC | GC % | GC1 % | GC2 % | GC3 % | CAI |
|---|---|---|---|---|---|---|
| *T. chatareus* | 58 | 45.8 | 52.1 | 45.6 | 39.8 | 0.7587 |
| *E. zonatum* | 55 | 42.7 | 48.1 | 45.6 | 34.4 | 0.8768 |
| *J. floridae* | 59 | 45.8 | 50.1 | 45.3 | 42.1 | 0.8011 |
| *P. orbicularis* | 60 | 50.8 | 55.9 | 45.6 | 51 | 0.7493 |
| *L. menadoensis* | 53 | 39.2 | 43.6 | 43.3 | 30.7 | 0.7396 |
| ***Mean±SD*** | **57±2.91** | **44.86±4.29** | **49.86±4.57** | **45.08±1.00** | **39.60±7.78** | **0.7851±0.05** |
| *G.gallus* | 60 | 44.5 | 41.8 | 42.9 | 48.7 | 0.8546 |
| *A.americana* | 60 | 49.1 | 51.9 | 42.7 | 52.7 | 0.7433 |
| *V.chalybeata* | 60 | 46 | 51.9 | 42.7 | 43.5 | 0.7251 |
| *F.peregrinus* | 59 | 42.7 | 41.5 | 44.4 | 42.4 | 0.7733 |
| *S.sharpei* | 60 | 45.8 | 46.7 | 44.1 | 46.7 | 0.7372 |
| ***Mean±SD*** | **59.80±0.44** | **45.62±2.34** | **46.76±5.12** | **43.36±0.82** | **46.80±4.14** | **0.7667±0.05** |
| *C.familiaris* | 56 | 36.6 | 36.2 | 37.6 | 35.9 | 0.7544 |
| *M.glis* | 55 | 36.7 | 36.3 | 39.5 | 34.3 | 0.7799 |
| *R..norvegicus* | 57 | 39 | 40.8 | 39.3 | 37 | 0.7247 |
| *D.novemcinctus* | 54 | 36.8 | 40.8 | 37.9 | 31.6 | 0.7988 |
| *O.cuniculus* | 53 | 34.8 | 35.6 | 37.6 | 31 | 0.7597 |
| ***Mean±SD*** | **55±1.58** | **36.78±1.49** | **37.94±2.62** | **38.38±0.94** | **33.96±2.61** | **0.7635±0.02** |

**Table 3:** Preferred codons of each amino acid in pisces, aves and mammals in MT-ND2

| | | **Pisces** | | **Aves** | | **Mammals** | |
|---|---|---|---|---|---|---|---|
| **Amino acid** | **Codon** | **Mean** | **Preferred codons** | **Mean** | **Preferred codons** | **Mean** | **Preferred codons** |
| **Ser** | TCA | 2.28 | **TCA** | 1.956 | **TCC** | 2.832 | **TCA** |
| | TCT | 1.032 | | 0.456 | | 0.924 | |
| | TCC | 1.476 | | 2.52 | | 1.536 | |
| | TCG | 0.68 | | 0.144 | | 0.048 | |
| | AGC | 0.888 | | 0.888 | | 0.456 | |
| | AGT | 0.168 | | 0.036 | | 0.204 | |
| **Phe** | TTC | 1.648 | **TTC** | 1.668 | **TTC** | 1.032 | **TTC** |
| | TTT | 0.912 | | 0.332 | | 0.968 | |
| **Leu** | TTA | 0.868 | **CTA** | 0.576 | **CTA** | 1.104 | **CTA** |
| | TTG | 1.1 | | 0.012 | | 0.048 | |
| | CTA | 2.028 | | 2.784 | | 2.88 | |
| | CTC | 1.44 | | 1.608 | | 0.972 | |
| | CTG | 0.36 | | 0.468 | | 0.168 | |
| | CTT | 1.152 | | 0.54 | | 0.864 | |
| **Tyr** | TAC | 1.272 | **TAC** | 1.508 | **TAC** | 1.172 | **TAC** |
| | TAT | 0.732 | | 0.496 | | 0.828 | |
| **Cys** | TGT | 0.012 | **TGC** | 0.016 | **TGC** | 0.004 | **TGC** |
| | TGC | 1.2 | | 2 | | 0.4 | |
| **Pro** | CCA | 1.272 | **CCC** | 1.88 | **CCA** | 2.12 | **CCA** |
| | CCC | 1.8 | | 1.56 | | 1.288 | |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| | CCG | 0 | | 0.08 | | 0.04 | |
| | CCT | 0.92 | | 0.472 | | 0.568 | |
| **His** | CAT | 0.376 | **CAC** | 0.596 | **CAC** | 0.248 | **CAC** |
| | CAC | 1.624 | | 1.404 | | 1.752 | |
| **Gln** | CAA | 1.876 | **CAA** | 1.588 | **CAA** | 1.756 | **CAA** |
| | CAG | 0.124 | | 0.412 | | 0.244 | |
| **Arg** | CGA | 2.88 | **CGA** | 2.264 | **CGA** | 2.626 | **CGA** |
| | CGC | 0.56 | | 1.064 | | 0.612 | |
| | CGG | 0.16 | | 0.2 | | 0 | |
| | CGT | 0.4 | | 0.464 | | 0.012 | |
| **Met** | ATG | 0.452 | **ATA** | 0.496 | **ATA** | 0.156 | **ATA** |
| | ATA | 1.548 | | 1.504 | | 1.844 | |
| **Ile** | ATC | 0.916 | **ATT** | 1.388 | **ATC** | 0.986 | **ATC** |
| | ATT | 1.084 | | 0.612 | | 0.714 | |
| **Thr** | ACA | 1.312 | **ACC** | 1.528 | **ACC** | 1.506 | **ACA** |
| | ACC | 1.68 | | 1.648 | | 0.984 | |
| | ACG | 0.08 | | 0.104 | | 0.188 | |
| | ACT | 0.92 | | 0.72 | | 0.58 | |
| **Asn** | AAC | 1.432 | **AAC** | 1.872 | **AAC** | 1.16 | **AAC** |
| | AAT | 0.568 | | 0.128 | | 0.54 | |
| **Asp** | GAC | 1.204 | **GAC** | 2 | **GAC** | 0.8 | **GAC** |
| | GAT | 0.804 | | 0.02 | | 0.008 | |
| **Val** | GTA | 1.408 | **GTC** | 1.912 | **GTA** | 1.314 | **GTA** |
| | GTC | 1.576 | | 1.664 | | 0.848 | |
| | GTG | 0.456 | | 0.288 | | 0.27 | |
| | GTT | 0.552 | | 0.136 | | 0.818 | |
| **Ala** | GCA | 1.232 | **GCC** | 1.312 | **GCC** | 1.288 | **GCC** |
| | GCC | 1.856 | | 1.912 | | 1.332 | |
| | GCG | 0.096 | | 0.16 | | 0.048 | |
| | GCT | 0.808 | | 0.608 | | 0.574 | |
| **Lys** | AAA | 1.592 | **AAA** | 1.936 | **AAA** | 1.58 | **AAA** |
| | AAG | 0.408 | | 0.064 | | 0.12 | |
| **Glu** | GAA | 1.764 | **GAA** | 1.58 | **GAA** | 1.224 | **GAA** |
| | GAG | 0.236 | | 0.42 | | 0.476 | |
| **Gly** | GGA | 1.48 | **GGA** | 1.712 | **GGC** | 2.088 | **GGA** |
| | GGC | 1.336 | | 1.952 | | 0.882 | |
| | GGG | 0.552 | | 0.272 | | 0.056 | |
| | GGT | 0.64 | | 0.064 | | 0.232 | |
| **Trp** | TGG | 0.14 | **TGA** | 0.288 | **TGA** | 0.09 | **TGA** |
| | TGA | 1.86 | | 1.712 | | 1.61 | |

**Table 4:** Correlation among the first two principle axes, nucleotide constraints and indices of natural selection in MT-ND2 gene

| | A | T | G | C | A3 | T3 | G3 | C3 | ENC | GC | GC1 | GC2 | GC3 | CAI | AR | GR |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| F1 | -0.253 | 0.407 | 0.544$^{*}$ | -0.249 | -0.227 | 0.587$^{*}$ | 0.454 | -0.414 | -0.271 | -0.005 | 0.135 | 0.327 | -0.268 | 0.663$^{**}$ | 0.175 | 0.035 |
| F2 | -0.559$^{*}$ | -.620$^{*}$ | .522$^{*}$ | 0.724$^{**}$ | -0.507 | -0.346 | 0.550$^{*}$ | 0.618$^{*}$ | 0.751$^{**}$ | 0.820$^{**}$ | 0.720$^{**}$ | 0.625$^{*}$ | 0.765$^{**}$ | -0.026 | -0.182 | -0.066 |

AR: aromaticity, GR: gravy score

Contents lists available at ScienceDirect

Genomics

journal homepage: www.elsevier.com/locate/ygeno

# Transcription factor gene *GATA2*: Association of leukemia and nonsynonymous to the synonymous substitution rate across five mammals

Tarikul Huda Mazumder, Arif Uddin, Supriyo Chakraborty *

*Department of Biotechnology, Assam University, Silchar 788011, Assam, India*

ARTICLE INFO

*Article history:*
Received 7 December 2015
Received in revised form 15 January 2016
Accepted 1 February 2016
Available online xxxx

*Keywords:*
*GATA2*
Transcription factor
Amino acid substitution
Mutation rate

ABSTRACT

*GATA2* gene encodes a member of the GATA family of zinc-finger transcription factors that play a pivotal role during the transition of primitive blood forming cells into white blood cells. Mutation in *GATA2* results in the loss of function or even gain of function, including abnormal proliferation of white blood cells that may predispose to acute myeloid leukemia. Our results showed that the codon usage in *GATA2* has been influenced by GC mutation bias where nature has highly favored fourteen most over represented codons but disfavored the ATA codon across five mammals. Purifying natural selection has affected *GATA2* gene in human and other mammals to maintain its protein function during the period of evolution. Our findings report an insight into the codon usage patterns in gaining the clues for codon optimization to alter the translational efficiency as well as for the functional conservation of gene expression and the significance of nucleotide composition in *GATA2* gene within mammals.

## 1. Introduction

*GATA2* is a DNA binding transcription factor which localizes predominantly in the nucleus of a cell and mainly expressed in hematopoietic progenitors as well as nonhematopoietic embryonic stem cells [1]. Literature suggests that mutations in the coding region of *GATA2* lead to negative regulation of hematopoietic stem/progenitor cell differentiation and causes several genetic disorders, including predisposition to acute myeloid leukemia [1–3]. Based on the homology model, it has been established that amino acids asparagine317, alanine318, lysine321, and arginine330 are directly implicated in the DNA binding side of *GATA2* zinc finger1 and mutations in these residues likely alter DNA affinity [4].

The degeneracy of codon usage in nature and unequal usage of synonymous codons for encoding the same amino acid during the translation of a gene into a protein are a well established phenomenon commonly known as codon usage bias. It is species specific and significantly differs among the genes of the same taxa [5–8].The codon usage patterns have been analyzed since the inception of the first molecular sequence databases [5]. The result of Grantham and his co-workers demonstrated that species specific genes share similar patterns of synonymous codon usage frequency as stated by the "genome hypothesis" [5–6]. Therefore, scanning the codon usage patterns of all the genes in an organism may obscure the underlying heterogeneity [7] and hence it is better to identify the trends of codon usage patterns within the genes of a species or between closely related species. Various factors that are responsible for codon usage bias in different organisms from lower prokaryotes to higher eukaryotes have been discussed earlier by several researchers, but till date the codon usage patterns within the genes of an organism during the course of evolution have been interpreted for varied explanations. In general, researchers reported that the compositional constraints under mutation pressure or natural selection have been considered as the major factors involved in the codon usage variation among different organisms [8–11]. Further, literature suggests that a gene can be characterized both in the presence of its amino acid sequence and the codon usage patterns shaped by the balance between mutational pressure and natural selection. Such selection pressure might have influenced the differentiation of codon bias between species and resulted in the non-uniform usage of synonymous codons within a gene [12]. Thus, the significance underlying codon bias study is not only to understand the evolution of a gene at the molecular level, but also to analyze the functional conservation of gene expression as well as genome characterization.

Several studies were carried out in *GATA2* gene mutation linked to leukemia in human [1,4,13–14], but the studies related to the factors influencing the extent of synonymous codon usage bias in this gene sequence in comparison to other mammals have not been done so far. The present study was based on bioinformatic approaches to elucidate the synonymous codon usage patterns in *GATA2* gene and the ratio of

* Corresponding author.
*E-mail address:* supriyoch_2008@rediffmail.com (S. Chakraborty).

http://dx.doi.org/10.1016/j.ygeno.2016.02.001
0888-7543/© 2016 Published by Elsevier Inc.

nonsynonymous and synonymous substitution per site across five different mammals during the process of evolution.

## 2. Materials and methods

### 2.1. Sequences

The complete nucleotide coding sequence (cds) for *GATA2* gene in FASTA format from five randomly chosen mammalian species (n = 5) namely *Homo sapiens*, *Mus musculus*, *Sus scrofa*, *Bos taurus* and *Rattus norvegicus* was retrieved from GenBank database of the National Center for Biotechnology Information (NCBI) (http://www.ncbi.nlm.nih.gov). In our analysis of codon usage bias, we selected only those nucleotide coding sequences of equal length, which have a perfect start and stop codons, devoid of any unknown bases (N) and exact multiple of three bases from five different mammals available in the databases (Table 1).

### 2.2. Codon usage analysis

The nucleotide distribution of AT and GC contents (Mean ± SD) at different synonymous positions of codon for *GATA2* gene in each coding sequence was analyzed in order to find out the extent of base composition bias across different mammals. In addition, the skewness for AT and GC contents was analyzed.

### 2.3. Effective number of codons

The expected effective number of codons (*ENC*) for each coding sequence of *GATA2* gene was calculated using the formula given by Wright (1990) as follows:

$$ENC = 2 + S + \frac{29}{S^2 + \left(1 - S^2\right)}$$

where, *S* corresponds to the given $GC_3$ values. *ENC* value generally ranges from 20 to 61. The lower ENC value (<35) indicates high codon usage bias in the gene and vice versa [15].

### 2.4. Relative synonymous codon usage

The relative synonymous codon usage (*RSCU*) values of different codons in the coding sequences of *GATA2* gene were calculated as per Comeron and Aguade (1998) using the following formula:

$$RSCU = \frac{g_{ij}}{\sum_{j}^{ni} g_{ij}} n_i$$

where, $g_{ij}$ is the relative codon usage frequency of the *i*th codon for the *j*th amino acid which is encoded by $n_i$ synonymous codons [16]. In our analysis, *RSCU* value greater than 1.0 represents positive codon usage bias indicating the usage of most abundant codons for the corresponding amino acid. *RSCU* value less than 1.0 represents a negative codon usage bias suggesting the usage of less-abundant codons for the corresponding amino acid. Moreover, synonymous codons with *RSCU* values greater than 1.6 are considered as over-represented codons and less than 0.6 as under-represented codons [17], respectively.

**Table 1**
Five mammalian species with accession number and codon bias indices.

| Mammals | Accession numbers | Length (bp) | *ENC* | *CAI* | GC skew | AT skew |
|---|---|---|---|---|---|---|
| *Homo sapiens* | gb\|M68891 | 1443 | 41 | 0.827 | −0.14 | 0.13 |
| *Mus musculus* | gb\|BC107009 | 1443 | 48 | 0.799 | −0.13 | 0.08 |
| *Sus scrofa* | gi\|47,523,099 | 1443 | 39 | 0.836 | −0.15 | 0.14 |
| *Bos taurus* | gi\|300,797,895 | 1443 | 31 | 0.841 | −0.14 | 0.18 |
| *Rattus norvegicus* | gb\|BC061745 | 1443 | 49 | 0.806 | −0.13 | 0.10 |

### 2.5. Codon adaptation index

Codon adaptation index (*CAI*) is used to the level of gene expression on the basis of extent of bias in coding sequence. The *CAI* value was measured as per Sharp and Li (1987) using the following formula:

$$CAI = \exp \frac{1}{L} \sum_{K=1}^{L} 1\text{n}\, w_{c(k)}$$

where, $w_{c(k)}$ is the relative adaptiveness (ω) value for the k-th codon and *L* is the number of codons in the gene [18].

### 2.6. Analysis of selective pressures

The degree of nonsynonymous substitution ($d_N$), synonymous substitution ($d_S$) and the ratio between them ($d_N/d_S$) were estimated as per Nielsen and Yang [19] for the protein coding DNA sequence of *GATA2* gene to investigate the effects of natural selection during the process of evolution.

### 2.7. Neutrality plot

A scatter plot of $GC_{12}$ against $GC_3$, depicts the roles of directional mutational pressure against natural selection. In this plot, regression coefficient of $GC_{12}$ on against $GC_3$ is the equilibrium condition mutation–selection pressure [20].

### 2.8. Software used for statistical analysis

All the above mentioned genetic parameters were estimated in a PERL program developed by SC (corresponding author) to measure the CUB and selection pressure on the selected coding sequences of *GATA2* gene across different mammals. Statistical analyses were carried out using the IBM SPSS version 21.0. Cluster analysis (Heat map) was performed using NetWalker software version 1.0 [21]. The genetic distance and phylogenetic analysis were performed using Mega 6.0 software [22]. No adjustment was done in the coding sequences of gene for comparisons.

## 3. Results

### 3.1. Nucleotide composition in GATA2

The overall nucleotide compositions in the complete coding sequences of *GATA2* gene across five mammalian species were analyzed (Table 2). The highest mean value of base C was observed among all the coding sequences of *GATA2* gene followed by G, A and T across the selected mammals. The percentage of overall GC (Mean ± SD) (65.2 ± 3.35) and AT (34.8 ± 3.35) content values for the coding sequences of *GATA2* showed a wide distribution of GC contents among the five mammals. In addition, we compared the values of nucleotide composition at the third codon position ($A_3$, $T_3$, $G_3$, $C_3$) of codon and observed that mean value of $C_3$ was the highest among the coding sequences of *GATA2* gene. The average percentages of GC contents at the third codon position $GC_3$ (77.2 ± 9.55) and $AT_3$ (22.8.0 ± 9.55) for *GATA2* revealed that $GC_3$ was higher than $AT_3$ across the selected mammals. Similarly, the overall mean percentage of GC contents at the first and second positions ($GC_1 + GC_2$) of codon for the coding sequences of *GATA2* (59.1 ± 0.37) varies significantly. Moreover, negative GC skew was observed in all the coding sequences (Table 1), suggesting the abundance of C over G [23]. Our analysis suggested that the codons ending with G/C base were mostly favored over A/T base in the coding sequences of *GATA2* gene across the five mammalian species.

Please cite this article as: T.H. Mazumder, et al., Transcription factor gene *GATA2*: Association of leukemia and nonsynonymous to the synonymous substitution rate across five mammals, Genomics (2016), http://dx.doi.org/10.1016/j.ygeno.2016.02.001

**Table 2**
Nucleotide composition analysis in the coding sequences of *GATA2* gene.

| *GATA2* gene | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Sl. | A | T | G | C | $A_3$ | $T_3$ | $G_3$ | $C_3$ | AT | GC | $GC_1$ | $GC_2$ | GC3 | $AT_3$ | $GC_{12}$ |
| No. | | | | | | | | | % | % | % | % | % | % | % |
| 1 | 288 | 220 | 400 | 535 | 54 | 58 | 142 | 227 | 35.2 | 64.8 | 60.1 | 57.6 | 76.7 | 23.3 | 58.9 |
| 2 | 291 | 249 | 391 | 512 | 61 | 85 | 134 | 201 | 37.4 | 62.6 | 60.3 | 57.8 | 69.6 | 30.4 | 59.0 |
| 3 | 281 | 212 | 403 | 547 | 48 | 50 | 146 | 237 | 34.2 | 65.8 | 60.7 | 57.2 | 79.6 | 20.4 | 59.0 |
| 4 | 250 | 175 | 437 | 581 | 20 | 19 | 179 | 263 | 29.5 | 70.5 | 61.7 | 58 | 91.9 | 8.1 | 59.9 |
| 5 | 300 | 247 | 388 | 508 | 72 | 82 | 131 | 196 | 37.9 | 62.1 | 60.5 | 57.8 | 68 | 32 | 59.2 |
| Mean | 282 | 221 | 404 | 537 | 51 | 59 | 146 | 225 | 34.8 | 65.2 | 60.7 | 57.7 | 77.2 | 22.8 | 59.1 |
| SD | 19.14 | 30.24 | 19.56 | 29.60 | 19.49 | 26.86 | 19.19 | 27.43 | 3.35 | 3.35 | 0.62 | 0.30 | 9.55 | 9.55 | 0.374 |

SD: standard deviation, $GC_{12}$: average of GC contents at first and second codon positions.

In order to investigate the relationship between the codon usage variation and the compositional constraints, we performed correlation analysis between the values of A, T, G, C and GC with $A_3$, $T_3$, $G_3$, $C_3$ and $GC_3$ values, respectively. Significant positive correlation as well as negative correlation was observed in different nucleotide compositions over all the coding sequences of the selected gene across five mammals. We preliminary inferred that nucleotide constraint under mutation pressure may influence the codon usage pattern in *GATA2* gene.

### 3.2. Codon usage patterns and GATA2 gene expression

In order to find out the relationship of the codon usage variation with GC constraints among the selected coding sequences of *GATA2* gene across five different mammalian species, we analyzed the correlation coefficients of codon usage with $GC_{3s}$ using heat map (Fig. 1). We observed that nearly all codons with G/C — ending base in the coding sequences of *GATA2* were positively correlated with $GC_3$ indicating that codon usage had been influenced by the GC bias and vice versa for the A/T — ending base. In addition our analysis revealed that the codon ATT (encoding isoleucine amino acid) was not favored by natural selection in the coding sequence of *GATA2* gene.

Gene expression level was predicted using the codon adaptation index (*CAI*) values [24–25], which ranged from 0.799 to 0.841 with a mean value of 0.822 and a standard deviation of 0.018. A highly significant positive correlation was observed between *CAI* and $GC_{3s}$ ($r = 0.896$, $p < 0.05$) as well as between *CAI* and GC ($r = 0.873$, $p < 0.05$) contents. But significant negative correlation ($r = -0.934$, $p < 0.05$) was observed between *CAI* and *ENC*. These results indicated that the gene expression level might play a role in shaping the codon usage pattern of *GATA2* gene and that the extent of codon usage bias raises with the level of gene expression.

Besides this, the correlation coefficient between codon usage and *CAI* showed that almost all G/C-ending codons are positively correlated with *CAI* suggesting that gene expression increases with the increase in usage of these G/C-ending codons.

### 3.3. Relative synonymous codon usage in GATA2

The relative synonymous codon usage values of 59 codons in the selected coding sequences of each *GATA2* gene across five mammals were analyzed excluding the codon ATG and TGG that encode amino acid methionine and tryptophan respectively. In our analysis, the overall *RSCU* values showed that 21 codons were most predominantly used among the 59 codons and the most recurrently used codons (*RSCU* > 1) were C-ending [14] compared to G-ending [7] for the *GATA2* gene. Our results further suggested that C ending codon was mostly favored as compared to G-ending codon in the coding sequences of the *GATA2* gene across the selected mammals.

Moreover, clustering analysis using heat map of *RSCU* values (Fig. 2) depicted that the codons GTG, ATC, CTG, TTC, AAG, CAG, GCC, TAC, CCC, GAC, GGC, TCC, CGG, and ACC were the over represented codons (*RSCU* > 1.6) across the selected mammals. The codon ATT encoding isoleucine amino acid showed the *RSCU* value zero because nature might have disfavored this codon in *GATA2* gene across the mammals.

### 3.4. Amino acid usage influencing codon bias

The frequency of amino acid usage in *GATA2* protein across mammals (Fig. 3) was analyzed. Our results showed that, four amino acids, namely alanine (A), glycine (G), proline (P) and serine (S) were more widely used and two amino acids isoleucine (I) and tryptophan (W) were least used. We performed the multiple amino acid sequence

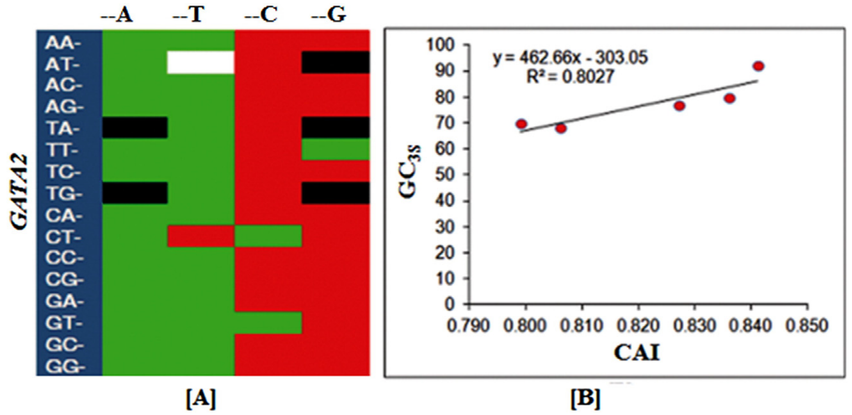


**Fig. 1.** Correlation coefficient between codon usage, $GC_3$ and *CAI* for GATA2 gene across mammals; [A] Heat maps of the correlation coefficients between codon usage and $GC_{3s}$. Red color coding represents the positive correlation, green as negative correlation, black fields are non-degenerate codons (ATG, TGG) and three termination codons (TAA, TAG, TGA), white color field is the codon (ATT) selected against by nature; [B] Correlation coefficient between *CAI* (a measure of gene expression) and $GC_{3s}$. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Please cite this article as: T.H. Mazumder, et al., Transcription factor gene *GATA2*: Association of leukemia and nonsynonymous to the synonymous substitution rate across five mammals, Genomics (2016), http://dx.doi.org/10.1016/j.ygeno.2016.02.001

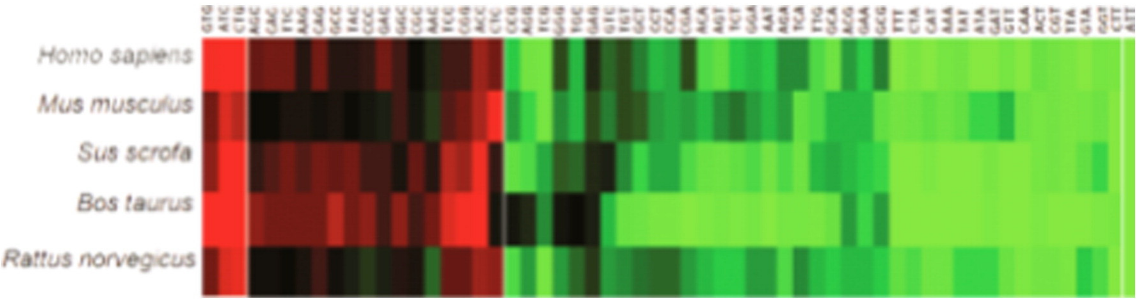


**Fig. 2.** Cluster analysis of *RSCU* values of *GATA2* gene across mammals. The heat map represents the *RSCU* value of a codon (shown in columns) corresponding to the *GATA2* gene across mammals (shown in rows). Green color indicates *RSCU* < 1, dark red *RSCU* > 1 and distinct red *RSCU* > 1.6. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

alignment of 480 amino acid residues in the *GATA2* protein (Fig. 4) in all the selected mammals. The results of our analysis showed that the amino acids at position 9, 21, 38, 39, 219, 226, 236 and 403 of protein i.e. G (glycine), D (aspartic acid), H (histidine), V (valine), T (threonine), S (serine), T (threonine) and N (asparagine) respectively in *GATA2* protein in human radically changed in comparison to other selected mammals during the process of evolution.

The solubility of *GATA2* protein across mammals was assessed through Gravy score [26]. The negative Gravy score was found in all the members, indicating that the protein is water soluble, which reflects its biological function as substrate transporter.

### 3.5. Selection pressure on the protein-coding DNA sequence of GATA2

The mean rate of synonymous substitution per synonymous site ($d_S$) for *GATA2* was higher in *H. sapiens*, *M. musculus*, *S. scrofa* and *B. taurus* (Table 3) but these data showed no statistically significant difference between the groups. But the rate of nonsynonymous substitution per site ($d_N$) for *GATA2* was also higher in *H. sapiens*, *M. musculus* and *B. taurus*, but relatively low in *S. scrofa* and *R. norvegicus*. However, all the nonsynonymous substitutions between the groups showed strong, statistically significant differences ($p < 0.001$). The $d_N/d_S$ ratio in the coding sequences of *GATA2* gene varied across mammals with a mean value of 0.106 and was lower than 0.5. Significant difference in nonsynonymous substitution rate indicates a divergent evolution in the mammals for *GATA2* gene.

### 3.6. Nucleotide distance on nonsynonymous substitution and phylogenetic analysis

We compared the values of non synonymous substitution per site ($d_N$) with the mean genetic p-distance in the coding sequence of *GATA2* gene across mammals (Fig. 5A). Our results showed that, the rate of deleterious nonsynonymous substitution rises remarkably with p-distance in *M. musculus*, *B. taurus*, *S. scrofa* and *R. norvegicus* when compared with *Homo sapiens.* It is evident from the figure (Fig. 5A) that the p-distance between any two mammals increases with the increase in nonsynonymous substitution.

A neighbor-joining tree based on nonsynonymous substitution ($d_N$) in the coding sequences of *GATA2* gene across mammals was constructed (Fig. 5B). There was a close relationship between the rate of nonsynonymous substitution of *GATA2* gene in *M. musculus* and *R. norvegicus* but distinctly different from *H. sapiens.*

### 3.7. Natural selection influences the codon bias of GATA2

A neutrality plot was constructed to quantify the extent of directional mutational pressure against selection in the codon usage bias in *GATA2* gene across the selected species (Fig. 6). In neutrality plots, when there exists a significant correlation between $GC_{12}$ and $GC_3$ and the slope of the regression line is close to 1, indicating that mutation bias supposed to be the main force in shaping the codon usage. Conversely, a lack of correlation between $GC_{12}$ and $GC_3$ indicates selection against mutation bias which results a narrow distribution of GC content [20]. In our analysis we compared the values of $GC_{12}$ and $GC_3$, and observed a positive correlation but not significant in the coding sequences of *GATA2* gene across the selected mammals. Moreover, the regression coefficient of $GC_{12}$ to $GC_3$ of *GATA2* is 0.029, indicating the relative neutrality is 2.9% while the relative constraint is 0.971 for $GC_3$ which suggest mutation pressure played a minor role while natural selection played a major role in codon usage pattern in *GATA2* genes.

## 4. Discussion

Several studies earlier reported that codon usage in mammals including human has been influenced by the variation of GC contents under mutation pressure. Moreover, the selection on codon bias is weak for nearly neutral synonymous mutations [27]. The mean *ENC* value in the coding sequences of *GATA2* gene was 41.60 ± 7.33, representing existence of relatively weak codon bias. The overall

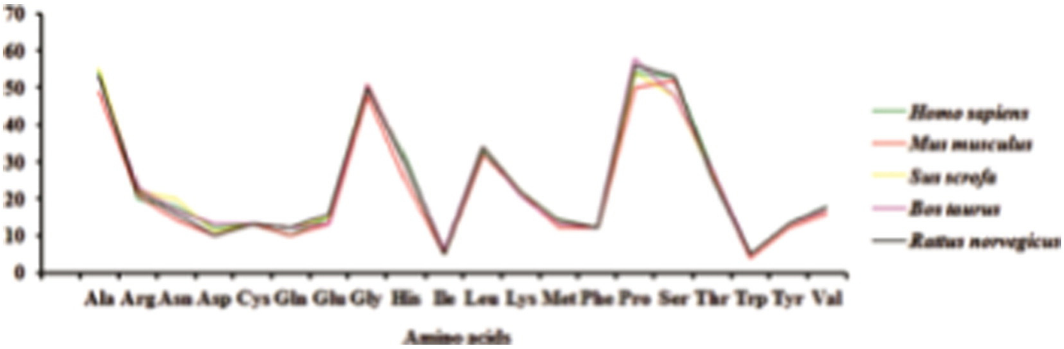


**Fig. 3.** Frequency of amino acid usage in *GATA2* gene across mammals; Five adjacent color-bars in a group representing five mammals indicate the usage of a particular amino acid with a small vertical line at the top of each bar as standard error. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

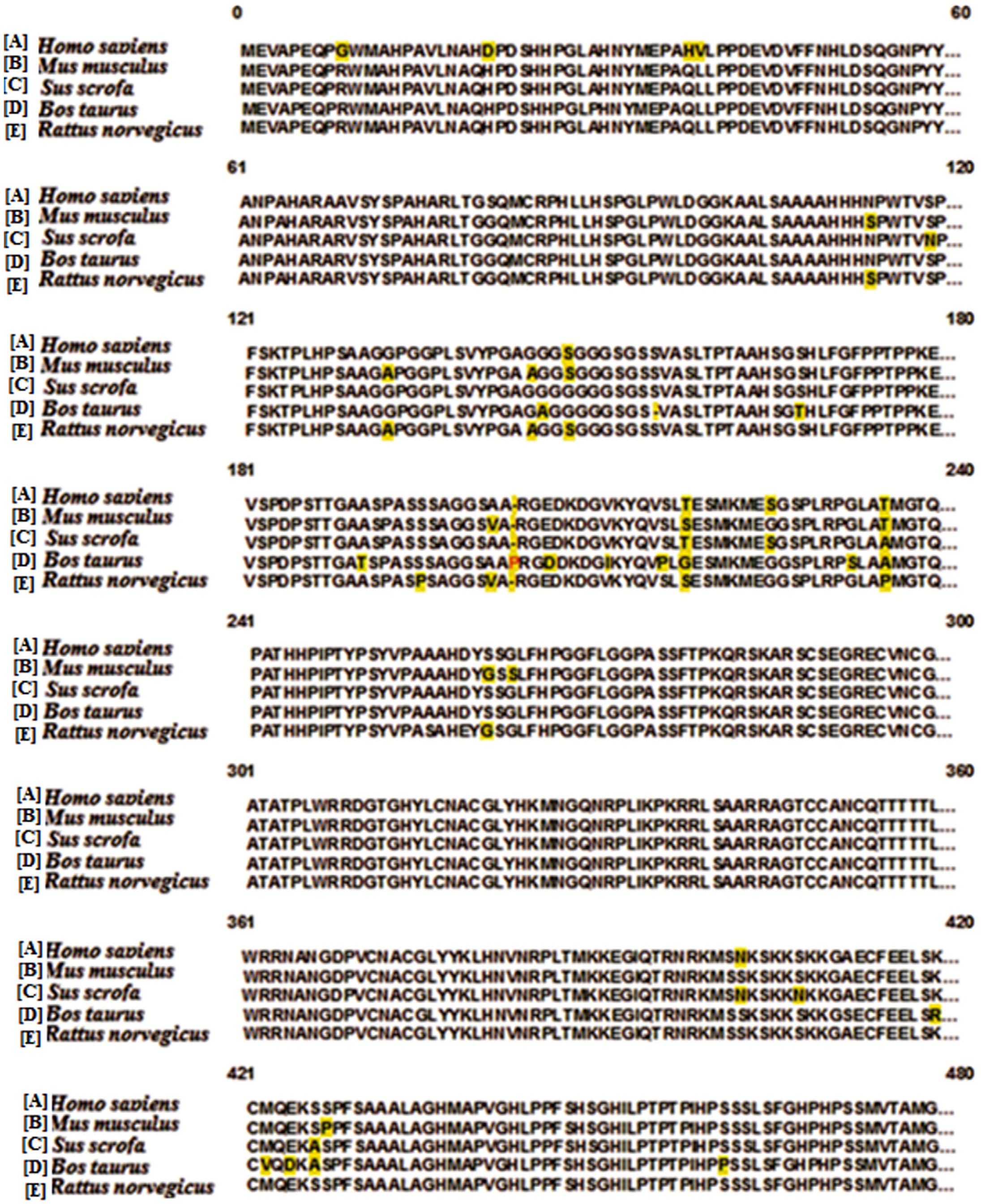


**Fig. 4.** Multiple sequence alignment of GATA2 protein for residues 0–480; Unique residues are highlighted at different positions of the complete amino acid sequence of the GATA2 protein across selected mammals where [A] represents *Homo sapiens*, [B] *Mus musculus*, [C] *Sus scrofa*, [D] *Bos taurus* and [E] *Rattus norvegicus*.

**Table 3**
Pairwise comparisons between different mammals for the number of substitutions per site summarized for *GATA* gene, with synonymous substitutions (dS) below the diagonal and and nonsynonymous substitutions (dN) above the diagonal in bold.

| | *Homo sapiens* | *Mus musculus* | *Sus scrofa* | *Bos taurus* | *Rattus norvegicus* |
|---|---|---|---|---|---|
| *Homo sapiens* | – | **0.022*** | **0.008** | **0.035*** | **0.006** |
| *Mus musculus* | 0.358* | – | **0.004** | **0.018** | **0.001** |
| *Sus scrofa* | 0.136* | 0.077 | – | **0.012** | **0.002** |
| *Bos taurus* | 0.109* | 0.067 | 0.035 | – | **0.011** |
| *Rattus norvegicus* | 0.085 | 0.020 | 0.041 | 0.040 | – |

* $p < 0.001$, bold: nonsynonymous substitution (dN)

nucleotide composition analysis in the complete coding sequences of *GATA2* gene across five mammals revealed that the GC content was higher than AT content and the codons ending with G/C base was mostly favored over A/T-ending base. In addition, significant correlation was observed between different nucleotide compositions, suggesting that nucleotide bias particularly GC constraint under mutation pressure might affect the codon usage patterns of *GATA2* gene.

We performed a heat map analysis of the correlation coefficients of codon usage with $GC_{3s}$ and our results revealed that codon usage patterns of *GATA2* gene have been influenced by GC bias. The cordon ATT (encoding Isoleucine amino acid) was not favored by natural selection in the coding sequence of *GATA2* gene. Gene expression level was measured using *CAI*. A significant positive correlation was observed between *CAI* and GC as well as between *CAI* and $GC_3$, but a significant negative

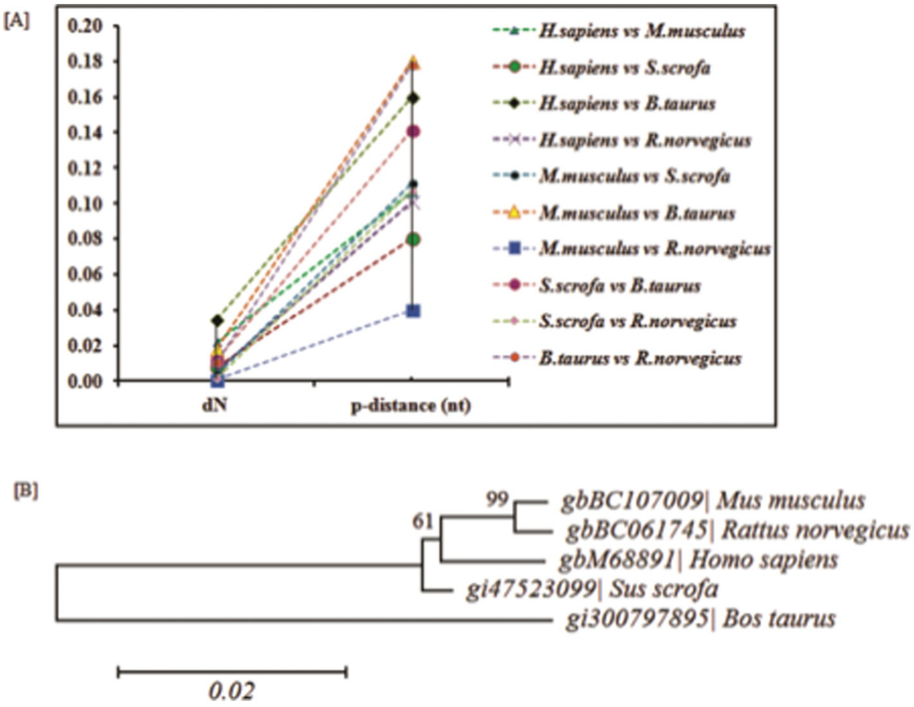


**Fig. 5.** Genetic variability of nonsynonymous mutation and phylogenetic tree; [A] Distribution of nonsynonymous mutation per site ($d_N$) and mean genetic p-distance in the coding sequence of *GATA2* gene across mammals. [B] Neighbor-Joining tree using $d_N$ distance based on codon alignment. The percentage of replicate trees in which the associated taxa clustered together in the bootstrap test (1000 replicates) is shown above the branches. The tree is drawn to scale, with branch lengths in the same units as those of the evolutionary distances used to infer the phylogenetic tree. The evolutionary distances were computed using the Nei-Gojobori method and are in the units of the number of nonsynonymous substitutions per nonsynonymous site. The analysis involved 5 coding sequences. All the positions containing gaps and missing data were eliminated. A total of 480 positions were present in the final dataset. Evolutionary analyses were conducted in MEGA6 [22].

correlation was found between *CAI* and *ENC*. Our results indicated that the gene expression level might play a role in codon usage patterns of *GATA2* gene.

The overall relative codon usage frequency of *GATA2* gene across mammals revealed that the C-ending codons were mostly favored in comparison to G-ending codons. Similar findings were reported earlier for different genes in mammalian species [28–29]. Moreover, cluster analysis of *RSCU* values (Fig. 2) indicated that fourteen codons were over represented (*RSCU* > 1.6) in the coding sequence of *GATA2* gene across mammals.

It was reported earlier that nucleotide bias may affect the amino acid composition of proteins [30–31] and that convergent amino acid composition may influence the protein sequences in the construction of phylogenetic trees [32]. The biasness in the nucleotide or amino acid composition may affect the evolution of protein structure because of the relationship between primary and secondary protein structure [33]. The change in charge distribution within a protein might be the outcome of amino acid bias. Such bias can alter the protein's secondary and tertiary structures. Moreover, these proteins may undergo positive selection at other positions in a protein to balance the nucleotide induced bias in the coding sequence of the gene encoding the protein. In general, amino acid substitutions in a protein are most commonly deleterious to the organism. But a few of these amino acid substitutions are neutral and hence do not affect the protein function much. These neutral amino acid substitutions become gradually adapted in the organism over time.

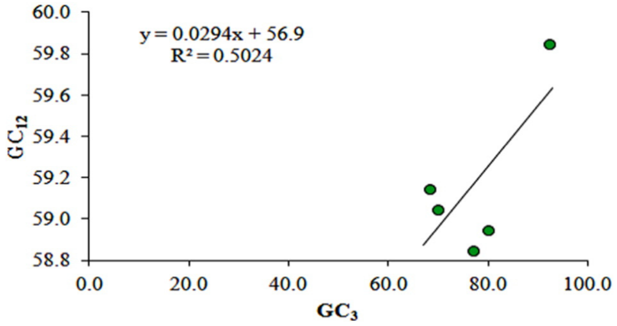


**Fig. 6.** Neutrality plots of *GATA-2* gene across mammalian species. Individual genes of different species are plotted based on the average GC content in the first and second codon position versus the GC content of the third codon position ($GC_3$).

The usage of amino acid frequency for *GATA2* across mammals (Fig. 3) revealed that four amino acids namely alanine (A), glycine (G), proline (P) and serine (S) were mostly used. Conversely least usage of two amino acids isoleucine (I) and tryptophan (W) was noted. In addition, the multiple sequence alignment of the amino acid residues showed that the amino acids namely glycine (G), aspartic acid (D), histidine (H), valine (V), threonine (T), serine (S), and asparagine (N) changed at different positions of human *GATA2* protein during the period of evolution when compared with other mammals. Therefore, in order to find out the selection pressure in the protein coding DNA sequence of *GATA2*, we estimated the values of nonsynonymous substitution ($d_N$) and synonymous substitution ($d_S$) per site as per Nielsen and Yang (2003). The ratio of nonsynonymous and synonymous substitutions ($d_N/d_S$) on gene sequence is a widely used measure for investigating the extent to which the natural selection has affected the gene during the process of evolution [34]. When the ratio of $d_N/d_S$ is greater than unity, it suggests that natural selection endorses alteration in protein sequences and the ratio less than unity is expected when natural selection suppresses protein changes [35]. In our analysis, we observed that the mean ratio of nonsynonymous substitution to synonymous substitution ($d_N/d_S$) was lower than 0.5, suggesting that the coding sequence of *GATA2* has undergone purifying selection to maintain its protein function. Besides this, a neighbor-joining tree using nonsynonymous substitution ($d_N$) distance based on codon alignment for *GATA2* gene revealed that there was a close relationship between nonsynonymous substitution rate in *M. musculus* and *R. norvegicus* but distinctly different from *H. sapiens*.

## 5. Conclusions

A majority of the frequently used codons was C-ending in the coding sequence of *GATA2* gene across mammals. Fourteen codons were mostly overrepresented and the codon ATT encoding isoleucine amino acid was selected against by nature in *GATA2* gene of all the mammals. The codon usage of *GATA2* gene was primarily affected by GC mutation bias and the gene expression level might play a pivotal role in shaping its codon usage patterns. The magnitude of $d_N/d_S$ ratio suggested that *GATA2* gene in different mammals was influenced by purifying natural selection in order to maintain its functionality. Different rates of amino acid changing mutations might be conservative mutations for maintaining the protein function and these differ in the level of selective constraint. Such mutations ultimately affect the rate of evolution across distant species. Our present findings certainly report a novel insight into the codon usage patterns in gaining the clues for codon optimization to alter the translational efficiency as well as for the functional conservation of gene expression and the significance of nucleotide composition in the evolution of *GATA2* gene within mammals.

## Conflict of interest

There is no conflict of interest in this research work.

## Acknowledgments

We are thankful to Assam University, Silchar, Assam, India, for providing the necessary facilities in carrying out this research work.

## References

[1] C.N. Hahn, C.E. Chong, C.L. Carmichael, E.J. Wilkins, P.J. Brautigan, X.C. Li, et al., Heritable GATA2 mutations associated with familial myelodysplastic syndrome and acute myeloid leukemia, Nat. Genet. 43 (10) (Oct 2011) 1012–1017.

[2] R.E. Dickinson, H. Griffin, V. Bigley, L.N. Reynard, R. Hussain, M. Haniffa, et al., Exome sequencing identifies GATA-2 mutation as the cause of dendritic cell, monocyte, B and NK lymphoid deficiency, Blood 118 (10) (Sep 8 2011) 2656–2658.

[3] S.J. Zhang, L.Y. Ma, Q.H. Huang, G. Li, B.W. Gu, X.D. Gao, et al., Gain-of-function mutation of GATA-2 in acute myeloid transformation of chronic myeloid leukemia, Proc. Natl. Acad. Sci. U. S. A. 105 (6) (Feb 12 2008) 2076–2081.

[4] P.A. Greif, A. Dufour, N.P. Konstandin, B. Ksienzyk, E. Zellmeier, B. Tizazu, et al., GATA2 zinc finger 1 mutations associated with biallelic CEBPA mutations define a unique genetic entity of acute myeloid leukemia, Blood 120 (2) (Jul 12 2012) 395–403.

[5] M. Nirenberg, P. Leder, M. Bernfield, R. Brimacombe, J. Trupin, F. Rottman, et al., RNA codewords and protein synthesis, VII. On the general nature of the RNA code, Proc. Natl. Acad. Sci. U. S. A. 53 (5) (May 1965) 1161–1168.

[6] R. Grantham, C. Gautier, M. Gouy, M. Jacobzone, R. Mercier, Codon catalog usage is a genome strategy modulated for gene expressivity, Nucleic Acids Res. 9 (1) (Jan 10 1981) r43–r74.

[7] Y. Prat, M. Fromer, N. Linial, M. Linial, Codon usage is associated with the evolutionary age of genes in metazoan genomes, BMC Evol. Biol. 9 (2009) 285.

[8] S.K. Behura, D.W. Severson, Comparative analysis of codon usage bias and codon context patterns between dipteran and hymenopteran sequenced genomes, PLoS One 7 (8) (2012), e43111.

[9] R. Grantham, C. Gautier, M. Gouy, R. Mercier, A. Pave, Codon catalog usage and the genome hypothesis, Nucleic Acids Res. 8 (1) (Jan 11 1980) r49–r62.

[10] S. Aota, T. Gojobori, F. Ishibashi, T. Maruyama, T. Ikemura, Codon usage tabulated from the GenBank genetic sequence data, Nucleic Acids Res. 16 (Suppl) (1988) r315–r402.

[11] L. Duret, D. Mouchiroud, Expression pattern and, surprisingly, gene length shape codon usage in *Caenorhabditis*, *Drosophila*, and *Arabidopsis*, Proc. Natl. Acad. Sci. U. S. A. 96 (8) (Apr 13 1999) 4482–4487.

[12] W.H. Li, Models of nearly neutral mutations with particular implications for nonrandom usage of synonymous codons, J. Mol. Evol. 24 (4) (1987) 337–345.

[13] A. Fasan, C. Eder, C. Haferlach, V. Grossmann, A. Kohlmann, F. Dicker, et al., GATA2 mutations are frequent in intermediate-risk karyotype AML with biallelic CEBPA mutations and are associated with favorable prognosis, Leukemia 27 (2) (Feb 2013) 482–485.

[14] H.A. Hou, Y.C. Lin, Y.Y. Kuo, W.C. Chou, C.C. Lin, C.Y. Liu, et al., GATA2 mutations in patients with acute myeloid leukemia-paired samples analyses show that the mutation is unstable during disease evolution, Ann. Hematol. 94 (2) (Feb 2015) 211–221.

[15] F. Wright, The "effective number of codons" used in a gene, Gene 87 (1) (Mar 1 1990) 23–29.

[16] J.M. Comeron, M. Aguade, An evaluation of measures of synonymous codon usage bias, J. Mol. Evol. 47 (3) (Sep 1998) 268–274.

[17] P.M. Sharp, W.H. Li, An evolutionary perspective on synonymous codon usage in unicellular organisms, J. Mol. Evol. 24 (1–2) (1986) 28–38.

[18] P.M. Sharp, W.H. Li, The codon adaptation index—a measure of directional synonymous codon usage bias, and its potential applications, Nucleic Acids Res. 15 (3) (Feb 11 1987) 1281–1295.

[19] R. Nielsen, Z. Yang, Estimating the distribution of selection coefficients from phylogenetic data with applications to mitochondrial and viral DNA, Mol. Biol. Evol. 20 (8) (Aug 2003) 1231–1239.

[20] N. Sueoka, Directional mutation pressure and neutral molecular evolution, Proc. Natl. Acad. Sci. 85 (8) (1988) 2653–2657.

[21] K. Komurov, S. Dursun, S. Erdin, P.T. Ram, NetWalker: a contextual network analysis tool for functional genomics, BMC Genomics 13 (2012) 282.

[22] K. Tamura, G. Stecher, D. Peterson, A. Filipski, S. Kumar, MEGA6: molecular evolutionary genetics analysis version 6.0. Mol. Biol. Evol. 30 (12) (Dec 2013) 2725–2729.

[23] E.R. Tillier, R.A. Collins, The contributions of replication orientation, gene direction, and signal sequences to base-composition asymmetries in bacterial genomes, J. Mol. Evol. 50 (3) (Mar 2000) 249–257.

[24] S.K. Gupta, T.K. Bhattacharyya, T.C. Ghosh, Synonymous codon usage in *Lactococcus lactis*: mutational bias versus translational selection, J. Biomol. Struct. Dyn. 21 (4) (Feb 2004) 527–536.

[25] H. Naya, H. Romero, N. Carels, A. Zavala, H. Musto, Translational selection shapes codon usage in the GC-rich genome of *Chlamydomonas reinhardtii*, FEBS Lett. 501 (2–3) (Jul 20 2001) 127–130.

[26] J. Kyte, R.F. Doolittle, A simple method for displaying the hydropathic character of a protein, J. Mol. Biol. 157 (1) (May 5 1982) 105–132.

[27] Z. Yang, R. Nielsen, Mutation-selection models of codon substitution and their use to estimate selective strengths on codon usage, Mol. Biol. Evol. 25 (3) (Mar 2008) 568–579.

[28] J.F. Dass, C. Sudandiradoss, Insight into pattern of codon biasness and nucleotide base usage in serotonin receptor gene family from different mammalian species, Gene 503 (1) (Jul 15 2012) 92–100.

[29] T.H. Mazumder, S. Chakraborty, Gaining insights into the codon usage patterns of TP53 Gene across eight mammalian species, PLoS One 10 (3) (2015), e0121709.

[30] G.A. Singer, D.A. Hickey, Nucleotide bias causes a genomewide bias in the amino acid composition of proteins, Mol. Biol. Evol. 17 (11) (Nov 2000) 1581–1588.

[31] X. Gu, D. Hewett-Emmett, W.H. Li, Directional mutational pressure affects the amino acid composition and hydrophobicity of proteins in bacteria, Genetica 102–103 (1–6) (1998) 383–391.

[32] P.G. Foster, D.A. Hickey, Compositional bias may affect both DNA-based and protein-based phylogenetic reconstructions, J. Mol. Evol. 48 (3) (Mar 1999) 284–290.

[33] T.C. Wood, W.R. Pearson, Evolution of protein sequences and structures, J. Mol. Biol. 291 (4) (Aug 27 1999) 977–995.

[34] S. Kryazhimskiy, J.B. Plotkin, The population genetics of dN/dS, PLoS Genet. 4 (12) (Dec 2008), e1000304.

[35] Z. Yang, J.P. Bielawski, Statistical methods for detecting molecular adaptation, Trends Ecol. Evol. 15 (12) (Dec 1 2000) 496–503.

Please cite this article as: T.H. Mazumder, et al., Transcription factor gene *GATA2*: Association of leukemia and nonsynonymous to the synonymous substitution rate across five mammals, Genomics (2016), http://dx.doi.org/10.1016/j.ygeno.2016.02.001

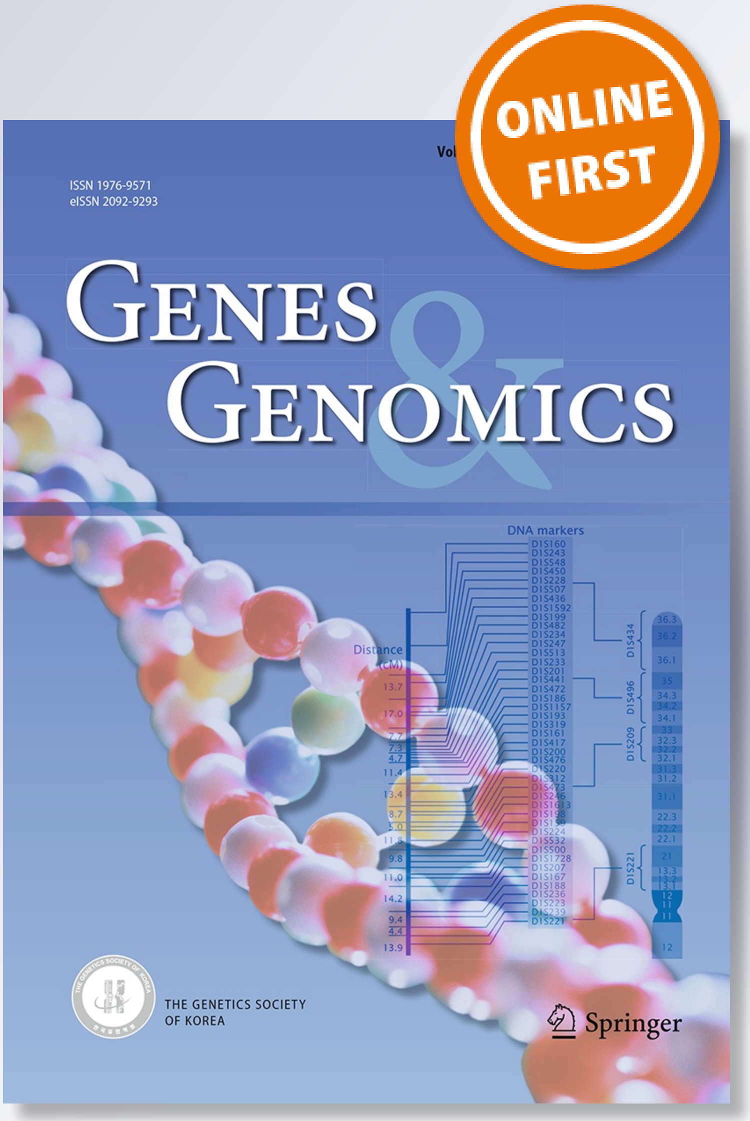
# *Prediction of gene expression and codon usage in human parasitic helminths*

**Gulshana A. Mazumder, Arif Uddin & Supriyo Chakraborty**

**Genes & Genomics**

ISSN 1976-9571

Genes Genom
DOI 10.1007/s13258-015-0381-3



Springer

Genes Genom
DOI 10.1007/s13258-015-0381-3

RESEARCH ARTICLE

# Prediction of gene expression and codon usage in human parasitic helminths

**Gulshana A. Mazumder[1] · Arif Uddin[1] · Supriyo Chakraborty[1]**

Received: 1 September 2015 / Accepted: 14 December 2015

**Abstract** Codon usage bias refers to the differences in the occurrence frequency of synonymous codons. To understand the patterns of codon usage in mitochondrial genes we used bioinformatic approaches to analyze the protein coding sequences of *W. bancrofti* and *S. haematobium* as no work was reported earlier. It was found that the ENC value ranged from 43 to 60 with a mean of 46.91 in *W. bancrofti* but varied from 49 to 60 with a mean of 45.17 in *S. haematobium*, respectively. In *W. bancrofti* a significant positive correlation was found between ENC and GC3% ($r = 0.826^{**}$, $p < 0.01$), but in *S. haematobium* significant correlation was found between ENC and GC3% ($r = 0.983^{**}$, $p < 0.01$). Principal component analysis suggests that the pattern of codon usage significantly differed between *W. bancrofti* and *S. haematobium*. Neutrality plot reveals that natural selection played a major role while mutation pressure played a minor role in codon usage pattern in the mitochondrial protein coding genes of *W. bancrofti* and *S. haematobium*. Various factors namely nucleotide composition, natural selection and mutation pressure affected the codon usage pattern.

**Keywords** Codon bias · Natural selection · Mutation pressure · Mitochondrial DNA

**Electronic supplementary material** The online version of this article (doi:10.1007/s13258-015-0381-3) contains supplementary material, which is available to authorized users.

✉ Supriyo Chakraborty
supriyoch_2008@rediffmail.com

[1] Department of Biotechnology, Assam University, Silchar, Assam 788011, India

## Introduction

Synonymous codon usage bias (SCUB) is the different frequency of synonymous codons encoding an amino acid in the coding DNA. In different organisms for protein expression the triplets coding for the same amino acid are not equally used. The bias ensures that the codons which are used more frequently, the optimal codons, can pair with the anticodons of the most abundant tRNA genes (Sun et al. 2009). It reveals a balance between natural selection (e.g. translational selection, gene length, and gene function) and mutation bias (such as GC content and mutation position of base) (Bulmer 1991; Sharp and Li 1986a). There are various factors related to SCUB, determined alone by mutational bias or by both mutation bias and natural selection as suggested by various other studies (Behura and Severson 2013). But in many mammals the bias has been proved to be mutational bias (Francino and Ochman 1999), some others suggested that natural selection may also determine the bias in eukaryotic organisms (Ingvarsson 2007; Powell and Moriyama 1997; Wang and Hickey 2007). Thus by understanding the codon usage bias we can show the pattern of codon usage in species, and it also further provides evidence about the evolution of organisms.

Mitochondrial DNA (mtDNA) is used as a marker in studying the molecular diversity in animals. It is found that amplification of the mtDNA occurs more easily because in the cell it occurs in a huge number of copies (Gissi et al. 2008). The mtDNA is used as a tool for tracing the ancestry of species because of its high mutational rate. Biologists compare the mtDNA sequences from different species and a relationship is developed from the data collected, which further provides the relationships among the species from which the mtDNAs were taken. If the species are related distantly, then the number of differences in sequence

Published online: 28 December 2015

becomes very large (Taylor and Turnbull 2005). MT-COX1 (mitochondrial fragment) was selected as a molecular tool in the molecular taxonomy and identification (Stock 2009). Moreover, mtDNA molecules are known to be located very close to the electron transport chain (ETC), where reactive oxygen species (ROS) is continuously generated (Richter et al. 1988).

The species analyzed in this study assume importance because they are relevant as vectors in the transmission of various animal diseases. Both species studied here are medically very important. About 90 % of lymphatic filariasis is caused by *Wuchereria bancrofti* (Melrose 2002), whereas a chronic disease known as schistosomiasis (also known as bilharziasis) is caused by *Schistosoma haematobium*. *W. bancrofti* affects over 120 million people worldwide (Melrose 2002). Infection caused by *W. bancrofti* is usually asymptomatic, this can be developed in the legs, arms, breasts and genitalia. As the time passes, thickening as well as hardening of the skin occurs which is referred to as elephantiasis, also there might be high levels of IgE (Immunoglobulin E) and antifilarial antibodies. Individuals of all ages are susceptible to this infection, whereas the humans are known to be the only host (Manguin et al. 2010; Bockarie et al. 2009). Treatment with 1 % sodium hypochlorite and 2 % glutaraldehyde mostly seem to be susceptible to *W. bancrofti*. For diagnosis, blood should be collected at night and a thick smear is stained with Giemsa or hematoxylin and eosin. By the use of therapeutic drugs, such as ivermectin, diethylcarbamazine (DEC), or albendazole the severe symptoms caused by this parasite can be avoided. On the other hand, many million people worldwide are affected by the parasitic disease, schistosomiasis. It is found to be endemic in some countries. Several surveys state that in Africa, urogenital schistosomiasis is due to the infection with *S. haematobium* (Chen et al. 2014). In human it is caused by the penetration of infectious larvae (cercariae) through the skin of people. Then in the body, adult schistosomes are developed from the larvae. For acute schistosomiasis (Katayama syndrome), the incubation period is found to be 14–84 days and characterized by headache, myalgia, diarrhea, fever and respiratory symptoms, but the chronic infection is thought to remain asymptomatic for many years (Chen et al. 2014). For the standard diagnostics of urogenital schistosomiasis, a filtration technique is typically used. Praziquantel is considered to be the most commonly used drug for the treatment of schistosomiasis and is typically effective against adult forms of the parasite (Chen et al. 2014).

Analysis of codon usage bias is a well-established technique for understanding the protein coding sequences of genomes. The study of codon bias is gaining renewed attention of scientists across the globe with the advent of whole genome sequencing of numerous organisms (Sharp and Li 1986b). In this study, one of the major objectives is to understand the patterns of codon usage bias among the protein coding genes of mitochondria in two parasitic species: *W. bancrofti* and *S. haematobium.* The goal of this study is to perform a comparative analysis of codon usage bias pattern among the sequenced genes of these two parasitic species belonging to two different phyla. Our results provide useful insights on the patterns of codon usage bias that facilitate better understanding of the structure and evolution of gene coding sequences of these species. On nuclear genomes various reports on synonymous codon usage bias have been focused, but only a few mitochondrial genes have been analyzed for codon usage bias. Mainly in vertebrates mitochondrial codon usage has been studied, while among invertebrates so far only some parasitic platyhelminthes had been surveyed for mitochondrial codon usage and bias (Sharp and Matassi 1994). Molecular evolutionary investigations suggest that codon usage bias varies both within and between genomes and may have a major significance in understanding genome evolution among related species at molecular level (Plotkin and Kudla 2011).

## Materials and methodology

### Sequences data

The coding sequences of mitochondrial genes from two species namely *W. bancrofti* and *S. haematobium* were retrieved from the Nucleotide Database of NCBI (http://wwwncbi.nlm.nih.gov/GeneBank). The species with their accession numbers are *W. bancrofti*- JQ316200.1 and *S. haematobium*- NC_008074.1.

### Estimation of DNA compositional properties

Using a in-house Perl programme developed by SC (corresponding author) the nucleotide compositions (A%, T%, G% and C%), nucleotide composition at the 3rd position, and the overall GC contents, GC1, GC2, GC3, AT1, AT2 and AT3 were estimated.

### Relative synonymous codon usage (RSCU)

The RSCU values of codons were estimated as

$$RSCU_{ij} = \frac{X_{ij}}{\frac{1}{ni}\sum_{j=1}^{ni} X_{ij}},$$

where $X_{ij}$ is the frequency of occurrence of the jth codon for ith amino acid (any $X_{ij}$ with a value of zero is arbitrarily assigned a value of 0.5) and $n_i$ is the number of codons for the ith amino acid (ith codon family).

If RSCU value of a codon >1, then the codon is used more frequently than expected. If RSCU value <1, it means that the codon is used less frequently. If RSCU = 1, it means that the codon is used randomly and equally with other synonymous codons for the same amino acid (Sharp and Li 1986b). If the RSCU value is <0.6, the codon is said to be under-represented and if the RSCU value of a codon is >1.6, the codon is said to be over-represented in the coding sequence (Gupta and Ghosh 2001).

### Effective number of codons (ENC)

ENC provides the range of codon preferences in a gene. Its value ranges from 20 to 61. Equal synonymous codon usage for amino acids is indicated when ENC equals to 61. When only a single codon is used, then ENC value is 20. High ENC value indicates low codon usage bias whereas low ENC indicates high codon usage bias (Wright 1990). It is calculated as:

$$ENC = 2 + \frac{9}{F_2} + \frac{1}{F_3} + \frac{5}{F_4} + \frac{3}{F_6},$$

where Fk (k = 2, 3, 4, 6) is the mean of Fk values for the k-fold degenerate amino acids.

### Codon adaptation index (CAI)

CAI is the most widely used measure for codon usage bias and for gene expression (Sharp and Li 1987). CAI values range from 0 to 1, with higher values indicating a higher percentage of the most abundant codons. The CAI of a coding sequence is calculated as

$$CAI = \exp\left(\frac{1}{L}\sum_{k=1}^{L} \ln \omega k\right),$$

where ωk is the relative adaptiveness of the kth codon and L is the number of synonymous codons in the gene.

### Heat map

The RSCU values of codons from different mitochondrial genes were used to generate heat map using Netwalker 1.0 software.

### Principal component analysis (PCA) and clustering

Principal component analysis was used to investigate the major trend in codon usage among different mitochondrial genes using RSCU value. The analysis was done by SPSS software (Jenkins and Holmes 2003).

### Statistical analysis

Correlation and regression analysis were performed to identify the relationship between the overall nucleotide compositions with the nucleotide compositions at 3rd codon position. In addition to this, correlation analysis of ENC with GC, GC3 and CAI was carried out as well as between GC12 and GC3. All the statistical analyses were done using the SPSS software.

## Results

### Compositional properties

The mitochondrial genes of *Wuchereria bancrofti* and *Schistosoma haematobium* with their gene length, CAI and overall GC (%), GC1 (%), GC2 (%) and GC3 (%) are shown in Tables 1 and 2 respectively. From the table it is seen that the average GC% and GC3 % in *W. bancrofti* were 25.29 ± 3.58 and 23.20 ± 4.36, respectively and in *S. haematobium*, the average overall GC% and GC3% are 26.35 and 20.48, respectively. The overall nucleotide composition and nucleotide composition at the third codon position and ENC values of *W. bancrofti* and *S. haematobium* coding sequences are provided in S1 and 2 (supplementary file). The ENC values ranged from 43 to 60 with a mean of 46.91 in *W. bancrofti* but in *S. haematobium* it ranged from 49 to 60 with a mean of 45.17. Higher ENC value means low codon usage bias. High ENC indicated that the codon usage bias is not very remarkable in these species. In *W. bancrofti*, the nucleobase T at the 3rd position of codons occurred more frequently while in the species *S. haematobium*, the base A or T at the 3rd position

**Table 1** Overall GC (%), GC1 (%), GC2 (%), GC3 (%), CAI, and gene length in *W. bancrofti*

| GC (%) | GC1 (%) | GC2 (%) | GC3 (%) | CAI | Gene length (nt) |
|---|---|---|---|---|---|
| 24.1 | 25.8 | 25.3 | 21.1 | 0.44 | 582 |
| 31.9 | 35.3 | 30.6 | 29.7 | 0.83 | 1638 |
| 29.6 | 36.3 | 31.6 | 20.9 | 0.51 | 702 |
| 25.7 | 28.3 | 24.8 | 24 | 0.6 | 774 |
| 27.2 | 30.3 | 30.6 | 20.7 | 0.64 | 1089 |
| 27.1 | 30.7 | 30.7 | 19.8 | 0.56 | 879 |
| 24.1 | 25.2 | 22.7 | 24.5 | 0.52 | 846 |
| 23.1 | 25.2 | 24.3 | 19.8 | 0.31 | 333 |
| 25.9 | 28.4 | 25.7 | 23.7 | 0.66 | 1227 |
| 19.2 | 17.5 | 10 | 30 | 0.51 | 240 |
| 25.5 | 26.8 | 21.5 | 28.3 | 0.81 | 1590 |
| 20.1 | 17.9 | 26.5 | 15.9 | 0.38 | 453 |

**Table 2** Overall GC (%), GC1 (%), GC2 (%),GC3 (%), CAI, and gene length in *S. haematobium*

| GC (%) | GC1 (%) | GC2 (%) | GC3 (%) | CAI | Gene length (nt) |
|---|---|---|---|---|---|
| 24 | 27.4 | 28 | 16.6 | 0.46 | 525 |
| 29.6 | 35.4 | 36.8 | 16.7 | 0.79 | 1542 |
| 29.5 | 36.7 | 31.2 | 20.6 | 0.56 | 597 |
| 26.4 | 30.6 | 27.5 | 21.2 | 0.58 | 666 |
| 27.9 | 23.6 | 32.6 | 27.4 | 0.76 | 1104 |
| 28.6 | 28.9 | 32.3 | 24.5 | 0.7 | 882 |
| 26.4 | 23.6 | 30.4 | 25.4 | 0.62 | 840 |
| 22.5 | 26 | 25.2 | 16.3 | 0.34 | 369 |
| 26.9 | 29.4 | 33.4 | 18 | 0.77 | 1266 |
| 24.5 | 26.4 | 28.7 | 18.4 | 0.16 | 261 |
| 28.3 | 31.4 | 31.3 | 22.3 | 0.79 | 1584 |
| 21.7 | 20.3 | 26.6 | 18.4 | 0.37 | 474 |

was mostly favored as shown in Fig. 1. The level of gene expression was measured by CAI. In Fig. 2, it is evident that the gene expression level is more in *S. haematobium* than *W. bancrofti*.

## Codon usage in *W. bancrofti* and *S. haematobium*

We performed correlation analysis between codon usage and GC3 to understand the general codon usage variation and GC bias. From Fig. 3a, b it was found that in *S. haematobium*, most of the AT ending codons were negative and most of the GC ending codons were positive which suggest GC ending codons increased with the increase in GC3 bias. But in *W. bancrofti*, most of the AT and GC ending codons were positive which suggest that both AT and GC contents increased with the increase in GC3 bias.

Based on RSCU value, in *W. bancrofti*, 20 codons which were used most frequently were TCT, AGT, TTT, TTA, TTG, TAT, TGT, CCT, CAT, CGT, ATG, ATT, ACT, AAT, GAT, GTT, GCT, AAG, GGT and TGG (Fig. 5). The nucleobase T at the 3rd position occurred more frequently. Total 16 overrepresented codons were identified and these were TCT, AGT, TTT, TTA, TTG, TAT, TGT, CCT, CGT, ATT, ACT, AAT, GAT, GTT, GCT and GGT as shown in Fig. 4a. But in *S. haematobium* 25 codons were used most frequently and these were TCA, TCT, AGT,TTT, TTA, TTG, TAT, TGT, CCA, CCT, CAT, CGT, ATA, ATT, ACT, AAT, GAT, GTA, GTT, GCT, AAA, GAA, GGA, GGT, and TGA. The nucleobase A/T at the 3rd position occurred more frequently. Total 13 overrepresented codons were detected and these were TTT, TTA, TAT, TGT, CCT, CGT, ATT, ACT, AAT, GAT, GTT, GCT, and GGT as shown in Fig. 4b. The composition of overall nucleotide confirms that compositional constraint influenced the codon usage pattern of these species.

**Fig. 1** Nucleotide composition (overall and at 3rd codon position) in *W. bancrofti* and *S. haematobium*

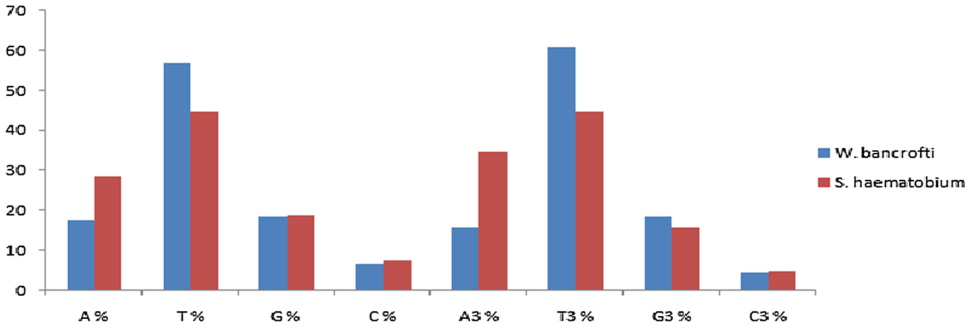
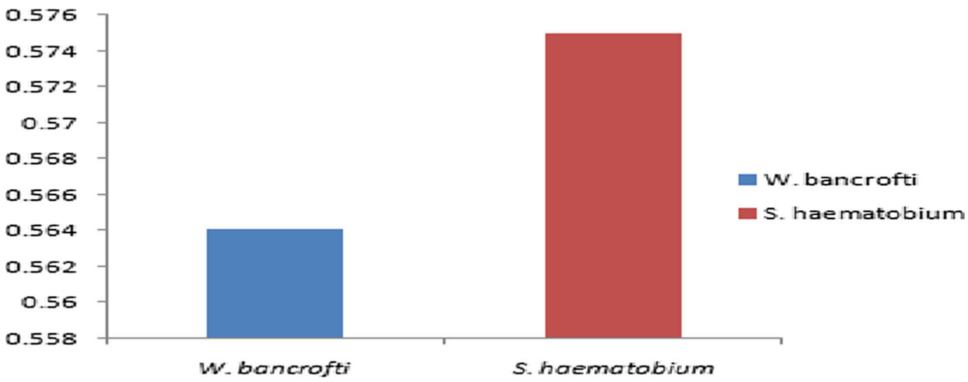


**Fig. 2** Comparison of CAI value for both *W. bancrofti* and *S. haematobium*



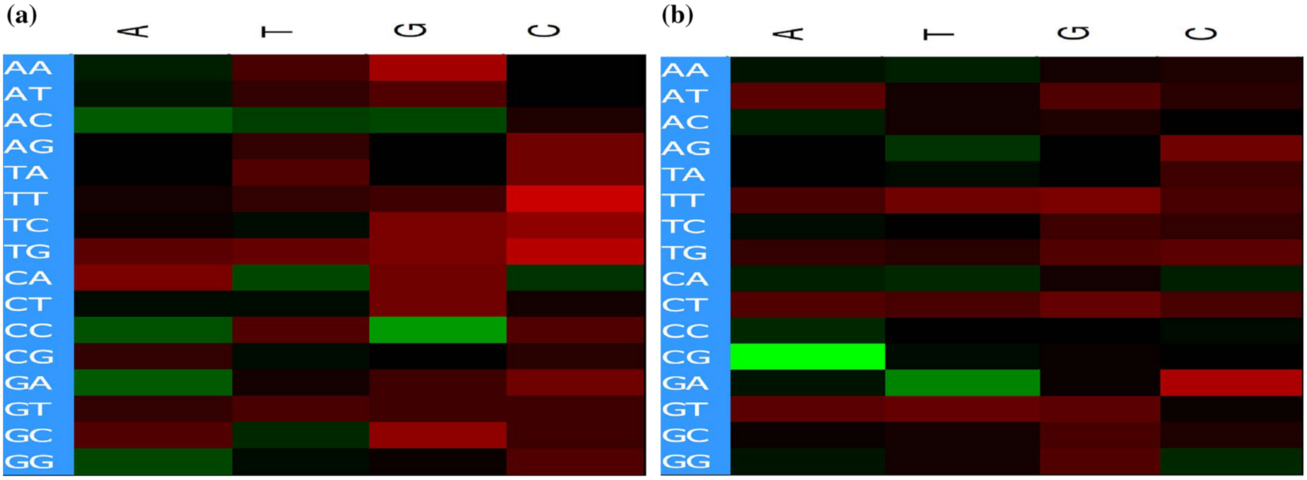


**Fig. 3** **a** Codon usage and GC3 relationship. Heat maps of correlation coefficient between codon usage and GC3. The type and degree of correlation indicated by color and intensity: *Red* indicates positive, *green* indicates negative correlation coefficient value. *Black* fields are stop and non degenerate codons. In *S. haematobium*, most of the AT ending codons were negative and most of the GC ending codons were positive. **b**. Heat map of correlation between codon usage and GC3 in *W. bancrofti* mitochondria. In *W. bancrofti*, most of the AT and GC ending codons were positive

### Trends of codon usage variation in mitochondrial protein coding genes between *W. bancrofti and S. haematobium*

PCA for *W. bancrofti* and *S. haematobium* are carried out in this study (Jenkins and Holmes 2003). PCA detected one major trend in axis 1 and the other major trend in axis 2 accounting for the total variation. The plots of axis 1 and axis 2 of the *W. bancrofti* and *S. haematobium* are shown in Fig. 5. From the figures it is evident, that in these species, the pattern of codon usage was distinct further suggesting that the codon usage was genetically quite different between *W. bancrofti* and *S. haematobium*.

### Effect of mutational bias on codon usage variation

To identify whether the evolution of codon usage bias in *W. bancrofti* and *S. haematobium* had been driven by mutation pressure alone or whether the translational selection had also played a major role, we compared the correlation between general nucleotide composition (A%, T%, G%, C%, GC%) and nucleotide composition at the third codon position (A3%, T3%, G3%, C3%, GC3%) of *W. bancrofti* and S. *haematobium* using the Pearson's correlation as shown in Tables 3 and 4 respectively. In *W. bancrofti*, a significant positive correlation was observed between ENC% and GC3 % (r = 0.826**, $p < 0.01$), GC1 and GC3 (r = 0.754**, $p < 0.01$), which suggests that mutational pressure had some contribution in codon usage bias in this species. Significant negative correlation was observed between ENC% and GC2% (r = −0.716**, $p < 0.01$). In *S. haematobium*, a significant positive correlation was observed between A% and A3% (r = 0.907**, $p < 0.01$) and T% and T3 % (r = 0.603*, $p < 0.05$) and significant negative correlation was observed between A% and T3% (r = −0.787**, $p < 0.01$). Significant positive correlation was also observed between ENC% and GC3% (r = 0.983**, $p < 0.01$), which suggests that the mutational pressure played a key role in codon usage of mitochondrial genes in *S. haematobium.*

However, no significant correlation between GC1 and GC3, GC2 and GC3 in *W. bancrofti* was found whereas significant positive correlation was observed between G% and G3% (r = 0.726**, $p < 0.01$) in *S. haematobium*. Furthermore, no significant correlation was observed between GC1and GC3, GC2 and GC3 and this suggests that natural selection contributed to codon usage bias.

### Role of natural selection versus mutation pressure on mitochondrial protein coding genes

The neutrality plot was drawn to determine the degree of natural selection against mutation pressure in the codon usage pattern of mitochondrial DNA in *W. bancrofti* and *S. haematobium.* Neutrality plot is the regression of GC12 (average of GC content at the 1st and 2nd codon position) on GC3. The regression coefficient of mitochondrial DNA in *W. bancrofti* is 0.352 which reveals that relative neutrality is 35.2 % while relative constraint is 64.8 % for GC3 in *W. bancrofti.* The GC12 was affected by mutation pressure and natural selection with a ratio of 0.352/0.648 = 0.543. In *S. haematobium*, the regression coefficient of GC12 on GC3 is 0.487 which indicates the relative neutrality of 48.7 % and the relative constraint of 51.3 %

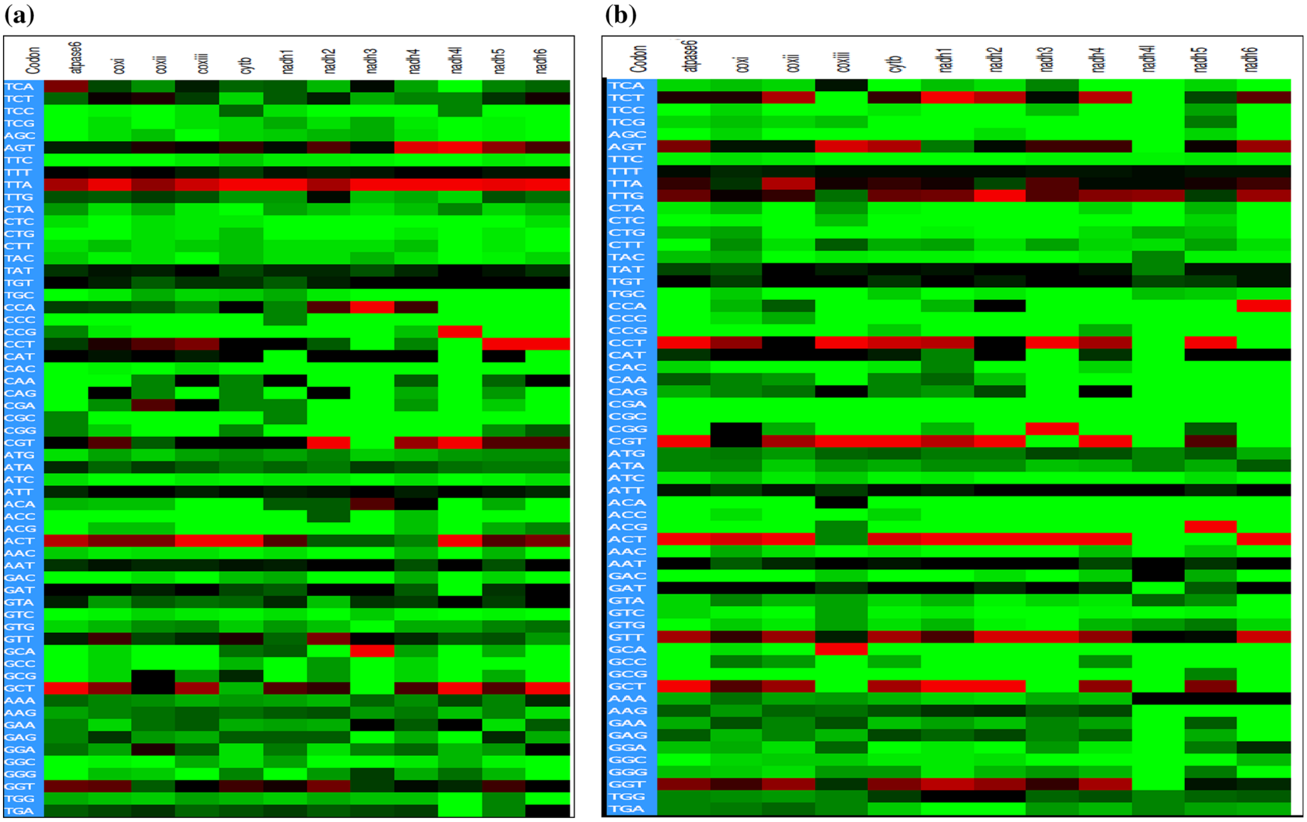


**Fig. 4** **a** Heat map of RSCU in mitochondrial protein coding genes of *W. bancrofti*. The *color* and the degree of color intensity represent the RSCU value. The *color* varies from *green* to *red* with low value of RSCU to high value respectively. *Light green* indicates RSCU value zero, *green* indicates >0.06, *dark black* RSCU value >1 and *red* indicates >1.6. **b** Heat map of RSCU in mitochondrial protein coding genes of *W. bancrofti*

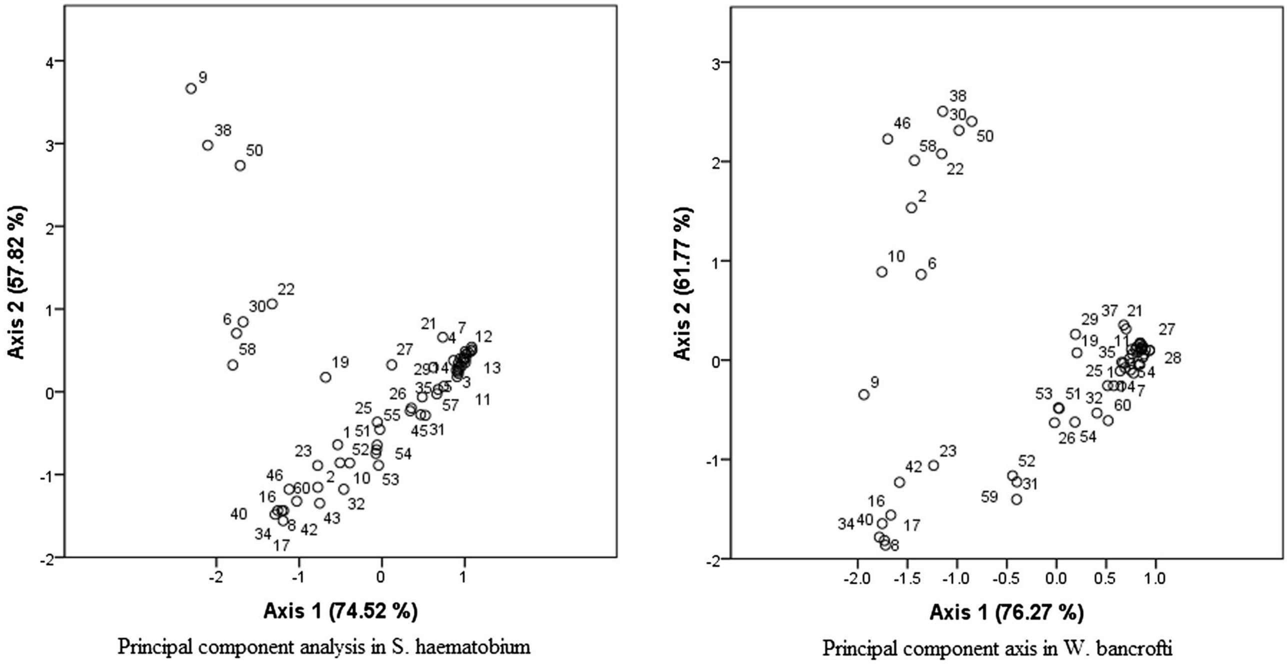


**Fig. 5** Principal component analysis in *W. bancrofti* and *S. haematobium*

for GC3. The GC12 was influenced by mutation pressure and natural selection with a ratio of 0.487/0.513 = 0.949. These results suggest that natural selection plays a major role in codon usage pattern while mutation pressure plays a minor role. In Fig. 6a, b, the points are in narrow range of distribution indicating that GC12 and GC3 were not solely governed by the mutational bias, rather natural selection operated upon the mitochondrial genes.

**Table 3** Correlation coefficients between overall nucleotide composition (A, T, G, C, and GC %) and nucleotide composition at 3rd position (A3, T3, G3, C3, and GC3 %) in *Wuchereria bancrofti*

| Nucleotide (%) | A3 (%) | T3 (%) | G3 (%) | C3 (%) | GC3 (%) |
|---|---|---|---|---|---|
| A | 0.445 | −0.144 | −0.479 | 0.176 | −0.224 |
| T | −0.350 | 0.210 | 0.208 | −0.246 | −0.004 |
| G | 0.072 | −0.116 | 0.124 | 0.082 | 0.129 |
| C | 0.247 | −0.245 | −0.008 | 0.318 | 0.170 |
| GC | 0.190 | −0.208 | 0.051 | 0.240 | 0.167 |

*, ** Significant at $p < 0.05$, and $p < 0.01$ respectively

## Effect of hydrophobicity and aromaticity of encoded protein on synonymous codon usage bias

We performed a correlation analysis to investigate whether other factors could explain the codon usage. The correlation analysis between the hydrophobicity (GRAVY Score) of protein and ENC value in *W. bancrofti* showed that the correlation coefficient was positive and significant (r = 0.671*, $p < 0.05$). But in *S. haematobium* correlation analysis between the hydrophobicity of each protein and NC value did not show any significant relationship. In *W. bancrofti* a significant correlation was found between aromaticity and GC (r = −0.885*), although no correlation with ENC, whereas in *S. haematobium* no significant correlation was found. These results indicated that the degree of hydrophobicity and the aromaticity of proteins were associated with codon usage variation only in *W. bancrofti*.

**Table 4** Correlation efficient between overall nucleotide composition (A, T, G, C, and GC %) and nucleotide composition at 3rd position (A3, T3, G3, C3, and GC %) in *Schistosoma haematobium*

| Nucleotide (%) | A3 (%) | T3 (%) | G3 (%) | C3 (%) | GC3 (%) |
|---|---|---|---|---|---|
| A | 0.907** | −0.787** | −0.443 | −0.094 | −0.370 |
| T | −0.516 | 0.603* | 0.035 | −0.097 | −0.028 |
| G | −0.244 | −0.075 | 0.726** | 0.046 | 0.544 |
| C | −0.406 | 0.481 | −0.248 | 0.247 | −0.038 |
| GC | −0.486 | 0.258 | 0.463 | 0.213 | 0.450 |

*, ** Significant at $p < 0.05$, and $p < 0.01$ respectively

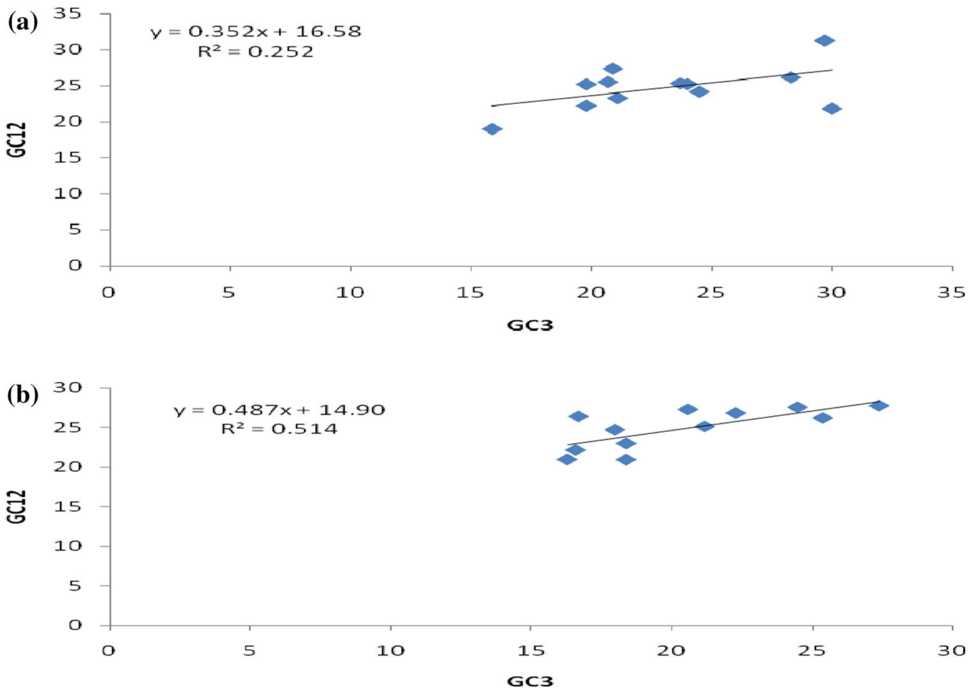


**Fig. 6** **a** Neutrality plot of GC12 versus GC3 in *W. bancrofti*. **b** Neutrality plot of GC12 versus GC3 in *S. haematobium*

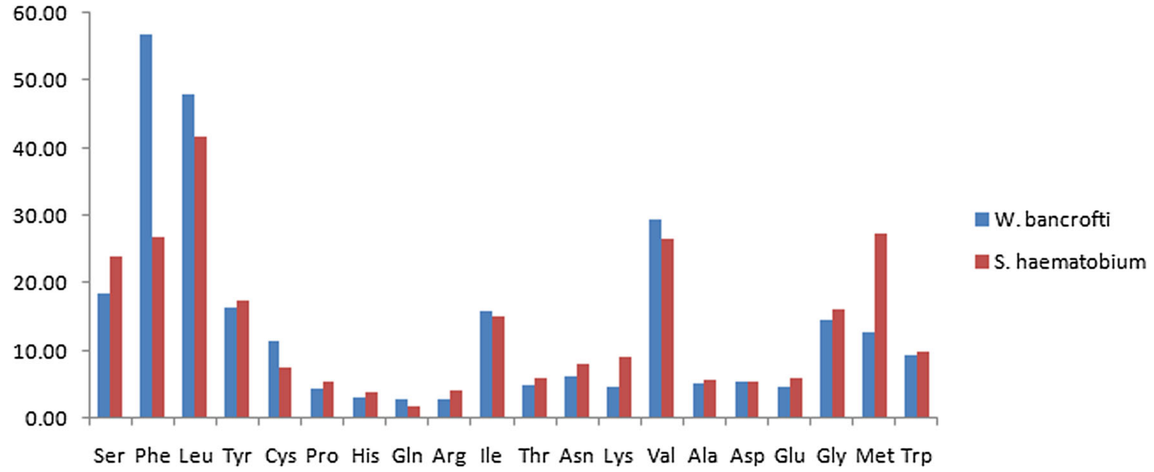


**Fig. 7** Distribution of amino acids in mitochondrial proteins of *W. bancrofti* and *S. haematobium*

### Contribution of amino acids to gene's codon usage bias

The usage of amino acids is different in the mitochondrial proteins. From the Fig. 7, it is observed that amino acid phenylalanine accounts for the greatest usage whereas glutamine and arginine account for the least usage in the gene products of *W. bancrofti*. In *S. haematobium,* leucine accounts for the greatest usage and glutamine accounts for the least. Thus it is evident that in both the species glutamine accounts for the least usage in mitochondrial proteins.

## Discussion

The present investigation highlights the codon usage patterns in a comparative study between two medically important parasitic species of humans, *W. bancrofti* and *S. haematobium*. As synonymous codon usage is not uniform during translation process, the identification of the codon usage pattern is important to understand the translational selection of codons in protein coding genes in these two species. Here we analyzed the synonymous codon usage bias in two parasitic species. In this study, we found that the most frequent codons end with T in both species. This finding may be the result of compositional constraint that occurred in codon usage pattern in these species. In *S. haematobium*, a significant positive correlation was observed between A% and A3% (r = 0.907**, $p < 0.01$) and T% and T3% (r = 0.603*, $p < 0.05$), and between G% and G3% (r = 0.726**, $p < 0.01$). Moreover, significant negative correlation was observed between the heterogeneous nucleotide comparison; A% and T3% (r = −0.787**, $p < 0.01$). This suggests that the mutational pressure played a key role in determining codon usage in *S. haematobium.* But in *W. bancrofti* only significant positive correlation was found between ENC% and GC3 % (r = 0.826**, $p < 0.01$). Uddin and Chakraborty (2014) analysed codon usage pattern in MT-ATP6 in some mammals and observed significant correlation between A% and A3%, C% and C3%, GC% and GC3% and significant negative correlation was observed between A% and GC3%, T% and G3%, T% and C3%. Their results further suggested that the mutational pressure was the prime factor for the pattern codon usage bias in these mammalian species (Uddin and Chakraborty 2014).

The ENC values ranged from 43 to 60 with a mean of 46.91 in *W. bancrofti* but from 49 to 60 with a mean of 45.17 in *S. haematobium,* respectively. This indicates that codon usage bias is not very remarkable in these two species. Yadav and Swati (2012), in their comparative genome analysis of six malarial parasites using codon usage bias based tools, calculated the ENC values in the coding sequences of *Plasmodium* species (*P. falciparum, P. vivax, P. knowlesi, P. berghei, P. chabaudii* and *P. yoelli*). The degree of codon usage bias in two species namely *P. vivax* (ENC = 55.54) and *P. knowlesi* (ENC = 55.28) was low as evident from high ENC value as compared to high codon usage bias in other four species such as *P. falciparum* (ENC = 37.89), *P. berghei* (ENC = 38.61), *P. chabaudii* (ENC = 39.71) and *P. yoelli* (ENC = 38.16)) indicated by low ENC (Yadav and Swati 2012). Thus it was evident that the ENC values of both *W. bancrofti* and *S. haematobium* were in the present study lower than *P. vivax and P. knowlesi,* but higher than that of other four parasitic species.

In both the species, *W. bancrofti* and *S. haematobium*, a highly significant positive correlation was observed between ENC% and GC3%. It suggested that mutation pressure had been a major factor in the codon usage bias in these species. This was also supported by Yadav and Swati (2012). They suggested that ENC of coding sequences and their corresponding GC3 values are used to demonstrate the role of dominant factors in shaping codon usage bias in *Plasmodium* species (Yadav and Swati 2012). ENC-GC3 plot clearly showed that the variation of GC3 was greater in *P. vivax* and its correlation with ENC was less compared with that of *P. falciparum*; resulting in variation of codon usage among them. The average ENC value for all the genes was less in *P. falciparum* suggesting that overall codon usage bias was more in *P. falciparum* compared to *P. vivax*. Correlation analysis between ENC and GC3 showed higher correlation coefficient for *P. falciparum* (0.63) compared with that of *P. vivax* (0.21). These observations clearly showed that the expression of genes in *P. falciparum* was more dependent on the composition biased mutational pressure than *P. vivax* (Yadav and Swati 2012). But from the neutrality plot it is seen that natural selection played major role while mutation pressure played minor role in codon usage pattern in mitochondrial protein coding genes in both *W. bancrofti* and *S. haematobium* supporting the result of Wei et al. on mitochondrial DNA in *B. mori* (Wei et al. 2014).

## Conclusion

This is the first work on the comparative analysis of the pattern of codon usage in two human parasitic species *W. bancrofti* and *S. haematobium* that are medically important. This work is useful for understanding the pattern of codon usage of mitochondrial genes in these species. Codon usage bias was not very remarkable in both the parasites. Natural selection and mutation pressure played important role in codon usage pattern while natural selection played a major role and mutation pressure a minor role in codon usage pattern. Other factors such as nucleotide composition, GRAVY (hydropathicity) of protein affected the codon usage pattern. However, further analysis would elucidate the role of any other factor that might be significant in determining the pattern of codon usage bias in these species.

**Acknowledgments** We are thankful to Assam University Silchar, for providing necessary facilities for this research work.

**Financial Disclosure** No fund was received from any funding agency of Govt. of India (DBT or DST) to undertake this research.

**Compliance with ethical standards**

**Conflict of interest** The authors declare no conflict of interests in this work.

## References

Behura SK, Severson DW (2013) Codon usage bias: causative factors, quantification methods and genome-wide patterns: with emphasis on insect genomes. Biol Rev 88:49–61

Bockarie MJ, Taylor MJ, Gyapong JO (2009) Current practices in the management of lymphatic filariasis. Expert Rev Anti Infect Ther 7:595–605

Bulmer M (1991) The selection-mutation-drift theory of synonymous codon usage. Genetics 129:897–907

Chen H, Sun S, Norenburg JL, Sundberg P (2014) Mutation and selection cause codon usage and bias in mitochondrial genomes of ribbon worms (*Nemertea*). PLoS One 9:e85631

Francino MP, Ochman H (1999) Isochores result from mutation not selection. Nature 400:30–31

Gissi C, Iannelli F, Pesole G (2008) Evolution of the mitochondrial genome of Metazoa as exemplified by comparison of congeneric species. Heredity 101:301–320

Gupta S, Ghosh T (2001) Gene expressivity is the main factor in dictating the codon usage variation among the genes in *Pseudomonas aeruginosa*. Gene 273:63–70

Ingvarsson PK (2007) Gene expression and protein length influence codon usage and rates of sequence evolution in *Populus tremula*. Mol Biol Evol 24:836–844

Jenkins GM, Holmes EC (2003) The extent of codon usage bias in human RNA viruses and its evolutionary origin. Virus Res 92:1–7

Manguin S, Bangs M, Pothikasikorn J, Chareonviriyaphap T (2010) Review on global co-transmission of human *Plasmodium* species and *Wuchereria bancrofti* by Anopheles mosquitoes. Infect Genet Evol 10:159–177

Melrose WD (2002) Lymphatic filariasis: new insights into an old disease. Int J Parasitol 32:947–960

Plotkin JB, Kudla G (2011) Synonymous but not the same: the causes and consequences of codon bias. Nat Rev Genet 12:32–42

Powell JR, Moriyama EN (1997) Evolution of codon usage bias in *Drosophila*. Proc Natl Acad Sci USA 94:7784–7790

Richter C, Park JW, Ames BN (1988) Normal oxidative damage to mitochondrial and nuclear DNA is extensive. Proc Natl Acad Sci USA 85:6465–6467

Sharp PM, Li WH (1986a) Codon usage in regulatory genes in *Escherichia coli* does not reflect selection for 'rare'codons. Nucleic Acids Res 14:7737–7749

Sharp PM, Li WH (1986b) An evolutionary perspective on synonymous codon usage in unicellular organisms. J Mol Evol 24:28–38

Sharp PM, Li WH (1987) The codon adaptation index-a measure of directional synonymous codon usage bias, and its potential applications. Nucleic Acids Res 15:1281–1295

Sharp PM, Matassi G (1994) Codon usage and genome evolution. Curr Opin Genet Dev 4:851–860

Stock S (2009) Molecular approaches and the taxonomy of insect–parasitic and pathogenic nematodes. Insect pathogens: molecular approaches and techniques. Cabi Publishing—CABI, Wallingford, pp 71–100

Sun Z, Wan DG, Murphy RW, Ma L, Zhang XS, Huang DW (2009) Comparison of base composition and codon usage in insect mitochondrial genomes. Genes Genom 31:65–71

Taylor RW, Turnbull DM (2005) Mitochondrial DNA mutations in human disease. Nat Rev Genet 6:389–402

Uddin A, Chakraborty S (2014) Mutation pressure dictates codon usage pattern in mitochondrial *ATP8* in some mammalian species. Int J Sci Res 3:1206–1212

Wang HC, Hickey DA (2007) Rapid divergence of codon usage patterns within the rice genome. BMC Evol Biol 7:S6

Wei L, He J, Jia X, Qi Q, Liang Z, Zheng H, Ping Y, Liu S, Sun J (2014) Analysis of codon usage bias of mitochondrial genome in *Bombyx mori* and its relation to evolution. BMC Evol Biol 14:262

Wright F (1990) The 'effective number of codons' used in a gene. Gene 87:23–29

Yadav MK, Swati D (2012) Comparative genome analysis of six malarial parasites using codon usage bias based tools. Bioinformation 8:1230–1239

# Proof Delivery Form

## Journal of Helminthology

**Date of delivery:** 23/02/2016

**Journal and vol/article ref:** jhl H3586

**Number of pages (not including this page):** 8

This proof is sent to you on behalf of Cambridge University Press. Please print out the file and check the proofs carefully. Please ensure you answer all queries.

Please EMAIL your corrections within **2** days of receipt to:

**Mrs Linda Antoniw: <antoniw@btinternet.com>**

**Authors are strongly advised to read these proofs thoroughly because any errors missed may appear in the final published paper. This will be your ONLY chance to correct your proof. Once published, either online or in print, no further changes can be made.**

**NOTE:** If you have no corrections to make, please also email to authorise publication.

- The proof is sent to you for correction of typographical errors only. Revision of the substance of the text is not permitted, unless discussed with the editor of the journal. Only **one** set of corrections are permitted.

- Please answer carefully any author queries.

- Corrections which do NOT follow journal style will not be accepted.

- A new copy of a figure must be provided if correction of anything other than a typographical error introduced by the typesetter is required.

- If you have problems with the file please email **jhlproduction@cambridge.org**

Please note that this pdf is for proof checking purposes only. It should not be distributed to third parties and may not represent the final published version.

**Important:** you must return any forms included with your proof. We cannot publish your article if you have not returned your signed copyright form

**Please do not reply to this email**

NOTE - for further information about **Journals Production** please consult our **FAQs** at http://journals.cambridge.org/production_faqs

*Journal of Helminthology*, page 1 of 9
© *Cambridge University Press 2016*
doi:10.1017/S0022149X16000092

# Expression levels and codon usage patterns in nuclear genes of the filarial nematode *Wucheraria bancrofti* and the blood fluke *Schistosoma haematobium*

**G.A. Mazumder, A. Uddin and S. Chakraborty***

Department of Biotechnology, Assam University, Silchar-788011, Assam, India

(Received 24 November 2015; Accepted 1 February 2016)

## Abstract

Synonymous codons are used with different frequencies, a phenomenon known as codon bias, which exists in many genomes and is mainly resolute by mutation and selection. To elucidate the genetic characteristics and evolutionary relationship of *Wucheraria bancrofti* and *Schistosoma haematobium* we examined the pattern of synonymous codon usage in nuclear genes of both the species. The mean overall GC contents of *W. bancrofti* and *S. haematobium* were 43.41 and 36.37%, respectively, which suggests that genes in both the species were AT rich. The value of the High Effective Number of Codons in both species suggests that codon usage bias was weak. Both species had a wide range of P3 distribution in the neutrality plot, with a significant correlation between P12 and P3. The codons were closer to the axes in correspondence analysis, suggesting that mutation pressure influenced the codon usage pattern in these species. We have identified the more frequently used codons in these species, most codons ending with an A or T. The nucleotides A/T and C/G were not proportionally used at the third position of codons, which reveals that natural selection might influence the codon usage patterns. The regression equation of P12 on P3 suggests that natural selection might have played a major role, while mutational pressure played a minor role in codon usage pattern in both species. These results form the basis of exploring the evolutionary mechanisms and the heterologous expression of medically important proteins of *W. bancrofti* and *S. haematobium.*

## Introduction

The genetic code is conserved among organisms, but the direction of the codon bias arising from unequal usage of codons always shifts between different organisms (Hershberg & Petrov, 2008). The combination of any three of the four bases (A, G, C and T) encodes each of the 20 amino acids. It is known that the 20 standard amino acids are encoded by 61 codons of the genetic code. Each amino acid can be encoded by at least one codon (e.g. Met and Trp); but some amino acids are encoded by up to six codons (e.g. Leu, Ser and Arg), which is due to the degeneracy of the genetic code. The codons that encode the same amino acid are referred to as synonymous codons. Codon usage bias (CUB) refers to the phenomenon whereby synonymous codons are used with unequal frequencies during translation of genes. CUB is a common phenomenon in a variety of organisms, from prokaryotes to eukaryotes (Bulmer, 1988). Codons that are used more repeatedly are often termed optimal or major codons, and those that are used less frequently are termed non-optimal or minor codons. Usually, these less repeatedly used codons correspond to less abundant tRNAs in the cell than those of optimal codons (Ikemura, 1981, 1985; Bulmer, 1987; Akashi, 2001), and the

*E-mail: supriyoch_2008@rediffmail.com

translational machinery mostly pauses there (Kurland, 1992). Various studies have suggested that synonymous codon usage pattern is species-specific and non-random. CUB varies both within and between genomes, and may play a significant role in our understanding of genome evolution among related species at the molecular level (Sharp and Matassi 1994).

Various factors that may influence codon bias include gene expression, gene length (Gupta & Ghosh, 2001; Hou & Yang, 2001; Jenkins & Holmes, 2003; Gustafsson *et al.*, 2004), base compositional mutational bias (Jia & Higgs, 2008) and natural selection (Kaufmann & Paules, 1996; Knight *et al.*, 2001; Lithwick & Margalit, 2005; Linacre & ToBe, 2009; Chen *et al.*, 2013). Kurland (1992) proposed that not only is the expression of individual genes increased by the codon usage bias, but also that the cell growth rate is changed under optimal conditions (Fennoy & Bailey-Serres 1993). Shackelton *et al.* (2006) found that codon usage bias was correlated with the overall genomic GC content, which indicates that the compositional constraint under mutation pressure, rather than natural selection, was the main factor for specific codons. Naya *et al.* (2001) studied the *Chlamydomonas reinhardtii* genome, which has a high GC content, and found no evidence that mutation pressure was responsible for determining the codon usage pattern (Shackelton *et al.*, 2006). It was also suggested by some researchers that codon usage variation is related to other factors, such as gene function and DNA replication, and selective transcription, protein secondary structure and environmental factors (Akashi, 1997; Powell & Moriyama, 1997; Moriyama & Powell, 1998).

Analysis of codon usage bias is a well-established technique for understanding the protein-coding sequences of genomes (Kurland, 1992). It also helps in understanding the dynamics of mRNA translation, design of transgenes, new gene discovery, and studies of molecular biology and evolution, etc. While numerous reports on synonymous codon usage bias have focused on nuclear genomes of other worms, no work has been reported on the nuclear genes of two medically important species: *Wuchereria bancrofti* and *Schistosoma haematobium*. In this context, one of the major objectives of this work was to understand the patterns of codon usage among the sequenced genes of *W. bancrofti* and *S. haematobium*. *Wuchereria bancrofti* is a parasitic roundworm in humans and the major cause of lymphatic filariasis, which affects 128 million people worldwide (Melrose, 2002). If it is left untreated, then the swelling and decreased flow of the lymph fluid will lead to lymphatic system infections and, over time, a condition called elephantiasis (permanent swelling of a limb) is reached. Humans are the host for *W. bancrofti* and they act as vectors (Bockarie *et al.*, 2009; Manguin *et al.*, 2010). The drug diethylcarbamazine (DEC) is used to kill the microfilaria and some of the adult worms, but it is no longer approved by the US Food and Drug Administration (FDA), because of its side-effects, such as dizziness, nausea, fever, headache, or pain in muscles or joints. A combination of both albendazole and ivermectin, or albendazole and diethylcarbamazine, is considered to be effective in eliminating microfilaria (Eberhard *et al.*, 1997). On the other hand, *S. haematobium*, which is found in some areas of the Middle East, is one of the main causes of urogenital schistosomiasis (Kumar *et al.*, 2010), which is caused by blood flukes (trematode worms) of the genus *Schistosoma* (Kumar *et al.*, 2010). It is endemic in around 78 countries. The common sign of urogenital schistosomiasis is haematuria (blood in the urine), in children it can cause anaemia. For the control of this disease, treatment with praziquantel is suggested by the World Health Organization (WHO, xxxx).

These two medically important worm species pose a serious threat to humans. The information on the synonymous codon usage patterns of *W. bancrofti* and *S. haematobium* is either scanty or unavailable. In this study, we investigated the codon usage profile of *W. bancrofti* and *S. haematobium* using a multivariate statistical analysis. Our results provide useful insights into the patterns of codon usage bias that facilitate better understanding of the structure and evolution of gene coding sequences of these species.

## Materials and methods

### *Sequence collection*

Coding DNA sequences (CDS) of *W. bancrofti* and *S. haematobium* were retrieved from GenBank (http://www.ncbi.nlm.nih.gov) using accession numbers. In these sequences, we only chose the CDS without anonymous bases. To minimize sampling errors, genes with incorrect initiation and termination codons or with internal termination codons were avoided. Finally, 121 CDS of *W. bancrofti* and 24 CDS of *S. haematobium* were left for analysis, and each corresponded to a unique gene in *W. bancrofti* and *S. haematobium*, respectively. The genes of *W. bancrofti* and *S. haematobium* are shown in supplementary tables S1 and S2, repectively.

### *Estimation of DNA compositional properties*

Using an in-house Perl program, nucleotide compositions (A%, T%, G% and C%), nucleotide composition at the third position (A3%, T3%, G3% and C3%), and overall GC contents, GC1 (P1), GC2 (P2), GC3 (P3), AT1, AT2 and AT3 were estimated.

The mutations that usually take place in the third position are mostly synonymous, whereas the mutations occurring in the first or the second positions are relatively small and known as non-synonymous mutations. This indicates that when there is no external pressure, mutations should occur in a random rather than in a specific direction. This will result in uniform base composition at three positions of codons. However, in the presence of selective pressure, preference for a particular base would occur in three different positions (Sueoka, 1988). In the neutrality plot, P12 has been used as the ordinate and P3 as abscissa. Each dot represents an independent gene. If all the points of the neutrality plot show narrow GC3 distribution, it indicates low mutation bias or high conservation of GC content throughout the whole genome. On the other hand, if the curve of the neutrality plot tends to be sloped or parallel to the horizontal axis, it indicates that the difference between P12 and P3 is very low. When selection pressure plays a major role in evolution, the neutrality plot is used to measure the degree of neutrality (Sueoka, 1988, 1999).

### *Relative synonymous codon usage (RSCU)*

The relative synonymous codon usage (RSCU) is the number of times a codon appears in a gene divided by the number of expected occurrences under equal codon usage. If the synonymous codons of an amino acid are used with equal frequencies, their RSCU values will equal 1 (Sharp *et al.*, 1986). It is calculated as:

$$RSCUij = \frac{X_{ij}}{\frac{1}{ni}\sum_{j=1}^{ni} X_{ij}}$$

where Xij is the observed number of the '*i*th' codon for the '*j*th' amino acid, '*ni*' is the total number of synonymous codons that encode the '*j*th' amino acid.

Correspondence analysis (CA) has been widely used to investigate the variation of codon usage among genes. CA is a multivariate statistical method in which the codon usage of 59 codons is plotted in a multidimensional space of 59 axes, and then the CA recognizes the axes that represent the most important factors contributing to codon usage variation among genes (Liu *et al.*, 2004; Wang & Hickey, 2007).

### *Effective number of codons (ENC)*

The effective number of codons (ENC) quantifies how far the codon usage of a gene departs from equal usage of synonymous codons. ENC values range from 20 (the number of amino acids), which means that the bias is at a maximum, and only one codon is used from each synonymous-codon group, to 61 (the number of sense codons), which indicates no codon-usage bias. A low ENC value means high codon usage bias, and a high ENC value means low codon bias. ENC is measured as:

$$ENC = 2 + S + 29/S^2 + (1 - S^2)$$

Q6

where *S* represents the given (G + C)3% value (Wright, 1990).

### *Codon adaptation index (CAI)*

In recognition of the role of selection in producing high codon bias, a parameter called the codon adaptation index (CAI) is calculated. The CAI measures the degree with which genes use the preferred codons. CAI values range from 0 to 1. A high CAI value indicates a higher percentage of the most abundant codons (Sharp *et al.*, 1988). It is calculated as:

$$CAI = \exp\left(\frac{1}{L}\sum_{k=1}^{L} \ln \omega k\right)$$

where ω*k* is the relative adaptiveness of the *k*th codon, and *L* is the number of synonymous codons in the gene. CAI is also a widely accepted measure of gene expression.

### *Data analysis*

Correlation and regression analyses were performed to identify the relationship between the overall nucleotide compositions with the nucleotide compositions at third codon position. All the statistical analyses were done using the SPSS software (SPSS Inc., Chicago, Illinois, USA).

## Results

### *Compositional properties*

Codon bias, or choice for one type of codon over another, can be influenced by the overall nucleotide composition of genomes (Jenkins & Holmes, 2003). Therefore, we first analysed the nucleotide composition of coding sequences from different nuclear genes of *W. bancrofti* and *S. haematobium.* The overall nucleotide composition and nucleotide composition at the third codon position of *W. bancrofti* and *S. haematobium* coding sequences are shown in supplementary tables S3 and S4, respectively. In both the species *W. bancrofti* and *S. haematobium*, unequal distribution of overall A, T, G, C and their composition at the third codon position, confirmed that the compositional constraint influences Q7 the codon usage pattern.

Linear regression analysis was performed between an observed value of ENC and base composition of third codon position (Wobble codon position). In *W. bancrofti*, the observed coefficients of regression analysis between ENC and A3 was 0.469; ENC and T3, 0.356; ENC and G3, 0.357; and ENC and C3, 0.479; while in *S. haematobium*, between ENC and A3 it was 0.704; ENC and T3, 0.558; ENC and G3, 0.795; and ENC and C3, 0.838. These results showed that the effective number of codons (ENC) was negatively affected by A3, T3, G3 and C3, since all the observed values are positive in both species.

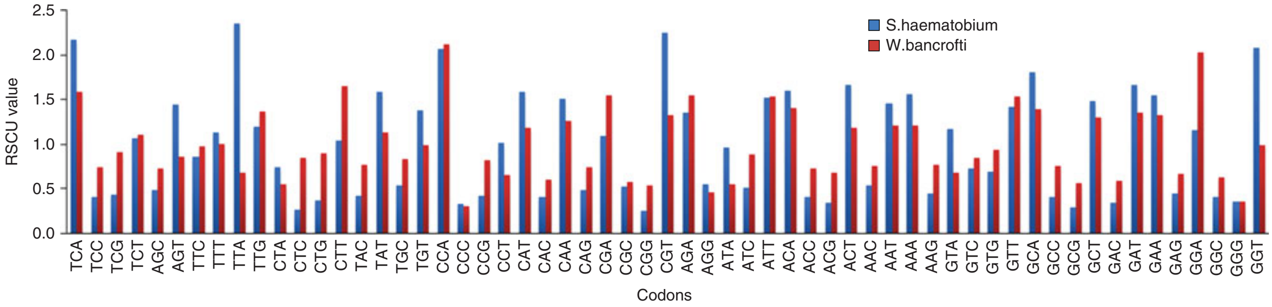


Fig. 1. Comparison of relative synonymous codon usage values in *Wuchereria bancrofti* and *Schistosoma haematobium*.

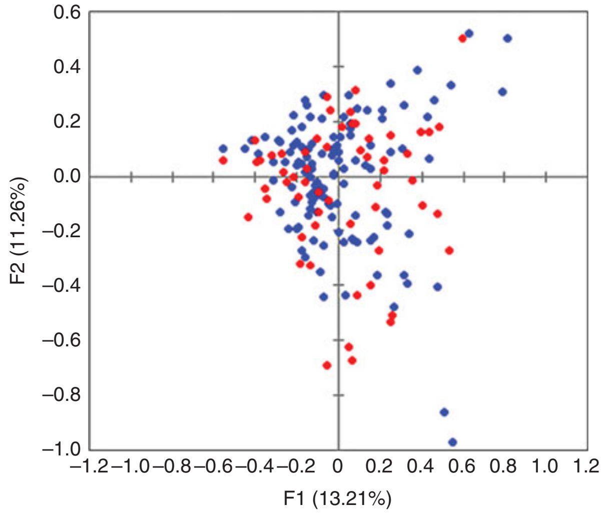


Fig. 2. Construction analysis of nuclear genes for *Wuchereria bancrofti*.

### *Codon usage in* W. bancrofti *and* S. haematobium

In *W. bancrofti*, the overall GC content had a mean value of 43.41%, and in *S. haematobium* it was 36.37%, which suggest that genes in both the species were AT rich. The GC content in the first and third codon position in *W. bancrofti* was 51.06 and 39.08%, respectively, while Q8 in *S. haematobium*, it was 46.57 and 26.50%, respectively. The overall GC(%), GC1(%), GC2(%), GC3(%) and ENC values of genes of *W. bancrofti* and *S. haematobium* are Q9 shown in supplementary tables S5 and S6, respectively. The mean ENC value was 56.24 in *W. bancrofti* but 49.29 in *S. haematobium*. Thus it can be stated that *W. bancrofti* had a weak codon bias, but *S. haematobium* had a moderate codon bias; this was also supported by the RSCU values in both the species, in which almost half of the codons were frequently used. In *S. haematobium*, 29 codons were used more frequently, but in *W. bancrofti* 23 codons were Q10 used more frequently (fig. 1). Among these, the codons which end with T at the third position in both the species were CAT, CGT, ATT, AAT, ACT, GTT and GAT. The codons ending with A were CAA, CGA, AGA, ACA, AAA, GCA, GAA and GGA. In *W. bancrofti*, other frequently used codons were TCA, TCT, TTT, TTG, CTT, TAT, CCA, AAT and GCT and in *S. haematobium*, these codons were CCT, GTA. The comparisons of RSCU values in both the species are shown in fig. 1. This indicates that the codon usage pattern in these two species is mostly contributed by compositional constraints, and the pattern of codon usage is different in the two species.

### *Effects of nucleotide composition in shaping codon bias*

Correspondence analysis was carried out for the nuclear genes in two species, *W. bancrofti* and *S. haematobium*. The first axis was found to account for 13.36% of the total observed variance while axis 2 contributed to 11.16% of the Q11 variance in *W. bancrofti*; these values were 27.18% and 11.83%, respectively, in *S. haematobium* (figs 2 and 3).

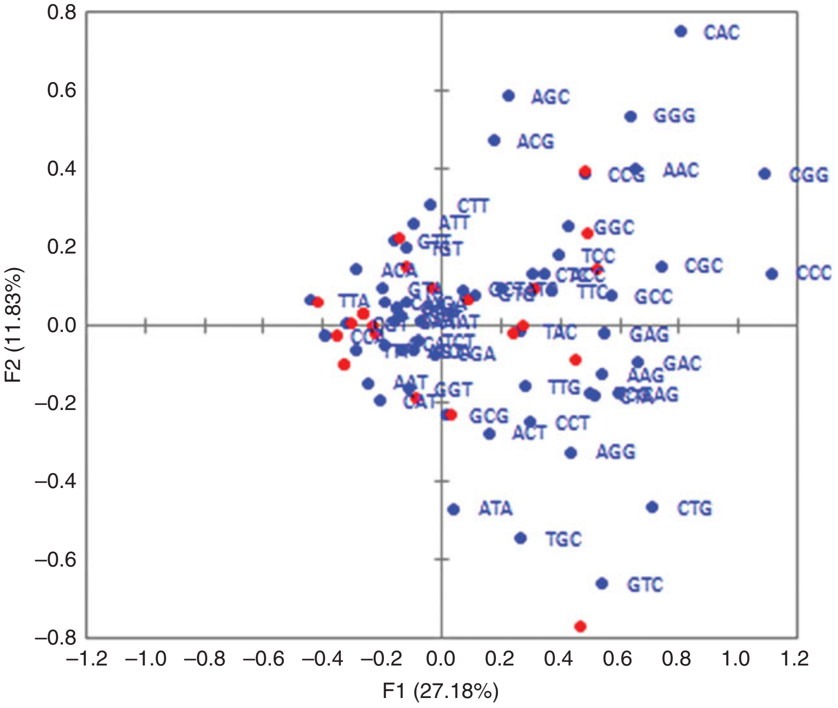


Fig. 3. Construction analysis of nuclear genes for *Schistosoma haematobium*.

Table 1. Correlation between axes of correspondence analysis and compositional constraints in *W. bancrofti*.

| | A | T | G | C | A3 | T3 | G3 | C3 | GC | GC1 | GC2 | GC3 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| F1 | − 0.511** | − 0.337** | 0.127 | 0.683** | − 0.612** | − 0.644** | 0.393** | 0.838** | 0.776** | 0.188* | 0.072 | 0.829** |
| F2 | 0.139 | 0.040 | − 0.585** | 0.278** | 0.206* | 0.355** | − 0.689** | 0.053 | − 0.165 | 0.024 | 0.219* | − 0.370** |

*$P < 0.05$, **$P < 0.01$.

The patterns of the scattering observed in the two species were distinctly non-identical, indicating a variation in the overall codon usage pattern in these two species. In both the species, it was found that AT-ending codons were more frequent near the origin, while GC-ending codons lie further away, suggesting that compositional constraints under mutational pressure might have played a part in shaping the codon usage pattern in these species. Furthermore, some of the genes were in a discrete distribution, indicating that other factors exist in shaping the codon usage bias, for example, natural selection (Wei *et al.*, 2014).

### *Effects of gene expression level*

To detect the effects of gene expression level on codon bias, correlation coefficients were estimated among the CAI (codon adaptation index), ENC (effective number of codons) and nucleotide composition. In *W. bancrofti*, the CAI showed a negative correlation with G3 ($r = -0.188$*), C3 ($r = -0.101$), GC ($r = -0.247$**), GC1 ($r = -0.097$), GC2 ($r = -0.111$), GC3 ($r = -0.184$*), Aromo ($r = -0.058$) and Gravy ($r = -0.164$), whereas it showed a positive correlation with gene length ($r = 0.644$**), A3 ($r = 0.075$), T3 ($r = 0.203$*) and ENC ($r = 0.010$) (supplementary table S7).

In *S. haematobium*, the CAI showed a negative correlation with A3 ($r = -0.119$), G3 ($r = -0.208$), C3 ($r = -0.185$), GC3 ($r = -0.216$) and ENC ($r = -0.226$), but a positive correlation with gene length ($r = 0.700$**), T3 ($r = 0.523$**), GC ($r = 0.091$), GC1 ($r = 0.211$), GC2 ($r = 0.416$*), Gravy ($r = 0.217$) and Aromo ($r = 0.104$) (supplementary table S8). These results suggest that, in these species, the nucleotide composition and the gene expression levels contribute to the codon usage bias (Jia *et al.*, 2015). In *W. bancrofti*, a positive correlation was found between CAI and ENC, which indicates that the gene expression level increases with decrease in codon usage bias, while in *S. haematobium*, the negative correlation between CAI and ENC suggests the gene expression level increases with increase in the codon usage bias.

### *Mutational pressure influences the codon bias*

Here, a set of indices including overall A, T, G, C and G + C content and the contents of bases in the third position (A3, T3, G, C3 and GC3) were considered as the indices of mutational pressure (Chen, 2013). The axis 1 and axis 2 values were correlated with nucleotide count at first and third positions. In *W. bancrofti*, significant correlation had been found among the axis 1 values and A, T, C, A3, T3, G3, GC, GC1 and GC3, while axis 2 values had shown significant correlation with G, C, A3, T3, G3, GC2 and GC3 (table 1). Also, in *S. haematobium*, a significant correlation was found among the axis 1 values and A, T, G, C, A3, T3, G3, C3, GC and GC3 (table 2). Thus the nucleotides A, T, C, GC, T3, C3 and GC3 were found to play a major role in determining the pattern of synonymous codon usage bias. These results suggest that compositional constraints under mutation pressure might play an important role in synonymous codon usage pattern in both the species, which strongly supports the findings of Butt *et al.* (2014).

Furthermore, in *W. bancrofti*, the neutrality plot showed that the genes had a wide range of GC3 value distributions, ranging from 15.5 to 65.1% (fig. 4). There was a negative correlation between GC12 and GC3 ($-0.063$, $P < 0.01$), suggesting that the effect of mutational pressure was present at all codon positions. But the effect of mutational pressure could be small, as revealed by the low magnitude of the correlation coefficient ($-0.063$). The linear regression model of GC12 on GC3 is $\mathrm{GC12} = 0.029\mathrm{GC3} + 46.75$, with $R^2 = 0.004$. This suggests that only 0.4% of the GC12 variance comes from GC3, and hence GC3 content is not likely to be the major determinant of GC12 variance in *W. bancrofti* coding sequences.

In *S. haematobium*, the neutrality plot analysis showed that the genes had a wide range of GC3 value distributions, ranging from 10.1 to 41.4% (fig. 5). A positive correlation between GC12 and GC3 ($r = 0.373$, $P < 0.01$) was observed and this indicates that the effect of mutational pressure was present at all codon positions. The linear regression model of $\mathrm{GC12} = 0.173\mathrm{GC3} + 36.68$ with $R^2 = 0.139$ suggests that

Table 2. Correlation between axes of correspondence analysis and compositional constraints in *S. haematobium*.

| | A | T | G | C | A3 | T3 | G3 | C3 | GC | GC1 | GC2 | GC3 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| F1 | − 0.645** | − 0.451* | 0.607** | 0.730** | − 0.856** | − 0.657** | 0.813** | 0.918** | 0.819** | 0.327 | 0.248 | 0.957** |
| F2 | − 0.19 | 0.08 | 0.137 | 0.011 | − 0.016 | 0.03 | − 0.177 | 0.156 | 0.099 | 0.143 | 0.132 | − 0.006 |

*$P < 0.05$, **$P < 0.01$.

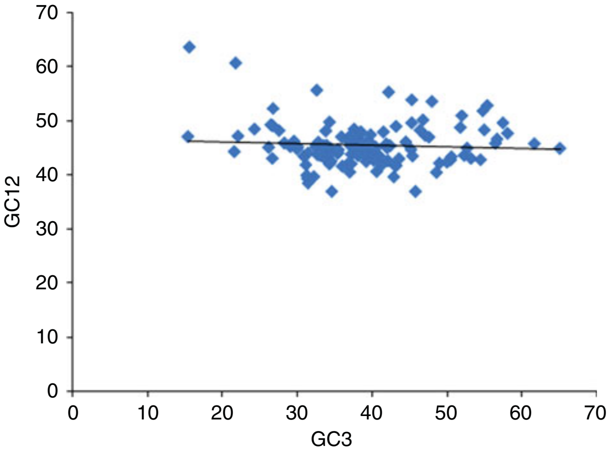


Fig. 4. Neutrality analysis of GC12 and GC3 for the coding sequence of *Wuchereria bancrofti*. GC12 is the average value of GC content in the first and second positions of the codons (GC1 and GC2) while GC3 refers to the GC content in the third position ($Y = -0.029x + 46.75$, $R^2 = 0.004$).

only 13.9% of the GC12 variance could be explained by GC3 in this organism. It reveals that other factors, apart from GC3, could play a role in determining the GC12 variance in *S. haematobium* coding sequences.

### *Natural selection affects the codon usage pattern*

It has been suggested that if synonymous codon usage pattern is influenced by mutational pressure only, then the frequency of nucleotides, A and T should be equal to that of C and G at the third position of codons (Zhang *et al.*, 2013). However, in the case of *W. bancrofti* and *S. haematobium* genes, AT and GC were not equal in the third position of codons (supplementary tables S3 and S4), indicating that natural selection could also influence overall synonymous codon usage bias.

It is not enough to distinguish the main determinant factor between natural selection and mutational pressure within a species using the ENC plot, so a neutrality plot was implemented here. In *W. bancrofti*, the slope of the regression line of the entire coding sequence was 0.029 (fig. 4). The results revealed that the effect of mutation pressure was only 2.9% (absolute value), while the magnitude of natural selection was 97.1% (Sueoka, 1988). In *S. haematobium*, the slope of the regression line of the entire coding sequence was 0.173 (fig. 5). The results revealed that the effect of mutational pressure was only 17.3%, while the natural selection was 82.7%. Hence, it is evident that natural selection played a major role in shaping the codon bias, whereas mutational pressure probably had a minor role.

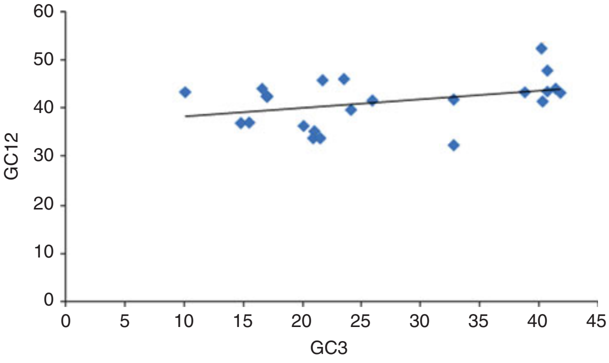


Fig. 5. Neutrality analysis of GC12 and GC3 for the coding sequences of *Schistosoma haematobium* ($Y = 0.173x + 36.68$, $R^2 = 0.139$).

### *Correlation between codon usage and skewness*

The GC skews, AT skews, purine skews, pyrimidine skews, amino skews and keto skews in both the species *W. bancrofti and S. haematobium* were all negative, except that of pyrimidine skew in *S. haematobium*, as shown in Table 3. Base composition is connected to the transcription process, which is exposed from skewness. In *W. bancrofti*, significant correlation of ENC had been found with GC skew (−0.189*), purine skew (−0.346**), pyrimidine skew (−0.372**), amino skew (0.220*) and keto skew (−0.440**), respectively. Therefore in *W. bancrofti* GC skew, purine skew, pyrimidine skew, amino skew and keto skew significantly affected the codon usage. In *S. haematobium*, ENC showed significant correlation with purine skew (−0.730**), pyrimidine skew (0.777**), amino skew (−0.647**) and keto skew (−0.698**). This suggests that in *S. haematobium*, the purine skew, pyrimidine skew, amino skew and keto skew played a significant role in codon usage bias.

## Discussion

In living organisms, synonymous codons are used with different frequencies and this phenomenon is known as codon bias. CUB is an important evolutionary relic that exists in an extensive variety of organisms, from prokaryotes to eukaryotes (Yang *et al.*, 2014). Various hypotheses are proposed to explain the origin of codon usage bias, out of which neutral theory and the selection–mutation–drift balance model are the most accepted ones (Yang *et al.*, 2014). The study of the pattern of codon usage has gained the attention of researchers since the beginning of whole-genome sequencing of many organisms (Plotkin & Kudla, 2011). Our investigation highlighted the comparative analysis of codon usage of

Table 3. Correlation between effective number of codons (ENC) and types of skewness.

| Skewness (*r*) | *W. bancrofti* (*P*) | *S. haematobium* (*P*) |
|---|---|---|
| GC | −0.189* (0.038) | −0.002 (0.991) |
| AT | −0.081 (0.376) | −0.056 (0.794) |
| Purine | −0.346** (0.000) | −0.730** (0.000) |
| Pyrimidine | −0.372** (0.000) | 0.777** (0.000) |
| Amino | −0.220* (0.015) | −0.647** (0.001) |
| Keto | −0.440** (0.000) | −0.698** (0.000) |

**P* < 0.05, ***P* < 0.01.

nuclear genes in two parasitic species, *W. bancrofti* and *S. haematobium*, which are considered to be medically important.

Compositional constraint could be one of the most essential factors in shaping the pattern of codon usage among genes and genomes. Here we analysed the synonymous codon usage bias in these two species, and found that the most frequent codons end with A and T in both the species. This finding might be the result of compositional constraint that occurred in codon usage pattern in these species. A previous study on five nematode species also revealed the preponderance of AT base pairs over GC base pairs, as well as the predominance of A/T-ending codons over G/C-ending codons in coding sequences (Fadiel *et al.*, 2001). Also the more frequently used codons end with A/T; this is supported by the AT-rich genome of the parasite *Plasmodium falciparum* (Peixoto *et al.*, 2004).

Correspondence analysis of the nuclear genes of both the species *W. bancrofti* and *S. haematobium* found that AT-ending codons lie near the axes whereas the GC-ending codons lie further away from the origin, which suggests that compositional constraints along with mutational pressure might be playing a significant role in shaping the codon usage pattern. This was also suggested by Wei *et al.* (2014), who found that, in the mitochondrial genome of *Bombyx mori*, the AT-ending codons were lying closer to Axis 1, which indicates that the mutational pressure might be correlated with the codon usage bias, and that lower GC-content genes were located closer to Axis 1, which again implies that GC content in mutational pressure probably influenced the bias (Wei *et al.*, 2014).

The mean ENC value was 56.24 in *W. bancrofti*, and 49.29 in *S. haematobium*, which indicates that codon usage bias was low in *W. bancrofti* and *S. haematobium*. Jia *et al.* (2015) also reported weak codon usage bias in the genes of *B. mori*. Low codon usage bias might be advantageous for efficient replication in different cell types having different preferences of codons (Jenkins & Holmes, 2003).

The regression equations from the neutrality plots in the cases of *W. bancrofti* and *S. haematobium* suggest that mutational pressure played a minor role while natural selection played the major role in shaping the codon usage pattern. The neutrality plot for *S. haematobium* revealed that the effect of mutational pressure was only 17.3%, while that of natural selection was 82.7%, thus, indicating a minor role of mutational bias but major role of natural selection in shaping the codon usage bias. Jia *et al.* (2015) implemented a neutrality plot in *B. mori* and found a significant positive correlation between GC12 and GC3, which suggested that the effect of mutation pressure was present. They also found that the mutation bias only played a minor role in shaping the codon bias, whereas natural selection probably dominated the codon bias (Jia *et al.*, 2015).

The frequencies of A and T are not equal to that of G and C at the third position of codons, which suggests that both mutational pressure and the natural selection affect the codon usage pattern, thus supporting the results of Butt *et al.* (2014) (Zhang *et al.*, 2013).

This is the first work on the comparative analysis of the pattern of codon usage in two medically important parasitic species. The codon usage bias was weak and the genes in both species were AT rich. The most frequent codons end with T at the third position, most probably suggesting the role of the compositional constraint under mutation pressure. Both mutation pressure and natural selection are two significant evolutionary factors that affect the codon usage pattern in these species. Apart from this, gene expression level and various compositional skewnesses influence the pattern of codon usage. Natural selection played the major role while mutational pressure played the minor role in shaping the pattern of codon usage. However, further analysis would elucidate the role of any other factor that might be responsible for codon usage bias in these species.

## Supplementary material

To view supplementary material for this article, please visit http://dx.doi.org/10.1017/S0022149X16000092

## Acknowledgements

We are thankful to Assam University, Silchar, Assam for providing the necessary facilities for this research work. No funding was received from DBT or DST, Government of India for carrying out this research work.

## References

**Akashi, H.** (1997) Codon bias evolution in *Drosophila*. Population genetics of mutation-selection drift. *Gene* **205**, 269–278.

**Akashi, H.** (2001) Gene expression and molecular evolution. *Current opinion in Genetics and Development* **11**, 660–666.

**Bockarie, M.J., Taylor, M.J.** ***et al.*** (2009) Current practices in the management of lymphatic filariasis.

**Bulmer, M.** (1987) Coevolution of codon usage and transfer RNA abundance. *Nature* **325**, 728–730.

**Bulmer, M.** (1988) Are codon usage patterns in unicellular organisms determined by selection-mutation balance? *Journal of Evolutionary Biology* **1**, 15–26.

**Butt, A.M., Nasrullah, I.,** ***et al.*** (2014) Genome-wide analysis of codon usage and influencing factors in chikungunya viruses. *PLoS ONE* **9**, e90905.

**Chen, H.-T., Gu, Y.-X.,** ***et al.*** (2013) Analysis of synonymous codon usage in dengue viruses. *Journal of Animal and Veterinary Advances* **12**, 88–98.

**Chen, Y.** (2013) A comparison of synonymous codon usage bias patterns in DNA and RNA virus genomes: quantifying the relative importance of mutational pressure and natural selection. *BioMed Research International*.

**Eberhard, M.L., Hightower, A.W.,** ***et al.*** (1997) Clearance of *Wuchereria bancrofti* antigen after treatment with diethylcarbamazine or ivermectin. *The American Journal of Tropical Medicine and Hygiene* **57**, 483–486.

**Fadiel, A., Lithwick, S.,** ***et al.*** (2001) Influence of intercodon and base frequencies on codon usage in filarial parasites. *Genomics* **74**, 197–210.

**Fennoy, S.L. & Bailey-Serres, J.** (1993) Synonymous codon usage in *Zea mays* L. nuclear genes is varied by levels of C- and G-ending codons. *Nucleic Acids Research* **21**, 5294–5300.

**Gupta, S. & Ghosh, T.** (2001) Gene expressivity is the main factor in dictating the codon usage variation among the genes in *Pseudomonas aeruginosa*. *Gene* **273**, 63–70.

**Gustafsson, C., Govindarajan, S., *et al.*** (2004) Codon bias and heterologous protein expression. *Trends in Biotechnology* **22**, 346–353. Q25

**Hershberg, R. & Petrov, D.A.** (2008) Selection on codon bias. *Annual Review of Genetics* **42**, 287–299.

**Hou, Z. & Yang, N.** (2001) Analysis of factors shaping *S. pneumoniae* codon usage. *Yi Chuan Xue Bao (Acta Genetica Sinica)* **29**, 747–752.

**Ikemura, T.** (1981) Correlation between the abundance of *Escherichia coli* transfer RNAs and the occurrence of the respective codons in its protein genes: a proposal for a synonymous codon choice that is optimal for the *E. coli* translational system. *Journal of Molecular Biology* **151**, 389–409.

**Ikemura, T.** (1985) Codon usage and tRNA content in unicellular and multicellular organisms. *Molecular Biology and Evolution* **2**, 13–34.

**Jenkins, G.M. & Holmes, E.C.** (2003) The extent of codon usage bias in human RNA viruses and its evolutionary origin. *Virus Research* **92**, 1–7.

**Jia, W. & Higgs, P.G.** (2008) Codon usage in mitochondrial genomes: distinguishing context-dependent mutation from translational selection. *Molecular Biology and Evolution* **25**, 339–351.

**Jia, X., Liu, S., *et al.*** (2015) Non-uniqueness of factors constraint on the codon usage in *Bombyx mori*. *BMC Genomics* **16**, 356. Q25

**Kaufmann, W.K. & Paules, R.S.** (1996) DNA damage and cell cycle checkpoints. *FASEB Journal* **10**, 238–247.

**Knight, R.D., Freeland, S.J., *et al.*** (2001) 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* **2**, research0010. Q25

**Kumar, V., Abbas, A., *et al.*** (2010) *Robbins and Cotran pathological basis of disease*. 8th edn. Elsevier, Sauders. Q25 Q28

**Kurland, C.** (1992) Translational accuracy and the fitness of bacteria. *Annual Review of Genetics* **26**, 29–50.

**Linacre, A. & ToBe, S.S.** (2009) Species identification using DNA loci. *Forensic Science in Wildlife Investigations* 61. Q29

**Lithwick, G. & Margalit, H.** (2005) Relative predicted protein levels of functionally associated proteins are conserved across organisms. *Nucleic Acids Research* **33**, 1051–1057.

**Liu, Q., Feng, Y., *et al.*** (2004) Synonymous codon usage bias in *Oryza sativa*. *Plant Science* **167**, 101–105. Q25

**Manguin, S., Bangs, M., *et al.*** (2010) Review on global co-transmission of human *Plasmodium* species and *Wuchereria bancrofti* by *Anopheles mosquitoes*. *Infection, Genetics and Evolution* **10**, 159–177. Q25

**Melrose, W.D.** (2002) Lymphatic filariasis: new insights into an old disease. *International Journal for Parasitology* **32**, 947–960.

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

**Peixoto, L., Fernandez, V., *et al.*** (2004) The effect of expression levels on codon usage in *Plasmodium falciparum*. *Parasitology* **128**, 245–251. Q25

**Plotkin, J.B. & Kudla, G.** (2011) Synonymous but not the same: the causes and consequences of codon bias. *Nature Reviews Genetics* **12**, 32–42.

**Powell, J.R. & Moriyama, E.N.** (1997) Evolution of codon usage bias in *Drosophila*. *Proceedings of the National Academy of Sciences, USA* **94**, 7784–7790.

**Shackelton, L.A., Parrish, C.R., *et al.*** (2006) Evolutionary basis of codon usage and nucleotide composition bias in vertebrate DNA viruses. *Journal of Molecular Evolution* **62**, 551–563. Q25

**Sharp, P.M. & Matassi, G.** (1994) Codon usage and genome evolution. *Current Opinion in Genetics and Development* **4**, 851–860.

**Sharp, P.M., Tuohy, T.M., *et al.*** (1986) Codon usage in yeast: cluster analysis clearly differentiates highly and lowly expressed genes. *Nucleic Acids Research* **14**, 5125–5143. Q25

**Sharp, P.M., Cowe, E., *et al.*** (1988) Codon usage patterns in *Escherichia coli*, *Bacillus subtilis*, *Saccharomyces cerevisiae*, *Schizosaccharomyces pombe*, *Drosophila melanogaster* and *Homo sapiens*; a review of the considerable within-species diversity. *Nucleic Acids Research* **16**, 8207–8211. Q25

**Sueoka, N.** (1988) Directional mutation pressure and neutral molecular evolution. *Proceedings of the National Academy of Sciences, USA* **85**, 2653–2657.

**Sueoka, N.** (1999) Two aspects of DNA base composition: G + C content and translation-coupled deviation from intra-strand rule of A=T and G=C. *Journal of Molecular Evolution* **49**, 49–62.

**Wang, H.C. & Hickey, D.A.** (2007) Rapid divergence of codon usage patterns within the rice genome. *BMC Evolutionary Biology* **7**, S6.

**Wei, L., He, J., *et al.*** (2014) Analysis of codon usage bias of mitochondrial genome in *Bombyx mori* and its relation to evolution. *BMC Evolutionary Biology* **14**, 262. Q25

**Wright, F.** (1990) The 'effective number of codons' used in a gene. *Gene* **87**, 23–29.

**Yang, X., Luo, X., *et al.*** (2014) Analysis of codon usage pattern in *Taenia saginata* based on a transcriptome dataset. *Parasites & Vectors* **7**, 1–11. Q25

**Zhang, Z., Dai, W., *et al.*** (2013a) Synonymous codon usage in TTSuV2: analysis and comparison with TTSuV1. *PLoS ONE* **8**, e81469. Q25

**Zhang, Z., Dai, W., *et al.*** (2013b) Analysis of synonymous codon usage patterns in torque teno sus virus 1 (TTSuV1). *Archives of Virology* **158**, 145–154. Q25

# Author Queries

*JOB NUMBER:* H3586
*JOURNAL:* JHL

**Q1** The distinction between surnames can be ambiguous, therefore to ensure accurate tagging for indexing purposes online (eg for PubMed entries), please check that the highlighted surnames have been correctly identified, that all names are in the correct order and spelt correctly.

**Q2** 'Kurland proposed' changed to 'Kurland (1992) proposed' - OK?

**Q3** Naya *et al.* (2001) – is not in the Reference list. Please give details.

**Q4** 'WHO [34]' Changed to '(WHO, xxxx)' Please give the year and the reference details for the Reference list.

**Q5** 'overall GC contents, GC1 (P1), GC2 (P2), GC3 (P3), AT1, AT2 and AT3 were estimated'. Please confirm that this is correct (i.e. that AT1, AT2 and AT3 are GC contents)

**Q6** '[36].' deleted here. Should a reference be cited instead of [36]?

**Q7** 'in both the species' – repetition, deleted. OK?

**Q8** 'was 51.06 and 39.08%, respectively, while in *S. haematobium,* it was 46.57 and 26.50%, respectively'. '%' added – OK?

**Q9** S1 and S2 changed to S5 and S6 to agree with the supplementary material. OK?

**Q10** 'In *W. bancrofti*, 23 codons were used most frequently but 29 codons in *S. haematobium*.' This repeated the previous sentence. Deleted and text run on with previous paragraph. OK?

**Q11** 'The first axis was found to account for 13.36% of the total observed variance while axis 2 contributed to 11.16% of the variance in *W. bancrofti*'. Please check these values against fig. 2, where they appear to be given as 13.21% and 11.26%, and advise.

**Q12** S5 changed to S7 – OK?

**Q13** S6 changed to S8 – OK?

**Q14** A3, T3, G, C3, and GC3. Should this be A3, T3, G**3**, C3, and GC3?

**Q15** Zhang *et al.* 2013 – there are two such references in the Reference list. Should this be a, or b or a, b?

**Q16** '[25]' deleted here. Should it be replaced by a reference citation?

**Q17** 'exposed from skewness' – can this be changed to 'exposed to skewness'?

**Q18** [59-61] deleted here. Should references be added in place of this?

**Q19** (0.220*) changed to (−0.220*) to agree with Table 3. OK?

**Q20** Wei. L (2014) changed to Wei *et al.* (2014) to agree with the Reference list. OK?

**Q21** The neutrality plot revealed. Changed to The neutrality plot for *S. haematobium* revealed. OK?

**Q22** thus supporting the results of Butt *et al.* (Zhang *et al.*, 2013). Should this be thus supporting the results of Butt *et al.* (2014) and Zhang *et al.* (2013)?

**Q23** Zhang *et al.* (2013) should this be 2013a, b or a, b?

**Q24** No funding was received from DBT or DST, Government of India for carrying out this research work. Please confirm that 'No funding' is correct. If so, should this be deleted as the usual practice is to report funding sources only?

**Q25** The journal style is to give all authors' names in the Reference list. Please provide the names of all authors for the following references: **Bockarie, Taylor, et al.** (2009), **Butt, Nasrullah, et al.** (2014), **Chen, Gu, et al**. (2013), **Eberhard, Hightower, et al.** (1997), **Fadiel, Lithwick, et al.** (2001), **Gustafsson, Govindarajan, et al.** (2004), **Jia, Liu, et al.** (2015), **Knight, Freeland, et al.** (2001), **Kumar, Abbas, et al.** (2010), **Liu, Q., Y. Feng, et al.** (2004), **Manguin, Bangs, et al.** (2010), **Peixoto, Fernandez, et al.** (2004), **Shackelton, Parrish, et al.** (2006), **Sharp, Tuohy, et al.** (1986), **Sharp, Cowe, et al.** (1988), **Wei, He, et al.** (2014), **Yang, Luo, et al.** (2014), **Zhang, Dai, et al.** (2013a), **Zhang, Dai, et al.** (2013b)

**Q26** **Bockarie, M.J., Taylor, M.J., et al.** (2009) Current practices in the management of lymphatic filariasis. This reference is incomplete. If it is a book, please give the publisher and place of publication. If a journal, please give the journal title, volume and page range.

**Q27** **Chen, Y**. (2013) A comparison of synonymous codon usage bias patterns in DNA and RNA virus genomes: quantifying the relative importance of mutational pressure and natural selection. *BioMed Research International*. Please complete this reference by giving the volumne and page range.

**Q28** **Kumar, V., Abbas, A., *et al.*** (2010) *Robbins and Cotran pathological basis of disease*. 8th edn. Elsevier, Sauders. Please give the place of publication

**Q29** **Linacre, A. & ToBe, S.S.** (2009) Species identification using DNA loci. *Forensic Science in Wildlife Investigations* 61. Please complete this reference. If it is a book, please give the page, range, publisher and place of publication. If a journal, please give the page range (is 61 the volume?).

**Q30** Are the suggested changes to Table 3 OK?