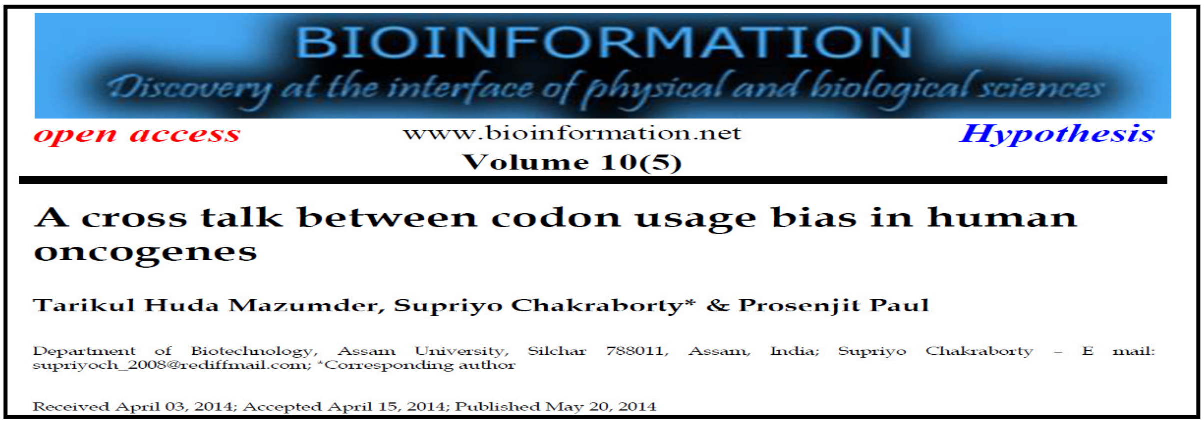
Appendix B – Lists of Publications

Published Research Articles

1. **Tarikul Huda Mazumder**, Supriyo Chakraborty and Prosenjit Paul. A cross talk between codon usage bias in human oncogenes. ***Bioinformation***. 2014; 10(5): 256–262. (**IF. 1.0**)



BIOINFORMATION

Discovery at the interface of physical and biological sciences

open access | www.bioinformation.net | Hypothesis

Volume 10(5)

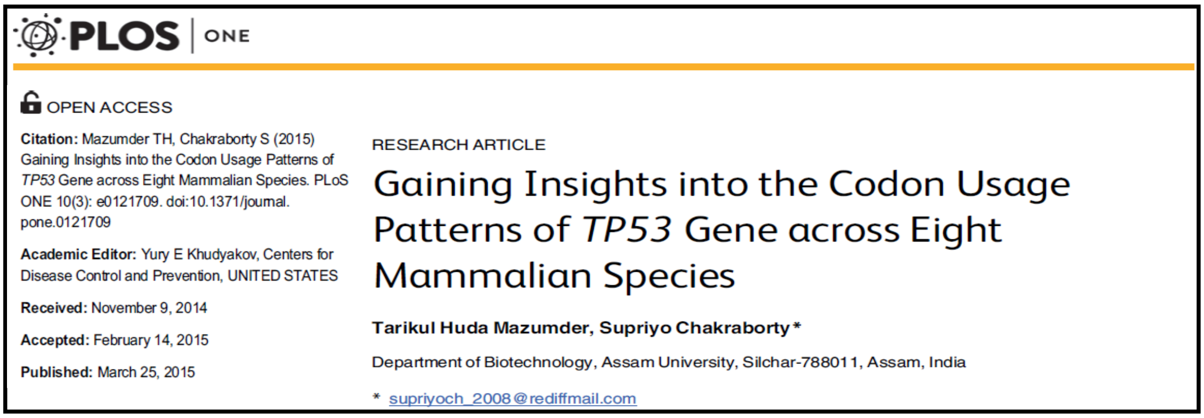
A cross talk between codon usage bias in human oncogenes

Tarikul Huda Mazumder, Supriyo Chakraborty* & Prosenjit Paul

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

Received April 03, 2014; Accepted April 15, 2014; Published May 20, 2014

2. **Tarikul Huda Mazumder** and Supriyo Chakraborty. Gaining Insights into the Codon Usage Patterns of *TP53* Gene across Eight Mammalian Species. ***PLoSOne.*** 2015 Mar 25; 10(3):e0121709. doi: 10.1371/journal.pone.0121709. (**IF. 3.5**)



PLOS | ONE

OPEN ACCESS

Citation: Mazumder TH, Chakraborty S (2015) Gaining Insights into the Codon Usage Patterns of *TP53* Gene across Eight Mammalian Species. PLoS ONE 10(3): e0121709. doi:10.1371/journal.pone.0121709

Academic Editor: Yury E Khudyakov, Centers for Disease Control and Prevention, UNITED STATES

Received: November 9, 2014

Accepted: February 14, 2015

Published: March 25, 2015

RESEARCH ARTICLE

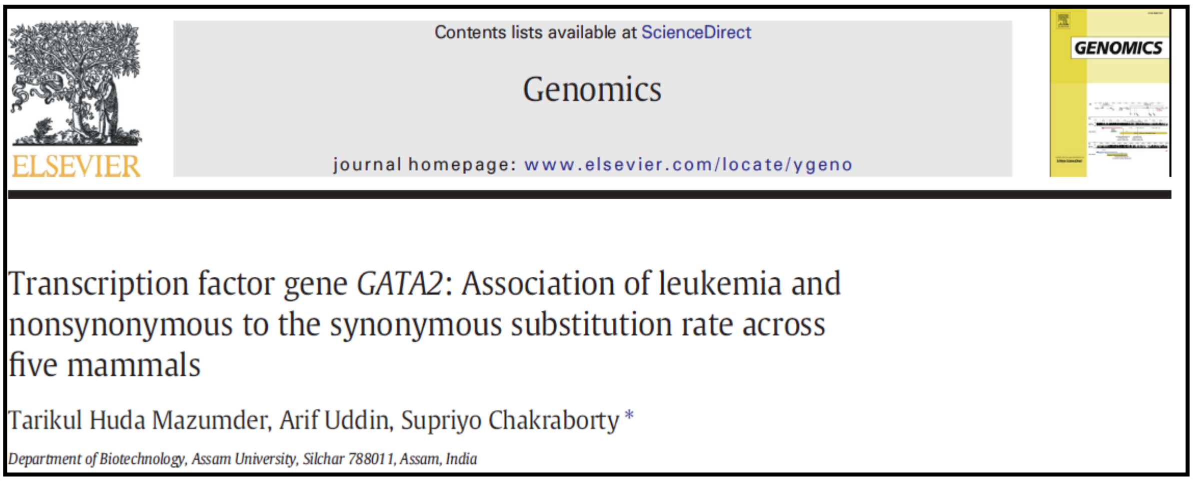
Gaining Insights into the Codon Usage Patterns of *TP53* Gene across Eight Mammalian Species

Tarikul Huda Mazumder, Supriyo Chakraborty*

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

* supriyoch_2008@rediffmail.com

3. **Tarikul Huda Mazumder,** Arif Uddin and Supriyo Chakraborty. Transcription factor gene *GATA2*: Association of leukemia and nonsynonymous to the synonymous substitution rate across five mammals. ***Genomics.*** 2016 Feb 2; pii: S0888-7543(16)30009-X. doi: 10.1016/j.ygeno.2016.02.001. (**IF. 2.3**)



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

4. **Tarikul Huda Mazumder** and Supriyo Chakraborty. Compositional dynamics and expression level of human tumor suppressor genes in cell. ***Funct Integr Genomics.*** [communicated]

Appendix C –

Workshop attended

- Participated in 2 days National Workshop on “Some Emerging Areas of Biotechnology & Bioinformatics”, organized by Biotechnology Hub (DBT Govt. of India), Karimganj College, Karimganj, Assam.
- Participated in 1 day National Workshop on “Clinical Trials-How & Why”, organized by Cachar Cancer Hospital & Research Centre, Silchar, Assam, India.
- Participated in 5 days National Workshop on “Basic Genetic Engineering Techniques for Gene Cloning”, organized by Biotechnology Hub (DBT Govt. of India), Department of Molecular Biology & Biotechnology, Tezpur University, Assam, India.
- Participated in 1 day National Workshop on “Computational Statistics in Biological Sciences”, organized by Department of Biotechnology, Assam University, Silchar, Assam, India.
- Participated in 1 weeks workshop on “Computer Aided Drug Designing: Basics to Molecular Dynamics Simulation”, organized by Bioinformatics Centre (DBT-BIF), Assam University, Silchar, Assam, India.
- Participated in 3 weeks National workshop on “Skill Development and Hands-on-training on Quality Control of Biologicals”, organized by National Institute of Biologicals, Noida and DBT, Govt. of India.

Conference attended

- Participated in the 41st Annual Conference of Indian Immunology Society at Madurai Kamaraj University, Madurai, 12-14 December 2014.
- Participated in a poster presentation on the topic entitled "*RUNX1*: a transcription factor gene associated with leukemia reflects codon bias across mammals", in the 1st International Conference on Novel Frontiers in Pharmaceutical & Health Sciences (INNOPHARM 1) hold during 10-11 October, 2015 at M.P. India.

Research Articles

BIOINFORMATION
Discovery at the interface of physical and biological sciences

open access www.bioinformation.net Hypothesis
Volume 10(5)

A cross talk between codon usage bias in human oncogenes

Tarikul Huda Mazumder, Supriyo Chakraborty* & Prosenjit Paul

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

Received April 03, 2014; Accepted April 15, 2014; Published May 20, 2014

Abstract:
Background: Oncogenes are the genes that have the potential to induce cancer. The extent and origin of codon usage bias is an important indicator of the forces shaping genome evolution in living organisms. **Results**: We observed moderate correlations between gene expression as measured by CAI and GC content at any codon site. The findings of our results showed that there is a significant positive correlation (Spearman's r= 0.45, P<0.01) between GC content at first and second codon position with that of third codon position. Further, striking negative correlation (r = -0.771, P < 0.01) between ENC with the GC3s values of each gene and positive correlation (r=0.644, P<0.01) in between CAI and ENC was also observed. **Conclusions**: The mutation pressure is the major determining factor in shaping the codon usage pattern of oncogenes rather than natural selection since its effects are present at all codon positions. The results revealed that codon usage bias determines the level of oncogene expression in human. Highly expressed oncogenes had rich GC contents with high degree of codon usage bias.

Keywords: Synonymous codon, Oncogene, Codon usage pattern.

Background:
Nature has gifted the genetic code that provides the basic instructions and information to direct efficient protein synthesis and folding. There are sixty-one codons that specify for only twenty amino acids found commonly in protein sequences; most of these amino acids (building blocks of protein) can be encoded by more than one codon (i.e., a triplet of nucleotides); such codons are described as being synonymous, and mostly differ by one nucleotide in the third position **[1].** The term codon bias or more preferably codon usage bias represents the unequal usage of synonymous codons for encoding amino acids which may vary significantly between genomes, between genes in the same genome, and within a single gene **[2-3].** Since the 1970s, the unequal use of synonymous codons has been confirmed in many organisms. To date, the codon usage patterns in many organisms have been interpreted for diverse reasons. Recently, it has been reported that two major factors are involved in the continuation of codon usage bias: weak natural selection and mutational pressure **[4]**. The selection associated with translational efficiency/accuracy is often termed as 'translation selection'. Moreover, scientific investigation also reported that synonymous codon usage pattern varied at distinct sites along a coding sequence, balances of strong versus weak base pair bonding, maintenance of DNA and RNA secondary structure, and translational efficiency and fidelity **[5].** That is why codon usage bias among different organisms or within the genes of the same organism has invited much attention and various works on the subject have been published in recent years.

Lavner and Kotlar (2005) suggested that there are three possible ways in which selection may act on codon bias in the human genome: (1) Increasing translation efficiency in highly expressed genes; (2) Regulating translation efficiency of some proteins that can be a disadvantage at high levels; and (3) Improving translation efficiency and reducing the rate of amino acid misincorporation in the production of biosynthetically expensive proteins **[3].** Many genomic analyses have been done on oncogenes but till date very little is known about the codon usage patterns and the factors that influence them. Codon usage patterns are important for bringing out molecular mechanism and evolutions of oncogenes. In this paper we have analyzed the key genetic factors playing crucial role in determining the codon usage pattern in fifty (50) oncogenes. To the best of our knowledge, it is the first systematic study to verify and insist that the synonymous codon usage pattern is one of the factors affecting the codon usage in oncogenes..

Methodology:

Retrieval of Sequence data

A list of human oncogenes was compiled from the web site (http://cbio.mskcc.org/CancerGenes/Select.action). Complete nucleotide coding sequence of each of the concerned gene, was obtained from NCBI nucleotide database website (http://www.ncbi.nlm.nih.gov). Codon usage bias was measured in the 50 oncogenes listed in **Table 1 (see supplementary material)**. The complete coding sequence (cds) of each oncogene was analyzed using PERL program developed by us.

Analysis of synonymous codon usage bias

We measured the non-uniform usage of synonymous codons for the oncogenes by analyzing several genetic indices given below-

Nucleotide composition

The frequency of the nucleotide G+C at the synonymous third codon position (GC3s) is a good indicator of the extent of base composition bias **[6].** The frequencies of the nucleotides A, C, U (T), G in the complete coding sequences of each oncogene and the occurrence of overall (G+C)% content at the different codon positions GC1%, GC2%, GC3% was calculated to study the relationship between codon usage variation and compositional constraints.

Effective number of codons

The effective number of codons (ENC) is generally used to measure the codon usage bias of a gene that is independent of the gene length and number of amino acids **[7]**. The ENC value ranges from 20-61. For a gene in which only one codon is used for each amino acid, this value would be 20 while all codons are used equally the value would be 61 **[7].** The ENC value closer to 20 indicates, strong codon usage bias in the gene and these biased genes are expressed highly **[8].** The ENC values for all cds sequence were computed as per Wright (1990) **[7].** In addition, to examine the influence of GC content on codon usage, the relationship of ENC and GC3s content of each gene was plotted according to the equation described by Wright (1990) **[7].**

Codon adaptation index

Codon adaptation index (CAI) is used to estimate the degree of bias toward codons in highly expressed genes and thus assesses the effective selection which helps in shaping the codon usage pattern **[9-10].** The CAI ranges from 0 to 1, for a gene in which all synonymous codons are used equally, the value would be 0 for no bias while only optimal codons are used, value will be 1 for strongest bias **[11]**. The CAI value was measured as per Sharp PM *et al.* **[12].**

Frequency of optimal codons

Frequency of optimal codons (Fop) is used to measure codon usage bias in a gene **[11].** Fop is calculated as the ratio of the number of optimal codons used to the total number of synonymous codons **[13].** The Fop value ranges from 0.36, for a gene in which codon usage pattern is uniform, to 1 for a gene in which codon usage is highly biased **[11].** We used the formula given by Lanver & Kotlar to calculate the Fop values for each of the cds selected for the present study **[3].**

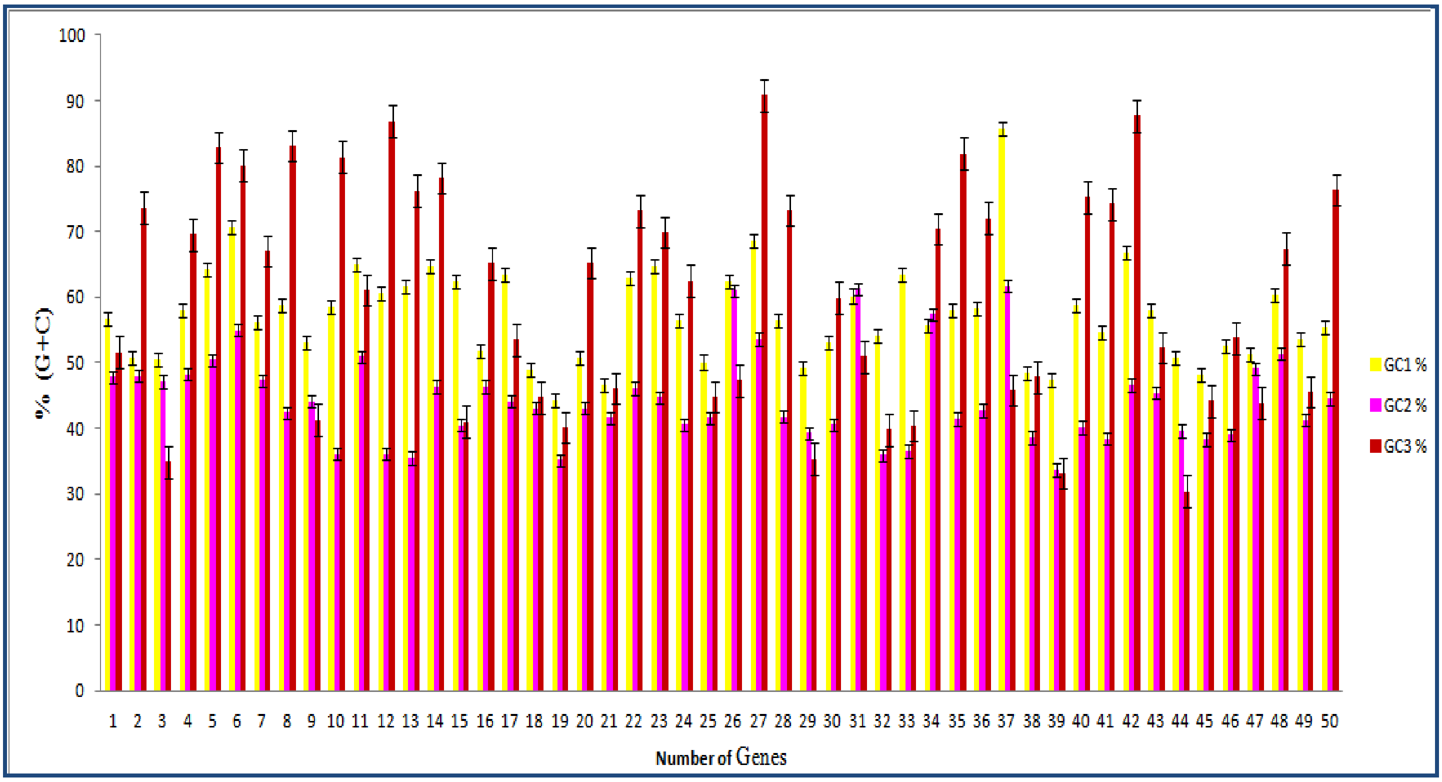


Figure 1: Percentage of GC content at three codon positions.

Relative synonymous codon usage

Relative synonymous codon usage (RSCU) is calculated by dividing the observed frequency of a codon by the expected if all synonymous codons for that amino acid were used equally **[14].** Thus, an RSCU value close to 1 indicates a lack of bias, RSCU >1 indicates a codon used more frequently than expected randomly, and RSCU <1 indicates a codon used less frequently than expected randomly **[14].**

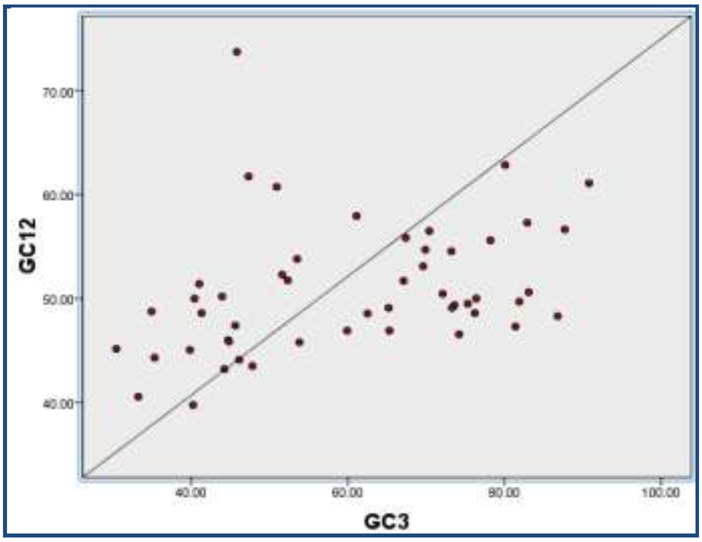


Figure 2: Correlation between GC content at first and second codon positions (GC1 & GC2) with that at synonymous third codon positions (GC3s). GC12: average GC content at first and second codon positions.

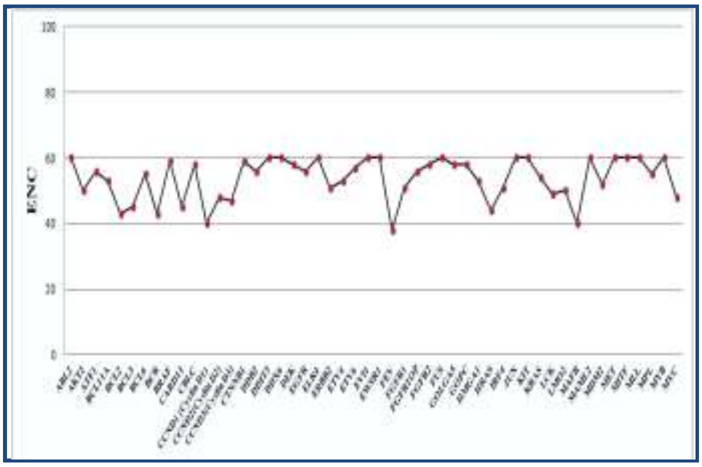


Figure 3: ENC distribution of 50 selected oncogenes.

Correlation analysis

Correlation analysis was used to identify the relationship between the pattern of synonymous codon usage and the genetic indices used for the present study. This analysis was implemented based on the Spearman's rank correlation analysis. All statistical analyses were carried out by using software SPSS.

Results:

In this present study, the selected oncogenic cds sequences were downloaded from NCBI nucleotide database using a perl program. The program was written in such a way that it selects only those cds sequences which have perfect start and stop signal and devoid of any unknown bases (N). We found fifty cds sequences in correct format for codon bias study. The extent of codon usage bias was determined in these fifty oncogenes **Table 1.** Two amino acids methionine and tryptophan coded by single codon ATG and TGG, respectively and three stop codons (TAA, TAG, and TGA) would not reveal any usage bias and therefore discarded from the calculation.

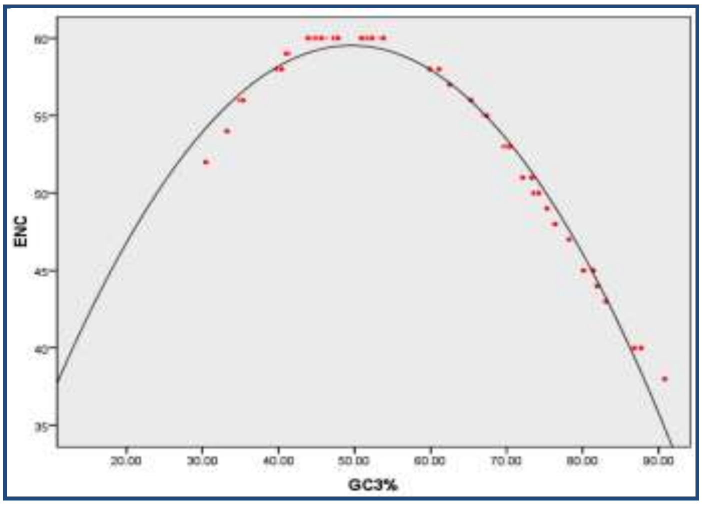


Figure 4: Distribution of ENC and GC content of the third codon position of 50 different oncogenes. The continuous curve represents the expected curve between ENC and GC contents under random codon usage.

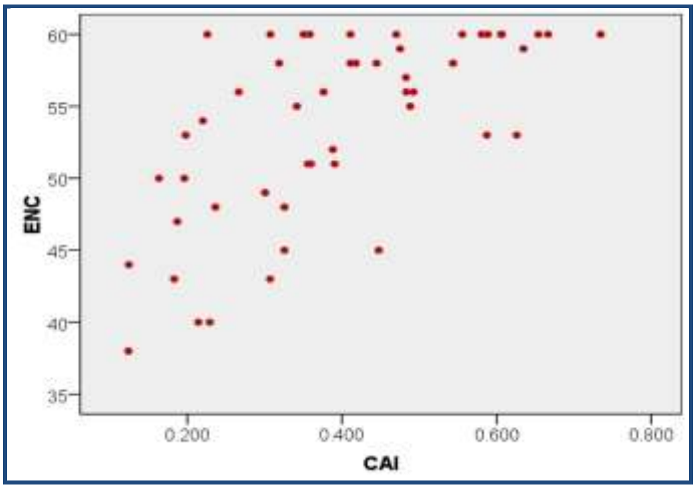


Figure 5: Correlation between Effective number of codons (ENC) and Codon adaptation index (CAI).

Codon usage bias and correlation with GC3s

The overall percentage of guanine and cytosine contents GC% and adenine and thymine contents AT% on the first, second, and third codon positions of the 50 target oncogenes of human were investigated **Table 2 (see supplementary material).** It can be assumed that the evolution of codon usage might be either controlled by natural selection or by mutation pressure. To determine the extent of the role of these two evolutionary forces on the codon usage pattern of human oncogene, we performed correlation analysis between different nucleotide constraints. First we calculated the GC content at different codon positions **(Figure 1)** and it was found that the GC content at each codon position varies among the genes. Finally, compared GC content at first codon position (GC1) and second codon positions (GC2) with that of third codon positions (GC3s) and observed a significant positive correlation (r=0.45, P<0.01) **(Figure 2)**, that reveals base compositions are prone to the result of mutation pressure rather than natural selection, since at all codon positions its effects are present.

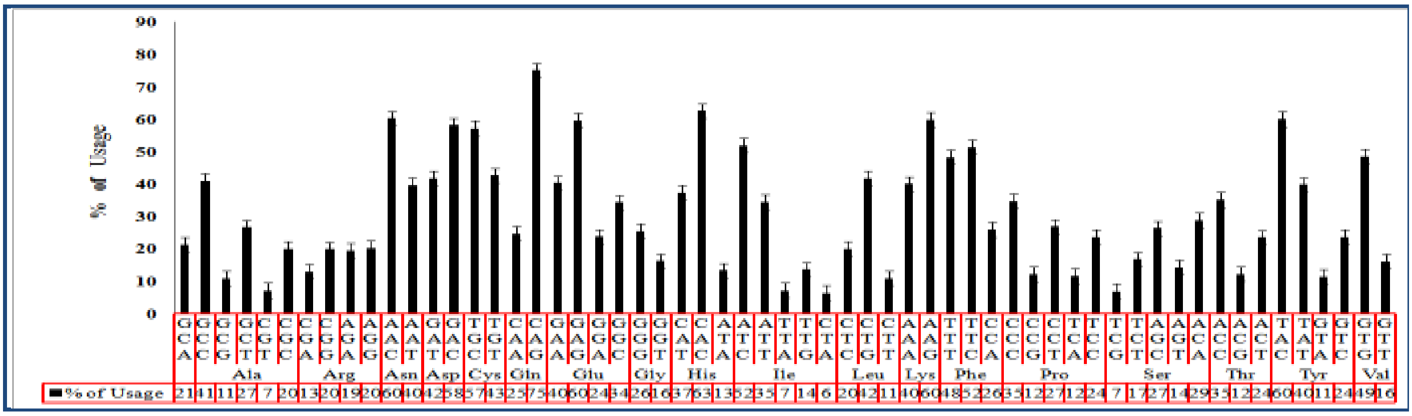


Figure 6: Frequency of highest and least used codons among the 50 cds selected for the present study.

Effective number of codons and its relationship with GC3s values

The average ENC value used by the oncogenes was found to be 53.74 with a range of 38 to 60. Thirty eight oncogenes had ENC values in the range 50-60, 11 in the range 40-50 and 1 between 38 and 40 **(Figure 3)**. Therefore, codon usage bias is in most cases little, although some variation is evident. Moreover the GC3 values were found to range from 0.3 to 0.1. We calculated the correlation coefficient between ENC and GC3s values. The results showed that the ENC value was strongly negatively correlated with the GC3s values of each gene (r = -0.771, P < 0.01). These calculations suggested that genes with higher GC3s values and lower ENC values had strong bias. Finally, we plotted the ENC against the GC3% values to investigate the general codon usage variation with different GC content of each gene **(Figure 4) [7]**. The continuous curve represents the expected positions of genes where GC3 values are the only determinant factor shaping the codon usage pattern. Most genes were found to be located on or above the reference line, representing that the codon usage pattern was only determined by GC3 values. Moreover, some genes located above the reference line, indicates that GC3 is not the only factor for shaping the codon usage pattern other factors like nucleotide composition, may be involved for these genes.

Level of oncogene expression and codon bias

The level of expression of oncogene was measured through codon adaptation index (CAI) values **[10, 15]**, which varied from 0.124 to 0.735 with the mean of 0.395 and standard deviation of 0.159. The CAI value indicates that most of the genes selected for the present study are highly expressive in nature. Moreover, a significant negative correlation was observed between CAI & GC3s (r=-0.489, P<0.01) and CAI & GC content (r=-0.463, P<0.01). Furthermore, significant positive correlation was also observed in between ENC and CAI (r= 0.644, P<0.01) **(Figure 5)**. The results revealed that codon usage pattern determines the level of all expression in human and highly expressed genes have high GC contents and a greater extent of codon usage bias. We also calculated the frequency of the occurrence of synonymous codons for the amino acids. The frequency was allied with statistical analysis to find out the highest and lowest frequently used codon (**Figure 6**). Relative synonymous codon usage (RSCU) values for each synonymous codon were calculated to find out the highest and least abundant codons. The results of our analysis indicate that the highest abundant codon is CTG for Leucine and GTT for Valine. Least abundant codons are GTC, ACT, TCG, CTA, and ATA for amino acid valine, threonine, serine, leucine and isoleucine, respectively **(Figure 7)**.

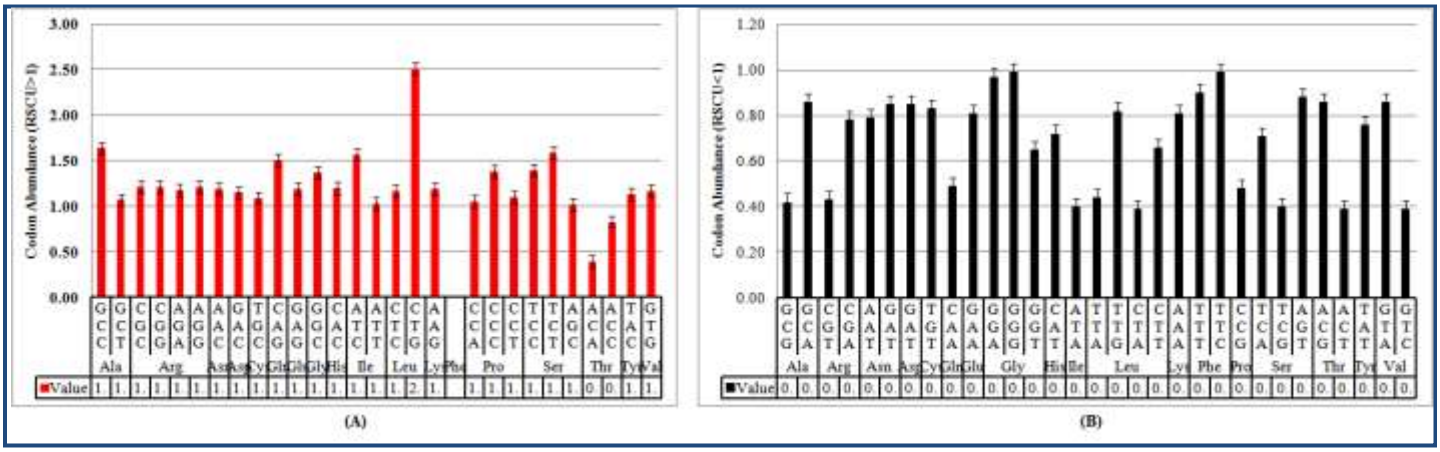


Figure 7: Relative synonymous codon usage and codon usage bias among the selected 50 cds. (A): Most abundant codons, RSCU > 1. (B): Least abundant codons, RSCU < 1.

Discussion:

In brief, we analyzed the codon usage pattern and the key genetic factors playing decisive role in determining the pattern of codon usage for the fifty oncogenes. Based on the hypothesis that gene expressivity and codon composition are strongly correlated, the codon adaptation index has been defined to provide an intuitively meaningful measure of the extent of the codon preference in a gene. The present study was carried out to analyze the CAI, Fop, ENC, RSCU, base composition for the oncogenes, and also to find out the level at which the above mentioned genetic factors are involved in the formation of codon usage pattern. As per our mentioned objectives in this present study, we selected fifty oncogenes from *Homo sapiens* for CUB analysis. The accurate coding sequences having correct initial and termination codons were retrieved using a program in perl, developed by us. After analyzing the cds sequences it was found that 70% of the cds selected for the study are rich in GC. We also predicted the heterogeneity of codon usage by analyzing the effective number of codons (ENC). We also measured the variation of codon usage bias among the oncogenes, further confirmed by the distributions of GC content at the third synonymous codon positions. These results indicate that apart from compositional constraints, other trends might influence the overall codon usage variation among the oncogenes. We calculated the CAI values for the oncogenes and it was found that seventy five percent of the cds selected from *Homo sapiens* qualify as highly expressed genes. We analyzed normalized AT and GC frequency at each codon site. Significant correlation was observed between gene expression as measured by CAI and GC content at any codon site. Among all GC3s showed highest correlation (-0.489) with gene expression. The frequency of the occurrence of each synonymous codon for the amino acids was calculated. The frequency was allied with statistical analysis to find out the highest and lowest frequently used codon. At the end of our frequency analysis we found that AAC, GAC, TGC, CAG, GAG, CAC, AAG and TAC are the codons used most frequently among cds sequence of oncogenes.

Conclusion:

The mutation pressure is the major determining factor in shaping the codon usage pattern of oncogenes rather than natural selection since its effects are present at all codon positions. The results revealed that codon usage bias determines the level of oncogene expression in human. Highly expressed oncogenes had rich GC contents with high degree of codon usage bias.

Acknowledgments:

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

References:

[1] Angov E, *Biotechnol J*. 2011 **6**: 650 [PMID: 21567958]
[2] Hooper SD & Berg OG, *Nucleic Acids Res*. 2000 **28**: 3517 [PMID: 10982871]
[3] Lavner Y & Kotlar D, *Gene* 2005 **345**: 127 [PMID: 15716084]
[4] Hershberg R & Petrov DA, *Annu Rev Genet*. 2008 **42**: 287 [PMID: 18983258]
[5] Cai MS *et al. Intervirology* 2009 **52**: 266 [PMID: 19672100]
[6] Zhou T *et al. Biosystems* 2005 **81**: 77 [PMID: 15917130]
[7] Wright F, *Gene* 1990 **87**: 23 [PMID: 2110097]
[8] Li ZP *et al. Virol Sin*. 2010 **25**: 329 [PMID: 20960179]
[9] Naya H *et al. FEBS Lett*. 2001 **501**: 127 [PMID: 11470270]
[10] Gupta SK *et al. J Biomol Struct Dyn*. 2004 **21**: 527 [PMID: 14692797]
[11] Stenico M *et al. Nucleic Acids Res*. 1994 **22:** 2437 [PMID: 8041603]
[12] Sharp PM & Li WH, *Nucleic Acids Res*. 1987 **15**: 1295 [PMID: 3547335]
[13] Ikemura T, *J Mol Biol*. 1981 **151**: 389 [PMID: 6175758]
[14] Sharp PM & Li WH, *Nucleic Acids Res*. 1986 **14**: 7749 [PMID: 3534792]
[15] Behura SK & Severson DW, *PLoS One* 2012 **7**: e43111 [PMID: 22912801]

Edited by P Kangueane
Citation: **Mazumder *et al.*** Bioinformation 10(5): 256-262 (2014)

License statement: This is an open-access article, which permits unrestricted use, distribution, and reproduction in any medium, for non-commercial purposes, provided the original author and source are credited

Supplementary material:

Table 1: The information of 50 Oncogenes used in this study with accession number and gene length

SL.NO.	GENES	ACCESSION NO.	GENE LENGTH Input CDS bp)
1	ABL2	DQ009672.1	3549
2	AKT2	BC063421.1	444
3	ATF1	BC029619.1	816
4	BCL11A	GU324937.1	2508
5	BCL2	AY220759.1	720
6	BCL3	M31732.1	1341
7	BCL6	EU883531.1	1953
8	BCR	U07000.1	3816
9	BRAF	EU600171.1	2301
10	CARD11	BC111719.1	3444
11	CBLC	BC006122.1	678
12	CCND1 (Cyclin D1)	M64349.1	888
13	CCND2(Cyclin D2)	M90813.1	870
14	CCND3(Cyclin D3)	M90814.1	879
15	CTNNB1	AB451264.1	2350
16	DDB2	AY220533.1	1284
17	DDIT3	AY880949.1	510
18	DDX6	BC039826.1	564
19	DEK	BC035259.1	1128
20	EGFR	U48722.1	1218
21	ELK4	BC063676.1	1218
22	ERBB2	AY208911.1	3768
23	ETV4	BC016623.1	1455
24	ETV6	BC043399.1	1359
25	EVI1	GQ352634.1	3156
26	EWSR1	BC011048.1	1968
27	FEV	BC023511.2	717
28	FGFR1	AY585209.1	2469
29	FGFR1OP	BC037785.1	450
30	FGFR2	M97193.1	2469
31	FUS	CR456747.1	1581
32	GOLGA5	BC023021.1	2196
33	GOPC	KF420123.1	1224
34	HMGA1	BC067083.1	324
35	HRAS	EF015887.1	513
36	IRF4	BC015752.1	1356
37	JUN	J04111.1	997
38	KIT	U63834.1	2931
39	KRAS	JX512447.1	570
40	LCK	M36881.1	1530
41	LMO2	BC034041.1	477
42	MAFB	BC036689.1	972
43	MAML2	AY040322.1	3462
44	MDM2	GQ848196.1	1401
45	MET	J02958.1	4227
46	MITF	BC065243.1	1260
47	MLL	AY373585.1	11910
48	MPL	M90103.1	1740
49	MYB	AF104863.1	1923
50	MYC	AY214166.1	1320

Table 2: GC content and the AT contents at different codon positions in the complete coding regions of 50 oncogenes

SL.NO.	GENES	GC %	GC1 %	GC2 %	GC3 %	AT1 %	AT2	AT3 %
1	ABL2	52.1	56.8	47.8	51.6	43.2	52.2	48.4
2	AKT2	57.4	50.7	48	73.6	49.3	52	26.4
3	ATF1	44.1	50.4	47.1	34.9	49.6	52.9	65.1
4	BCL11A	58.6	57.9	48.3	69.6	42.1	51.7	30.4
5	BCL2	65.8	64.2	50.4	82.9	35.8	49.6	17.1
6	BCL3	68.6	70.7	55	80.1	29.3	45	19.9
7	BCL6	56.8	56.1	47.3	67.1	43.9	52.7	32.9
8	BCR	61.4	58.8	42.4	83.1	41.2	57.6	16.9
9	BRAF	46.2	53.1	44.1	41.3	46.9	55.9	58.7
10	CARD11	58.7	58.5	36.1	81.4	41.5	63.9	18.6

11	CBLC	59	65	50.9	61.1	35	49.1	38.9
12	CCND1 (Cyclin D1)	61.1	60.5	36.1	86.8	39.5	63.9	13.2
13	CCND2(Cyclin D2)	57.8	61.7	35.5	76.2	38.3	64.5	23.8
14	CCND3(Cyclin D3)	63.1	64.8	46.4	78.2	35.2	53.6	21.8
15	CTNNB1	48	62.3	40.5	41	37.7	59.5	59
16	DDB2	54.4	51.9	46.3	65.2	48.1	53.7	34.8
17	DDIT3	53.7	63.5	44.1	53.5	36.5	55.9	46.5
18	DDX6	45.6	48.9	43.1	44.7	51.1	56.9	55.3
19	DEK	39.9	44.4	35.1	40.2	55.6	64.9	59.8
20	EGFR	53	50.7	43.1	65.3	49.3	56.9	34.7
21	ELK4	44.7	46.6	41.6	46.1	53.4	58.4	53.9
22	ERBB2	60.8	63	46.1	73.2	37	53.9	26.8
23	ETV4	59.8	64.7	44.7	69.9	35.3	55.3	30.1
24	ETV6	53.2	56.5	40.6	62.5	43.5	59.4	37.5
25	EVI1	45.5	50.1	41.6	44.8	49.9	58.4	55.2
26	EWSR1	56.9	62.5	61	47.3	37.5	39	52.7
27	FEV	71	68.6	53.6	90.8	31.4	46.4	9.2
28	FGFR1	57.1	56.4	41.8	73.3	43.6	58.2	26.7
29	FGFR1OP	41.3	49.3	39.3	35.3	50.7	60.7	64.7
30	FGFR2	51.2	53.2	40.6	59.9	46.8	59.4	40.1
31	FUS	57.4	60.2	61.3	50.9	39.8	38.7	49.1
32	GOLGA5	43.3	54.2	35.9	39.8	45.8	64.1	60.2
33	GOPC	46.8	63.5	36.5	40.4	36.5	63.5	59.6
34	HMGA1	61.1	55.6	57.4	70.4	44.4	42.6	29.6
35	HRAS	60.4	57.9	41.5	81.9	42.1	58.5	18.1
36	IRF4	57.7	58.2	42.7	72.1	41.8	57.3	27.9
37	JUN	64.5	85.8	61.7	45.8	14.2	38.3	54.2
38	KIT	44.9	48.4	38.6	47.8	51.6	61.4	52.2
39	KRAS	38.1	47.4	33.7	33.2	52.6	66.3	66.8
40	LCK	58.1	58.8	40.2	75.3	41.2	59.8	24.7
41	LMO2	55.8	54.7	38.4	74.2	45.3	61.6	25.8
42	MAFB	67	66.7	46.6	87.7	33.3	53.4	12.3
43	MAML2	51.9	58.1	45.4	52.3	41.9	54.6	47.7
44	MDM2	40.3	50.7	39.6	30.4	49.3	60.4	69.6
45	MET	43.5	48.1	38.3	44.2	51.9	61.7	55.8
46	MITF	48.5	52.6	39	53.8	47.4	61	46.2
47	MLL	48.1	51.3	49.1	43.9	48.7	50.9	56.1
48	MPL	59.7	60.3	51.4	67.4	39.7	48.6	32.6
49	MYB	48.6	53.5	41.3	45.6	46.5	58.7	54.4
50	MYC	58.8	55.5	44.5	76.4	44.5	55.5	23.6

Supplementary material contains two tables (Table 3 and Table 4). Table 3 contains the frequency of optimal codons (Fop) in the complete coding region of 50 oncogenes. Table 4 contains relative synonymous codon usage values (RSCU) of 50 cds selected in this study.

RESEARCH ARTICLE

Gaining Insights into the Codon Usage Patterns of *TP53* Gene across Eight Mammalian Species

Tarikul Huda Mazumder, Supriyo Chakraborty*

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

* supriyoch_2008@rediffmail.com

Abstract

TP53 gene is known as the "guardian of the genome" as it plays a vital role in regulating cell cycle, cell proliferation, DNA damage repair, initiation of programmed cell death and suppressing tumor growth. Non uniform usage of synonymous codons for a specific amino acid during translation of protein known as codon usage bias (CUB) is a unique property of the genome and shows species specific deviation. Analysis of codon usage bias with compositional dynamics of coding sequences has contributed to the better understanding of the molecular mechanism and the evolution of a particular gene. In this study, the complete nucleotide coding sequences of *TP53* gene from eight different mammalian species were used for CUB analysis. Our results showed that the codon usage patterns in *TP53* gene across different mammalian species has been influenced by GC bias particularly GC_3 and a moderate bias exists in the codon usage of *TP53* gene. Moreover, we observed that nature has highly favored the most over represented codon CTG for leucine amino acid but selected against the ATA codon for isoleucine in *TP53* gene across all mammalian species during the course of evolution.

OPEN ACCESS

Citation: Mazumder TH, Chakraborty S (2015) Gaining Insights into the Codon Usage Patterns of *TP53* Gene across Eight Mammalian Species. PLoS ONE 10(3): e0121709. doi:10.1371/journal.pone.0121709

Academic Editor: Yury E Khudyakov, Centers for Disease Control and Prevention, UNITED STATES

Received: November 9, 2014

Accepted: February 14, 2015

Published: March 25, 2015

Copyright: © 2015 Mazumder, Chakraborty. This is an open access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Data Availability Statement: All relevant data are within the paper.

Funding: The authors have no support or funding to report.

Competing Interests: The authors have declared that no competing interests exist.

Introduction

TP53 gene encodes tumor protein p53 which is known as the "guardian of the genome" as it plays a vital role in maintaining genomic stability by preventing mutation in the genome [1]. The p53 primarily acts as transcription factor and stands out as a key player in restricting tumor cell invasion that includes the ability to induce cell cycle arrest, DNA repair, senescence and apoptosis [2]. Mutation in p53 results in abnormal proliferation of cells that leads to the formation of tumor development and so *TP53* gene is cataloged as tumor suppressor gene [3].

The nucleus of a cell is the main store house of tumor protein p53 where it binds to DNA. When any damage occurs in the DNA of a cell by some external agents like toxic chemicals, radiation, exposure to sun light or ultra violet rays, p53 plays the crucial role in activating other genes and inhibits cell cycle to repair the damage [4]. In case of failure of DNA repair, the tumor protein p53 prevents the cell from dividing and provokes signals to a wide variety of genes that contribute to TP53 mediated cell death *i.e.*, apoptosis [5].

Unequal usage of synonymous codons that encode the same amino acid during translation of a gene into protein is known as codon usage bias (CUB). Some codons in a synonymous group are used more frequently whereas others less frequently in the genome of an organism [6,7]. CUB is a unique property of the genome and it may vary between genes from the same genome or within a single gene [8,9].

The advent of whole genome sequencing in different organisms and the easily accessible nucleotide database from NCBI (GenBank) have attracted much attention of the scientific community to study CUB in gaining clues for understanding the molecular evolution of genes and genome characterization.

Previously, several studies were conducted on synonymous codon usage bias in a wide variety of organisms including prokaryotes and eukaryotes [10–16], and till date in many organisms the codon usage patterns have been interpreted for diverse reasons. Many genomic factors such as gene length, GC-content, recombination rate, gene expression level, or modulation in the genetic code are associated with CUB in different organisms [17–21]. In general, compositional constraints under natural selection or mutation pressure are considered as major factors in the codon usage variation among different organisms [8,22–25]. Moreover, studies revealed that mutation pressure, natural or translational selection, secondary protein structure, replication and selective transcription, hydrophobicity and hydrophilicity of the protein and the external environment play a major role in the codon usage pattern of organisms [26]. In unicellular and multicellular organisms it was observed that, preferred synonymous codons/optimal codons with abundant tRNA gene copy number rise with gene expression level within the genome that supports selection on high codon bias confirmed by positive correlation between optimal codons and tRNA abundance [18,22,27]. Urrutia and Hurst (2003) reported weak correlation between gene expression level and codon usage bias within human genome though not related with tRNA abundance [19]. However, Comeron (2004) observed that in human genome, highly expressed genes have preference towards codon bias favoring codons with most abundant tRNA gene copy number compared to less highly expressed genes [28].

The study of codon usage bias acquires significance in biology not only in the context of understanding the process of evolution at molecular level but also in designing transgenes for increased expression, discovering new genes [29] based on nucleotide compositional dynamics, detecting lateral gene transfer and for analyzing the functional conservation of gene expression [30]. Codon usage bias may be superimposed on the effect of natural selection. The amount of protein produced from the mRNA transcript may vary significantly since the translational properties of alternate synonymous codons are not equivalent [31]. Several studies have further shown that codon usage bias is associated with highly expressed genes as some codons are used more often than others in the coding sequences [32]. Moreover, literature suggested that a gene can be epitomized not only by the sequence of its amino acid but also by its codon usage patterns shaped by the balance between mutational bias and natural selection [33]. As a consequence of selection pressure within a gene, differentiation in codon bias may arise between species of the same genus.

The present study was undertaken in order to perform a comparative analysis of codon bias and compositional dynamics of codon usage patterns in *TP53* gene across eight different mammalian species using nucleotide chemistry (GC contents) and several genetic indices namely effective number of codons (*ENC*), relative synonymous codon usage (*RSCU*), and relative codon usage bias (*RCBS*) etc. Our analysis has given a novel insight into the codon usage patterns of *TP53* gene that would facilitate better understanding of the structural, functional as well as evolutionary significance of the gene among the mammalian species.

Results and Discussion

Codon usage patterns in *TP53* genes across mammalian species

Correlation coefficient between codon usage and GC bias was analyzed using heat map (Fig. 1) in order to find out the relationship between the codon usage variation and the GC constraints among the selected coding sequences of *TP53* genes. In our analysis, nearly all codons ending with G/C base showed positive correlation with GC bias and nearly all A/T—ending codons showed negative correlation with GC bias. But, 8 G/C—ending codons (ATC, ACG, TAC, TTG, TCC, CAC, GTG, GGG) showed negative correlation with GC bias whereas 6 A/T-ending codons (AAT, ATT, TGT, CGA, GTA, GGA) showed positive correlation with AT bias although statistically not significant ($p>0.05$). Two G-ending codons *i.e.* TCG for serine and CTG for leucine amino acid showed strong positive correlation ($p<0.01$) with GC_{3s}, indicating that codon usage has been influenced by GC bias due to GC_{3s}. Interestingly, we observed that the codon

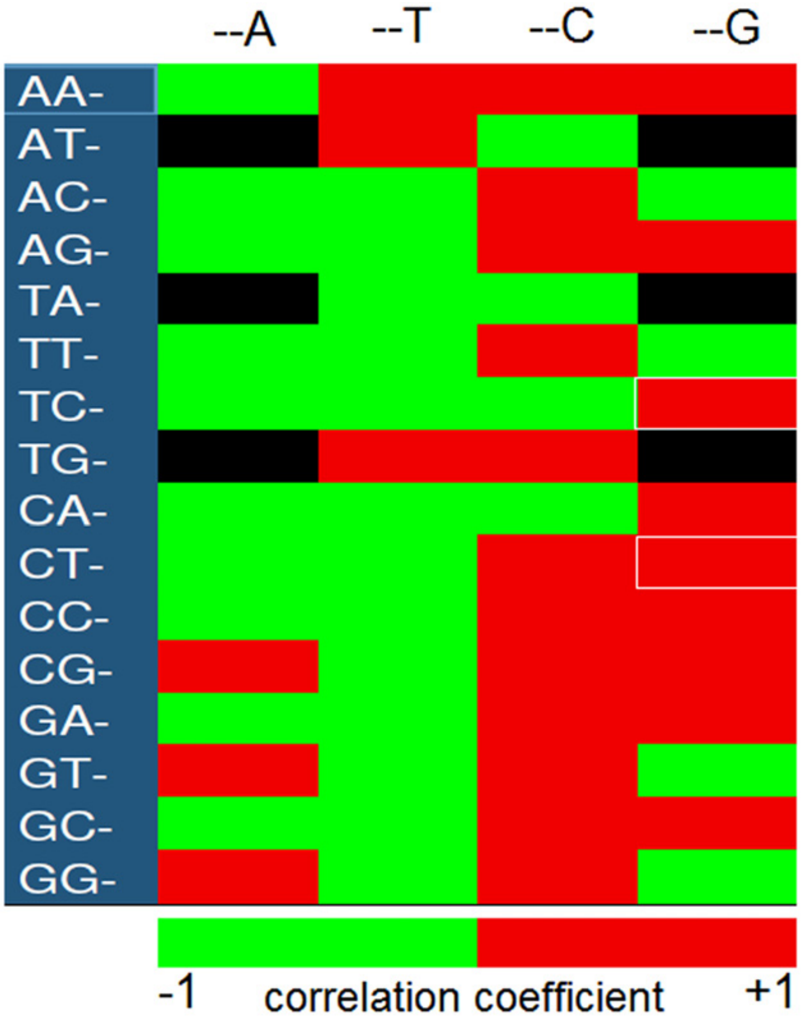


Fig 1. Heat maps of correlation coefficient of codons with GC_3. The color coding red represents the positive correlation, green as negative correlation. The black fields are stop codons (TAA, TAG, TGA) and non-degenerate codons (ATG, TGG) together with ATA codons for isoleucine which was altogether absent in *TP53* gene across mammalian species. White marked rectangular boxes are the codons that showed strong positive correlation with GC_{3s}.

doi:10.1371/journal.pone.0121709.g001

Table 1. Nucleotide composition analysis in the coding sequences of *TP53* gene.

Sl. No.	A	T	G	C	A_3	T_3	G_3	C_3	AT %	GC %	GC_1 %	GC_2 %	GC_3 %	AT_3 %	GC_{12} %
1	262	224	308	358	55	69	122	138	42.2	57.8	56.8	49.0	67.7	32.3	52.9
2	267	234	294	366	56	79	117	135	43.2	56.8	56.8	48.6	65.1	34.9	52.7
3	261	219	321	360	53	66	134	134	41.3	58.7	57.4	49.4	69.3	30.7	53.4
4	266	229	300	351	55	78	115	134	43.2	56.8	56.5	48.7	65.2	34.8	52.6
5	264	204	321	384	57	58	125	151	39.9	60.1	59.8	49.9	70.6	29.4	54.9
6	289	244	302	341	73	89	112	118	45.3	54.7	57.4	48.0	58.7	41.3	52.7
7	282	232	301	367	68	82	112	132	43.5	56.5	58.1	49.5	61.9	38.1	53.8
8	276	234	307	365	62	86	115	131	43.1	56.9	59.1	49.0	62.4	37.6	54.1
Mean	270.9	227.5	307	361.5	59.9	75.9	119	134.1	42.7	57.3	57.7	49.0	65.1	34.9	53.4
SD	10.30	12.05	9.79	12.6	7.18	10.6	7.60	9.06	0.016	0.016	0.011	0.005	0.040	0.040	0.008

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

doi:10.1371/journal.pone.0121709.t001

ATA encoding isoleucine amino acid was not favored by natural selection in *TP53* genes across mammalian species during the course of evolution. Thus, scanning the codon usage pattern provides the basis of the mechanism for synonymous codon usage bias and has both practical as well as theoretical significance in gaining clues of understanding molecular biology [34].

G/C-ending codons are favored by *TP53* gene across mammalian species

We analyzed the nucleotide composition of coding sequences from *TP53* genes (Table 1) which revealed that mean value of C (361.50) was the highest followed by G (306.75), A (270.88) and T (227.50) among all the selected mammals. The mean percentage of GC and AT compositions was 57.3% and 42.7% respectively. Thus, the overall nucleotide composition suggested that the nucleotide C and G occurred more frequently compared to A and T in the coding sequences of *TP53* gene across the mammalian species. The nucleotide composition at the third position of codon (A_3,T_3,G_3,C_3) showed that the mean values of C_3 and G_3 were the highest followed by T_3 and A_3. The GC_3 values (ranged from 58.7%-70.6%, mean = 65.1%, SD = 0.040) was compared with that of AT_3 values (ranged from 29.4%-41.3%, mean = 34.9%, SD = 0.040) in the coding sequences of *TP53* genes. The average percentage of GC contents at the first and second codon positions (GC_{12}) was found in the range of 52.6% to 54.9% with a mean value of 53.4% and a standard deviation (SD) of 0.008. Therefore, nucleotide composition analysis suggested that GC—ending codons might be preferred over AT—ending codons in the coding sequences of *TP53* genes across the selected mammalian species. Further, we calculated the occurrence of frequently used optimal codons (Fop) for each amino acid as suggested by Lavner and Kotler (2005) [14]. The frequency was allied with statistical analysis to find out the highest and lowest frequently used codon. Our results showed that the most frequently used codons were G/C—ending for the corresponding amino acid (Fig. 2) in *TP53* genes across mammalian species.

Relative synonymous codon usage in *TP53* gene across mammals

The relative synonymous codon usage values of 59 codons for *TP53* gene across eight mammalian species were analyzed excluding the codons ATG (methionine) and TGG (tryptophan). In our calculation RSCU value greater than 1.0 represents that the particular codon is used more

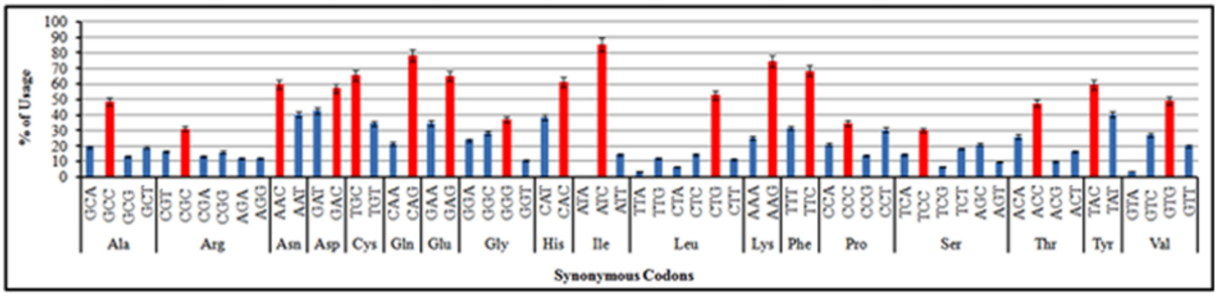


Fig 2. Overall frequency of optimal and non optimal codon used in *TP53* genes among mammals. Red color coding represents optimal used codons with corresponding amino acid.

doi:10.1371/journal.pone.0121709.g002

frequently and less than 1.0 represents the less frequently used codon for the corresponding amino acid. The RSCU value greater than 1.6 indicates over represented codon for the corresponding amino acid. The overall RSCU values in the selected coding sequences of *TP53* gene revealed that 25 codons were most frequently used among the 59 codons and the most predominantly used codons were G/C—ending compared to A/T—ending (Table 2). Besides, it was observed that C—ending codon was mostly favored compared to G—ending codon in the coding sequence of *TP53* gene among the selected mammalian species. Our results showed marked similarities as reported by Dass *et al*., (2012) in serotonin receptor gene family from different mammalian species [35]. Further, clustering analysis of RSCU values (Fig. 3) depicted that the codon GCC, CGC (except *Rattus norvegicus*), ATC (except *Tupaia chinensis*), CTG, ACC (except *Rattus norvegicus*, *Macaca mulatta*), GTG (except *Felis catus*) were displayed as the over represented codons (RSCU>1.6). The highest RSCU value was found for the codon CTG for leucine amino acid in all *TP53* genes across mammalian species. The codon ATA showed the RSCU value zero because natural selection has not favored this codon in *TP53* gene across mammalian species.

Codon usage patterns of *TP53* gene correspond to phylogeny of mammalian species

We have performed a neighbor joining tree analysis based on Kimura 2-parameter (K2P) distances of the coding sequences in *TP53* gene across mammalian species (Fig. 4). We observed that codon usage patterns in *TP53* genes have significant similarities among the closely related mammalian species. The gene *TP53* in *H. sapiens* showed resemblance to the *TP53* gene in *M. mulatta*, Similarly, *TP53* of *F. catus* resembled to that of *C. lupus* and *M.unguiculatus* with *R. norvigicus*. Generally, genes with similar functions exhibit similar patterns of codon usage frequency [36]. Our analysis further suggested that the coding sequence of *TP53* gene share similar patterns of codon usage bias across eight mammalian species.

Selection pressure over *TP53* gene across mammalian species

The ENC values of the coding sequences ranged from 52 to 59 with a mean of 55.5±2.33 indicating relatively smaller variation in the codon usage of *TP53* gene across eight mammalian species. However, the GC_{3s} values ranged from 0.59 to 0.71 with a mean value of 0.65±0.040. Significant negative correlation (Pearson r = -0.979, p<0.01) was observed between ENC and GC_{3s}. Moreover, a plot of ENC *vs* GC_{3s} revealed that the ENC values had negative correlation with the GC_3 content (Fig. 5) and comparatively lower ENC was linked to higher GC_{3s} values. All the selected coding sequences of *TP53* gene across the selected mammalian species had

Table 2. Overall relative synonymous codon usage patterns (RSCU) for *TP53* gene among eight mammalian species.

Amino Acid	Codon	N	RSCU[a]	Amino Acid	Codon	N	RSCU[a]
Ala	GCA	37	0.78	Leu	TTA	9	0.2
	GCC*	91	1.95		TTG	31	0.71
	GCG	26	0.53		CTA	17	0.38
	GCT	36	0.76		CTC	39	0.88
Arg	CGT	34	0.98		CTG*	143	3.18
	CGC*	65	1.86		CTT	30	0.68
	CGA	27	0.78	Lys	AAA	42	0.51
	CGG	33	0.96		AAG*	125	1.5
	AGA	24	0.71	Phe	TTT	29	0.63
	AGG	25	0.73		TTC*	61	1.37
Asn	AAC*	68	1.2	Pro	CCA	70	0.84
	AAT	44	0.81		CCC*	115	1.4
Asp	GAT	67	0.86		CCG	44	0.55
	GAC*	82	1.15		CCT*	100	1.22
Cys	TGC*	60	1.31	Ser	TCA	45	0.86
	TGT	32	0.69		TCC*	94	1.81
Gln	CAA	23	0.43		TCG	21	0.4
	CAG*	82	1.57		TCT*	58	1.09
Glu	GAA	84	0.7		AGC*	66	1.26
	GAG*	157	1.31		AGT*	30	0.57
Gly	GGA	41	0.95	Thr	ACA*	47	1.05
	GGC*	49	1.13		ACC*	84	1.91
	GGG*	65	1.5		ACG	18	0.4
	GGT	18	0.43		ACT	29	0.65
His	CAT	32	0.78	Tyr	TAC*	45	1.19
	CAC*	50	1.23		TAT	32	0.81
Ile	ATA	0	0	Val	GTA	5	0.15
	ATC*	61	2.58		GTC*	38	1.1
	ATT	11	0.43		GTG*	69	1.98
					GTT	28	0.8

[a] mean values of RSCU based on the synonymous codon usage frequencies of *TP53* gene, N: Total number of preferred codons, *RSCU>1.

doi:10.1371/journal.pone.0121709.t002

a higher predominance of G/C—ending codons. It suggested that GC_{3s} values determined the codon usage pattern in the coding sequences of *TP53* gene [33]. Nabiyouni *et al*., (2013) reported that eukaryotic organisms with very high GC-contents have high GC_3-composition while organisms with low GC-content have low GC_3-composition in the genome [37]. We also calculated GC_3 skew values which ranged from 0.000 to -0.094, indicating that GC_3 composition at the third position of codon might have played an important role in the codon usage bias [38]. Negative GC skew was observed in all the coding sequences of *TP53* gene which revealed that the abundance of C over G [39]. In addition, lower values of the frequency of optimal codons (FOP) and the effective number of codons (ENC) along with higher GC contents suggested that a moderate bias exists in the usage of synonymous codons [33] for *TP53* gene in different mammalian species. Predominant codon usage bias was observed in *TP53* gene of *M.unguiculatus* compared to other mammalian species (Table 3).

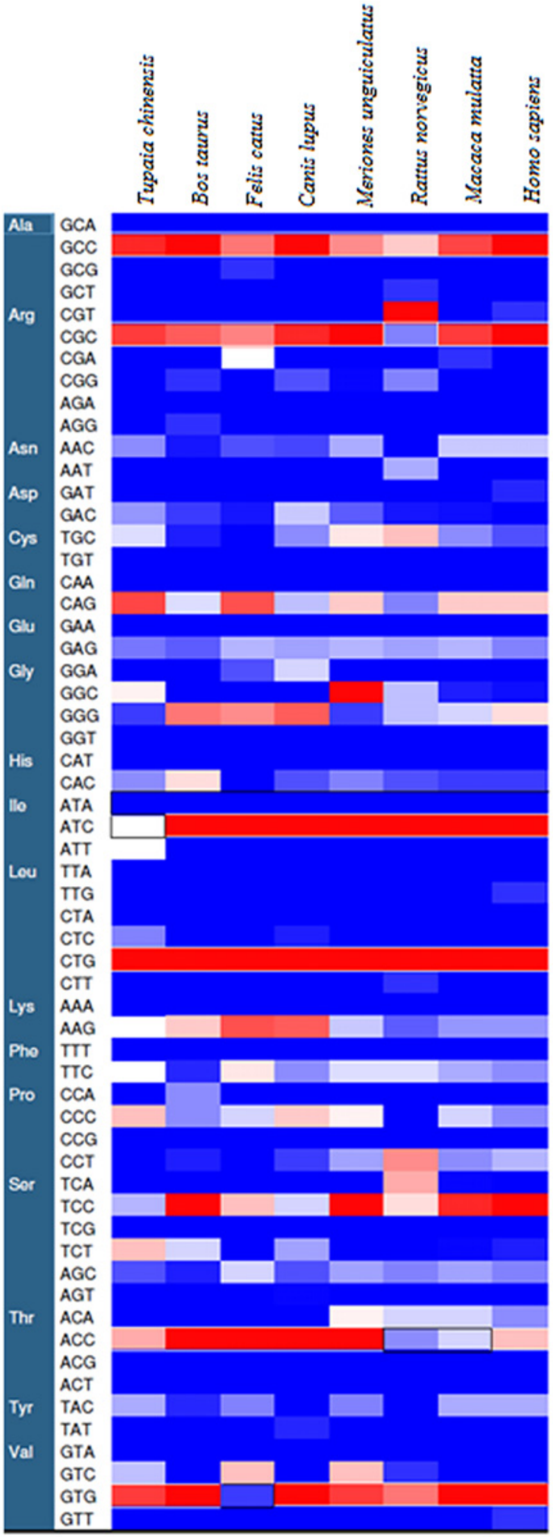


Fig 3. Clustering of RSCU values of each codon among *TP53* gene across mammals. Each rectangular box on the map represents the RSCU value of a codon (shown in rows) corresponding to the *TP53* gene across mammalian species (shown in columns). The intensity of color coding indicates different RSCU values: intensity towards blue RSCU<1, white RSCU>1 and red RSCU>1.6.

doi:10.1371/journal.pone.0121709.g003

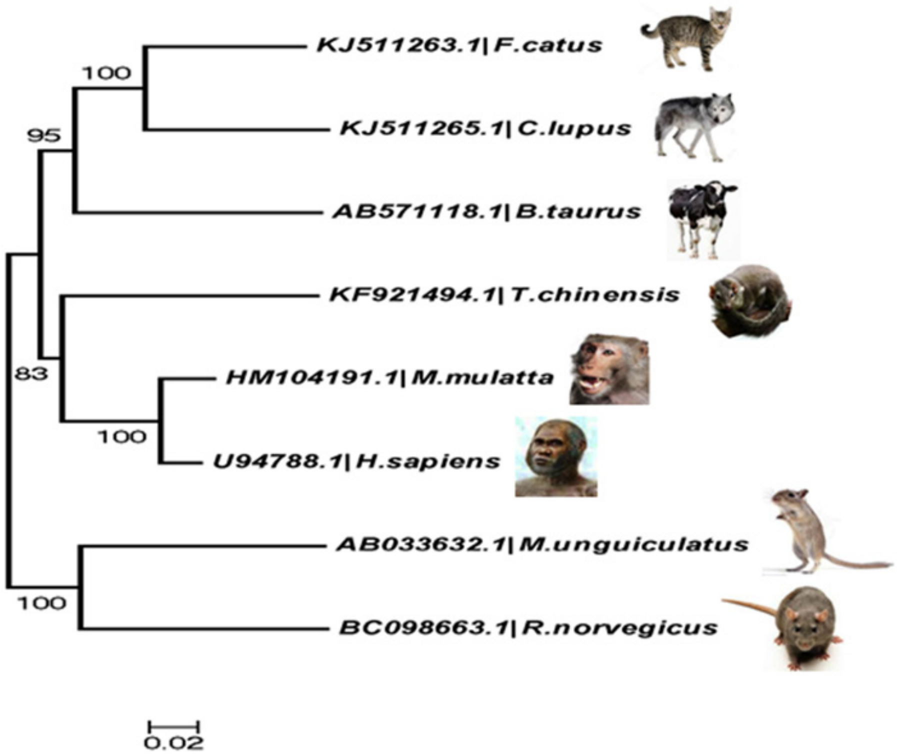


Fig 4. Phylogenetic analysis of the Kimura 2- parameter (K2P) distances of the selected coding sequences among *TP53* genes of different mammalian species. The percentage of replicate trees in which the associated taxa clustered together in the bootstrap test (1000 replicates) is shown next to the branches. 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 Kimura 2-parameter method and are in the units of the number of base substitutions per site. The analysis involved 8 cds sequences. All positions containing gaps and missing data were eliminated. There were a total of 1200 positions in the final dataset. Evolutionary analyses were performed in MEGA6.

doi:10.1371/journal.pone.0121709.g004

RCBS value of a gene can be used as an effective measure of predicting gene expression and its value depends on the patterns of codon usage along with nucleotide compositional bias of a gene [20]. The distribution of *RCBS* values for *TP53* gene across eight mammalian species is shown in figure below (Fig. 6). The *RCBS* values ranged from 0.006 to 0.065 with a mean value of 0.039 and a standard deviation (SD) of 0.021. In our analysis, low mean *RCBS* value suggested that there exists a low codon bias for *TP53* gene associated with low expression level [20].

Conclusions

In brief, our results showed that codon usage in *TP53* gene in mammals has been influenced by GC bias, mainly due to GC_{3s}. The majority of frequently used codons were G/C ending in which C—ending codons were mostly favored compared to G—ending codons for the corresponding amino acid. The most over-represented codon was CTG encoding the amino acid leucine in the *TP53* gene of all the selected mammalian species. We further observed that the codon ATA encoding isoleucine was selected against by nature in *TP53* genes across the mammalian species under study during the course of evolution. The codon usage pattern for *TP53* in *H. sapiens* showed resemblance to that of *M. mulatta*; similarly, *F. catus* to *C. lupus* and *M. unguiculatus* to *R. norvigicus*. Moderate codon bias was observed for the *TP53* gene in different mammalian species.

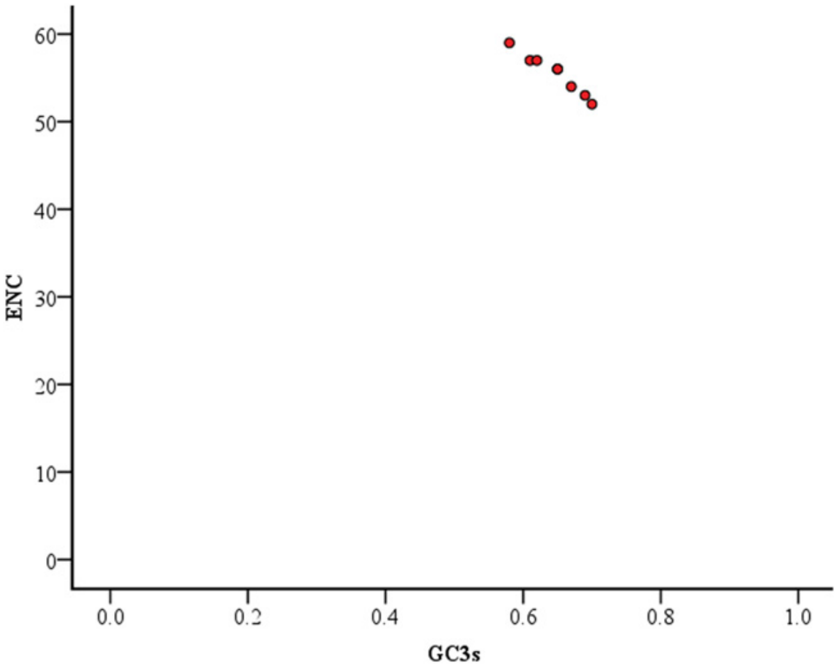


Fig 5. ENC vs GC3s values for *TP53* gene. Red dots represent the ENC and GC_{3s} values of the coding sequences for *TP53* gene across mammalian species.

doi:10.1371/journal.pone.0121709.g005

The codon usage patterns in the coding sequence of *TP53* gene across different mammalian species showed significant similarities, suggesting that the evolutionary pattern might be similar. According to Yang and Nielsen (2008), codon bias in mammals is mainly influenced by mutation bias and the selection on codon bias is weak for nearly neutral synonymous mutations [40]. From the outstanding work of Grantham *et al*., (1980–1981) on "genome hypothesis" it was evident that species specific genes share similar spectrum of codon usage frequency [41,42]. The present study revealed that specific gene of closely related species with similar functions exhibit similar patterns of codon bias across different mammals as evident from the previous work of Dass *et al*., (2012) [35]. To the best of our knowledge, this is the first report on the codon usage pattern in *TP53* gene across the mammalian species. Since our analysis has

Table 3. Codon usage bias indices for *TP53* gene across mammalian species.

MAMMALS	RCBS	ENC	GC_{3s}	FOP	Highest RSCU	GC Skew	GC_3 Skew
Tupaia chinensis	0.006	54	0.68	0.305	CTG (Leu)	-0.08	0.061
Bos taurus	0.015	56	0.65	0.305	CTG (Leu)	-0.11	0.071
Felis catus	0.03	53	0.69	0.305	CTG (Leu)	-0.06	0
Canis lupus	0.042	56	0.65	0.305	CTG (Leu)	-0.08	0.076
Meriones unguiculatus	0.043	52	0.71	0.306	CTG (Leu)	-0.09	0.094
Rattus norvegicus	0.054	59	0.59	0.306	CTG (Leu)	-0.06	0.026
Macaca mulatta	0.063	57	0.62	0.305	CTG (Leu)	-0.10	0.081
Homo sapiens	0.065	57	0.62	0.305	CTG (Leu)	-0.09	0.065

RCBS-Relative codon usage bias, ENC-Effective number of codons, GC_{3s}-GC contents at third positions of codon, FOP-Frequency of optimal codons, RSCU-Relative synonymous codon usage.

doi:10.1371/journal.pone.0121709.t003

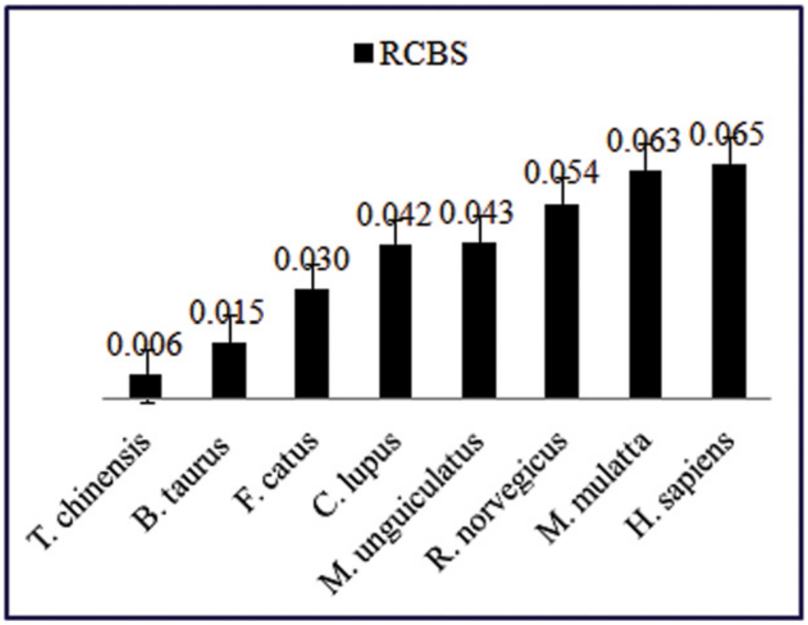


Fig 6. Distribution of RCBS for *TP53* gene across eight mammalian species.

doi:10.1371/journal.pone.0121709.g006

given better insights into the codon usage, it may have theoretical value in further understanding the molecular evolution of *TP53*gene.

Materials and Methods

Sequence Data

The complete nucleotide coding sequences (cds) for TP53 gene having perfect start and stop codon, devoid of any unknown bases (N) and perfect multiple of three bases, were retrieved from National Center for Biotechnology Information (NCBI) GenBank database (http://www.ncbi.nlm.nih.gov). Finally, we selected eight coding sequences for TP53 gene that fulfill the above mentioned criteria in different mammalian species and used in our CUB analysis (Table 4).

Table 4. Eight mammalian species with accession number and length (bp) of coding sequences for *TP53* gene.

CDS NO.	MAMMALS	ACCESSION NO.	Length (bp)
1	*Tupaia chinensis*	KF921494.1	1152
2	*Bos taurus*	AB571118.1	1161
3	*Felis catus*	KJ511263.1	1161
4	*Canis lupus*	KJ511265.1	1146
5	*Meriones unguiculatus*	AB033632.1	1173
6	*Rattus norvegicus*	BC098663.1	1176
7	*Macaca mulatta*	HM104191.1	1182
8	*Homo sapiens*	U94788.1	1182

CDS NO: Coding sequence number.

doi:10.1371/journal.pone.0121709.t004

Prediction of Base Composition Bias

The occurrence of overall frequency of the nucleotide (G+C) at first (GC_1), second (GC_2) and third (GC_3) position of synonymous codons were calculated to quantify the extent of base composition bias. Moreover, we analyzed the skewness for AT, GC and GC_{3s} of each coding sequence to estimate the base composition bias particularly in relation to transcription processes.

Effective Number of Codons (ENC) Analysis

ENC is generally used to quantify the codon usage bias of a gene that is independent of the gene length and number of amino acids [43]. This measure was computed as per Wright (1990) to estimate the extent of CUB exhibited by the coding sequences of *TP53* gene across the selected mammalian species:

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

Where, F_k ($_k$ = 2, 3, 4 or 6) is the average of the F_k values for k-fold degenerate amino acids. The F value denotes the probability that two randomly chosen codons for an amino acid with two codons are identical.

The values of ENC ranged from 20 indicating strong codon bias in the gene using only one synonymous codon for the corresponding amino acid, to 61indicating no bias in the gene using all synonymous codons equally for the corresponding amino acid [43].

Frequency of Optimal Codon (Fop) Analysis

Fop is a measure of codon usage bias in a gene [44]. Fop values represent the ratio of the number of optimal codons used to the total number of synonymous codons [22]. The Fop value ranges from 0.36 for a gene showing uniform codon usage bias to 1 for a gene showing strong codon usage bias [45]. Fop value for each selected coding sequence was calculated using the formula given by Lavner and Kotler (2005) [14].

Relative Synonymous Codon Usage (RSCU) Analysis

RSCU is defined as the observed frequency of a codon divided by the expected frequency if all codons are used equally for any particular amino acid [46]. RSCU values of codons for each of the selected coding sequence of *TP53* gene was calculated as follows:

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

Where, g_{ij} is the observed number of the *i*th codon for the *j*th amino acid which has n_i kinds of synonymous codons [26].

Computation of Gene Expression

Gene expression was estimated through RCBS which can be defined as the overall score of a gene indicating the influence of relative codon bias (RCB) of each codon in a gene [20]. The RCBS value of each coding sequence of *TP53* gene was calculated as follows:

$$w_c^{RCB} = \frac{O_C - E[O_C]}{E[O_C]}$$

where, *Oc* is the observed number of counts of codon *c* of the query sequence and $E[O_c]$ is the expected number of codon occurrences given the nucleotide distribution at three codon positions (b1b2b3) [20].

$$RCB = \exp(\frac{1}{O_{tot}}\sum_{c\in C}\log w_c^{RCB}) - 1$$

[47]

Where, O_{tot} is the total number of codons

Software Used

The above mentioned genetic indices were estimated in a PERL program developed by SC (corresponding author) to measure the CUB on the selected coding sequences of *TP53* genes in different mammalian species. All statistical analyses were carried out using the SPSS software. Cluster analysis (Heat map) of correlation coefficient of codons with GC3 and the RSCU values of codons among the eight mammalian species were clustered using a hierarchical clustering method implemented in NetWalker software [48].

Acknowledgments

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

Author Contributions

Conceived and designed the experiments: SC THM. Performed the experiments: SC THM. Analyzed the data: THM. Wrote the paper: THM.

References

1. Lane DP (1992) Cancer. p53, guardian of the genome. Nature 358: 15–16. PMID: 1614522
2. Rivlin N, Brosh R, Oren M, Rotter V (2011) Mutations in the p53 Tumor Suppressor Gene: Important Milestones at the Various Steps of Tumorigenesis. Genes Cancer 2: 466–474. doi: 10.1177/1947601911408889 PMID: 21779514
3. McBride OW, Merry D, Givol D (1986) The gene for human p53 cellular tumor antigen is located on chromosome 17 short arm (17p13). Proc Natl Acad Sci U S A 83: 130–134. PMID: 3001719
4. Vousden KH, Lu X (2002) Live or let die: the cell's response to p53. Nat Rev Cancer 2: 594–604. PMID: 12154352
5. Carnero A, Hudson JD, Hannon GJ, Beach DH (2000) Loss-of-function genetics in mammalian cells: the p53 tumor suppressor model. Nucleic Acids Res 28: 2234–2241. PMID: 10871344
6. Hooper SD, Berg OG (2000) Gradients in nucleotide and codon usage along Escherichia coli genes. Nucleic Acids Res 28: 3517–3523. PMID: 10982871
7. 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. doi: 10.1371/journal.pone.0043111 PMID: 22912801
8. Duret L, Mouchiroud D (1999) Expression pattern and, surprisingly, gene length shape codon usage in Caenorhabditis, Drosophila, and Arabidopsis. Proc Natl Acad Sci U S A 96: 4482–4487. PMID: 10200288
9. Supek F, Vlahovicek K (2005) Comparison of codon usage measures and their applicability in prediction of microbial gene expressivity. BMC Bioinformatics 6: 182. PMID: 16029499
10. McInerney JO (1998) Replicational and transcriptional selection on codon usage in Borrelia burgdorferi. Proc Natl Acad Sci U S A 95: 10698–10703. PMID: 9724767
11. Ermolaeva MD (2001) Synonymous codon usage in bacteria. Curr Issues Mol Biol 3: 91–97. PMID: 11719972
12. Jenkins GM, Holmes EC (2003) The extent of codon usage bias in human RNA viruses and its evolutionary origin. Virus Res 92: 1–7. PMID: 12606071

13. 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: 155–161. PMID: 15041183

14. Lavner Y, Kotlar D (2005) Codon bias as a factor in regulating expression via translation rate in the human genome. Gene 345: 127–138. PMID: 15716084

15. Liu H, Huang Y, Du X, Chen Z, Zeng X, Chen Y, et al. (2012) Patterns of synonymous codon usage bias in the model grass Brachypodium distachyon. Genet Mol Res 11: 4695–4706. doi: 10.4238/2012.October.17.3 PMID: 23096921

16. Ahn I, Jeong BJ, Son HS (2009) Comparative study of synonymous codon usage variations between the nucleocapsid and spike genes of coronavirus, and C-type lectin domain genes of human and mouse. Exp Mol Med 41: 746–756. doi: 10.3858/emm.2009.41.10.081 PMID: 19561398

17. Palidwor GA, Perkins TJ, Xia X (2010) A general model of codon bias due to GC mutational bias. PLoS One 5: e13431. doi: 10.1371/journal.pone.0013431 PMID: 21048949

18. Duret L (2000) tRNA gene number and codon usage in the C. elegans genome are co-adapted for optimal translation of highly expressed genes. Trends Genet 16: 287–289. PMID: 10858656

19. Urrutia AO, Hurst LD (2003) The signature of selection mediated by expression on human genes. Genome Res 13: 2260–2264. PMID: 12975314

20. Roymondal U, Das S, Sahoo S (2009) Predicting gene expression level from relative codon usage bias: an application to Escherichia coli genome. DNA Res 16: 13–30. doi: 10.1093/dnares/dsn029 PMID: 19131380

21. Karlin S, Mrazek J (1996) What drives codon choices in human genes? J Mol Biol 262: 459–472. PMID: 8893856

22. Ikemura T (1981) Correlation between the abundance of Escherichia coli transfer RNAs and the occurrence of the respective codons in its protein genes: a proposal for a synonymous codon choice that is optimal for the E. coli translational system. J Mol Biol 151: 389–409. PMID: 6175758

23. Li WH (1987) Models of nearly neutral mutations with particular implications for nonrandom usage of synonymous codons. J Mol Evol 24: 337–345. PMID: 3110426

24. Xu C, Cai X, Chen Q, Zhou H, Cai Y, Ben A (2011) Factors affecting synonymous codon usage bias in chloroplast genome of oncidium gower ramsey. Evol Bioinform Online 7: 271–278. doi: 10.4137/EBO.S8092 PMID: 22253533

25. Nair RR, Nandhini MB, Sethuraman T, Doss G (2013) Mutational pressure dictates synonymous codon usage in freshwater unicellular alpha—cyanobacterial descendant Paulinella chromatophora and beta—cyanobacterium Synechococcus elongatus PCC6301. Springerplus 2: 492. doi: 10.1186/2193-1801-2-492 PMID: 24255825

26. Butt AM, Nasrullah I, Tong Y (2014) Genome-wide analysis of codon usage and influencing factors in chikungunya viruses. PLoS One 9: e90905. doi: 10.1371/journal.pone.0090905 PMID: 24595095

27. Akashi H (1995) Inferring weak selection from patterns of polymorphism and divergence at "silent" sites in Drosophila DNA. Genetics 139: 1067–1076. PMID: 7713409

28. Comeron JM (2004) Selective and mutational patterns associated with gene expression in humans: influences on synonymous composition and intron presence. Genetics 167: 1293–1304. PMID: 15280243

29. Carbone A, Zinovyev A, Kepes F (2003) Codon adaptation index as a measure of dominating codon bias. Bioinformatics 19: 2005–2015. PMID: 14594704

30. Lithwick G, Margalit H (2005) Relative predicted protein levels of functionally associated proteins are conserved across organisms. Nucleic Acids Res 33: 1051–1057. PMID: 15718304

31. Miyasaka H (2002) Translation initiation AUG context varies with codon usage bias and gene length in Drosophila melanogaster. J Mol Evol 55: 52–64. PMID: 12165842

32. Sueoka N (1988) Directional mutation pressure and neutral molecular evolution. Proc Natl Acad Sci U S A 85: 2653–2657. PMID: 3357886

33. Mandlik V, Shinde S, Singh S (2014) Molecular evolution of the enzymes involved in the sphingolipid metabolism of Leishmania: selection pressure in relation to functional divergence and conservation. BMC Evol Biol 14: 142. doi: 10.1186/1471-2148-14-142 PMID: 24951280

34. Hassan S, Mahalingam V, Kumar V (2009) Synonymous codon usage analysis of thirty two mycobacteriophage genomes. Adv Bioinformatics: 316936. doi: 10.1155/2009/316936 PMID: 20150956

35. Dass JF, Sudandiradoss C (2012) Insight into pattern of codon biasness and nucleotide base usage in serotonin receptor gene family from different mammalian species. Gene 503: 92–100. doi: 10.1016/j.gene.2012.03.057 PMID: 22480817

36. Ma J, Zhou T, Gu W, Sun X, Lu Z (2002) Cluster analysis of the codon use frequency of MHC genes from different species. Biosystems 65: 199–207. PMID: 12069729

37. Nabiyouni M, Prakash A, Fedorov A (2013) Vertebrate codon bias indicates a highly GC-rich ancestral genome. Gene 519: 113–119. doi: 10.1016/j.gene.2013.01.033 PMID: 23376453

38. Tatarinova TV, Alexandrov NN, Bouck JB, Feldmann KA (2010) GC3 biology in corn, rice, sorghum and other grasses. BMC Genomics 11: 308. doi: 10.1186/1471-2164-11-308 PMID: 20470436

39. 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: 249–257. PMID: 10754068

40. Yang Z, Nielsen R (2008) Mutation-selection models of codon substitution and their use to estimate selective strengths on codon usage. Mol Biol Evol 25: 568–579. doi: 10.1093/molbev/msm284 PMID: 18178545

41. Grantham R, Gautier C, Gouy M, Mercier R, Pave A (1980) Codon catalog usage and the genome hypothesis. Nucleic Acids Res 8: r49–r62. PMID: 6986610

42. Grantham R, Gautier C, Gouy M, Jacobzone M, Mercier R (1981) Codon catalog usage is a genome strategy modulated for gene expressivity. Nucleic Acids Res 9: r43–74. PMID: 7208352

43. Wright F (1990) The 'effective number of codons' used in a gene. Gene 87: 23–29. PMID: 2110097

44. Ikemura T (1985) Codon usage and tRNA content in unicellular and multicellular organisms. Mol Biol Evol 2: 13–34. PMID: 3916708

45. Stenico M, Lloyd AT, Sharp PM (1994) Codon usage in Caenorhabditis elegans: delineation of translational selection and mutational biases. Nucleic Acids Res 22: 2437–2446. PMID: 8041603

46. Sharp PM, Li WH (1986) Codon usage in regulatory genes in Escherichia coli does not reflect selection for 'rare' codons. Nucleic Acids Res 14: 7737–7749. PMID: 3534792

47. Fox JM, Erill I (2010) Relative codon adaptation: a generic codon bias index for prediction of gene expression. DNA Res 17: 185–196. doi: 10.1093/dnares/dsq012 PMID: 20453079

48. Komurov K, Dursun S, Erdin S, Ram PT (2012) NetWalker: a contextual network analysis tool for functional genomics. BMC Genomics 13: 282. doi: 10.1186/1471-2164-13-282 PMID: 22732065

Genomics xxx (2015) xxx–xxx

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.

© 2016 Published by Elsevier Inc.

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 ith codon for the jth 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

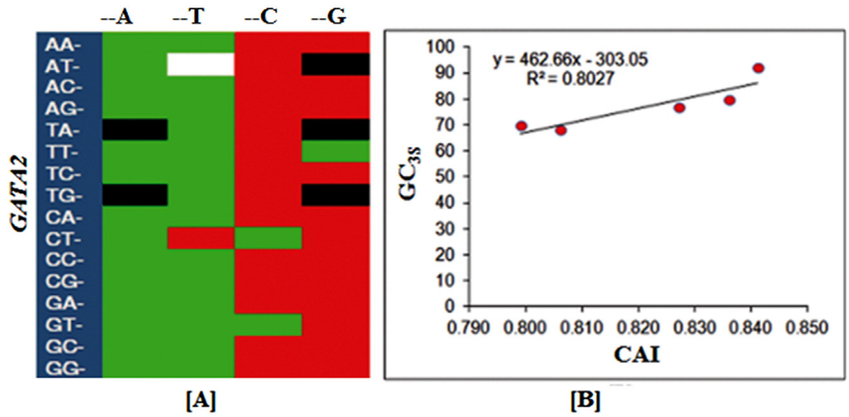


Fig. 1. Correlation coefficient between codon usage, GC_3 and *CAI* for GATA2 gene across mammals; [A] Heat maps of the correlation coefficients between codon usage and GC_{3s}. Red color coding represents the positive correlation, green as negative correlation, black fields are non-degenerate codons (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.)

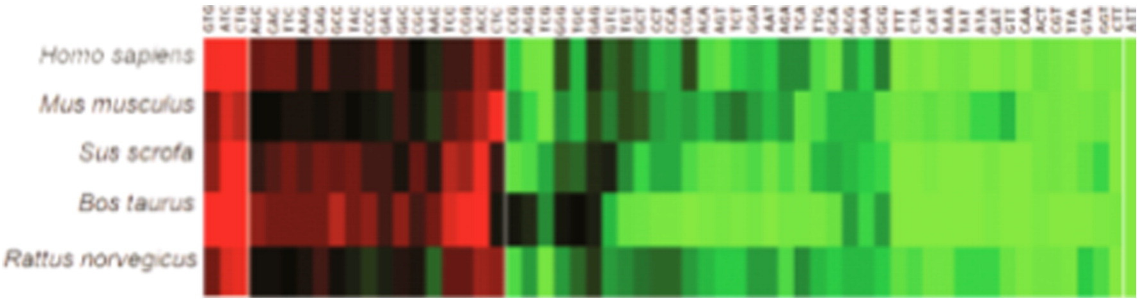


Fig. 2. Cluster analysis of *RSCU* values of *GATA2* gene across mammals. The heat map represents the *RSCU* value of a codon (shown in columns) corresponding to the *GATA2* gene across mammals (shown in rows). Green color indicates *RSCU* < 1, dark red *RSCU* > 1 and distinct red *RSCU* > 1.6. (For 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

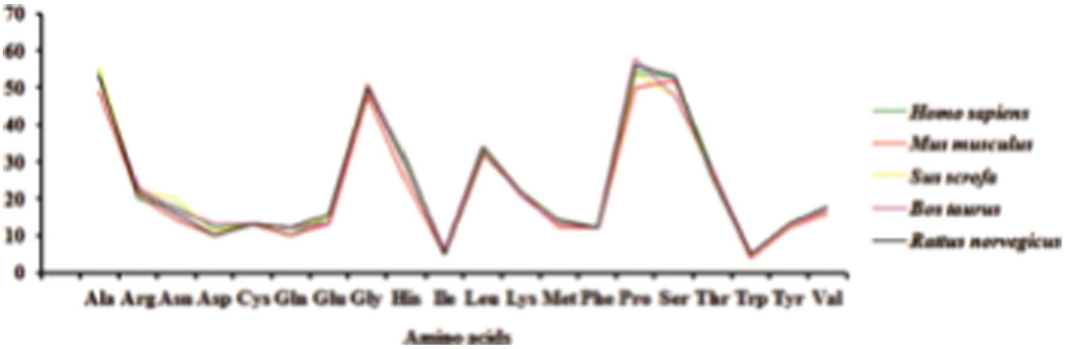


Fig. 3. Frequency of amino acid usage in *GATA2* gene across mammals; Five adjacent color-bars in a group representing five mammals indicate the usage of a particular amino acid with a small vertical line at the top of each bar as standard error. (For interpretation of the references to color in this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

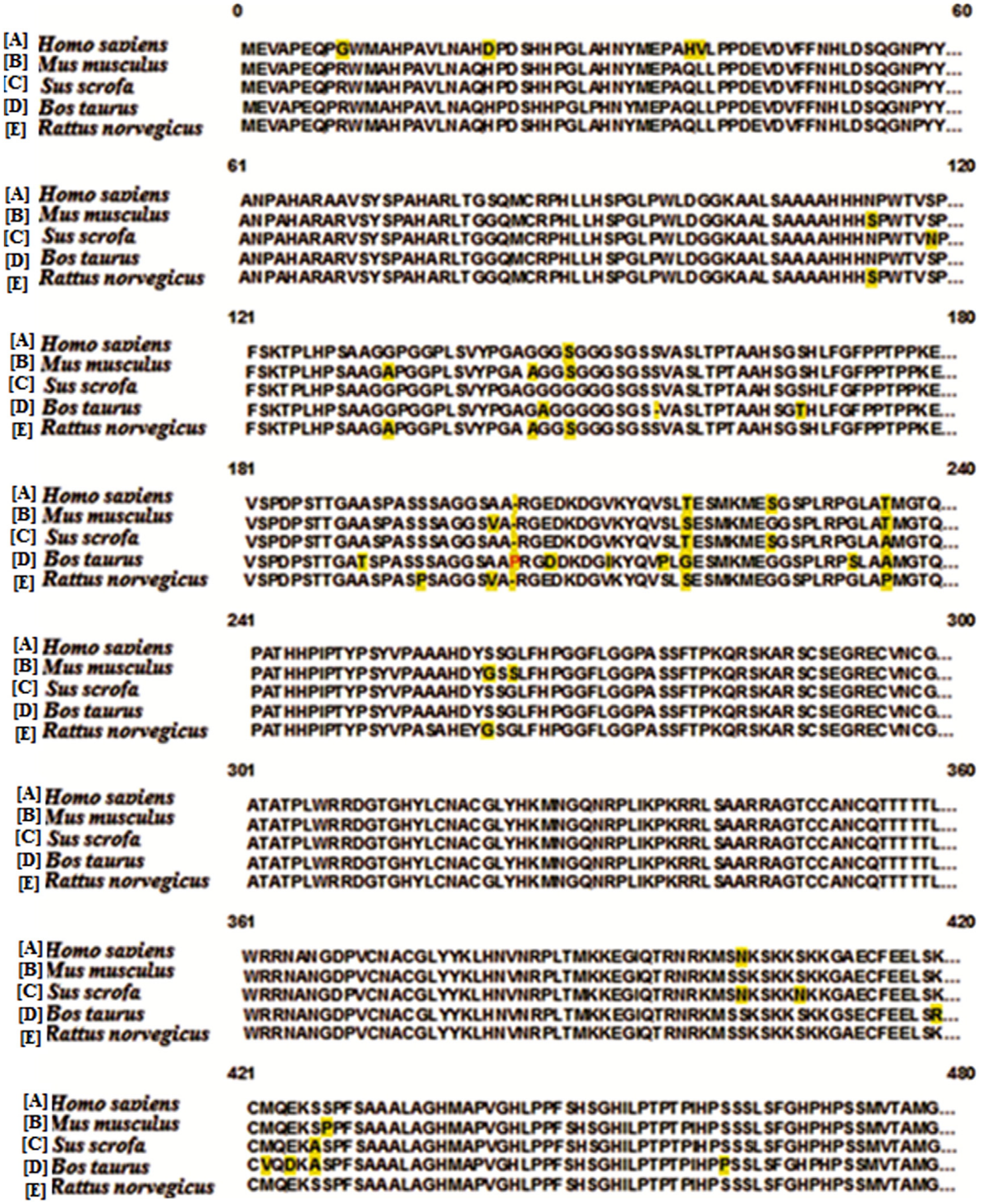

0
60
[A] Homo sapiens MEVAPEQPGWMAHPAVLNAHDPDSHHPGLAHNYMEPAHVLPPDEVDVFFNHLDSQGNPYY...
[B] Mus musculus MEVAPEQPRWMAHPAVLNAQHPDSHHPGLAHNYMEPAQLLPPDEVDVFFNHLDSQGNPYY...
[C] Sus scrofa MEVAPEQPRWMAHPAVLNAQHPDSHHPGLAHNYMEPAQLLPPDEVDVFFNHLDSQGNPYY...
[D] Bos taurus MEVAPEQPRWMAHPAVLNAQHPDSHHPGLPHNYMEPAQLLPPDEVDVFFNHLDSQGNPYY...
[E] Rattus norvegicus MEVAPEQPRWMAHPAVLNAQHPDSHHPGLAHNYMEPAQLLPPDEVDVFFNHLDSQGNPYY...
61
120
[A] Homo sapiens ANPAHARAAVSYSPAHARLTGSQMCRPHLLHSPGLPWLDGGKAALSAAAAHHHNPWTVSP...
[B] Mus musculus ANPAHARARVSYSPAHARLTGGQMCRPHLLHSPGLPWLDGGKAALSAAAAHHHSPWTVSP...
[C] Sus scrofa ANPAHARARVSYSPAHARLTGGQMCRPHLLHSPGLPWLDGGKAALSAAAAHHHNPWTVNP...
[D] Bos taurus ANPAHARARVSYSPAHARLTGGQMCRPHLLHSPGLPWLDGGKAALSAAAAHHHNPWTVSP...
[E] Rattus norvegicus ANPAHARARVSYSPAHARLTGGQMCRPHLLHSPGLPWLDGGKAALSAAAAHHHSPWTVSP...
121
180
[A] Homo sapiens FSKTPLHPSAAGGPGGPLSVYPGAGGGSGGGSGSSVASLTPTAAHSGSHLFGFPPTPPKE...
[B] Mus musculus FSKTPLHPSAAGAPGGPLSVYPGAAGGSGGGSGSSVASLTPTAAHSGSHLFGFPPTPPKE...
[C] Sus scrofa FSKTPLHPSAAGGPGGPLSVYPGAGGGGGGGSGSSVASLTPTAAHSGSHLFGFPPTPPKE...
[D] Bos taurus FSKTPLHPSAAGGPGGPLSVYPGAGAGGGGGSGS-VASLTPTAAHSGTHLFGFPPTPPKE...
[E] Rattus norvegicus FSKTPLHPSAAGAPGGPLSVYPGAAGGSGGGSGSSVASLTPTAAHSGSHLFGFPPTPPKE...
181
240
[A] Homo sapiens VSPDPSTTGAASPASSSAGGSAA-RGEDKDGVKYQVSLTESMKMESGSPLRPGLATMGTQ...
[B] Mus musculus VSPDPSTTGAASPASSSAGGSVA-RGEDKDGVKYQVSLSESMKMEGGSPLRPGLATMGTQ...
[C] Sus scrofa VSPDPSTTGAASPASSSAGGSAA-RGEDKDGVKYQVSLTESMKMESGSPLRPGLAAMGTQ...
[D] Bos taurus VSPDPSTTGATSPASSSAGGSAAPRGDDKDGIKYQVPLGESMKMEGGSPLRPSLAAMGTQ...
[E] Rattus norvegicus VSPDPSTTGAASPASPSAGGSVA-RGEDKDGVKYQVSLSESMKMEGGSPLRPGLAPMGTQ...
241
300
[A] Homo sapiens PATHHPIPTYPSYVPAAAHDYSSGLFHPGGFLGGPASSFTPKQRSKARSCSEGRECVNCG...
[B] Mus musculus PATHHPIPTYPSYVPAAAHDYGSSLFHPGGFLGGPASSFTPKQRSKARSCSEGRECVNCG...
[C] Sus scrofa PATHHPIPTYPSYVPAAAHDYSSGLFHPGGFLGGPASSFTPKQRSKARSCSEGRECVNCG...
[D] Bos taurus PATHHPIPTYPSYVPAAAHDYSSGLFHPGGFLGGPASSFTPKQRSKARSCSEGRECVNCG...
[E] Rattus norvegicus PATHHPIPTYPSYVPASAHEYGSGLFHPGGFLGGPASSFTPKQRSKARSCSEGRECVNCG...
301
360
[A] Homo sapiens ATATPLWRRDGTGHYLCNACGLYHKMNGQNRPLIKPKRRLSAARRAGTCCANCQTTTTTL...
[B] Mus musculus ATATPLWRRDGTGHYLCNACGLYHKMNGQNRPLIKPKRRLSAARRAGTCCANCQTTTTTL...
[C] Sus scrofa ATATPLWRRDGTGHYLCNACGLYHKMNGQNRPLIKPKRRLSAARRAGTCCANCQTTTTTL...
[D] Bos taurus ATATPLWRRDGTGHYLCNACGLYHKMNGQNRPLIKPKRRLSAARRAGTCCANCQTTTTTL...
[E] Rattus norvegicus ATATPLWRRDGTGHYLCNACGLYHKMNGQNRPLIKPKRRLSAARRAGTCCANCQTTTTTL...
361
420
[A] Homo sapiens WRRNANGDPVCNACGLYYKLHNVNRPLTMKKEGIQTRNRKMSNKSKKSKKGAECFEELSK...
[B] Mus musculus WRRNANGDPVCNACGLYYKLHNVNRPLTMKKEGIQTRNRKMSSKSKKSKKGAECFEELSK...
[C] Sus scrofa WRRNANGDPVCNACGLYYKLHNVNRPLTMKKEGIQTRNRKMSNKSKKNKKGAECFEELSK...
[D] Bos taurus WRRNANGDPVCNACGLYYKLHNVNRPLTMKKEGIQTRNRKMSSKSKKSKKGSECFEELSR...
[E] Rattus norvegicus WRRNANGDPVCNACGLYYKLHNVNRPLTMKKEGIQTRNRKMSSKSKKSKKGAECFEELSK...
421
480
[A] Homo sapiens CMQEKSSPFSAAALAGHMAPVGHLPPFSHSGHILPTPTPIHPSSSLSFGHPHPSSMVTAMG...
[B] Mus musculus CMQEKSPPFSAAALAGHMAPVGHLPPFSHSGHILPTPTPIHPSSSLSFGHPHPSSMVTAMG...
[C] Sus scrofa CMQEKASPFSAAALAGHMAPVGHLPPFSHSGHILPTPTPIHPSSSLSFGHPHPSSMVTAMG...
[D] Bos taurus CVQDKASPFSAAALAGHMAPVGHLPPFSHSGHILPTPTPIHPPSSLSFGHPHPSSMVTAMG...
[E] Rattus norvegicus CMQEKSSPFSAAALAGHMAPVGHLPPFSHSGHILPTPTPIHPSSSLSFGHPHPSSMVTAMG...

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

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

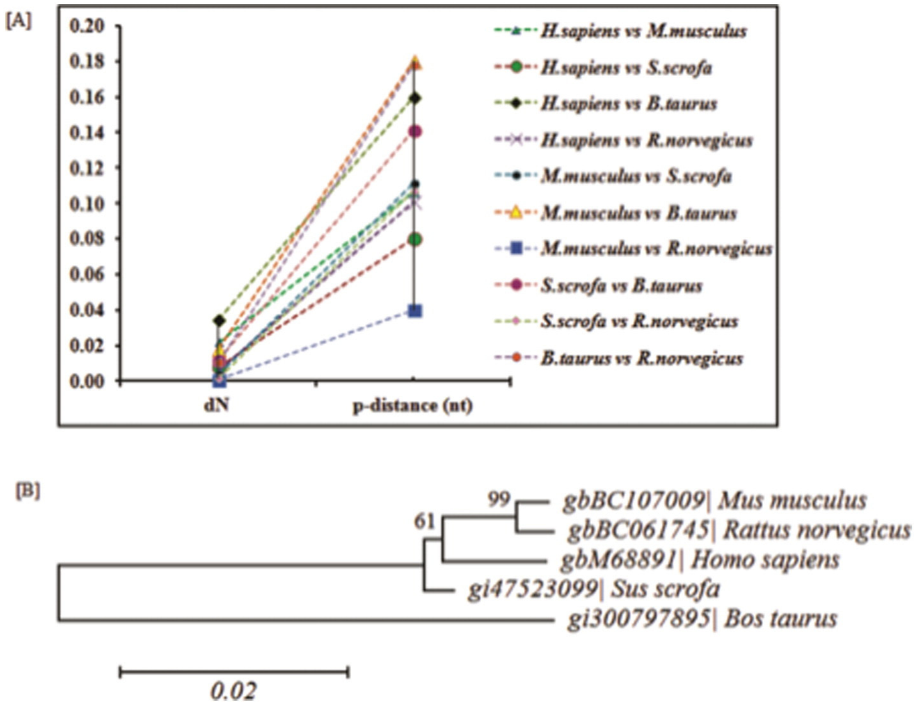


Fig. 5. Genetic variability of nonsynonymous mutation and phylogenetic tree; [A] Distribution of nonsynonymous mutation per site (d_N) and mean genetic p-distance in the coding sequence of *GATA2* gene across mammals. [B] Neighbor-Joining tree using d_N distance based on codon alignment. The percentage 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.

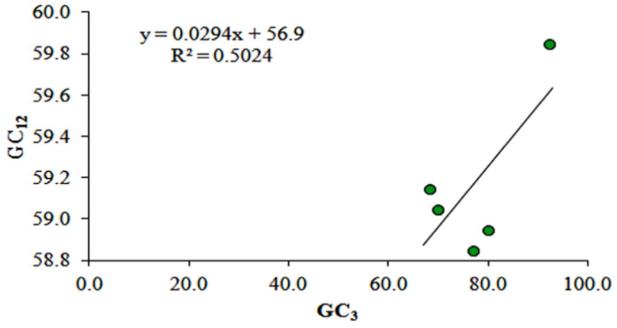


Fig. 6. Neutrality plots of *GATA-2* gene across mammalian species. Individual genes of different species are plotted based on the average GC content in the first and second codon position versus the GC content of the third codon position (GC_3).

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.