Bioinformation 17(8): 731-740 (2021) ISSN 0973-2063 (online) 0973-8894 (print)

BIOMEDICAL INFORMATICS

www.bioinformation.net
Volume 17(8)

Research Article

BIOINFORMATION
Discovery at the interface of physical and biological sciences

OPEN ACCESS GOLD

Codon usage bias analysis of genes linked with esophagus cancer

Hemashree Bordoloi[1,2] & SR Nirmala[3]

[1]Deptartment of Electronics and Communication Engineering, Gauhati University, Assam, India; [2]Department of Electronics and Communication Engineering, Assam Don Bosco University, Assam, India; [3] School of Electronics and Communication Engineering, KLE Technological University, Karnataka, India; *Corresponding author; Hemashree Bordoloi, Email: hemashree.bordoloi@dbuniversity.ac.in; SR Nirmala – nirmalasr3@gmail.com

Received June 26, 2021; Revised August 30, 2021; Accepted August 30, 2021, Published August 31, 2021

DOI: 10.6026/97320630017731

Declaration on Publication Ethics:
The author's state that they adhere with COPE guidelines on publishing ethics as described elsewhere at https://publicationethics.org/. The authors also undertake that they are not associated with any other third party (governmental or non-governmental agencies) linking with any form of unethical issues connecting to this publication. The authors also declare that they are not withholding any information that is misleading to the publisher in regard to this article.

Author responsibility:
The authors are responsible for the content of this article. The editorial and the publisher have taken reasonable steps to check the content of the article in accordance to publishing ethics with adequate peer reviews deposited at PUBLONS.

Declaration on official E-mail:
The corresponding author declares that official e-mail from their institution is not available for all authors

Abstract:
Esophageal cancer involves multiple genetic alternations. A systematic codon usage bias analysis was completed to investigate the bias among the esophageal cancer responsive genes. GC-rich genes were low (average effective number of codon value was 49.28). CAG and GTA are over-represented and under-represented codons, respectively. Correspondence analysis, neutrality plot, and parity rule 2 plot analysis confirmed the dominance over mutation pressure in modulating the codon usage pattern of genes linked with esophageal cancer.

Keywords: Natural selection; mutation pressure; compositional constraints; RSCU

Background:
Due to the degeneracy of the genetic code, we observe significant variation in synonymous codon usage in the coding sequences **[1].** Mutation is the prime source of synonymous codon usage variation where selection pressure decides their selection and adaptation **[2, 3].** Supek et al 2014, revealed the role of synonymous mutation in cancer progression **[4].** Later on, it became evident that CUB and synonymous mutations have a non-trivial effect on human diseases **[5].** Reports on the synonymous variant of genes having a role in metabolic processes revealed that CUB information may increase the diagnostic accuracy in a variety of diseases **[6-8].** Cancer is a genetic disorder that results from catastrophic mutations causing genetic alternations **[9].** Among the different cancer types, esophageal cancer (EA) ranks as seventh and sixth serious malignancy with respect to prognosis and mortality rate, respectively **[10].** The 5-year survival rate of EC patients is within the range of 15% to 25% **[10].** The incidence rate is higher in developing countries from the Asian region (highest in China) **[11].** According to the Indian Council for Medical Research (ICMR), the number of EC patients is increasing each passing year in India **[12].** Several independent research done on the molecular changes linked to EC have identified key regulatory genes of EC **[11, 13].** Therefore, it is of interest to document the codon usage bias analysis of genes linked with esophagus cancer.

Materials and methods:
Sequence data:
We identified 82 human genes with a role in EC. The complete coding sequence of these genes was retrieved from the NCBI nucleotide database (www.ncbi.nlm.nih.gov). The sequences were pre-processed with an in-house Perl script to check sequences taken for CUB analysis starts with a proper initiation, ends with termination codons, and are perfect multiple of 3 bases.

Effective number of codons (ENc):
Wright proposed ENc in 1990. It quantifies the degree of CUB in a given sequence **[14].** The values of ENc values can vary from 20

(extreme bias where only one codon is used per amino acid) to 61 (without bias where codons are used in equal probability). ENc was calculated as follows:

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

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

Relative synonymous codon usage (RSCU):
RSCU is the observed frequency of a codon divided by the expected frequency **[15].** If all synonymous codons encoding the same amino acid are used equally, RSCU values are close to 1.0, indicating a lack of bias**.** Moreover, the codon with RSCU value greater than 1.6 is treated as over-represented codon, whereas the codon with RSCU value lower than 0.6 is considered as under-represented codon. RSCU value of a codon is estimated as:

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

Where *Xij* is the frequency of occurrence of the j^{th} codon for i^{th} amino acid (any Xij with a value of zero is arbitrarily assigned a value of 0.5) and *ni* is the number of codons for the i^{th} amino acid (i^{th} codon family).

GC content analysis:
To quantify the variation in base frequencies that occur in varying numbers at each codon site we calculated GC1s, GC2s & GC3s *i.e.* frequency of GC content at first, second, and third codon positions, respectively.

$$GC_n = \frac{N}{N_{GC}}$$

Where, GC_n is the frequency of use of G or C on the n^{th} codon position, N is the total number of codons in the coding sequence of the gene and N_{GC} is the sum of codons with G or C on the n^{th} codon position.

Neutrality plot:
Reports suggest that there is a bias in the rate of mutations at three different codon positions, particularly high at the synonymous third codon position. Theoretically, mutation should occur randomly if there is no external pressure. The preference of bases in three different codon positions is not the same in the presence of selection pressure **[16].** Neutrality plot, a graphical plot of GC12 against GC3 depicts the roles of directional mutational pressure against natural selection.

Parity rule 2 (PR2) plot:
PR2 bias plot was generated by plotting the [G3 /(G3 + C3)] vs [A3 /(A3 + T3)] [17]. In the PR2 plot, the center region coordinate is (0.5, 0.5). If a gene shows its presence on the center then it suggests that A = T and G = C and no bias is the composition of the gene.

Correspondence analysis:
Correspondence analysis has been successfully used to explore codon usage variation among genes **[18].** It is a commonly used multivariate statistical technique, in which all genes were plotted in a 59-dimensional space, according to the usage of the 59 sense codons (excluding codons for Met, Trp, and stop codons). The plot was then used to identify the axes, which represent the most prominent factors contributing to variation among genes.

Software used:
Codon-pair context quantifications were performed using Anaconda 2 **[19].** The association of codon pairs was assessed using the chi-square test of independence. All the statistical analyses were done using the SPSS software (version 16.0). Acua software was used to find the nucleotide composition at synonymous positions, compositional skewness, CAI, and ENc values **[20].** RSCU value was calculated using INCA software **[21]**. Correspondence analysis was done using past3 **[22]**. CodonW was used to calculate GRAVY and Aromo values (http://codonw.sourceforge.net/).

Table 2: Interrelationships of overall nucleotide composition with their usage frequency at synonymous third codon position in the genes responsible for esophageal cancer.

	A3	T3	G3	C3	AT3	GC3
A	0.995	0.97	0.93	0.899	0.674	-0.681
T	0.982	0.982	0.971	0.948	0.812	-0.808
G	0.934	0.939	0.995	0.983	-0.828	0.828
C	0.926	0.944	0.982	0.994	-0.749	0.753
AT	0.851	0.45	-0.503	-0.71	0.882	-0.885
GC	-0.851	-0.45	0.503	0.71	-0.882	0.885

Codon usage bias:
ENc value was used to determine the overall codon usage bias. From **Table 3** we observed that ENc value of the EC-related genes ranges from 33.778 to 56.976, with a mean ± standard deviation of 49.28 ± 5.72. This low bias (ENc<35) in codon selection might have a link with DNA replication as these genes are actively synthesized in different cell types. Moreover, we observed a significant negative correlation between ENc and GC3 (r=-0.56, p<0.05). This further suggests the impact of GC nucleotide composition on the CUB. Genes with high GC3 composition showed the highest CUB. However, the correlation coefficient did not reach the extreme value i.e. 1, which further suggests that GC3 is not the sole determinant of CUB confirms the variation of nucleotides and thus CUB in EC-responsive genes. RSCU analysis further showed the dominance in the use of a specific group of codons in EC-responsible genes. 28 codons out of 60 codons showed RSCU score greater than 1, 7 codons RSCU > 1.6, whereas, 9 codons RSCU < 0.6 (**Table 4**).

Results:
Nucleotide Composition:
Nucleotide composition analysis was performed to identify the compositional variation present in the genes responsible for EC. From our analysis, it was observed that the EC-responsive genes are GC-rich (52.01%), similar to the overall genomic composition. Positional bias analysis revealed that in 59% gene G3>A3, whereas in 7% gene A3 is equal to G3 and in 34% gene A3>G3. Similarly, in 72% gene GC3>AT3, only in 1% gene AT3> GC3, whereas, in 27% gene AT3>GC3. This indicates the dominance of GC composition at synonymous codon positions. AT3 was found to be highest in the RBBP6 gene and GC3 in the SOX17 gene. **Table 1** represents the nucleotide composition of different genes responsible for EC. Furthermore, we observed a strong significant correlation between homogenous and heterogeneous nucleotide compositions (**Table 2**). These intricate correlation patterns suggest the influence of mutation (major) and selection pressure (minor) is shaping the CUB of EC-responsive genes. Our findings corroborated with the results reported elsewhere **[23].**

Table 1: Nucleotide composition of genes responsible for esophagus cancer

Sl. No.	Gene	A3	T3	G3	C3	AT3 %	GC3 %
1	AF325503.1_cds_AAG42321.1_1	31	24	49	44	12.304	20.805
2	NM_032566.3_cds_NP_115955.1_1	26	22	22	15	18.605	14.341
3	NM_032411.3_cds_NP_115787.1_1	31	24	49	44	12.304	20.805
4	NM_001114387.2_cds_NP_001107859.1_1	137	84	127	70	17.582	15.672
5	NM_182606.4_cds_NP_872412.3_1	138	84	128	71	17.536	15.719
6	NM_002810.4_cds_NP_002801.1_1	108	37	151	81	12.787	20.459
7	NM_001330692.2_cds_NP_001317621.1_1	109	37	153	81	12.773	20.472
8	NM_001203258.2_cds_NP_001190187.1_1	117	74	144	95	14.772	18.484
9	NM_004689.4_cds_NP_004680.2_1	198	105	212	200	14.106	19.181
10	NM_006846.4_cds_NP_006837.2_1	370	164	360	170	16.714	16.588
11	NM_001127698.2_cds_NP_001121170.1_1	378	169	371	176	16.651	16.651
12	NM_003979.4_cds_NP_003970.1_1	92	76	103	86	15.642	17.598
13	NM_001288661.2_cds_NP_001275590.1_1	170	79	208	113	14.536	18.739
14	NM_001313.5_cds_NP_001304.1_1	171	81	205	115	14.66	18.615
15	NM_001014809.3_cds_NP_001014809.1_1	186	90	258	152	13.392	19.893
16	NM_001288662.1_cds_NP_001275591.1_1	167	79	206	114	14.462	18.812
17	NM_001127699.2_cds_NP_001121171.1_1	323	136	317	140	16.685	16.612
18	NM_014360.4_cds_NP_055175.2_1	27	39	84	89	9.167	24.028
19	NM_030916.3_cds_NP_112178.2_1	98	70	188	154	10.959	22.309
20	NM_002318.3_cds_NP_002309.1_1	191	147	266	170	14.538	18.753
21	NM_020998.3_cds_NP_066278.3_1	136	140	224	225	12.672	20.615
22	NM_005429.5_cds_NP_005420.1_1	116	97	116	90	16.905	16.349
23	NM_006010.6_cds_NP_006001.5_1	59	27	59	37	15.665	17.486
24	NM_001320810.2_cds_NP_001307739.1_1	169	78	230	108	14.05	19.226
25	NM_001282599.2_cds_NP_001269528.1_1	174	75	227	108	14.188	19.088
26	NM_004389.4_cds_NP_004380.2_1	265	111	351	178	13.834	19.463
27	NM_001164883.2_cds_NP_001158355.1_1	256	100	337	167	13.782	19.512
28	NM_001282600.2_cds_NP_001269529.1_1	154	70	212	101	13.879	19.393
29	NM_001282598.2_cds_NP_001269527.1_1	275	117	360	187	13.901	19.397
30	NM_001174147.2_cds_NP_001167618.1_1	88	73	124	117	13.317	19.934
31	NM_002316.4_cds_NP_002307.2_1	86	72	122	115	13.3	19.949
32	NM_001174146.2_cds_NP_001167617.1_1	86	72	128	120	12.94	20.311
33	NM_182644.3_cds_NP_872585.1_1	172	111	152	104	17.469	15.802
34	NM_005233.6_cds_NP_005224.2_1	317	180	301	185	16.836	16.463
35	NM_001282597.3_cds_NP_001269526.1_1	280	119	369	185	13.941	19.357
36	NM_001271082.2_cds_NP_001258011.1_1	68	39	122	82	11.432	21.795
37	NM_203401.2_cds_NP_981946.1_1	42	16	57	34	12.889	20.222
38	NM_057167.4_cds_NP_476508.2_1	784	394	1115	678	13.212	20.11
39	NM_203399.2_cds_NP_981944.1_1	42	16	57	34	12.889	20.222
40	NM_001145454.3_cds_NP_001138926.1_1	40	28	57	49	12.952	20.19

41	NM_005563.4_cds_NP_005554.1_1	42	16	57	34	12.889	20.222
42	NM_004369.4_cds_NP_004360.2_1	837	429	1196	715	13.279	20.044
43	NM_033120.4_cds_NP_149111.1_1	80	50	163	158	9.587	23.673
44	NM_001198754.2_cds_NP_001185683.1_1	25	7	25	12	15.238	17.619
45	NM_001198756.1_cds_NP_001185685.1_1	25	7	25	12	15.238	17.619
46	NM_001198755.1_cds_NP_001185684.1_1	25	7	25	12	15.238	17.619
47	NM_031498.2_cds_NP_113686.1_1	25	7	25	12	15.238	17.619
48	NM_057165.5_cds_NP_476506.3_1	307	170	451	309	12.843	20.463
49	NM_057164.5_cds_NP_476505.3_1	260	142	370	264	12.922	20.379
50	NM_057166.5_cds_NP_476507.3_1	686	334	970	580	13.224	20.096
51	AF268198.1_cds_AAK27795.1_1	26	22	22	15	18.605	14.341
52	NM_017671.5_cds_NP_060141.3_1	197	140	199	141	16.568	16.716
53	NM_052972.3_cds_NP_443204.1_1	64	53	98	132	11.207	22.031
54	NM_173452.2_cds_NP_775628.1_1	51	56	98	83	12.341	20.877
55	NM_017729.4_cds_NP_060199.3_1	109	68	219	200	9.883	23.395
56	NM_032626.6_cds_NP_116015.2_1	45	26	29	18	19.888	13.165
57	NM_003665.4_cds_NP_003656.2_1	52	57	102	88	12.111	21.111
58	NM_018703.4_cds_NP_061173.1_1	631	300	481	346	17.643	15.672
59	NM_133180.3_cds_NP_573441.2_1	126	87	270	240	9.807	23.481
60	NM_006910.5_cds_NP_008841.2_1	645	308	489	350	17.717	15.598
61	NM_022454.4_cds_NP_071899.1_1	59	50	154	151	8.755	24.498
62	NM_181892.4_cds_NP_871621.1_1	46	32	37	33	17.45	15.66
63	NM_181890.3_cds_NP_871619.1_1	46	31	35	35	17.342	15.766
64	NM_001300795.2_cds_NP_001287724.1_1	40	28	23	27	19.048	14.006
65	NM_181893.3_cds_NP_871622.1_1	45	34	34	36	17.556	15.556
66	NM_001001420.3_cds_NP_001001420.1_1	135	95	116	119	16.452	16.81
67	NM_181887.3_cds_NP_871616.1_1	46	31	35	35	17.342	15.766
68	NM_001001419.3_cds_NP_001001419.1_1	135	95	116	119	16.452	16.81
69	NM_181891.3_cds_NP_871620.1_1	46	31	35	35	17.342	15.766
70	NM_005903.7_cds_NP_005894.3_1	135	95	116	119	16.452	16.81
71	NM_003340.6_cds_NP_003331.1_1	46	31	35	35	17.342	15.766
72	NM_181889.2_cds_NP_871618.1_1	46	31	35	35	17.342	15.766
73	NM_181888.3_cds_NP_871617.1_1	46	31	35	35	17.342	15.766
74	NM_181886.3_cds_NP_871615.1_1	46	31	35	35	17.342	15.766
75	NM_006825.4_cds_NP_006816.2_1	129	79	226	168	11.498	21.78
76	NM_174911.5_cds_NP_777571.1_1	55	40	114	101	10.182	23.044
77	NM_005416.3_cds_NP_005407.1_1	46	20	45	58	12.941	20.196
78	NM_001097589.2_cds_NP_001091058.1_1	46	20	45	58	12.941	20.196
79	AF228422.1_cds_AAK00708.1_1	27	14	23	19	16.27	16.667
80	NM_001286246.2_cds_NP_001273175.1_1	78	44	106	119	11.686	21.552
81	NM_001286245.2_cds_NP_001273174.1_1	80	44	106	120	11.776	21.462
82	NM_178502.4_cds_NP_848597.1_1	78	44	106	119	11.686	21.552

Table 3: Effective number of codon (ENc) values of the esophageal cancer responsible genes. ENc ranged from 33.78 to 56.98, indicates low codon usage bias

Sl. No.	Gene	ENc
1	AF325503.1_cds_AAG42321.1_1	33.778
2	NM_032566.3_cds_NP_115955.1_1	34.154
3	NM_032411.3_cds_NP_115787.1_1	36.597
4	NM_001114387.2_cds_NP_001107859.1_1	36.698
5	NM_182606.4_cds_NP_872412.3_1	36.738
6	NM_002810.4_cds_NP_002801.1_1	36.882
7	NM_001330692.2_cds_NP_001317621.1_1	38.852
8	NM_001203258.2_cds_NP_001190187.1_1	39.961
9	NM_004689.4_cds_NP_004680.2_1	40.079
10	NM_006846.4_cds_NP_006837.2_1	40.772
11	NM_001127698.2_cds_NP_001121170.1_1	41.831
12	NM_003979.4_cds_NP_003970.1_1	42.167
13	NM_001288661.2_cds_NP_001275590.1_1	43.134
14	NM_001313.5_cds_NP_001304.1_1	43.196
15	NM_001014809.3_cds_NP_001014809.1_1	43.196
16	NM_001288662.1_cds_NP_001275591.1_1	43.275
17	NM_001127699.2_cds_NP_001121171.1_1	43.484
18	NM_014360.4_cds_NP_055175.2_1	43.716
19	NM_030916.3_cds_NP_112178.2_1	45.137
20	NM_002318.3_cds_NP_002309.1_1	45.859
21	NM_020998.3_cds_NP_066278.3_1	47.927
22	NM_005429.5_cds_NP_005420.1_1	48.393
23	NM_006010.6_cds_NP_006001.5_1	48.622
24	NM_001320810.2_cds_NP_001307739.1_1	48.724
25	NM_001282599.2_cds_NP_001269528.1_1	48.735
26	NM_004389.4_cds_NP_004380.2_1	48.822
27	NM_001164883.2_cds_NP_001158355.1_1	49.295
28	NM_001282600.2_cds_NP_001269529.1_1	49.369
29	NM_001282598.2_cds_NP_001269527.1_1	49.403
30	NM_001174147.2_cds_NP_001167618.1_1	49.466
31	NM_002316.4_cds_NP_002307.2_1	49.921
32	NM_001174146.2_cds_NP_001167617.1_1	50.079
33	NM_182644.3_cds_NP_872585.1_1	50.313
34	NM_005233.6_cds_NP_005224.2_1	50.313
35	NM_001282597.3_cds_NP_001269526.1_1	50.313
36	NM_001271082.2_cds_NP_001258011.1_1	50.313
37	NM_203401.2_cds_NP_981946.1_1	50.313
38	NM_057167.4_cds_NP_476508.2_1	50.313
39	NM_203399.2_cds_NP_981944.1_1	50.313
40	NM_001145454.3_cds_NP_001138926.1_1	50.415
41	NM_005563.4_cds_NP_005554.1_1	50.54
42	NM_004369.4_cds_NP_004360.2_1	50.637
43	NM_033120.4_cds_NP_149111.1_1	50.925
44	NM_001198754.2_cds_NP_001185683.1_1	50.936
45	NM_001198756.1_cds_NP_001185685.1_1	51.095
46	NM_001198755.1_cds_NP_001185684.1_1	51.118
47	NM_031498.2_cds_NP_113686.1_1	51.402
48	NM_057165.5_cds_NP_476506.3_1	51.633
49	NM_057164.5_cds_NP_476505.3_1	51.633
50	NM_057166.5_cds_NP_476507.3_1	51.636
51	AF268198.1_cds_AAK27795.1_1	51.67
52	NM_017671.5_cds_NP_060141.3_1	51.67
53	NM_052972.3_cds_NP_443204.1_1	51.67
54	NM_173452.2_cds_NP_775628.1_1	51.67
55	NM_017729.4_cds_NP_060199.3_1	51.753
56	NM_032626.6_cds_NP_116015.2_1	51.915
57	NM_003665.4_cds_NP_003656.2_1	52.024
58	NM_018703.4_cds_NP_061173.1_1	52.216
59	NM_133180.3_cds_NP_573441.2_1	52.216
60	NM_006910.5_cds_NP_008841.2_1	52.355
61	NM_022454.4_cds_NP_071899.1_1	53.056
62	NM_181892.4_cds_NP_871621.1_1	53.899
63	NM_181890.3_cds_NP_871619.1_1	53.965
64	NM_001300795.2_cds_NP_001287724.1_1	54.461
65	NM_181893.3_cds_NP_871622.1_1	54.461
66	NM_001001420.3_cds_NP_001001420.1_1	54.461
67	NM_181887.3_cds_NP_871616.1_1	54.739
68	NM_001001419.3_cds_NP_001001419.1_1	54.739
69	NM_181891.3_cds_NP_871620.1_1	54.753
70	NM_005903.7_cds_NP_005894.3_1	54.857
71	NM_003340.6_cds_NP_003331.1_1	54.857
72	NM_181889.2_cds_NP_871618.1_1	54.857
73	NM_181888.3_cds_NP_871617.1_1	54.889
74	NM_181886.3_cds_NP_871615.1_1	54.946
75	NM_006825.4_cds_NP_006816.2_1	55.051
76	NM_174911.5_cds_NP_777571.1_1	55.443
77	NM_005416.3_cds_NP_005407.1_1	55.446
78	NM_001097589.2_cds_NP_001091058.1_1	55.82
79	AF228422.1_cds_AAK00708.1_1	55.993
80	NM_001286246.2_cds_NP_001273175.1_1	56.004
81	NM_001286245.2_cds_NP_001273174.1_1	56.247
82	NM_178502.4_cds_NP_848597.1_1	56.976

Table 4: Average RSCU score of the codons. Codons with green colored values represents over-represented codons, red color represents frequently used codons. Nine codons namely GTA, TCG, ATA, TTA, CCG, CGT, ACG, GCG and CTA were found to be under-represented (RSCU<0.6).

Codon	RSCU	Codon	RSCU	Codon	RSCU	Codon	RSCU	Codon	RSCU	Codon	RSCU
GCA	0.9584	AAC	1.1415	GGA	0.8584	CTC	1.1065	CCC	1.1091	ACC	1.3616
GCC	1.5344	AAT	0.8584	GGC	1.5346	CTG	2.2892	CCG	0.4461	ACG	0.4826
GCG	0.5144	GAC	1.0895	GGG	0.8641	CTT	0.8382	CCT	1.2354	ACT	0.7823
GCT	0.9926	GAT	0.9104	GGT	0.7427	TTA	0.4302	AGC	1.733	TAC	1.0778
AGA	1.65054	TGC	0.8757	CAC	1.049	TTG	0.793	AGT	0.7685	TAT	0.8489
AGG	0.8393	TGT	1.0266	CAT	0.8280	AAA	0.8134	TCA	0.7666	GTA	0.3634
CGA	0.6499	CAA	1.1750	ATA	0.4091	AAG	1.1865	TCC	1.2710	GTC	0.8839
CGC	0.8451	CAG	2.8249	ATC	1.4112	TTC	1.1310	TCG	0.3958	GTG	1.8612
CGG	1.4224	GAA	1.6304	ATT	1.1796	TTT	0.8689	TCT	1.0647	GTT	0.8913
CGT	0.4461	GAG	2.3695	CTA	0.542	CCA	1.2092	ACA	1.3733		

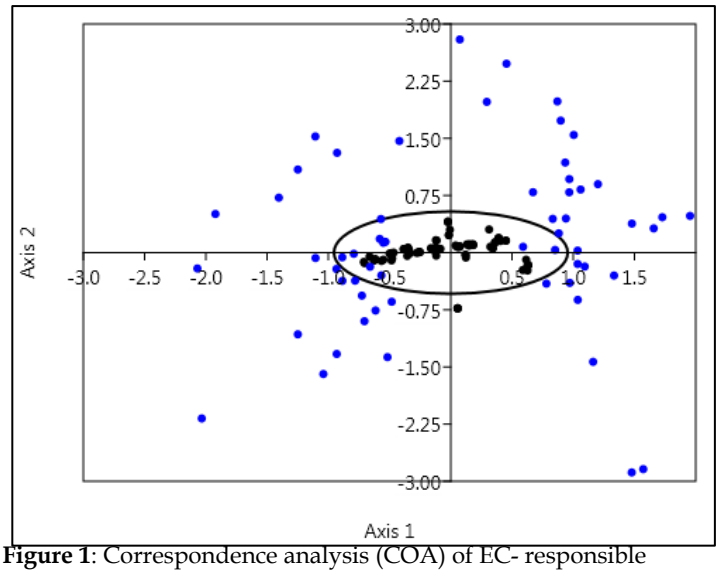

Figure 1: Correspondence analysis (COA) of EC- responsible genes based on the RSCU score of 59 codons. Black color indicates genes, blue-colored dots represent synonymous codons. Axis 1 and axis 2 contributes 27.5% and 11.6% of the total variation, respectively

Variation in codon usage:

To explore the variation in synonymous codon usage we did correspondence analysis based on the RSCU score of 59 codons across the 82 EC-responsible genes. Axis 1 and axis 2 were found to be the major contributors i.e. 27.5% and 11.6% of the total CUB variation. In **Figure 1** black dot represents the genes and the blue dots represent codons. From **Figure 1** it is evident that the majority of the codons are close to the two axes, whereas, genes were found to be near the center region. These findings altogether confirm the relative contribution of nucleotide composition under the influence of mutation pressure (major) in shaping the observed codon usage pattern **[24].** Next, we did cluster analysis to validate the findings of our correspondence analysis (**Figure 2**). The findings of our cluster analysis corroborated well with the COA result.

Neutrality plot analysis:

To evaluate the contribution of two major forces in shaping the codon usage pattern of EC-responsible genes we plotted GC12 vs GC3. We observed a significant positive correlation between them ($r=0.684$, $p<0.01$). The regression coefficient of GC12 on GC3 is 0.0246 (**Figure 3**). Previous reports suggest that if the regression coefficient in the neutrality plot is greater than 0.5 then there is a significant contribution of mutation pressure **[25].** However, here we observed the opposite thereby our neutrality analysis suggests the dominance of selection pressure in shaping the CUB of EC-responsible genes.

Parity rule 2 (PR2) bias plot analysis

Mutation forces the random use of nucleotides at the synonymous codon position whereas selection pressure does not force the equal use of nucleotides **[26]**. In the PR2 plot, if the genes coincide in the center region then it suggests the dominance of mutation pressure whereas deviation from the center indicates the contribution of selection pressure. From **figure 4** it is seen that the majority of the EC-responsible genes are far away from the center region. The average coordinates of A3/(A3+T3) and G3/(G3+C3) was 0.6308 and 0.5729, respectively. Therefore, in support of COA, our PR2 plot analysis further confirms the supremacy of selection pressure in shaping the CUB of EC-responsive genes.

Gene expression and its relation with various skews:

CAI was used to estimate the expression value of EC-responsible genes. The genes showed CAI value within the range of 0.705 to 0.863, which suggests higher expression. GRAVY analysis revealed that 5% of genes are hydrophobic and 95% of genes are hydrophilic. The aromaticity score can determine stability of the gene. A high aromaticity value signifies a more stable gene structure. 21% of genes show a high aromaticity value. CAI showed a positive correlation of 0.361 and 0.190 ($p<0.05$) with AT and GC skewness, respectively. Where as, CAI showed a negative correlation with GRAVY and aromaticity score of the EC-responsible genes (GRAVY: $r= -0.404$, $p<0.05$; aromaticity: $r= -0.357$, $p<0.05$). Amino acid composition analysis revealed that leucine is the highest used amino acid whereas tryptophan is the lowest used amino acid. Serine, alanine, glutamine, and lysine are the frequently used amino acids. Altogether, these findings suggest that the skews played a significant role in modulating the CUB and thereby the gene expression.

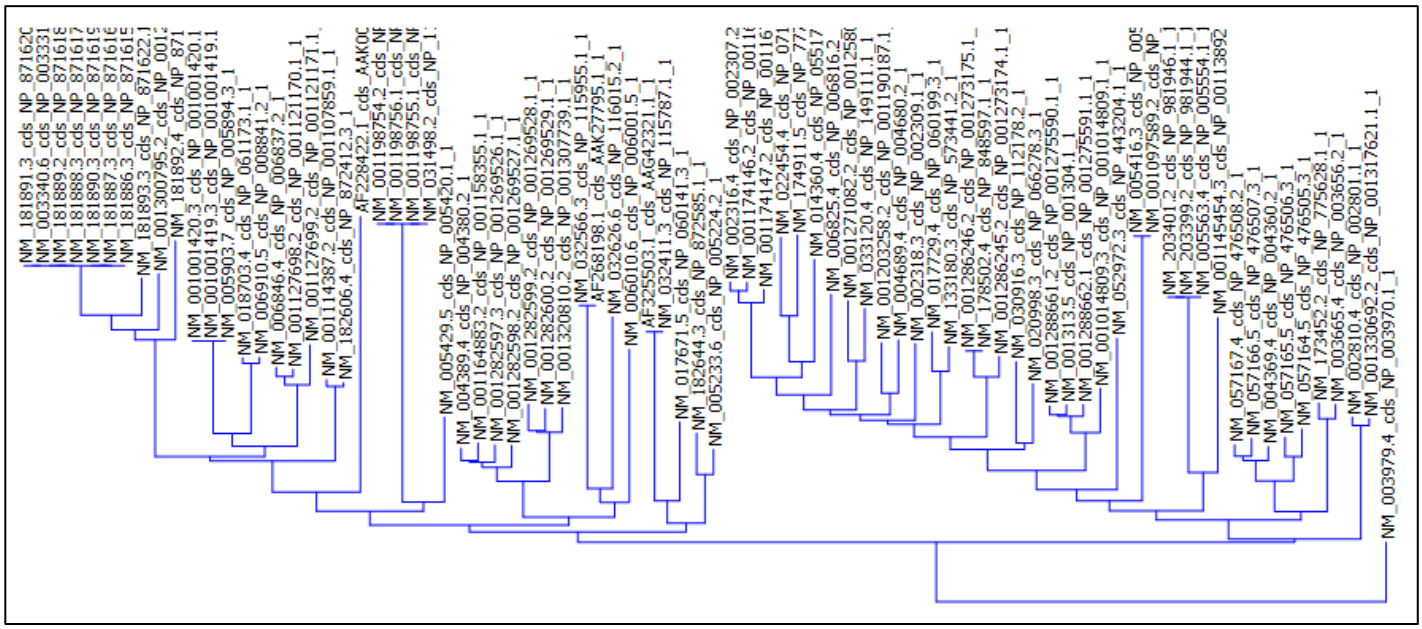

Figure 2: Cluster analysis for genes responsible for esophageal cancer. The gene cluster was constructed based on the neighbor joining method.

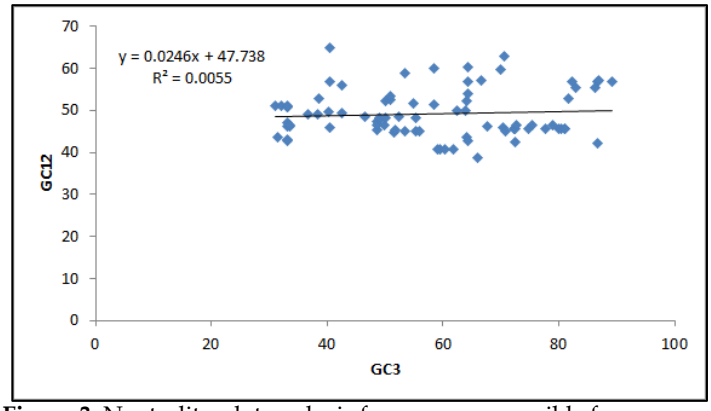

Figure 3: Neutrality plot analysis for genes responsible for esophageal cancer

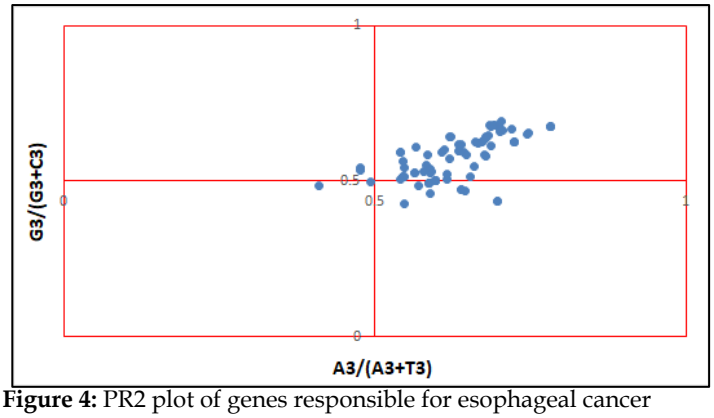

Figure 4: PR2 plot of genes responsible for esophageal cancer

Codon context analysis:

Recently it became evident that not only preference of single codon selection but also the codon pair influences the genes expression and mRNA structure [**27, 28**]. From our analysis, it was observed that 56.73% of codon pairs are over-represented and 26.26% of codon pairs are under-represented. 16.99% of codon pairs are absent in the genes. Heat map for the codon context analysis is shown in **Figure 5**. The top 10 over-represented codon pairs are GAG-GAG, GAG-CUG, GAG-AAG, CUG-GAG, AAG-AAA, CAG-AAG, GAU-GAC, GAA-GAA, GUG-GAG, and AAA-GAA. Similarly, the top 10 under-represented codon pairs are ACG-GAU, ACG-CGU, ACG-CGG, ACG-CGC, ACG-CAA, ACG-AGG, ACC-CAA, ACA-UCA, ACA-CUA, and ACA-CCA.

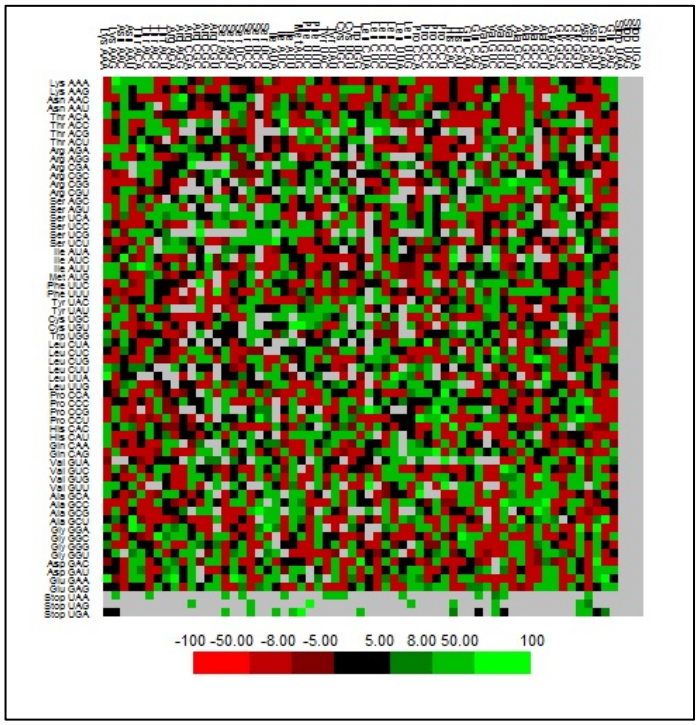

Figure 5: Heat map representing the codon pair context pattern of the esophagus cancer-related genes. Green, black and red-colored dots represents over, absent, and under-represented codon pairs, respectively

Discussion:

A diverse array of mechanisms regulates protein biogenesis. This phenomenon is more complex in multicellular organisms. A vast range of studies reported the role of transcriptional regulation in disease progression. Here, in the present study, we have done a detailed analysis of the nucleotide composition and their variation resulting in CUB between the genes responsible for the

development of esophageal cancer. Next, we did a systematic comparison with the results obtained by other researchers on the same genome or other genomes to identify similarities and differences. This will help to get new molecular insights into esophageal cancer. EC-related genes showed higher use of GC as compared to AT at the synonymous positions as well as in overall gene compositions. Previous reports suggest that the GC-rich genome prefers to use GC-ending codons whereas the AT-rich genome prefers to use AT-ending codons [**29**]. Therefore, our findings revealed that the nucleotide composition of the EC-related genes followed a similar pattern, corroborated well with the existing CUB reports on human genes **[30]**. GC composition pattern of a gene greatly influences its codon usage pattern **[31**]. EC-responsive genes showed an average ENc value of 49.28, which is significantly higher than 35. Extensive research on CUB proposed that ENc>35 should be considered as low CUB [**32, 33**]. In support of the ENc pattern observed in the CNS responsible genes, we observed low CUB of the EC-responsive genes [**30**]. The overall low CUB observed in multicellular organisms might be related to the replication process as different cell types have different codon preferences **[34]**. Transcription factor SOX-17 showed the highest CUB whereas, fermitin family homolog 1 showed the least CUB. Interestingly, a significant positive correlation was observed between the GC3 and ENc ($r=0.56$, $p<0.05$), suggest genes with higher GC3 have a lower CUB. A similar finding was reported by Malaker et al 2020 in a study conducted on mammalian genes **[23]**. Furthermore, GC showed strong positive correlations with GC endings codons **(Table 2)**. Therefore, our finding confirmed the relative contribution of GC compositional constraints under the strong mutational pressure in shaping the codon usage pattern of human genes. RSCU analysis revealed the over-representation of GAA, AGA, AGC, GTG, CTG, GAG, and CAG. Similarly the codons GTA, TCG, ATA, TTA, CCG, CGT, ACG, GCG, and CTA as under-represented. 66.66% and 44.44% GC-ending codons were found in over and under-represented codon groups. This confirms the significance of AT nucleotides in modulating the CUB in association with GC compositional constraints. Correspondence analysis identified the contribution of forces behind the observed CUB of EC-responsive genes, where mutation pressure showed dominance over selection pressure. Although a few codons and genes showed scatted distribution *i.e.* away from the center and major axes, respectively. Cluster analysis supported the COA analysis. In support of the findings reported by Zhang et al **[35]** here we hypothesized that the observed variation might be due to the prevalence of different EC-responsive genes in different cell types. Mutation pressure causes the proportional use of nucleotides i.e. A=T and G=C. Neutrality and PR2 plot revealed the disproportional use of nucleotides at synonymous codon positions. EC-responsible genes showed higher use of A/G nucleotides, suggest the dominance of selection pressure. Uddin et al observed higher use C at the synonymous codon position for CNS genes **[30]**. Therefore, the PR2 bias pattern observed in the present study can be used as fingerprints for the EC-responsive genes, which requires further validation and systematic comparison with other diseases. Estimation of the gene expression revealed that all the EC-responsive genes are highly expressive in nature (based on CAI value). A few previous reports confirmed the role of various skews in gene expression **[36]**. Similarly the EC-responsive genes studied here showed significant relationship with AT-skew, GC-skew, GRAVY, and aromaticity. Furthermore, findings of our codon context analysis revealed that GAN-NNG as the frequent contexts present in EC-responsive genes. Compositional properties of a gene affect the CUB and gene function **[37].**

Conclusion:

We show that CAG and GTA are over-represented and under-represented codons, respectively in genes linked with esophageal cancer. Correspondence analysis, neutrality plot, and parity rule 2 plot analysis confirmed the dominance over mutation pressure in modulating the codon usage pattern of genes linked with esophageal cancer.

Acknowledgments:

Authors are also grateful to the Department of Electronics Communication and Engineering, Gauhati University, Assam, India for providing the necessary research facilities to carry out this work.

Conflicts of interest:

The authors declare that no conflict of interest exists for this work.

References:

[1] Quax TEF *et al. Molecular Cell,* 2015. **59**:149.[PMID: 26186290]
[2] Lynch M, *Proceedings of the National Academy of Sciences*, 2010. **107**:961.[PMID: 20080596]
[3] Lynch M *TRENDS in Genetics*, 2010. **26**: 345.[PMID: 27739533]
[4] Supek F *et al. Cell,* 2014. **156**: 1324.[PMID: 24630730]
[5] Hodgman *et al. Nucleic acids research*, 2020. **48**:11030.[PMID: 33045750]
[6] Sauna ZE & C Kimchi-Sarfaty, *Nature Reviews Genetics*, 2011. **12**:683.[PMID: 21878961]
[7] Fornasiero EF & S O Rizzoli, *BMC genomics*, 2019. **20**: 1.[PMID: 31288782]
[8] Miller JE *et al. Proceedings of the Pacific Symposium*. 2018. 23:365. [PMID: 29218897]
[9] Son H *et al. Scientific reports*, 2017. **7**: 1. [PMID: 29079855]
[10] Zhou W *et al. Medicine*, 2020. **99**: e20340. [PMID: 32443386]
[11] Wang X *et al. Oncology letters*, 2018. **15**: 8983. [PMID: 29844815]
[12] Mathur P *et al.* JCO *Global Oncology*, 2020. **6**:1063. [PMID: 32673076]
[13] Yu VZ *et al.* AACR. 2016:1158.
[14] Wright F *Gene*, 1990. **87**:23. [PMID: 2110097]
[15] Sharp PM & WH Li *Journal of molecular evolution*, 1986. **24**: 28. [PMID: 3104616]
[16] Sueoka N *Proceedings of the National Academy of Sciences*, 1988. **85**: 2653. [PMID: 3357886]
[17] Mazumder GA *et al. Infection, Genetics and Evolution*, 2018. **57**:128. [PMID: 29066170]
[18] Greenacre MJ 1984 Academic Press, BOOK title needed London.http://www.carme-n.org/?sec=books5
[19] Moura G *et al. Genome biology*, 2005. **6**:R28. [PMID: 15774029]
[20] Vetrivel U et al. *Bioinformation*, 2007. **2**:62. [PMID: 18188422]
[21] Supek F& K Vlahoviček *Bioinformatic*s, 2004. **20**: 2329. [PMID: 15059815]
[22] Hammer Øet *al. Palaeontologia electronica*, 2001. **4: 1**:
[23] Malakar AK *et al. Genomics*, 2020. **112**: 1319. [PMID: 31377427]
[24] Wei L *et al. BMC evolutionary biology*, 2014. **14**: 1 [PMID: 25515024]
[25] Deb B*et al. Archives of virology*, 2020. **165**: 557. [PMID: 32036428]
[26] Sueoka N, *Cold Spring Harb Symp Quant Biol*. 1961.26:35. [PMID: 13918160]
[27] Moura G *et al. Plos one,* 2007. **2**: e847. [PMID: 17786218]

[28] Moura G *et al. Genome biology*, 2005. **6**: 1. [PMID: 15774029]
[29] Sun Y*et al. Genome biology and evolution*, 2017. **9**:2560. [PMID: 27540085]
[30] Uddin A & S. Chakraborty, *Molecular neurobiology*, 2019. **56**:1737. [PMID: 29922982]
[31] Li J *et al.* G3: *Genes, Genomes, Genetics*, 2015. **5**: 2027. [PMID: 26248983]
[32] Wang L *et al. PLoS One,* 2018. **13**: p. e0194372. [PMID: 29584741]
[33] He Z*et al. Viruses,* 2019. **11**: 752. [PMID: 31416257]
[34] Jenkins GM & EC Holmes, *Virus research*, 2003. **92**: 1. [PMID: 12606071]
[35] Zhang Z *et al. Archives of virology*, 2013. **158**: 145. [PMID: 23011310]
[36] Fujimori S*et al. BMC genomics*, 2005. **6**: 1. [PMID: 15733327]
[37] Garcia JA *et al. Molecular phylogenetics and evolution*, 2011. **61**: 650. [PMID: 21864693]

Edited by P Kangueane
Citation: Bordoloi & Nirmala, Bioinformation 17(8): 731-740 (2021)