International Journal of Health and Allied Sciences

Manuscript 1016

Research Article A new Informatics Framework for Evaluating the Codon Usage Metrics, Evolutionary Models and Phylogeographic reconstruction of Tomato yellow leaf curl virus (TYLCV) in different regions of Asian countries

Mamathashree Mn

Kuralyanapalya Putta Honnappa Suresh

Sharanagouda S Patil

Uma Bharathi Indrabalan

Mallikarjun S Beelagi

See next page for additional authors

Research Article A new Informatics Framework for Evaluating the Codon Usage Metrics, Evolutionary Models and Phylogeographic reconstruction of Tomato yellow leaf curl virus (TYLCV) in different regions of Asian countries

Authors

Mamathashree Mn, Kuralyanapalya Putta Honnappa Suresh, Sharanagouda S Patil, Uma Bharathi Indrabalan, Mallikarjun S Beelagi, Sushma Pradeep, Krishnamoorthy Paramanandham, Siju Susan Jacob, Chandrashekar Srinivasa, Shiva Prasad Kollur, Raghu Ram Achar, Ashwini Prasad, Shashanka K Prasad, and Chandan Shivamallu

ORIGINAL STUDY

A New Informatics Framework for Evaluating the Codon Usage Metrics, Evolutionary Models and Phylogeographic Reconstruction of Tomato Yellow Leaf Curl Virus (TYLCV) in Different Regions of Asian Countries

Mamathashree Mn [a], Kuralyanapalya Putta Honnappa Suresh [a,b,c,d,e,f,g,*], Sharanagouda S. Patil [a], Uma B. Indrabalan [a], Mallikarjun S. Beelagi [a], Sushma Pradeep [b], Krishnamoorthy Paramanandham [a], Siju S. Jacob [a], Chandrashekar Srinivasa [c], Shiva P. Kollur [d,e], Raghu R. Achar [f], Ashwini Prasad [g], Shashanka K. Prasad [b], Chandan Shivamallu [b]

[a] ICARNational Institute of Veterinary Epidemiology and Disease Informatics (NIVEDI), Yelahanka, Bengaluru 560064, India
[b] Department of Biotechnology and Bioinformatics, School of Life Sciences, JSS Academy of Higher Education and Research, Mysore, 570015, Karnataka, India
[c] Department of Biotechnology, Davangere University, Shivagangotri, Davangere, 577002, Karnataka, India
[d] School of Agriculture, Geography, Environment, Ocean and Natural Sciences, The University of the South Pacific, Laucala Campus, Suva, Fiji
[e] Department of Sciences, Amrita School of Arts and Sciences, Amrita Vishwa Vidyapeetham, Mysuru Campus, Mysuru, 570 026, Karnataka, India
[f] Department of Biochemistry, School of Life Sciences, JSS Academy of Higher Education and Research, Mysore, Karnataka, India
[g] Department of Microbiology, School of Life Sciences, JSS Academy of Higher Education and Research, Mysore, Karnataka, India

Abstract

Tomato yellow leaf curl virus (TYLCV) is a major devastating viral disease, majorly affecting the tomato production globally. The disease is majorly transmitted by the Whitefly. The *Begomovirus* (TYLCV) having a six major protein coding genes, among them the C1/AC1 is evidently associated with viral replication. Owing to immense role of C1/AC1 gene, the present study is an initial effort to elucidate the factors shaping the codon usage bias and evolutionary pattern of TYLCV-C1/AC1 gene in five major Asian countries. Based on publicly available nucleotide sequence data the Codon usage pattern, Evolutionary and Phylogeographic reconstruction was carried out. The study revealed the presence of significant variation between the codon bias indices in all the selected regions. Implying that the codon usage pattern indices (eNC, CAI, RCDI, GRAVY, Aromo) are seriously affected by selection and mutational pressure, taking a supremacy in shaping the codon usage bias of viral gene. Further, the tMRCA age was 1853, 1939, 1855, 1944, 1828 for China, India, Iran, Oman and South Korea, respectively for TYLCV-C1/AC1 gene. The integrated analysis of Codon usage bias, Evolutionary rate and Phylogeography analysis in viruses signifies the positive role of selection and mutational pressure among the selected regions for TYLCV (C1/AC1) gene.

Keywords: Tomato yellow leaf curl virus (TYLCV), Tomato, C1/AC1 protein gene, Asian countries, Codon usage bias, Evolutionary characters, Phylogeography analysis, Positive selection, tMRCA

Received 21 March 2022; revised 15 July 2022; accepted 20 September 2022.
Available online 16 December 2022

* Corresponding author.
E-mail addresses: sureshkp97@rediffmail.com (K.P.H. Suresh), mallikbeelagi@gmail.com (M.S. Beelagi), sushmap@jssuni.edu.in (S. Pradeep), p.krishnamoorthy@icar.gov.in (K. Paramanandham), siju.jacob@icar.gov.in (S.S. Jacob), chandans@jssuni.edu.in (C. Shivamallu).

https://doi.org/10.55691/2278-344X.1016

1. Introduction

Tomato yellow leaf curl virus (TYLCV) is a major noxious *Begomovirus,* which is a prime threat to the tomato production extensively [1]. Heretofore, TYLCV is one of the major devastating virus, which has a wide host range yet it's quite severe in tomato (*Solanum lycopersicum*), and also infects other plant hosts such as eggplant (*Solanum melongena*), potato (*Solanum tuberosum*), tobacco (*Nicotiana tabacum*), bean (*Phaseolus vulgaris*), sweet pepper (*Capsicum annuum*), chili pepper (*Capsicum chinense*) and lisianthus *(Eustoma grandiflorum)* [2]. Tomato is one of major crop grown worldwide, and TYLCV have alarming impact on its productions. The classical disease symptoms in tomato are, chlorotic (yellow) leaf edges, upward leaf cupping, leaf mottling, reduced leaf size, and flower drop, early infected plants won't bear fruit and their growth will be severely stunted, which causes 100% yield reduction [3,4]. The virus is phloem limited in its hosts, and its extensively transmitted by sweet potato whitefly (*Bemisia tabaci*), in a persistent and circulate manner [5–9].

The manopartite ssDNA virus TYLCV belong to the genus *Begomovirus* of the family *Geminiviridae.* with circular genome (2787 nt in size) [10]. TYLCV has a characteristic twinned morphology [11]. Like other gemini-viruses, TYLCV capsid (total MW 3,330,000) consists of two joined, incomplete icosahedra, with a T = 1 surface lattice containing a total of 22 capsomeres, each containing five units of a 260 amino acid coat protein (CP) of 30.3 kDa [12]. The TYLCV genome encodes for six open-reading frames (ORF) that can code for products larger than 10 kDa, with two large genes on the viral strand (V1 and V2), and four on the complementary sense strand (C1, C2, C3, C4) [13]. V1 encodes for coat protein (which protect the viral DNA by encapsulation) [10]. V2 encodes pre-coat protein or movement-like protein (which is associated with viral movement and act has suppressor of RNA silencing) [4]. C1/AC1 encodes viral replication-associated protein (Rep-essential for virus replication), C2 encodes for transcriptional activator protein (TrAP- act as post-transcriptional gene silencing suppressor), C3 encodes for replication enhancer protein (Ren-act as virus accumulation enhancer) and C4 encodes for symptom induction and movement determinant [2,14]. TYLCV DNA also contains an intergenic region (29 nucleotide-long), stem-loop structure with the conserved nano-nucleotide TAATATTAC sequence, which act as cleaving site during replication of the viral genome, according to the rolling circle model [15].

During the gene translation (gene to protein), several triplet codons are preferentially used over the other synonymous codons is known as codon usage bias. Molecular-evolutionary studies imply that the codon usage bias is ubiquitous across genome and contributes to genome evolution caused mainly by mutation pressure and protein gene expression [16]. The codon usage bias depends on various factors such as mutational pressure, natural selection, gene expression level, gene length, composition bias (G + C % content and GC skew), recombination rate, RNA stability, position of codon in the gene and structure of virus [17]. The codon usage is also widely influenced by the hosts and virus, which has an impact for the viral existence (for immunity, aptness, host resistance, and evolution). The synonymous codon usage is non-random with mutational and natural selection pressure is a prime factor in deviation from the equal usage of synonymous codons [18].

Codon usage patterns analysis in viruses can provide important insights into the virus molecular evolution, gene expression & its regulations and protein synthesis [19,20]. Nevertheless, the codon usage patterns of TYLCV and its hosts, as well as the association between them, have not been elucidated so far. Therefore, pinning down the codon usage bias of TYLCV provides a basic knowledge on regulation of gene expression and molecular evolution of the virus. Owing to the importance of TYLCV-C1/AC1 protein in virus replication, the present study is a primary effort to elucidate the factors shaping the codon usage bias and evolutionary pattern of TYLCV-C1/AC1 gene in five major Asian countries such as China, India, Iran, Oman and South Korea using available sequence data.

2. Materials and methods

2.1. Data assembly and sequence editing

The C1/AC1 coding sequences (CDs) of five major Asian countries such as China, India, Iran, Oman and South Korea were collected from the GenBank database, National Center for Biotechnology Information (NCBI) (https://www.ncbi.nlm.nih.gov/nucleotide/). The C1/AC1 coding sequences for all the five regions (245, 13, 40, 31 and 30 respectively) were downloaded in FASTA format and used for further analysis. The multiple alignment and editing of sequences were performed using the MEGA-X

(Molecular Evolutionary Genetics Analysis) software.

2.2. Nucleotide composition

The Nucleotide component parameters of TYLCV (gene C1/AC1) genomic sequences was analysed by online software, codonW. The frequency of occurrence of Adenine, Cytosine, Guanine, and Thymine was obtained from the R software. The frequency of nucleotide bases at third codon position (A3, C3, G3, and T3) was obtained from the MEGA-X software. Additionally, GC frequency at first (GC1), second (GC2) and third (GC3) positions were calculated and mean frequency of GC contents at the first two codon position (GC12) was calculated with R software.

2.3. Analysis of codon usage

2.3.1. Relative synonymous codon usage (RSCU)

The codon usage pattern is primly determined by the relative synonymous codon usage (RSCU) values. It's determine the bias existing among the genes and within the genes (varies with their sizes and composition of codons that encodes a respective amino acid). RSCU is a ratio of observed frequency observations to the expected frequency observation of a specific amino acid on every codon. The stronger bias in the codon usage is specified with the RSCU estimates. The codon with RSCU value > 1.0 has positive codon usage bias, and the value < 1.0 has relative negative codon usage bias. When RSCU value is equal to 1.0, it means that this codon is chosen equally and randomly. The RSCU values of gene C1/AC1 sequences were calculated using MEGA-X software. The amino acids encoded by single codon AUG (Met) and UGG (Trp), and the termination codons UAA, UAG, and UGA were excluded from the analyses.

2.3.2. Effective Number of Codons (eNC)

The Effective Number of Codons (eNC) value is a quiet and definite compute of codon usage bias of a gene/genome, which reflects the codon deviation from random selection. The major focus of eNC is to estimate the degrees of departure from the equal use of synonymous codons of coding regions of C1/AC1 gene. The eNC values varies between 20 and 60, eNC values varies from 20 to 60. The closer the value to 21, specifies higher codon bias (amino acids are encoded by only one synonymous codon); the closer the value to 60, specifies lower codon bias (amino acids are encoded by all synonymous codons equally); higher the values than 60, specifies no codon bias and high-risk codon bias if the values lower than 20. Genes, whose codon usage is reliance on mutation bias, will lie down just below the curve of the predicted values. The eNC values for C1/AC1 genes were calculated using software codonW.

2.4. Analysis of natural selection and mutation bias

2.4.1. Codon adaptation index (CAI)

The Codon adaptation index (CAI) estimates the relative adaptiveness of a gene towards C1/AC1 gene codon usage bias, which varied from 0 to 1. High CAI value specifies a preferred adaptiveness. The CAI values for TYLCV (C1/AC1) gene and reference datasets of synonymous codon usage pattern of the tomato (*S. lycopersicum*) were estimated using CAIcal (http://genomes.urv.es/CAIcal/).

2.4.2. Relative codon de-optimization index analysis (RCDI)

The Relative codon de-optimization index analysis (RCDI) is a quantification of comparison between codon usage frequencies of a gene and a reference genome [21]. If the RCDI value is close to 1, indicates the codon usage bias of virus is similar to host, which considered as higher translation rate [22]. The RCDI values for TYLCV (C1/AC1) gene and reference datasets of synonymous codon usage pattern of the tomato (*S. lycopersicum*) were estimated using CAIcal (http://genomes.urv.es/CAIcal/).

2.4.3. Relative dinucleotide abundance frequency

The dinucleotide biases generally affect the codon bias, and its mainly influenced by the natural selection and mutational pressure [23]. The relative abundance of dinucleotides in the coding regions of TYLCV (C1/AC1) genomes was assessed using the method described by Karlin and Burge [24], it represents a total of 16 dinucleotides composition of gene C1/AC1. The dinucleotide frequency is a ratio of,

$$P_{XY}=F_{XY}/F_X\ F_Y$$

where, F_X denotes the frequency of the nucleotide X, F_Y denotes the frequency of the nucleotide Y, $F_X\ F_Y$ the expected frequency of the dinucleotide XY and F_{XY} the frequency of the dinucleotide XY, for each dinucleotide were calculated. As a conservative criterion, for $P_{XY} > 1.23$ (or < 0.78), the XY pair is considered to be of over-represented (or under-represented) relative abundance compared with a random association of mononucleotides. However, dinucleotide abundant frequencies, $P_{XY} \geq 1.50$ is assessed as extremely overrepresented, and P_{XY}

≤0.50 as underrepresented. The dinucleotide abundance frequency was calculated in R software.

2.4.4. eNC-plot mapping

eNC-plot (eNC *vs* G3s) used to estimate the influence of mutation or selection pressure on codon bias of a gene, which indicates the relationship between eNC values plotted against GC3 (GC at third position) values. eNC plot was generated with the ggplot2 library of R software. The eNC values recline on or near the standard curve indicates that the bias is affected by mutational pressure, while the values recline below the curve indicates influence of natural selection in framing the codon usage bias.

2.4.5. Parity rule 2 bias plot

To estimate the compositional bases bias in purines and pyrimidines usage, the Chargaff's second parity rule (PR2) is used. The relation between the nucleotide bases purines (A and G) pyrimidine (T and C) at the third codon position used to plot the PR2 bias, was plotted as (G3/(G3+C3)) on X-axis and AT bias (A3/(A3+T3)) on the Y-axis. According to Chargaff's second parity rule the ratio between the base compositions i.e. A = T and G = C should be 1:1, then it is concluded that there is no deviance between natural selection pressure and mutation pressure. The PR2 bias plot was generated with library ggplot2 in R software. The PR2 estimates degree of deviation indicates that the bias might be due to natural selection, mutation pressure, or both. If the PR2 estimates are found evenly plotted, then the bias is entirely due to mutation pressure.

2.4.6. Neutrality plot mapping

The neutrality plot analysis determines how two variables (GC12 and GC3) related, they plotted against one another (GC12: ordinates & GC3: abscissa) to investigate the mutation−selection equilibrium (dominant factors) in framing the codon usage bias. The regression coefficient of GC3s is significant or close to 1, indicates the mutation pressure is the major strength to framing the codon usage, if the slope is ≤ 0 indicates lower mutational pressure and the natural selections (slope = 0) playing a role in codon usage bias. The neutrality plot was generated with R software.

2.4.7. General average hydropathicity (GRAVY) and aromacity (Aromo)

Biochemical parameters of proteins, i.e., aromaticity and hydropathicity, is contributes to framing of the codon usage bias. The average of the hydropathicity scores of the amino acids in the coding sequence is known as General Average Hydropathicity Score (GRAVY). Its commonly varies from −2.0 to +2.0, the hydrophilic proteins specify the negative values, while the hydrophobic proteins for positive. The aromatic amino acids (tryptophan, tyrosine, and phenylalanine) distribution in the proteins, indicated as AROMO values. The GRAVY and AROMA were estimated using the software CodonW.

2.5. Correlation analysis

Correlation is carried out to understand the statistical relationship of variables. The relationship between codon usage bias indices (eNC, CAI, RCDI, GRAVY, Aromo) of TYLCV (C1/AC1) gene were estimated by the Spearman's rank correlation method in all the selected five major Asian countries [21].

2.6. Evolutionary analysis

2.6.1. Data sequence and alignment

The nucleotide sequence of TYLCV (C1/AC1) gene of all five different Asian countries were downloaded from NCBI website (nih.gov). A total of 245, 13, 40, 31 and 30 sequences of China, India, Iran, Oman and South Korea respectively were downloaded in FASTA format and used for further analysis. The multiple sequence alignment and sequence editing were individually aligned using the MEGA-X software by incorporating the MUSCLE algorithm.

2.6.2. Evolutionary rate and coalescent analysis

The phylogenetic analysis is prime to understand the evolutionary tie between the individuals. The basic criteria of construction of phylogenetic analysis based on the choices of statistical best-fit models. Accordingly, the phylogenetic models were selected based on the Akaike Information Criteria (AIC) and Bayesian Information Criteria (BIC) obtained from the jModelTest2 tool. To build the input analysis, the BEAUti interface of the BEAST software was used. The four molecular models such as Relaxed Clock log-normal (RCLN), Relaxed Clock Exponential (RCE), Strict clock and Random Local Clock (RLC) were considered along with Coalescent: Bayesian skyGrid and Coalescent: Extended Bayesian sky plot trees [25]. To measure the evolutionary rate and time to most recent common ancestor, an algorithm Markov Chain Monte Carlo (MCMC) in the Bayesian Analysis by Sampling Tree (BEAST) was used [26]. The MCMC chain length was frequently changed until all the constraints had an effective sample size (ESS) extent value of >200.

The log files generated by the BEAST software were analysed by using the BEAST integrated Tracer tool. Additionally, phylogenetic tree was constructed by Bayesian stochastic search variable selection (BSSVS) by tracking the geographical locations state [27]. To estimate the most significant route dispersal, the SPREAD software was used [28]. The visual representation of spatiotemporal diffusion was generated in Google Earth (https://earth.google.com) using the Continuous Tree module in SPREAD software.

2.6.3. Determination of selection pressure or positive selection

To evaluate the positive selection pressure, the Datamonkey Adaptive Evolution (DAE) server is used to eliminate the superfluous sequences in a dataset of evolutionary rate analysis. The selection pressure is determined mainly by considering the ratios of non-synonumous (dN) to synonymous (dS) substitutions. The Fixed-Effects Likelihood (FEL) was employed for the evaluation of dN, dS, and dN/dS (ω) rate per site of a coding alignment sequences. Further, the FEL strategy expects that the determination of positive selection for each site is consistent along the whole phylogeny [29].

3. Results

3.1. The characteristics of synonymous codon usage in TYLCV (C1/AC1)

To evaluate the influencing factors of codon usage bias for the TYLCV (C1/AC1) gene, bellow listed traits were estimated.

3.1.1. Data assembly and sequence editing of TYLCV-C1/AC1

The coding sequences of C1/AC1 gene of five different Asian regions such as China (n = 245), India (n = 13), Iran (n = 40), Oman (n = 31) and South Korea (n = 30) of TYLCV virus were obtained from the GenBank database (NCBI). The sequences with homogeneity higher than 99% were eliminated from the study. The TYLCV (C1/AC1) coding sequences were multi-aligned, edited, and estimation of nucleotides were carried out using MEGA-X software.

3.1.2. Nucleotide content and composition analysis of TYLCV-C1/AC1

The nucleotide composition is the major deciding factor of codon usage bias pattern [21]. While, the nucleotide content frequency (A, T, G, C; third position nucleotides A3, T3, G3, C3; G + C contents GC, GC1, GC2, GC3) of TYLCV (C1/AC1) gene was assess to understand the nucleotide composition impact on codon usage bias. The nucleotide bases composition of TYLCV (C1/AC1) for all the five Asian countries have been calculated and represented in the Table 1 and Figs. 1 and 2.

1. China: A (33.17%), T (24.65%), G (19.41%) and C (22.78%), nucleotide bases at third codon position A3 (41.96%), C3 (33.78%), G3 (20.16%) and, T3 (35.30%) and The GC (42.31%), GC1 (47.57%), GC2 (36.75%), and GC3 (40.87%).
2. India: A (32.69%), T (25.05%), G (19.36%) and C (22.90%), nucleotide bases at third codon position A3 (40.76%), C3 (32.07%), G3 (20.40%) and, T3 (37.79%) and The GC (42.33%), GC1 (47.16%), GC2 (38.17%), and GC3 (39.68%).
3. Iran: A (32.61%), T (24.95%), G (19.49%) and C (22.95%), nucleotide bases at third codon position A3 (40.89%), C3 (32.21%), G3 (21.36%) and, T3 (36.31%) and The GC (42.56%), GC1 (47.12%), GC2 (38.06%), and GC3 (40.61%).
4. Oman: A (31.92%), T (24.46%), G (19.65%) and C (23.96%), nucleotide bases at third codon position A3 (39.02%), C3 (34.43%), G3 (21.32%) and, T3 (35.21%) and The GC (43.70%), GC1 (47.10%), GC2 (39.29%), and GC3 (42.72%).

Table 1. The nucleotide indices of TYLCV (C1/AC1) isolates with mean and standard deviations.

Nucleotides	China	India	Iran	Oman	South Korea
A	33.17% ± 0.20	32.69% ± 0.87	32.61% ± 0.54	31.92% ± 0.45	33.38% ± 0.20
T	24.65% ± 0.13	25.05% ± 0.43	24.95% ± 0.64	24.46% ± 0.22	24.69% ± 0.09
G	19.41% ± 0.15	19.36% ± 0.27	19.49% ± 0.56	19.65% ± 0.45	19.21% ± 0.08
C	22.78% ± 0.21	22.90% ± 0.31	22.95% ± 0.66	23.96% ± 0.27	22.72% ± 0.19
A3	41.96% ± 0.59	40.76% ± 2.54	40.89% ± 1.45	39.02% ± 1.81	42.40% ± 0.79
T3	35.30% ± 0.31	37.79% ± 2.10	36.31% ± 1.13	35.21% ± 0.57	35.74% ± 0.34
G3	20.16% ± 0.65	20.40% ± 1.14	21.36% ± 1.66	21.32% ± 1.33	19.89% ± 0.67
C3	33.78% ± 0.41	32.07% ± 0.80	32.21% ± 1.40	34.43% ± 1.01	33.44% ± 1.00
GC	42.31% ± 0.26	42.33% ± 0.52	42.56% ± 0.93	43.70% ± 0.52	42.05% ± 0.22
GC1	47.57% ± 0.40	47.16% ± 0.23	47.12% ± 0.60	47.10% ± 0.57	47.61% ± 0.32
GC2	36.75% ± 0.34	38.17% ± 1.34	38.06% ± 1.07	39.29% ± 0.39	36.48% ± 0.82
GC3	40.87% ± 0.48	39.68% ± 0.59	40.61% ± 1.56	42.72% ± 1.32	40.31% ± 0.99

Table 2. The RSCU frequency of TYLCV (C1/AC1) gene of five major Asian countries. Overrepresented codons (>1.6) are highlighted in light red and underrepresented codons (<0.6) are in green.

	Relative synonymous Codon Usage Frequency				
Codon	China	India	Iran	Oman	South Korea
UUU	0.68	0.82	0.67	0.79	0.75
UUC	1.32	1.18	1.33	1.21	1.25
UUA	1.81	1.65	1.68	1.59	1.79
UUG	0.21	0.29	0.44	0.4	0.21
CUU	0.81	0.95	0.85	0.99	0.82
CUC	1.59	1.43	1.52	1.68	1.55
CUA	1.18	1.14	1.13	0.93	1.2
CUG	0.39	0.53	0.38	0.41	0.43
AUU	1.55	1.44	1.51	1.43	1.57
AUC	0.84	1.02	0.95	1.07	0.78
AUA	0.62	0.54	0.54	0.5	0.65
GUU	1.16	0.91	1.12	0.99	1.17
GUC	1.37	1.43	1.31	1.83	1.4
GUA	0.71	0.78	0.64	0.67	0.54
GUG	0.77	0.87	0.92	0.52	0.88
UCU	1.68	1.77	1.76	1.46	1.7
UCC	1.12	1.2	1.12	1.31	1.11
UCA	1.13	1.04	1.19	1.24	1.15
UCG	0	0.22	0.05	0.09	0
CCU	0.73	1.07	1.04	0.92	0.76
CCC	0.72	0.69	0.56	0.9	0.74
CCA	2.18	1.75	1.9	1.55	2.12
CCG	0.37	0.49	0.51	0.63	0.37
ACU	0.33	0.67	0.77	1.15	0.44
ACC	1.65	1.49	1.56	1.59	1.56
ACA	2.01	1.8	1.65	1.22	2
ACG	0.01	0.04	0.02	0.04	0
GCU	0.73	1.2	0.94	1.04	0.63
GCC	1.51	1.22	1.27	1.38	1.59
GCA	1.28	1.21	1.23	1.2	1.23
GCG	0.47	0.37	0.56	0.38	0.55
UAU	1.19	1.11	1.23	1.35	1.2
UAC	0.81	0.89	0.77	0.65	0.8
CAU	1.6	1.39	1.51	1.46	1.59
CAC	0.4	0.61	0.49	0.54	0.41
CAA	1.64	1.44	1.45	1.27	1.72
CAG	0.36	0.56	0.55	0.73	0.28
AAU	1.28	1.38	1.34	1.33	1.29
AAC	0.72	0.62	0.66	0.67	0.71
AAA	1.07	1.22	1.06	1.26	1.09
AAG	0.93	0.78	0.94	0.74	0.91
GAU	0.94	1.02	1	0.89	0.96
GAC	1.06	0.98	1	1.11	1.04
GAA	1.17	1.21	1.27	1.03	1.24
GAG	0.83	0.79	0.73	0.97	0.76
UGU	1.19	1.08	0.95	0.35	1.21
UGC	0.81	0.92	1.05	1.65	0.79
CGU	0.48	0.35	0.2	0.03	0.56
CGC	0	0	0.15	0.03	0.02
CGA	0.02	0.14	0.26	0.13	0.03
CGG	0.92	0.56	0.74	0.75	1
AGU	1.1	0.96	0.99	0.98	0.99
AGC	0.97	0.82	0.89	0.92	1.05
AGA	2.3	2.65	2.4	3.04	2.26
AGG	2.29	2.3	2.25	2.03	2.13
GGU	0.68	0.87	0.56	0.49	0.69
GGC	1.06	0.8	0.8	0.5	0.96
GGA	1.35	1.45	1.64	1.81	1.34
GGG	0.91	0.88	1	1.2	1.01

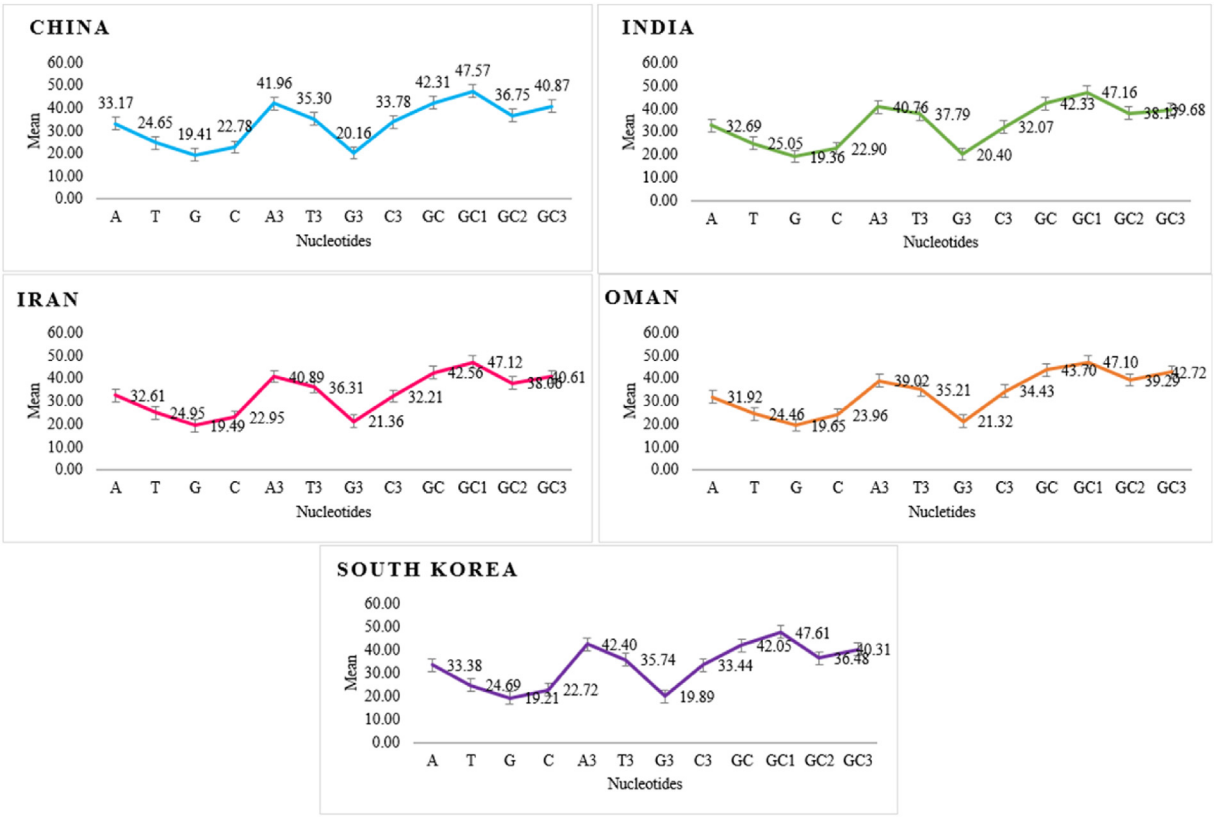

Fig. 1. Graphical representation of overall nucleotide composition for gene C1/AC1 of TYLCV virus in five major countries of Asia. Error bars are indicative of standard deviation.

5. South Korea: A (33.38%), T (24.69%), G (19.21%) and C (22.72%), nucleotide bases at third codon position A3 (42.40%), C3 (33.44%), G3 (19.89%) and, T3 (35.74%) and The GC (42.05%), GC1 (47.61%), GC2 (36.48%), and GC3 (40.31%).

3.1.3. Relative synonymous codon usage (RSCU) of TYLCV-C1/AC1

To explore the extent of codon usage bias in TYLCV, all RSCU values of different codon in all the five major countries of Asia were calculated. Among the 64 codons only 59 codons were used to estimate the RSCU values, the other 5 codons were excluded because ATG and TGG codes only for a single amino acid, and TAG, TAA, TGA stop codons that don't code for any amino acid. Out of 64 codons only 59 codons considered to assess the RSCU values (ATG and TGG-single amino acid coding codons, and TAG, TAA, TGA-stop codons were excluded). The frequency value of synonymous codon is segregated based on RSCU values ranged from 0.6 (<0.6-underrepresented) to1.6 (>1.6-over-represented). The estimated over and underrepresented codons are highlighted light-red and green colours, respectively. The RSCU values > 1.0 are known as positively biased or high-frequency, while, RSCU values < 1.0 is known as negatively biased or low-frequency (Table 3). The overall RSCU frequencies of TYLCV (C1/AC1) gene are represented in the Heat map (Fig. 4).

The estimated RSCU values of,

1. China: Among the 59 codons, 28 codons were positively biased and 31 codons were negatively biased. The 8 over-represented and 12 under-represented codons were observed. while, the codons UCG and CGC were having “0” value, indicates those codons were never used and CCA (2.18), ACA (2.01), AGA (2.3) and AGG (2.29) codons were highly over expressed.
2. India: The 28 codons were positively biased and 31 codons were negatively biased. The 6 over-represented and 12 under-represented codons were observed. while, the codon CGC was never used and the codons AGA (2.65) and AGG (2.3) were highly over expressed.
3. Iran: The 30 codons were positively biased and 29 codons were negatively biased. 7 over-represented and 14 under-represented codons were observed. while, the codons GAU, GAC, and GGG having the value of “1” indicates they were

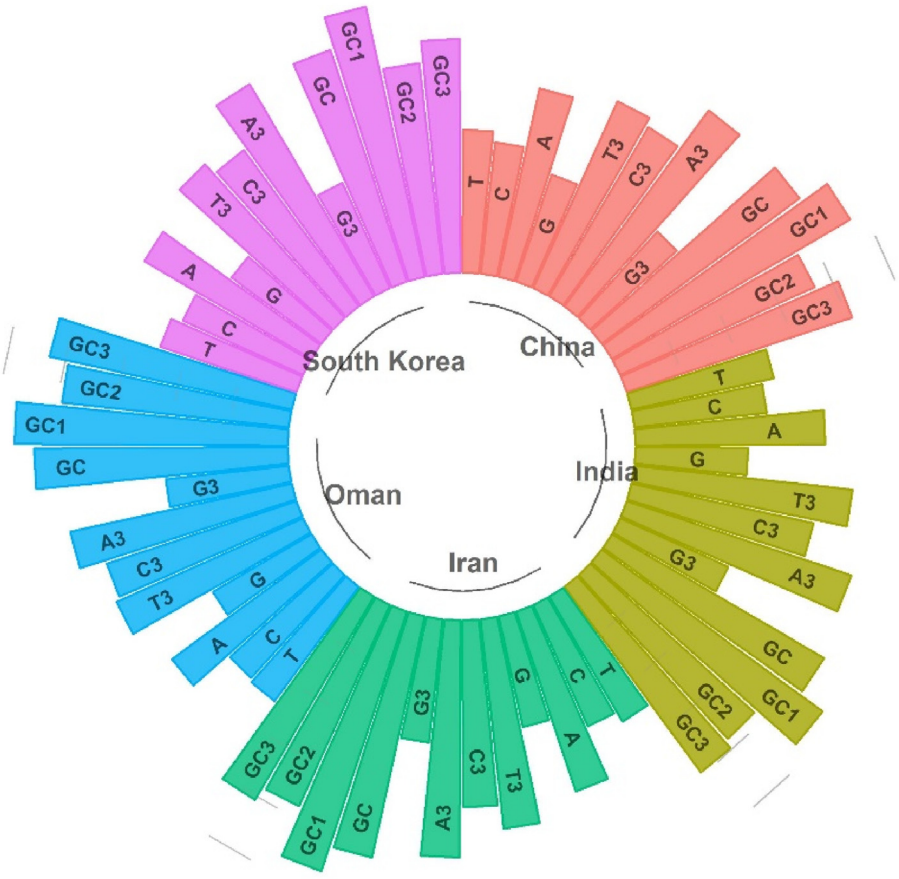

Fig. 2. Mononucleotide abundance frequencies of TYLCV (C1/AC1) gene for five major Asian countries.

having no bias and the codons AGA (2.4) and AGG (2.25) were highly over expressed.

4. Oman: The 28 codons were positively biased and 31 codons were negatively biased. 6 over-represented and 14 under-represented codons were observed. while, the codons AGA (3.04) and AGG (2.03) were highly over expressed.
5. South Korea: The 29 codons were positively biased and 30 codons were negatively biased. 7 over-represented and 13 under-represented codons were observed. while, the codons UCG and ACG were having "0" value that indicates those codons were never used, codon CGG having the value of "1" indicates that was having no bias and the codons CCA (2.12) and ACA (2.0) were highly over expressed.

3.1.4. Effective Number of Codons (eNC)

The extent of codon usage bias is measured by the effective Number of Codons (eNC). eNC value is calculated for all the five countries for the TYLCV (C1/AC1) gene. The eNC values observed for China, India, Iran, Oman, and South Korea was ranged as 51.62–57.13 (SD ± 0.56) with mean of 53.21, 51.12–56.55 (SD ± 1.62) with mean of 53.79, 49.35–56.75 (SD ± 2.08) with mean of 53.60, 50.36–55.03 (SD ± 1.15) with mean of 52.51 and 52.54–54.77 (SD ± 0.77) with mean of 53.66.

Table 3. Relative dinucleotide abundance frequencies of gene across the regions.

Dinucleotides	China	India	Iran	Oman	South Korea
AA	1.07	1.10	1.13	1.12	1.09
AC	0.71	0.80	0.72	0.65	0.70
AG	1.17	1.11	1.13	1.16	1.16
AT	1.02	0.96	0.97	1.05	1.02
CA	1.19	1.18	1.08	1.17	1.15
CC	1.20	1.10	1.20	1.12	1.24
CG	0.62	0.62	0.69	0.61	0.63
CT	0.87	0.98	0.94	0.96	0.87
GA	1.00	0.95	1.00	1.01	1.02
GC	0.96	0.97	0.98	1.02	0.91
GG	1.39	1.32	1.36	1.37	1.35
GT	0.74	0.82	0.76	0.69	0.79
TA	0.73	0.74	0.75	0.65	0.72
TC	1.25	1.20	1.24	1.32	1.26
TG	0.82	0.96	0.81	0.88	0.85
TT	1.29	1.20	1.28	1.24	1.26

3.2. Analysis of natural selection and mutation bias

To determine the influencing factors for mutational pressure or selection pressure among five different countries for TYLCV (C1/AC1) gene, bellow listed traits were estimated.

3.2.1. Relative abundance of dinucleotides frequencies of TYLCV-C1/AC1

The relative abundance of 16 dinucleotides frequencies of gene C1/AC1 was estimated for all five countries are represented in Table 2 and Fig. 3. All the five regions showed the significant variations in the 16 dinucleotide frequencies. Generally, the dinucleotides frequencies ranges between 0.78 (underrepresented) and 1.23 (overrepresented).

1. China: Out of 16 dinucleotides bases, three dinucleotides GG (1.39), TC (1.25) and TT (1.29) were overrepresented i.e > 1.23. And, four dinucleotides, AC (0.71), CG (0.62), GT (0.74) and TA (0.73) were underrepresented i.e < 0.78.
2. India: The dinucleotide, GG (1.32) was overrepresented and two dinucleotides CG (0.62) and TA (0.74) were underrepresented.
3. Iran: Three dinucleotides, GG (1.36), TC (1.24) and TT (1.28) were overrepresented, and four dinucleotides, AC (0.72), CG (0.69), GT (0.76) and TA (0.75) were underrepresented.
4. Oman: Three dinucleotides, GG (1.37), TC (1.32) and TT (1.24) were overrepresented, and four dinucleotides, and four dinucleotides, AC (0.65), CG (0.61), GT (0.69) and TA (0.65) were underrepresented.
5. South Korea: Four dinucleotides, CC (1.24), GG (1.35), TC (1.26) and TT (1.26) were overrepresented, and four dinucleotides, and four dinucleotides, AC (0.70), CG (0.63), GT (0.79) and TA (0.72) were underrepresented.

3.2.2. Effective Number of Codons (eNC) plot

Effective Number of Codons (eNC) values of all the five selected regions were plotted in a single frame to understand the synonymous codon usage pattern, which majorly affected by the selection or mutational pressure. Figure 5, wherein the eNC values at ordinate and GC3 values at abscissa was plotted along with the standard curve to obtain the role of mutational pressure in framing the codon usage bias. Each different colour specifies the different regions/countries selected.

3.2.3. Parity rule 2 bias plot

The relationship of pyrimidines (C & T) and purine (A & G) bias at third codon position was indicated in PR2 bias plot. There is no bias in selection and mutation pressure if all the plot values lie on the center, where both ordinate and abscissa meet are 0.5. PR2 signifies the direction and degree of codon bias. The PR2 plot constructed by plotting the G3/(G3+C3) on X-axis and A3/(A3+T3) on Y-axis for TYLCV (C1/AC1) gene of each selected five countries. The average value of GC and AT for China (0.34 and 0.54), India (0.38 and 0.51), Iran

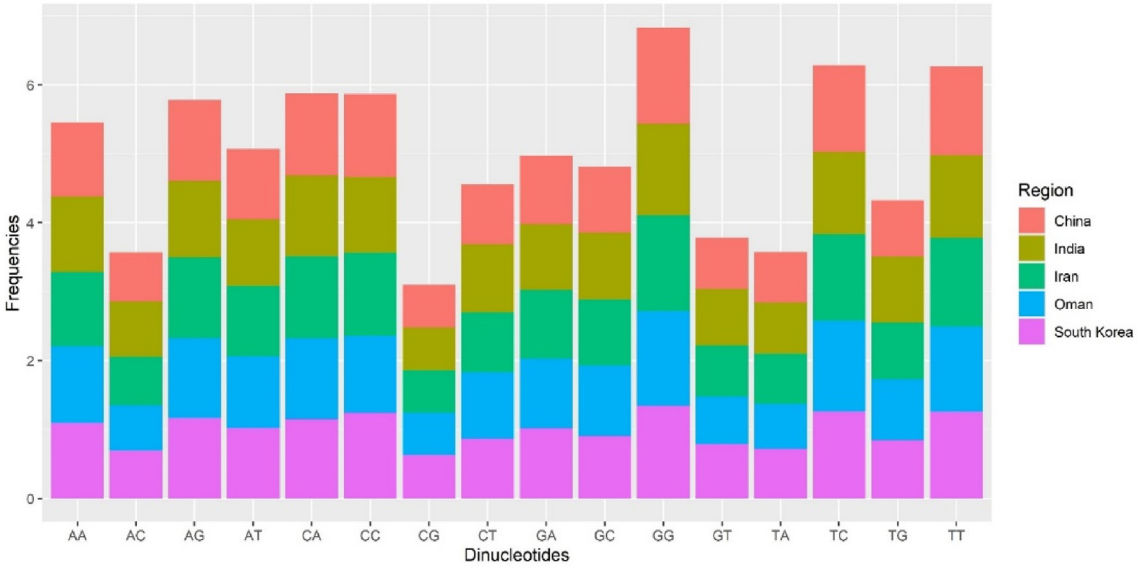

Fig. 3. Dinucleotide abundance frequencies of TYLCV (C1/AC1) gene for five major Asian countries.

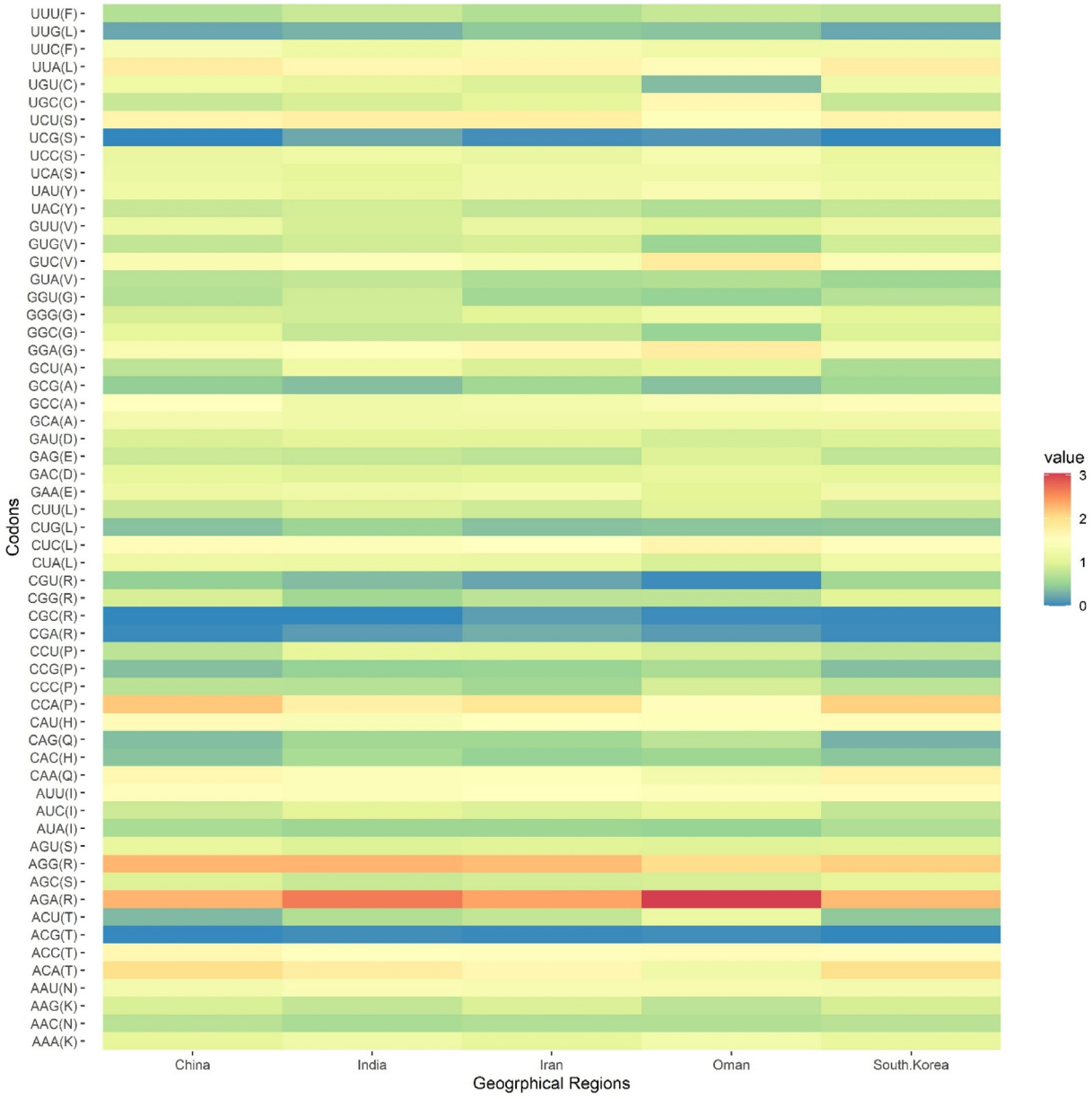

Fig. 4. The overall RSCU frequencies of TYLCV (C1/AC1) gene presented in the Heat map. The red and blue indicates the over and under representations respectively.

(0.39 and 0.53), Oman (0.38 and 0.52), South Korea (0.37 and 0.54) was observed. Here, in all the five regions, domination of AT over the GC was observed and also, the values indicate that the purines and pyrimidines are equally preferred in codon usage bias. The PR2 bias plot specifies that all the five regions were situated away from the origin, suggest that the occurrence of bias at the third position of AT and GC for framing the TYLCV (C1/AC1) codons. Consequently, the plot indicating the natural selection over the mutational pressure (Fig. 6).

3.2.4. Neutrality plot

The neutrality plot determines the relationship between GC12 (mean of GC1 and GC2) and GC3 of TYLCV (C1/AC1) gene, clarifies the dimensions of natural selection and mutational pressure impact over framing the codon usage pattern. The neutrality plot was obtained with plotting GC3 against GC12, the slope of the regression line indicates the evolutionary rate of natural selection and mutational pressure, which considered as selection–mutational equilibrium coefficient. The neutrality plot analysis in the study resulted as.

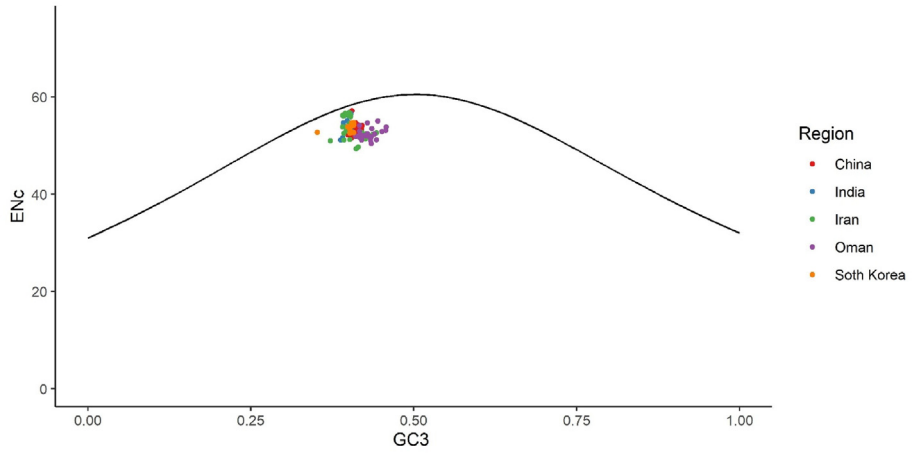

Fig. 5. The ENC-GC3 plot analysis of all the five major countries for TYLCV (C1/AC1) gene. Indicating the bias influenced by the GC3, that affecting the codon usage pattern.

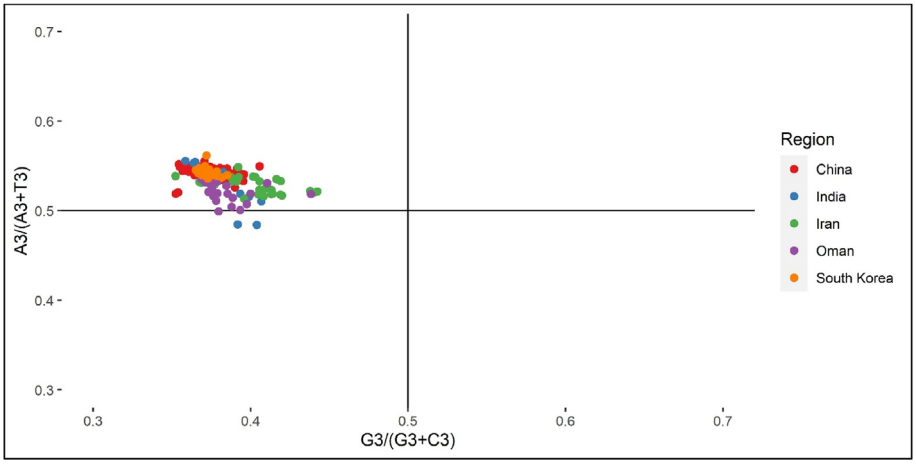

Fig. 6. The Parity rule (PR2) plot of all the five major countries. Indicating the AT and GC bias at the third codon position that affects the codon usage bias in TYLCV (C1/AC1) gene.

1. China: Negative regression line and negative significant R-value with y = 0.465−0.106, R^2 = 0.02, with comparative neutrality was indicating 2.0% mutation pressure (minor factor) and 98% comparative constraint indicating natural selection (dominant factor).
2. India: Negative regression line and negative significant R-value with y = 0.429−0.0061, R^2 < 0.01, with comparative neutrality was indicating <1.0% mutation pressure (minor factor) and 99% comparative constraint indicating natural selection (dominant factor).
3. Iran: Positive regression line and positive significant R-value with y = 0.305 + 0.297, R^2 = 0.54, with comparative neutrality was indicating 54% mutation pressure (minor factor) and 46% comparative constraint indicating natural selection (dominant factor).
4. Oman: Positive regression line and positive significant R-value with y = 0.397 + 0.082, R^2 = 0.19, with comparative neutrality was indicating 19% mutation pressure (minor factor) and 81% comparative constraint indicating natural selection (dominant factor).
5. South Korea: Negative regression line and negative significant R-value with y = 0.608−0.466, R^2 = 0.69, with comparative neutrality was indicating 69% mutation pressure (minor factor) and 31% comparative constraint indicating natural selection (dominant factor).

In all the regions obtained results indicated that natural selection is assertive over mutation pressure and has greater effect on the codon usage of TYLCV (C1/AC1) (Fig. 7).

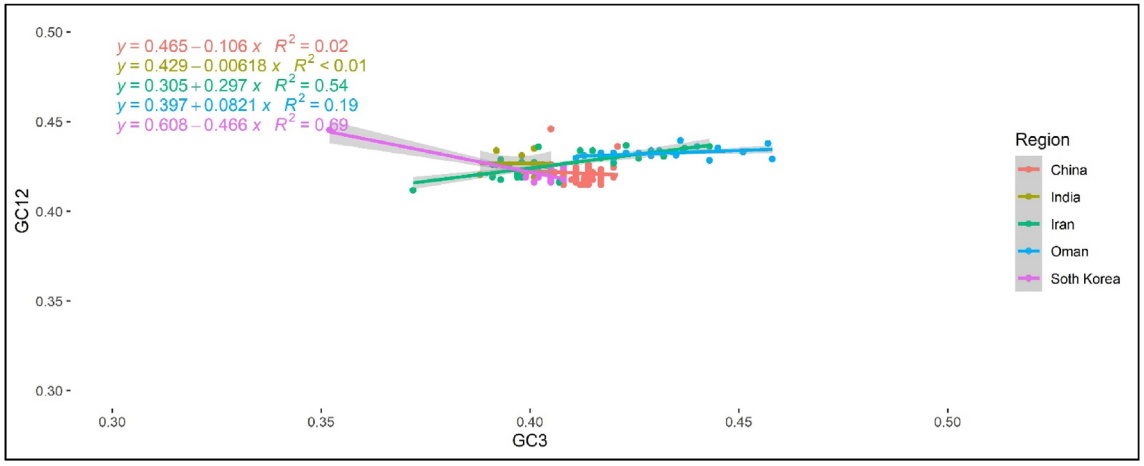

Fig. 7. Neutrality regression plot of TYLCV (C1/AC1) gene indicating the correlation between GC12 and GC3 for codon usage bias.

3.2.5. General average hydropathicity (GRAVY) and aromacity (Aromo)

The general average hydropathicity (GRAVY) and aromacity (Aromo) is determined for TYLCV (C1/AC1) gene of all the five selected Asian regions.

1. China: The negative significant value of GRAVY ranged between −0.63 and −0.49 with mean of −0.60 indicating hydrophilic nature of TYLCV (C1/AC1) gene and aromacity ranged between 0.09 and 0.117 with mean of 0.114 indicating significant positive AROMA of TYLCV (C1/AC1) gene.
2. India: The negative significant value of GRAVY ranged between −0.64 and −0.59 with mean of −0.623 indicating hydrophilic nature of TYLCV (C1/AC1) gene and aromacity ranged between 0.10 and 0.11 with mean of 0.110 indicating significant positive AROMA of TYLCV (C1/AC1) gene.
3. Iran: The negative significant value of GRAVY ranged between −0.711 and −0.45 with mean of −0.61 indicating hydrophilic nature of TYLCV (C1/AC1) gene and aromacity ranged between 0.104 and 0.121 with mean of 0.113 indicating significant positive AROMA of TYLCV (C1/AC1) gene.
4. Oman: The negative significant value of GRAVY ranged between −0.67 and −0.54 with mean of −0.60 indicating hydrophilic nature of TYLCV (C1/AC1) gene and aromacity ranged between 0.09 and 0.11 with mean of 0.107 indicating significant positive AROMA of TYLCV (C1/AC1) gene.
5. South Korea: The negative significant value of GRAVY ranged between −0.66 and −0.58 with mean of −0.611 indicating hydrophilic nature of TYLCV (C1/AC1) gene and aromacity ranged between 0.08 and 0.117 with mean of 0.114 indicating significant positive AROMA of TYLCV (C1/AC1) gene.

3.3. Correlation analysis

Correlation coefficients are indicators of the strength of the linear relationship between two different variables. The values ranges between −1 and +1 and quantifies the direction and strength of the linear association between the two variables. The correlation between two variables can be positive (greater than zero) or negative (less than zero). The spearman correlation analysis in five major Asian Countries are represented in Table 4.

1. China: The negative significant values were observed between eNC-CAI, RCDI-CAI, eNC-RCDI and GRAVY-RCDI. The negative non-significant were observed between eNC-Aromo and Aromo-RCDI. The positive significant observed between Aromo-CAI and eNC-GRAVY. The positive non-significant were observed between GRAVY-CAI and GRAVY-Aromo.
2. India: The negative significant values were observed between eNC-Aromo and eNC-RCDI. The negative non-significant were observed between GRAVY-CAI, Aromo-CAI, RCDI-CAI and eNC-GRAVY. The positive significant observed between RCDI-Aromo. The positive non-significant were observed between eNC-CAI, GRAVY-Aromo, GRAVY-RCDI.
3. Iran: The negative significant values were observed between RCDI-CAI, eNC-RACDI,

Table 4. The spearman correlation analysis in five regions.

China	CAI	ENC	Gravy	Aromo
CAI				
ENC	-0.36****			
Gravy	0.03	0.15*		
Aromo	0.20**	-0.04	0.05	
RCDI	-0.19**	-0.64****	-0.31****	-0.11
India	CAI	ENC	Gravy	Aromo
CAI				
ENC	0.04			
Gravy	-0.42	-0.33		
Aromo	-0.32	-0.80**	0.38	
RCDI	-0.34	-0.89****	0.38	0.81***
Iran	CAI	ENC	Gravy	Aromo
CAI				
ENC	0.23			
Gravy	0.1	0.56***		
Aromo	0.58****	0.42**	0.51***	
RCDI	-0.48**	-0.80****	-0.47**	-0.32*
Oman	CAI	ENC	Gravy	Aromo
CAI				
ENC	-0.23			
Gravy	-0.27	-0.43*		
Aromo	-0.33	0.42*	-0.42*	
RCDI	-0.64***	-0.43*	0.28	0.08
South Korea	CAI	Nc	Gravy	Aromo
CAI				
ENC	0.36			
Gravy	-0.49**	-0.43*		
Aromo	-0.13	0.3	0.37*	
RCDI	-0.72****	-0.79****	0.37*	-0.18

****p-value: <0.0001, ***p-value: <0.001 **p-value: < 0.01 *p-value: < 0.05.

GRAVY-RACDI and RCDI-Aromo. The positive significant observed between Aromo-CAI, eNC-GRAVY, eNC-Aromo and GRAVY-Aromo. The positive non-significant were observed between eNC-CAI and GRAVY-CAI.

4. Oman: The negative significant values were observed between RCDI-CAI, eNC-GRAVY, eNC-RCDI and GRAVY-Aromo. The negative non-significant were observed between eNC-CAI, GRAVY-CAI and Aromo-CAI. The positive significant observed between eNC-Aromo. The positive non-significant were observed between RCDI-GRAVY and RCDI-Aromo.
5. South Korea: The negative significant values were observed between GRAVY-CAI, RACDI-CAI, eNC-GRAVY and eNC-RCDI. The negative non-significant were observed between Aromo-CAI and RCDI-Aromo. The positive significant observed between GRAVY-Aromo and GRAVY-RCDI. The positive non-significant were observed between eNC-CAI and Aromo-eNC.

3.4. Evolutionary characteristic analysis

3.4.1. Data sequence and alignment

The nucleotide sequence data of TYLCV (C1/AC1) gene from five different Asian countries were downloaded from the NCBI database (https://www.ncbi.nlm.nih.gov) to analyse the evolutionary characteristics. The nucleotide sequence of China (245), India (13), Iran (40), Oman (31) and South Korea (30) were aligned and edited by MEGA-X software.

3.4.2. Evolutionary rate analysis

The evolutionary rate and time analysis specifies the significant changes in evolutionary rate over the period in samples. The complete coding sequence of TYLCV (C1/AC1) gene for five different Asian countries were used to evaluate the time of Most Recent Common Ancestor (tMRCA) and substitution rate (s/s/y) using the Bayesian-based coalescent method. The DNA substitution model selection was done using the jModelTest2 tool, as a preliminary criterion. The best fit substitution model is gleaned from Akaike Information Criteria (AIC) and Bayesian Information Criteria (BIC) score obtained from the jModelTest2 tool. The GRT + G + I was the observed best fit substitution model for both China and Oman, HKY + I was observed for both India and South Korea, and HKY + G for Iran. Using the BEAUti tool, the required parameters and priors, MCMC chain length, clock rate was specified and XML format files were generated. The BEAST were performed using the generated XML files and tree files (.trees) and logarithmic (.log) files were obtained. The evaluated evolutionary rate and tMRCA are presented in the Table 5.

To each dataset, the MCMC chain cycle 1−10 million generation were run to accomplish

Table 5. Estimation of evolutionary and tMRCA of TYLCV using Datamonkey server.

Region	Substitution rate mean	95% Higher posterior density (HPD)	
		Lower	Upper
China	4.19×10^{-3}	3.03×10^{-3}	5.075×10^{-3}
India	2.68×10^{-4}	3.07×10^{-10}	8.83×10^{-4}
Iran	8.04×10^{-4}	2.60×10^{-4}	1.32×10^{-3}
Oman	1.90×10^{-4}	1.48×10^{-7}	3.91×10^{-4}
South Korea	1.18×10^{-5}	1.67×10^{-14}	6.08×10^{-5}

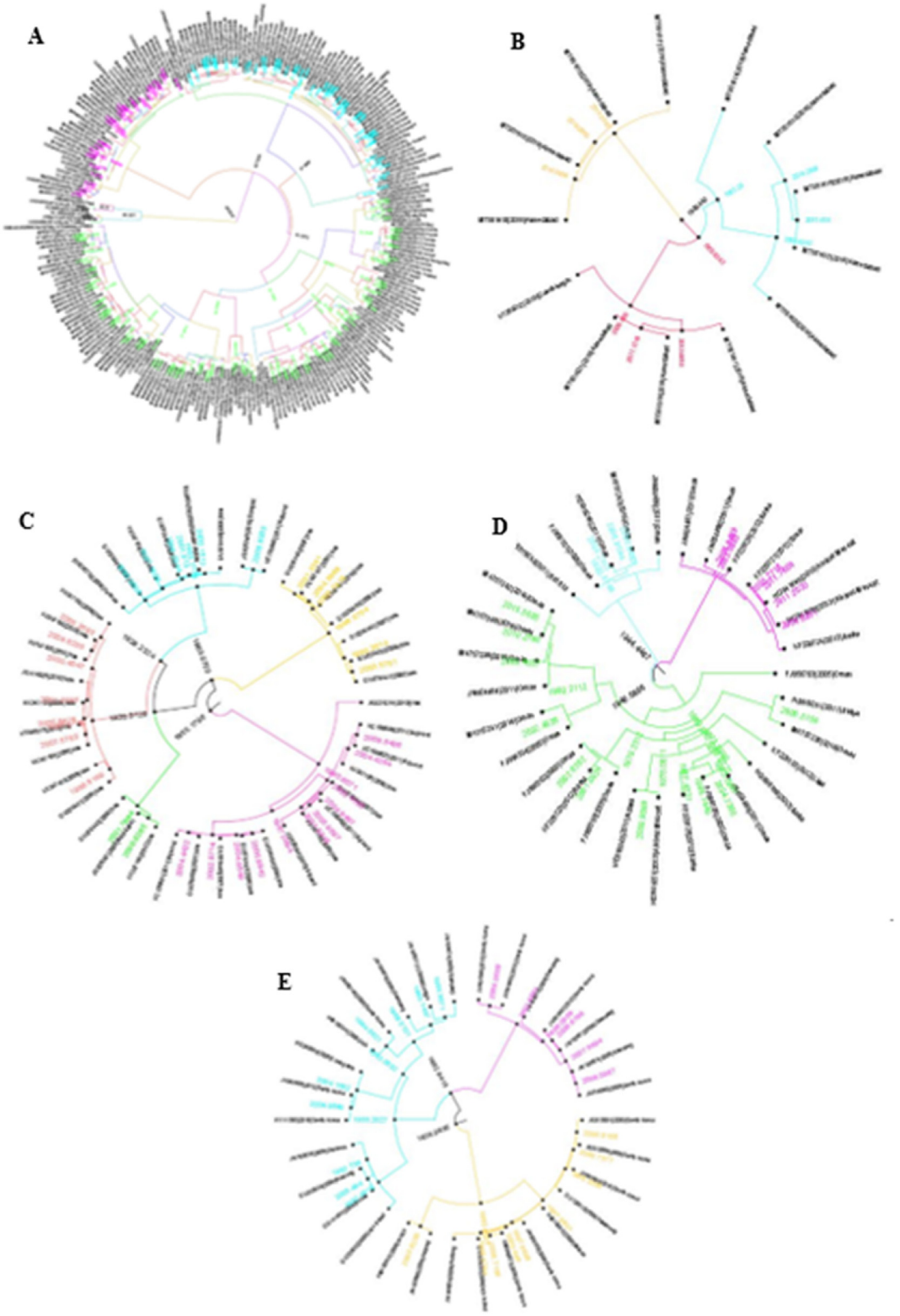

Fig. 8. The phylogenetic tree of selected countries for TYLCV (C1/AC1) gene. A- China, B- India, C- Iran, D- Oman and E− South Korea. Each tree nodes indicates the tMRCA in years.

prominent concurrence of statistical parameters. The 95% HPD (Highest Posterior Density) interlude of the divergence parameter for tMRCA was obtained from the tMRCA/tree height, and the substitution rate was obtained from the mean rate/clock rate. The log files were generated from BEAST and the Tracer tool was used to visualize the achieved statistical scores.

The evolutionary rate for the different Asian countries- China, India, Iran, Oman and South Korea for the TYLCV (C1/AC1) gene are 4.19×10^{-3}, 2.68×10^{-4}, 8.04×10^{-4}, 1.90×10^{-4} and 1.18×10^{-5} respectively. The recorded tMRCA ages for China- 1853 year with 95% HPD (lowest 3.03×10^{-3}, Highest 5.075×10^{-3}), India- 1939 year with 95% HPD (lowest 3.07×10^{-10}, Highest 8.83×10^{-4}), Iran- 1855 years with 95% HPD (lowest 2.60×10^{-4}, Highest 1.32×10^{-3}), Oman- 1944 year with 95% HPD (lowest 1.48×10^{-7}, Highest 3.91×10^{-4}), and South Korea- 1828 year with 95% HPD (lowest 1.67×10^{-14}, Highest 6.08×10^{-5}) Table 5. The high evolutionary rate was observed in South Korea (188 years with higher tMRAC rate of 1.18×10^{-5}) for TYLCV (C1/AC1) gene, indicating that the TYLCV (C1/AC1) gene evolving faster rate in South Korea compare to other regions.

For spatial diffusion visualization, map had variation in the colours to metamorphose between migration path (High and Low) of virus that had spread across Asian countries. The concentric circles in the map specifies the proportional to the number of isolates spread from the specific region and conversely for the recently transmitted isolates (Figs. 8 and 10).

3.4.3. Selection pressure analysis

Selection pressure was determined from the Datamonkey server using the FEL algorithm. The selection pressure results indicate that the China having three positive sites (6, 334, 348) and ninety negative sites were observed. In India one positive (110) and ninety-three negative sites, in Iran twenty positive (54, 77, 79, 82, 122, 154, 171, 174, 175, 181, 197, 223, 250, 256, 257, 288, 289, 301, 319, 320) and nineteen negative sites, in Oman three positive (206, 212, 285) and twenty-two negative sites, and in South Korea zero positive sites and thirteen negative sites were observed. The overall dN/dS (ω) rate ratio varied as 0.323, 0.216, 0.923, 0.776, and 0.289 for China, India, Iran, Oman, and South Korea respectively (Table 6 and Fig. 9).

4. Discussion

The codon usage bias analysis provides the adequate knowledge on genetic features and molecular evolution of the organisms. The major shaping factors for codon usage bias are gene length, nucleotide composition, selection and mutational pressure [30–33]. The TYLCV is an emerging complex virus, causing a prominent

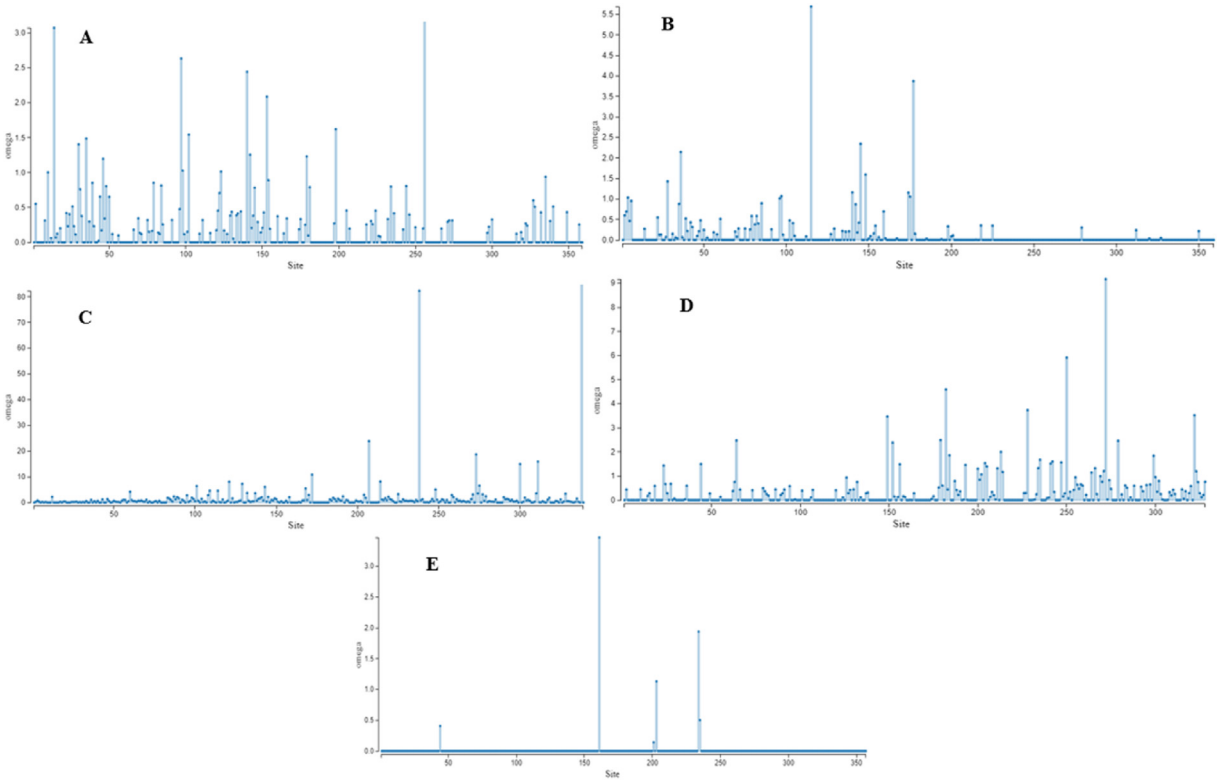

Fig. 9. The dN/dS (ω) rate ratio of TYLCV (C1/AC1) gene in all the selected major Asian countries. A- China, B- India, C- Iran, D- Oman and E– South Korea.

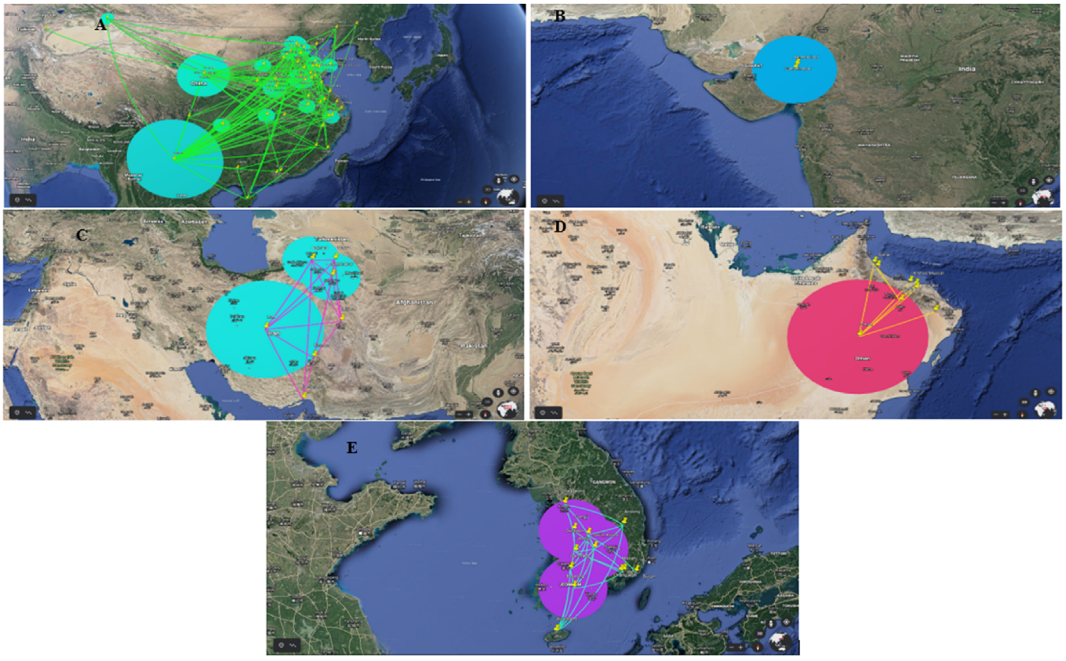

Fig. 10. The phylogeography of the TYLCV (C1/AC1) genes obtained from the tree. The plot indicates the estimate of special demography with a maximum radius of the circle has maximum isolates and the light coloured lines showing recent transitions of the disease. A- China, B- India, C- Iran, D- Oman and E– South Korea.

Table 6. The selection pressure analysis using the FEL method.

Regions	Positively diversifying site	Negatively diversifying site	dN/Ds (ω) substitution ratio
China	3	90	0.323
India	1	93	0.216
Iran	20	19	0.923
Oman	3	22	0.776
South Korea	No evidence	13	0.323

economic loss in tomato production [34]. The present study we focused on TYLCV (C1/AC1) gene of five Asian countries to observe the variation of pattern and factors influencing the codon usage bias. Chen et al. [35], observed that the codon usage evolution of virus evolves at different periods. Hence, recognizing this codon usage pattern would be a way to elucidate the phylogenetic relationship and by enhancing the codons that helps in target gene expression. The codon usage bias analysis mainly carried out by the RSCU, eNC, neutrality plot, parity rule plot analysis etc.

The present study, observed that the tomato TYLCV (C1/AC1) gene favour using codon ending with A or T more frequently over G or C because, the dicot genomes have low GC-content and prefers the A/T ending codons usually. The results concur with the studies on dicot genome such as, Anwar et al. [36]; Wang et al. [37]; Sablok et al. [38]; Kawabe and Miyashita [39]; Murray et al. [40] and Camiolo et al. [41] observed that the significant over-representation of codons ending with A/T in dicot than in monocot species. The results indicated the abundance of A nucleotide frequency and third position A frequency were high in all the regions. The GC analysis showed that the higher frequency of GC content and South Korea had high GC1 frequency (47.61%), Oman had high GC2 (39.29%) and GC3 (42.72%) frequencies. Whereas, the dinucleotides abundancy analysis revealed that the variation among the five regions and relatively large abundance in South Korea region for TYLCV (C1/AC1) gene. To determine the optimal codons for TYLCV (C1/AC1) gene, RSCU analysis was performed. The study revealed, A/T codon endings were favourably preferred than expected (over represented) due to high degree of uniformity among the selected regions having high codon usage bias and G/C codon endings were under-represented with lower codon usage [25]. The observed nucleotide compositional frequencies varying among the regions indicating they are biased and contributing to shaping the codon usage pattern of TYLCV (C1/AC1) gene [42].

The eNC indicates the overall codon usage bias, the Indian region showing the higher eNC value of each five region varies as China (53.21 ± 0.56) India (53.79 ± 1.6), Iran (53.60 ± 2.08), Oman (52.51 ± 1.15) and South Korea (53.66 ± 0.77) for TYLCV (C1/AC1) gene indicates lower the eNC value higher the gene

expression. the codon usage bias is due to the random codon preferences mainly affected by the mutational pressure. The low codon usage bias could be beneficial for effective replication with more selection possibilities [43]. Since, eNC value ranged as 49.35–57.13, indicates moderate level of codon bias. The study suggests that the codon usage bias facilitates the TYLCV (C1/AC1) gene replication and transcription. In addition, the eNC plotting was carried against GC3 content to know the extent motional pressure (GC3) influence for codon usage bias. The eNC plot observed that the eNC values were below the expected eNC curve in all the regions, indicating the selection pressure is the driving force for codon usage bias over mutational pressure for TYLCV (C1/AC1) gene [44].

To determine the dimensions of mutational and sectional pressure in relation to GC3 and GC12 on codon usage bias neutrality plot was analysed. The regression coefficient value < 0.5 indicates natural selection while, >0.5 is indicated as mutational pressure known as mutational–selection equilibrium [45]. The results revealed that selection pressure is a driving force in China (98%), India (99%) and Oman (81%) and Iran (54%) and South Korea (69%) experiencing the mutational pressure. The mutation in viral genome is due to errors in the replication process and high mutational rate is proportional to the population size, thus, this mutational variation can lead to the wide genetic diversities [46].

To quantify the selection and mutational pressure over purines and pyrimidines choices in codon usage bias, the parity rule-2 analysis was carried out. The PR2 value of all the five Asian countries revealed, that the domination of AT over the GC, indicated that the purines and pyrimidines are equally preferred in codon usage bias. The PR2 bias plot indicates that all the regions were pointed away from the origin, suggesting that the occurrence of bias at the third position of AT and GC for framing the TYLCV (C1/AC1) genes. The asymmetry between purines and pyrimidines indicates selection, mutation and other factors such as gene expression, gene length etc. were the shaping factors of codon usage bias [47].

To elucidate the factors influencing physical properties of amino acids such as hydropathicity (GRAVY) and aromacity (Aromo) were determined in all the five countries. The GRAVY and Aromo are indicative of consequences of section and translational pressures [34]. The study resulted that the hydropathicity is negatively and aromacity is positively significant among the five regions. The study indicates that GRAVY and Aromo are contributing to the physical properties of amino acids, which lead the way to shaping the codon usage bias.

To know the relationship between codon frequencies of a TYLCV (C1/AC1) gene with respect to codon usage bias indices, the correlation was analysed between eNC, CAI, RCDI, GRAVY, Aromo of each five countries. The study resulted that there was significant variation between the codon bias indices in all the selected regions. Which, imply that the codon usage pattern of TYLCV (C1/AC1) gene is seriously affected by eNC, CAI, RCDI, GRAVY, Aromo parameters. Hence, the selection and mutational pressure, other factors including geographic origin and translational selection taking a positive role in shaping the codon usage bias of viral gene [21–23,35,36].

After the codon usage pattern is observed, the evolutionary analysis (tMRCA) of TYLCV (C1/AC1) gene was calculated in selected five major Asian countries using the BEAST software. Based on the AIC/BIC scores of TYLCV (C1/AC1) gene, the best fit model for each region were determined using the jModelTest tool [23]. The results specified the substitutional model GRT + G + I has the lowest AIC/BIC value for China and Oman, HKY + I for India and South Korea, and HKY + G for Iran. The best fit models were selected with clock type and prior tree. The strict clock and random local clock were used with coalescent: Bayesian sky grid and coalescent: Extended Bayesian sky line plot prior trees. Sky grid method is well suited starting point among all the nonparametric coalescent-based models. Which is flexible that allows multiple loci and the changes occurs at the specified points in real-time [37]. However, the skyline method extracts information about past population genetics in the nonparametric method. Which splits the time between the root of the tree (tMRCA) and present into section, and estimates the diverse, effective population for each section [38].

The quantification of selection pressure reveals the evolutionary pattern of genetic loci that undergone specific adaptation. Which also a major evolutionary biological concept. Several statistical algorithms have been developed to estimate the selection pressure that measure the evolutionary pressure on protein coding sites. dN/dS is the most commonly used statistical approach for quantifying the selection pressure (Positive selection pressure) [39]. This method identifies the selection pressure by comparing the rate of the substitution at a silent site (dS) against non-silent sites (dN). Which is interpreted as negative selection (dN/dS or $\omega < 1$), neutral (dN/dS or $\omega = 1$), and positive selection (dN/dS or $\omega > 1$) [39]. Based on this we identified the

single positive selection site in India, three positive selection site in China and Oman, twenty sites in Iran and zero positive sites in South Korea for TYLCV (C1/AC1) gene. These suggest the influence of codon sites in the evolutionary process. The high negative selection indicates the elimination of deleterious variations from the populations while maintaining the function of transcription and replication of the TYLCV (C1/AC1) genome.

5. Conclusion

TYLCV is a major devastating viral threat to tomato production worldwide, causing a major yield loss. The TYLCV (C1/AC1) protein gene plays a vital role in viral replication. The present study is an initial effort to present a vision on codon usage pattern of TYLCV (C1/AC1) protein gene representing five major Asian Countries (China, India, Iran, Oman and South Korea) and significant difference was found amongst the gene sequences of different geographical regions. The selection and mutational pressure including geographic origin and translational selection taking a beneficial impact in shaping the codon usage bias of viral gene. The results from this study will assist to analyse the phylogenetic and evolutionary pattern of TYLCV gene. Having a comprehensive understanding of TYLCV codon usage and molecular evolutionary pattern is necessary for developing the strategies to understand the epidemiology and in turn help in prevention and control of the disease in tomato production.

Authors contributions

Mamathashree MN: Conceptualization, Data curation, Data editing and analysis, Drafting, editing and finalising the manuscript. Suresh KP and Krishnamoorthy Para manandham: Methodology and formal analysis Supervision, Sharanagouda S.Patil, Shiva Prasad Kollur and Chandan Shivamallu helped in Intrepretation and discussion. Uma Bharathi Indrabalan and Siju Susan Jacob: Conceptualization, Methodology, Data analysis. Mallikarjun S Beelagi and Sushma Pradeep and Chandrashekar Srinivasa: Data analysis.

All the authors reviewed, proof read and approved the manuscript.

Funding

Not applicable.

Ethical statements

Not applicable.

Conflict of interest

The authors declare that they have no conflict of interest for this study.

Supplementary materials

Supplementary file (S1): Contains overall compositions frequencies- A, T, G, C, A_3, T_3, G_3, C_3, GC, GC_1, GC_2, GC_3, eNC, Gravy, Aromo, CAI and RCDI of 359 TYLCV (C1/AC1) isolates with accession number and regions from which they have isolated.

Acknowledgement

MM, KPHS, SSP, UMI, MSB, KP and SSJ thank ICAR National Institute of Veterinary Epidemiology and Disease Informatics for the research support provided. SP, RRA, AP, SKP and CS acknowledge JSS AHER for the support and infrastructure provided. SPK acknowledges the Director Amrita Vishwa Vidyapeetham, Mysuru campus for laboratory infrastructural facilities.

References

[1] Moriones E, Navas-Castillo J. Tomato yellow leaf curl virus, an emerging virus complex causing epidemics worldwide. Virus Res 2000;71:123−34.
[2] Díaz-Pendón JA, Cañizares MC, Moriones E, Bejarano ER, Czosnek H, Navas-Castillo J. Tomato yellow leaf curl viruses: ménage à trois between the virus complex, the plant and the whitefly vector. Mol Plant Pathol 2010;11:441−50.
[3] Abhary M, Patil BL, Fauquet CM. In: Czosnek H, editor. Molecular biodiversity, taxonomy, and nomenclature of Tomato Yellow Leaf Curl-like Viruses in *Tomato Yellow Leaf Curl Virus Disease: Management, Molecular Biology, Breeding for Resistance*. Dordrecht: Springer; 2007. p. 85−118.
[4] Yan Z, Pérez-de-Castro A, Diez MJ, Hutton SF, Visser RGF, Wolters AMA, Bai y, et al. Resistance to tomato yellow leaf curl virus in tomato germplasm. Front Plant Sci 2018;9:1198.
[5] Cohen S, Harpaz I. Periodic, rather than continual acquisition of a new tomato virus by its vector, the tobacco whitefly (*Bemisia tabaci Gennadius*). Entomol Exp Appl 1964;7:155−60.
[6] Cohen S, Nitzany FE. Transmission and host range of the tomato yellow leaf curl virus. Phtyopathology 1966;56: 1127−31.
[7] Ghanim M, Medina V. Localization of tomato yellow leaf curl virus in its whitefly vector Bemisia tabaci. In: Czosnek H, editor. Tomato Yellow Leaf Curl Virus Disease. Dordrecht: Springer; 2007. p. 171−83.
[8] Czosnek H. In: Mahy BWJ, van Regenmortel MHV, editors. Tomato yellow leaf curl virus" in *Encylopedia of Virology*. Oxford: Academic Press; Elsevier Ltd); 2008. p. 138−45.
[9] Marchant WG, Gautam S, Hutton SF, Srinivasan R. Tomato yellow leaf curl virus-resistant and -susceptible tomato genotypes similarly impact the virus population genetics. Front Plant Sci 2020;13(6).
[10] Navot N, Pichersky E, Zeiden Zamir D, Czosnek H. Tomato yellow leaf curl virus: a whitefly-transmitted geminivirus with a single genomic component. Virology 1991;185:151−61.
[11] Czosnek H, Ber R, Antignus Y, Cohen S, Navot N, Zamir D. Isolation of tomato yellow leaf curl virus, a geminivirus. Phytopathology 1988;78(5):508−12.

[12] Zhang W, Olson NH, Baker TS, Faulkner L, Agbandje-McKenna M, Boulton MI, Davies JW, McKenna R. Structure of the maize streak virus geminate particle. Virology 2001; 279(2):471–7.
[13] Gronenborn B. The tomato yellow leaf curl virus genome and function of its proteins. In: Tomato Yellow Leaf Curl Virus Disease; 2007. p. 67–84.
[14] Scholthof KBG, Adkins S, Czosnek H, Palukaitis P, Jacquot E, Hohn T, et al. Top 10 plant viruses in molecular plant pathology. Mol Plant Pathol 2011;12(9):938–54.
[15] Laufs J, Traut W, Heyraud F, Matzeit V, Rogers SG, Schell J, et al. In vitro cleavage and joining at the viral origin of replication by the replication initiator protein of tomato yellow leaf curl virus. Proc Natl Acad Sci U S A 1995;92(9): 3879–83.
[16] Behura SK, Severson DW. Codon usage bias: causative factors, quantification methods and genome-wide patterns: with emphasis on insect genomes. BIO Rev 2013;88(1):49–61.
[17] Qin H, Wu WB, Comeron JM, Kreitman M, Li WH. Intragenic spatial patterns of codon usage bias in prokaryotic and eukaryotic genomes. Genetics 2004;168(4):2245–60.
[18] Indrabalan UB, Suresh KP, Shivamallu C, Patil SS. An extensive evaluation of codon usage pattern and bias of structural proteins p30, p54 and, p72 of the African swine fever virus (ASFV). Virus Dis 2021. https://doi.org/10.1007/s13337-021-00719-x.
[19] Butt AM, Nasrullah I, Tong Y. Genome-wide analysis of codon usage and influencing factors in chikungunya viruses. PLoS One 2014;9(3):e90905.
[20] Wei Y, Wang J, Xia X. Coevolution between stop codon usage and release factors in bacterial species. Mol Biol Evol 2016; 33(9):2357–67.
[21] Yao X, Fan Q, Yao B, Lu P, Rahman SU, Chen D, et al. Codon usage bias analysis of bluetongue virus causing livestock infection. Front Microbiol 2020;11:655.
[22] Gomez MM, Volotao EDM, Assandri IR, Peyrou M, Cristina J. Analysis of codon usage bias in potato virus Y non-recombinant strains. Virus Res 2020;286:1–7.
[23] Beelagi MS, Indrabalan UB, Patil SS, Suresh KP, Kollur SP, Prasad A, et al. Insight of codon usage bias and evolutionary rate among the genes C, E, prM and NS5 of the kyasanur forest disease virus. Int J Res Pharm Sci 2021;12(3):2028–46.
[24] Chen Y, Li X, Chi X, Wang S, Ma Y, Chen J. Comprehensive analysis of the codon usage patterns in the envelope glycoprotein E2 gene of the classical swine fever virus. PLoS One 2017;12(9).
[25] Anwar AM, Aljabri M, El-Soda M. Patterns of genome-wide codon usage bias in tobacco, tomato and potato. Biotechnol Biotechnol Equip 2021;35(1):657–64.
[26] Wang L, Xing H, Yuan Y, Wang X, Saeed M, Tao J, et al. Genome-wide analysis of codon usage bias in four sequenced cotton species. PLoS One 2018;13(3):e0194372.
[27] Sablok G, Wu X, Kuo J, Nayak KC, Baev V, Varotto C, et al. Genomics combinational effect of mutational bias and translational selection for translation efficiency in tomato (*Solanum lycopersicum*) Cv. Micro-Tom. Genomics 2013; 101(5):290–5.
[28] Kawabe A, Miyashita NT. Patteren of codon usage bias in three dicot and four monocot plant species. Genes Genet Syst 2003;78(5):343–52.
[29] Murray EE, Lotzer J, Eberle M. Codon usage in plant genes. Nucleic Acids Res 1989;17(2):477–98.
[30] Camiolo S, Melito S, Porceddu A. New insights into the interplay between codon bias determinants in plants. DNA Res 2015;22(6):461–70.
[31] Jenkins GM, Holmes EC. The extent of codon usage bias in human RNA viruses and its evolutionary origin. Virus Res 2003;92(1):1–7.
[32] Deb B, Uddin A, Chakrabortry S. Codon usage pattern and its influencing factors in deferentgenomes of hepadnaviruses. Arch Virol 2020;165(3):557–70.
[33] Zhang R, Zhang L, Wang W, Zhang Z, Du H, Qu Z, et al. Differences in codon usage bias between photosynthesis-related genes and genetic system-related genes of chloroplast genomes in cultivated and wild solanum species. Int J Mol Sci 2018;19(10):3142.
[34] Singh KN. Tyagi AA detailed analysis of codon usage patterns and influencing factors in Zika virus. Arch Virol 2017; 162(7):1963–73. https://doi.org/10.1007/s00705-017-3324-2.
[35] Schubert AM, Catherine P. Evolution of the sequence composition of Flaviviruses. Infect Genet Evol 2010;10:129–36.
[36] Cristina J, Moreno P, Moratorio G, Musto H. Genome-wide analysis of codon usage bias in Ebolavirus. Virus Res 2015; 196:87–93.
[37] Gill MS, Lemey P, Bennett SN, Biek R, Suchard MA. Understanding past population dynamics: bayesian coalescent-based modelling with covariates. Syst Biol 2016;65(6): 1041–56.
[38] Heller R, Chikhi L, Siegismund HR. The confounding effects of population structure on bayesian skyline plot inferences of demographic history. PLoS One 2013;8(5):e62992.
[39] Kryazhimskiy S, Plotkin JB. The population Genetics of dN/dS. PLoS Genet 2008;4(12):e1000304.
[40] Mueller S, Papamichail D, Coleman JR, Skiena S, Wimmer E. Reduction of the rate of poliovirus protein synthesis through large-scale codon deoptimization causes attenuation of viral virulence by lowering specific infectivity. J Virol 2006;80(19): 9687–96.
[41] Butt AM, Nasrullah I, Qamar R. Evolution of codon usage in Zika virus genomes is host and vector specific. Emerg Microb Infect 2016;5:1–14.
[42] Khandia R, Singhal S, Kumar U, Ansari A, Tiwari R, Dhama K, et al. Analysis of nipah virus codon usage and adaptation to hosts. Front Microbiol 2019;10:1–18.
[43] Karlin S, Burge C. Dinucleotide relative abundance extremes: a genomic signature. Trends Genet 1995;11:283–90.
[44] Suresh KP. Evolutionary analysis and detection of positive selection of hemagglutinin and neuraminidase genes of H5N1 avaian influenza from chicken, duck and goose across Asia. Explor Anim Med Res 2020;10(2):169–78.
[45] Das B, Mohapatra JK, Pande V, Subramaniam S, Sanyal A. Evolution of foot-and-mouth disease virus serotype A capsid coding (P1) region on a timescale of three decades in an endemic context. Infect Genet Evol 2016;41:36–46.
[46] Lemey P, Rambaut A, Drummond AJ, Suchard MA. Bayesian phylogeography finds its roots. PLoS Comput Biol 2009;5: e1000520.
[47] Bielejec F, Rambaut A, Suchard MA, Lemey P. SPREAD: spatial phylogenetic re-construction of evolutionary dynamics. Bioinformatics 2011;27:2910–2.