ORIGINAL ARTICLE

INTERNATIONAL JOURNAL OF RESEARCH IN PHARMACEUTICAL SCIENCES

Published by JK Welfare & Pharmascope Foundation Journal Home Page: www.pharmascope.org/ijrps

Insight of Codon usage bias and Evolutionary rate among the genes C, E, prM and NS5 of the Kyasanur Forest Disease virus

Mallikarjun S Beelagi[1], Uma Bharathi Indrabalan[2], Sharanagouda S Patil[2], Suresh K P[2], Shiva Prasad Kollur[3], Ashwini Prasad[4], Chandrashekar Srinivasa[5], Chandan Shivamallu[*1]

[1]Department of Biotechnology and Bioinformatics, Faculty of Life Sciences, JSS Academy of Higher Education & Research, Mysuru-570015, India
[2]ICAR-National Institute of Veterinary Epidemiology and Disease Informatics (NIVEDI), Yelahanka, Bengaluru-560064, India
[3]Department of Sciences, Amrita School of Arts and Sciences, Mysuru, Amrita Vishwa Vidyapeetham, Karnataka – 570 026, India
[4]Department of Microbiology, Faculty of Life Sciences, JSS Academy of Higher Education & Research, Mysuru-570015, India
[5]Department of Studies in Biotechnology, Davangere University, Shivagangotri, Davangere Karnataka-577 007, India

Article History:

Received on: 15 Apr 2021
Revised on: 19 May 2021
Accepted on: 31 May 2021

Keywords:

KFD virus,
Codon usage bias,
Evolutionary characteristics Analysis,
Positive Selection,
tMRCA

ABSTRACT

Kyasanur Forest Disease was first evolved in the Kyasanur forest, Karnataka. The transmission of the virus has occurred from the monkey to the human by the tick vector. On the early day of viral spread, the disease was restricted to the surrounded region of Kyasanur forest, Shimoga district. But in the present days, the disease has been spreading to neighboring districts and states as well. So, this study involves estimation of codon bias among the gene C, gene E, gene prM, and gene NS5 of the KFD virus and rate of evolution with phylogenetic analysis. The codon usage analysis has revealed the moderate codon bias among all the selected genes and the role of mutation pressure in genes- C and E and natural selection in genes- prM and NS5. Also, the tMRCA age was 1942, 1982, 1975, and 1931 of genes- C, E, prM, and NS5, respectively, of the KFD virus. The integrated analysis of codon usage bias and evolutionary rate analysis signifies that both mutational pressure and natural selection among the selected genes of the KFD virus.

*Corresponding Author

Name: Chandan Shivamallu
Phone:
Email: chandans@jssuni.edu.in

ISSN: 0975-7538

DOI: https://doi.org/10.26452/ijrps.v12i3.4811

Production and Hosted by

Pharmascope.org

INTRODUCTION

Kyasanur forest disease (KFD) virus is a tick-borne viral disease that belongs to the family *Flaviviridae*, genus *flavivirus*. The first case of KFDV was reported in 1957 at Shimoga district, Karnataka, India (Kasabi, 2011). An outbreak of disease caused high mortality among monkeys in the Kyasanur forest, and viral transmission was followed to birds and humans through ticks, a disease found mainly with people who often travel to the forest and have direct contact with forest areas (*Haemphysalis spinigera*) (Author n.d.). A virus has found a high rate of susceptibility in both hosts, humans and monkeys. Since the exploration of the KFDV, the dis-

ease was reported in and around five districts (Shimoga, Chikkamagalore, Uttara-Kannada, Dakshina-Kannada, and Udupi) of Karnataka state, India. Also, the recent studies stated the presence of KFDV in neighbouring states of Karnataka, including Maharashtra, Kerala, Tamil Nadu, and Goa, which indicates the risk factor or prevalence of viral spread in the nearby geographical locality.

KFDV contains a positive sense of RNA genomes that are an approximate length of 11kb and encoded a single 3416 amino acid polyproteins. During the post-translational process, the polyproteins are cleaved into three structural (C, M/prM, and E) and seven non-structural (NS1, NS2a, NS2b, NS3, NS4a, NS4b, and NS5)proteins (Yadav, 2020; Dodd, 2011). Viral is well known as a biphasic syndrome with hemorrhagic fever, severe headache, arthralgia, myalgia. The first phase of the viral features is known to have thrombocytopenia, leukopenia, and elevated transaminases. Neurological symptoms are the second phase of the syndrome (Dodd, 2011; Taming the BEAST, 2018; Gupta *et al.*, 2020).

Capsid polyproteins are highly basic in nature, containing 25.5% of arginine and lysine residues which is similar to other flaviviruses. It takes a major role to build nucleocapsids by interacting with RNA. Also, the C terminus of the capsid gene contributes to transmit the signal sequence to the synthesis of the prM, which shows a significant difference among other tick-borne *flaviviruses*. Glycoprotein E (Envelope)possesses a major functional part as an inducing haemagglutination inhibition, counteracting, and as an antibody during the period of natural infection or immunization. Hence, envelope proteins have been a hot spot to drug target or vaccination (Venugopal *et al.*, 1994). Non-structural NS5 of the flaviviruses is the largest matured viral protein that is synthesized during the period of infection. Also, it secretes N-terminal methyltransferase (MTase) that produces cap1 structure at the 5' end of newly created viral genomes, promoting the translational process and protect viral RNA from immune detection (Fajardo, 2020). So the protein NS5 is a vital replication enzyme for both methyltransferase and RNA polymerase (Yadav, 2020).

The nonhomogenous use of synonymous codons has been broadly reported in several genes and genomes of different organisms or species. The rate of synonymous codon usage frequency in an individual sample species is known as the codon usage bias (CUB) (Wu *et al.*, 2020; Zhou, 2016). The amount of bias is extremely variable among each different sample species; also, codon usage bias explains the feature of molecular evolution. Natural selection and mutational pressure are two major factors of codon usage bias. Also, other factors such as dinucleotide composition, nucleotide composition, GC, GC1, GC2, GC3, aromaticity (AROMO), and hydropathy (GRAVY) are influencing elements of shaping the codon usage pattern. So the CUB can be a potential analytical approach for determining the major cause of the bias (Comeron and Aguadé, 1998; Rahman and Ur, 2017).

Exploring the distinct diversity of a subpopulation from its ancestral populations has remained an essential quest in population genetics. When the ancestry population acquires random genetic mutation over the period, it produces a distinct diversity in its progeny populations or subpopulations, and interpretation of the divergence period between populations has become a significant study of population evolution (Zhou and Teo, 2016). The coalescent theory has proven to be an effective approach for genetic problems in terms of both modelling biological aspects and generating rich statistical data. The theory has been extensively implemented in evolutionary studies. It is capable of analyzing the rigorous statistical population data and hypothesizing rational simulation datasets (Donnelly and Tavare, 1995). Several methodologies have been developed based on the coalescent theory, and these can be categorized corresponding to the type of input genetic data and hypothesis of the population demography. For instance, tMRCA (time of the Most Recent Common Ancestor) works on mainly three methods, one class of method recognizes multiple neutral loci of each $\sim$1000 in multiple populations, the second method interprets the tMRCA from full information of the chromosomes, the third one infers the tMRCA based on the extent linkage equilibrium (LD) (Zhou and Teo, 2016). BEAST: Bayesian Evolutionary Analysis By Sampling Tree is a fast, user-friendly software that has become a widespread platform for resolving the evolutionary analysis and phylogenetic time-tree. BEAST relatively allows the Bayesian Markov chain Monte Carlo (MCMC) algorithm or method for phylogenetic reconstruction, which is already the most embraced and core algorithm. Also, it creates a platform to analyze multiple data partitions at the same time, which is helpful to estimate the single multilocus coalescent analysis (Drummond and Rambaut, 2007). BEAUti is an analysis engine that is integrated into the BEAST software, which allows creating the modelling file without any GUI programming. BEAUti has the ability to checkpoint and restart analysis, template-based GUI enhancement, and extensible XML format, and the tool tracer, which is an in-built tool of BEAST software, helps

to visualize the log file in the graphical format generated after the execution of BEAST (Zhou and Teo, 2016). The tree annotator tool was used to burn in the tree file, and Figtree software used to further phylogenetic tree visualization and to illustrate the year of tMRCA. Whereas the system requires a Java platform to execute the BEAST software in Linux or ubuntu.

The proportion of substitution rate at the non-synonymous and synonymous site are quantified for the estimation of selection and evolutionary pressure on the protein-coding regions; the ratio of *dN/dS* is the most often used method. The *dN/dS* compares the rate of substitution at a silent site (dS) to the substitution non-silent sites (*dN*), which may experience the selection (Kryazhimskiy and Plotkin, 2008).

Although the evolutionary analysis of the whole KFDV genome was already analyzed in the previous studies (Kasabi, 2011; Yadav, 2020; Dodd, 2011; Mehla, 2009), in the current study, a strategy to analyze the pattern of codon usage bias, with phylogenetic background, and the variations among volutionary characters using the tMRCA & evolutionary rate methods have been employed to understand the common ancestors, selection pressure, and most abundant codon frequency, respectively among the Capsid (C), pre Membrane (prM), Envelope (E), and non-structural RNA polymerase NS5 of the KFD virus.

MATERIALS AND METHODS

Codon Usage Bias Analysis

Data Collection and Sequence Editing

The complete nucleotide coding sequence (CDs) of gene C, gene E, prM, and NS5 of *Homo sapiens*, monkeys, and ticks were downloaded individually from the NCBI database (National Center for Biotechnology Information (nih.gov))in FASTA format, which had 30, 48, 34, and 33, respectively. The CDs sequence was aligned using MEGA-X (Molecular Evolutionary Genetics Analysis)software, MUSCLE algorithm.

Quantitative Analysis of Nucleotide Composition and Analysis of Dinucleotide Abundance Frequency

The nucleotide composition of gene C, gene E, prM, and NS5, include nucleotide at 3^{rd} position, and GC, GC1, GC2, GC3, GC12 (mean of GC at the first and second position) were estimated using the MEGA-X program. The frequency of GC and dinucleotide content, mononucleotides were calculated using the R program by adopting the "SeqinR" add-on package.

A certain dinucleotide contributes to a codon usage bias, and it represents a total of 16 dinucleotide compositions of gene C, gene E, prM, and NS5 genes. The dinucleotide abundance frequency is calculated using by following Karlin and Burge method (Tao, 2009).

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

In the above formula, the frequency of nucleotide X and Y are represented as F_x and F_y, di nucleotide frequency is represented as F_{XY}. If the value of P_{XY} is >1.23, it represents the over representation of dinucleotides, and the P_{XY} value < 0.78 are underrepresented dinucleotides. Extreme dinucleotide abundance is $P_{XY} \geq 1.50$, and $P_{XY} \leq 0.50$ is extremely underrepresented. Framework work of relative dinucleotide abundance analysis identifies both natural selection and mutational pressure (Khandia *et al.*, 2019). In order to calculate the abundance frequency, the dinucleotide frequency was calculated in R studio software using the "SeqinR" package.

Analysis of Relative Synonymous Codon Usage (RSCU)

The RSCU analysis was performed to determine the resemblance impact and study the feature of synonymous codon usage between genes- C, E, prM, and NS5of the KFD virus (Rahman and Ur, 2017; Takaaki and Tatsuo, 2018). The method is implemented by assigning the numerical values to each codon of the gene or genome and analyzing each codon bias by comparing to the referencing standard random expectations of the codon usages. Therefore, the RSCU index is the division of the observed frequency of codon usage by the probable frequency under the hypothesis of equivalent synonymous codon usage of amino acids (Behura and Severson, 2013). The individual RSCU value was calculated using the simple formula:

$$RSCU_{ij} = \frac{X_{ij}}{\left[\left(\frac{1}{n_i}\right)\sum_{j=1}^{ni}\ (X_{ij})\right]}$$

In the above formula, X_{ij} denotes the number of the j^{th} codon for the i^{th} amino acid, and synonymous codons that codes for the i^{th} amino acids are represented by *n*.

Parity rule 2 bias plot analysis

The Parity rule 2 (PR2) bias was determined by the values of GC bias [G3/(G3+C3)] on the abscissa and AT bias[A3/(A3+T3)] on the ordinate. The exploration of this method helps to examine the magnitude between mutational pressure and natural

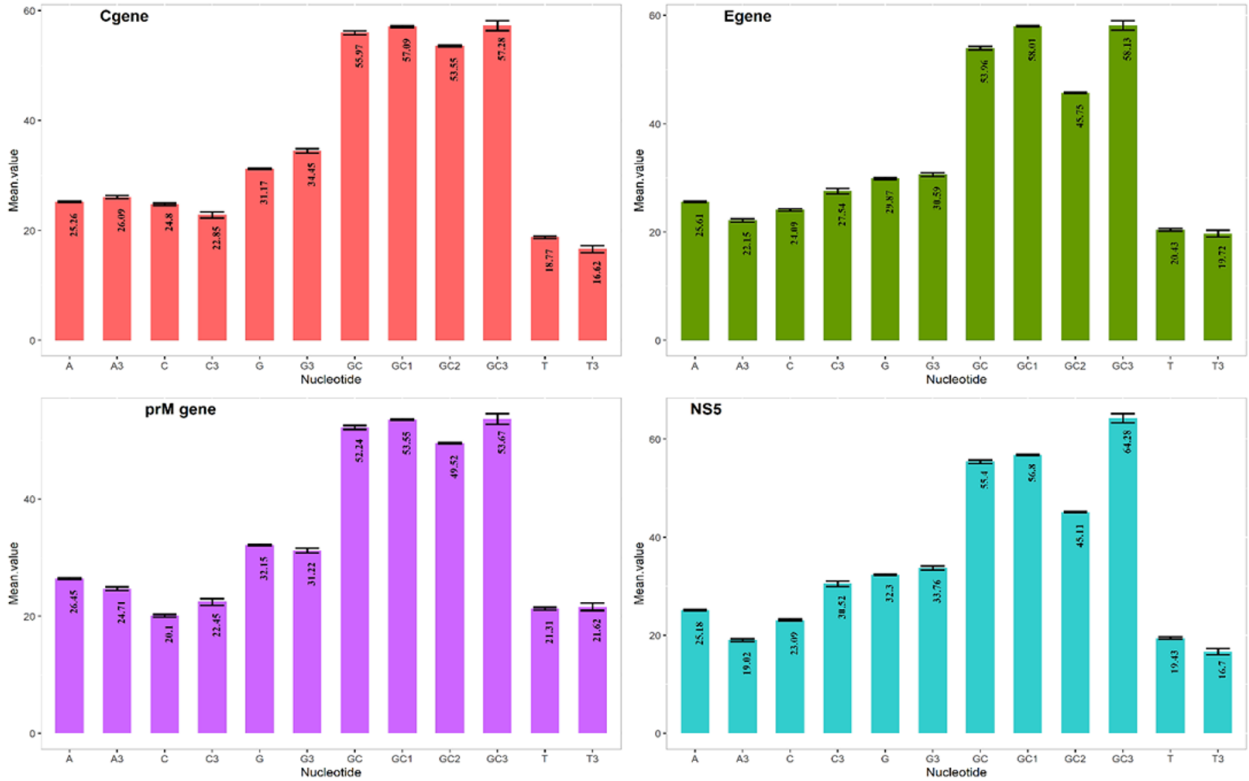

Figure 1: Graphical representation of overall nucleotide composition among genes- C, E, prM, NS5 of KFD virus. Error bars are indicative of standard deviation.

Table 1: Nucleotide composition of gens- C, E, prM, and NS5 of the KFD virus (in frequency)

Nucleotide composition	Genes of KFDV			
	Gene C	Gene E	Gene prM	Gene NS5
T	18.77% ± 0.30	20.42% ± 0.26	21.30% ± 0.19	19.42% ± 0.22
C	24.80% ± 0.20	24.09% ± 0.24	20.09% ± 0.38	23.09% ± 0.20
A	25.25% ± 0.35	25.61% ± 0.08	26.44% ± 0.41	25.17% ± 0.12
G	31.16% ± 0.40	29.86% ± 0.07	32.14% ± 0.25	32.30% ± 0.13
T3	16.61% ± 7.99	19.71% ± 0.78	21.61% ± 1.58	16.69% ± 0.63
C3	22.85% ± 1.68	27.53% ± 0.66	22.44% ± 1.68	30.51% ± 054
A3	26.09% ± 5.52	22.15% ± 0.23	24.70% ± 1.21	19.02% ± 0.30
G3	34.44% ± 4.08	30.59% ± 0.20	31.22% ± 1.87	33.76% ± 0.40
GC	55.97% ± 0.54	53.95% ± 0.20	52.24% ± 0.58	55.39% ± 0.33
GC1	57.08% ± 1.42	58.00% ± 0.19	53.55% ± 1.01	56.79% ± 0.15
GC2	53.45% ± 2.19	45.74% ± 0.09	49.51% ± 1.14	45.10% ± 0.14
GC3	57.28% ± 2.59	58.12% ± 0.68	53.67% ± 1.36	64.28% ± 0.89

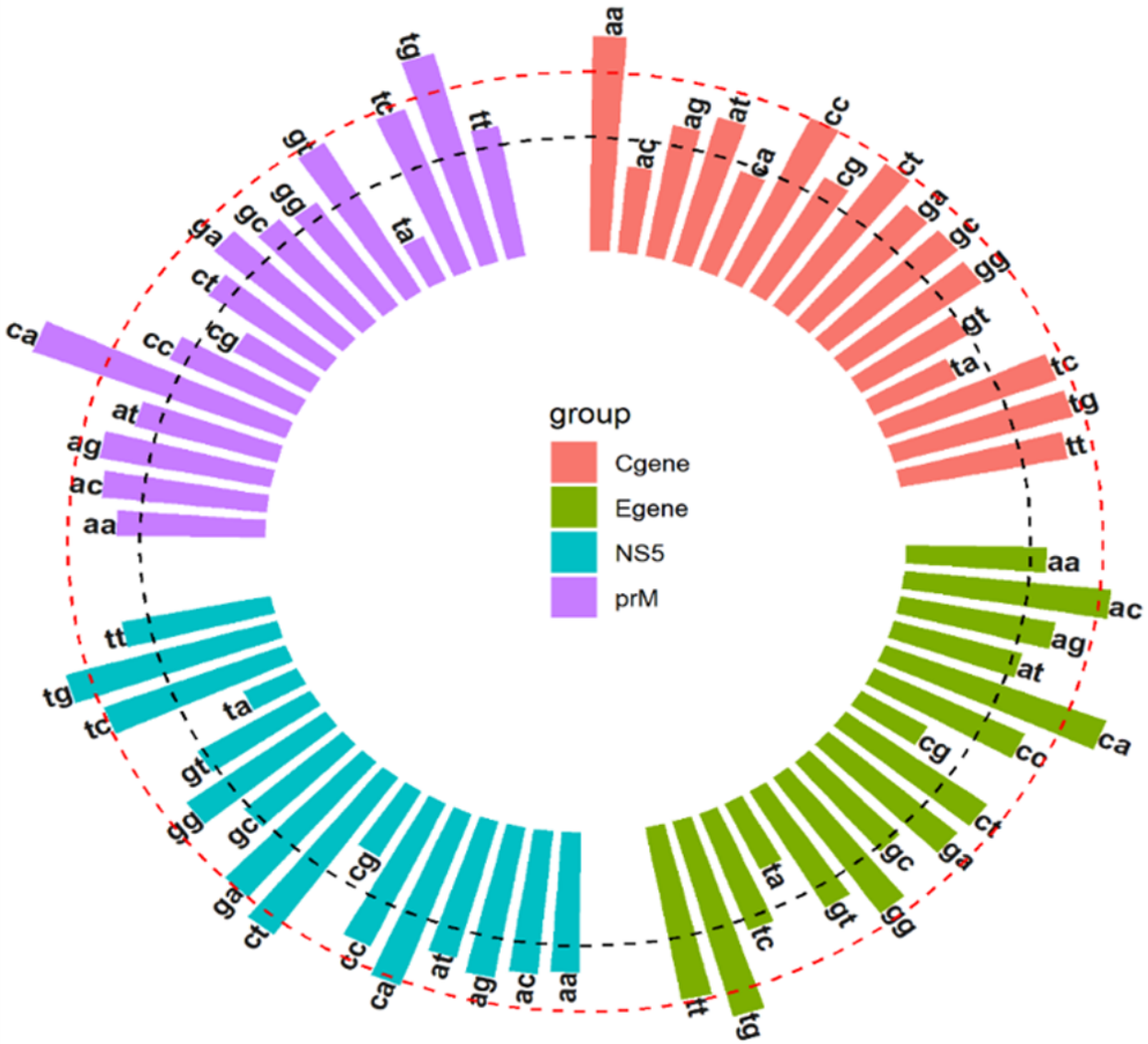

Figure 2: Illustration of dinucleotide abundance frequency of genes- C, E, prM, and NS5 of KFD virus. Red and black lines are shown as an indication of over and under-represented dinucleotide frequencies, respectively.

selection (Tao and Yao, 2020). In the PR2 plot, 0.5 is the focal point of both coordinates, is the position where G=C and A=T (Chen *et al.*, 2014). Therefore, points situated on the focal point is an indication of unbiased and nonconformity between natural selection and mutation pressure (Pan *et al.*, 2020).

Analysis of Effective Number of Codons (eNC)

An effective number of codons is aimed to measure the distant between codon usage of a gene and standard expected synonymous codon usage. So, the notch of codon usage bias is measurable by the size of eNC values (Wu *et al.*, 2020; Wang and Zhang, 2020). The range of the eNC values lies between 20 and 61. In which only one codon is used to code for each amino acid is known as an extremely biased gene, eNC value will be closer to 20. In contrast, the value of the unbiased gene will be 61. The method also suggests that the more the extent of codon favourable in a gene with, the lesser the ENC value (Xu, 2017). Values between 35 to 55 are an indication of slight bias in a gene (Wang and Zhang, 2020; Tao and Yao, 2020). The eNC values and eNC plot was illustrated using the following formulas:

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

In the formula, F_i (i = 2, 3, 4, 6) represents the average of F_i values for *i* fold amino acids, where F_i can be calculated using the below formula:

$$F_{i=}\frac{n\sum_{j=1}^{i}(n_{j/n})^2 - 1}{n-1}$$

Codon	Relative Synonymous Codon Usage Frequency			
	Gene C	Gene E	GenePrM	GeneNS5
AAA	1.090909	0.5	0.75	0.75
AAC	0.666667	1.684211	1.5	1.44
AAG	0.909091	1.5	1.25	1.25
AAT	1.333333	0.315789	0.5	0.56
ACA	0.444444	1.52381	2	1.302326
ACC	1.777778	1.142857	1.142857	1.395349
ACG	1.333333	0.666667	0.857143	0.837209
ACT	0.444444	0.666667	0	0.465116
AGA	1.285714	1.363636	2	1.463415
AGC	0	1.241379	0.666667	1.222222
AGG	1.285714	1.090909	2.5	1.756098
AGT	0	1.241379	1.333333	1
ATA	0	0.882353	0.428571	0.315789
ATC	2.25	1.058824	0.857143	2.052632
ATT	0.75	1.058824	1.714286	0.631579
CAA	1	0.142857	2	0.545455
CAC	0	1.333333	0.5	0.8
CAG	1	1.857143	0	1.454545
CAT	2	0.666667	1.5	1.2
CCA	1.5	2.181818	1.6	1.333333
CCC	0.5	0.545455	0.8	1.466667
CCG	1.5	0.727273	1.6	0.666667
CCT	0.5	0.545455	0	0.533333
CGA	1.714286	1.363636	0.5	0.731707
CGC	0.428571	0.272727	0.5	1.170732
CGG	0.428571	1.636364	0	0.585366
CGT	0.857143	0.272727	0.5	0.292683
CTA	0.8	0.782609	0	0.375
CTC	0.8	1.304348	1.8	1.725
CTG	2	1.826087	3	1.95
CTT	0.4	0.782609	1.2	0.825
GAA	0	0.857143	1.111111	0.727273
GAC	1	1.454545	1	1.333333
GAG	2	1.142857	0.888889	1.272727
GAT	1	0.545455	1	0.666667
GCA	0.727273	0.727273	0.923077	0.848485
GCC	1.454545	1.454545	1.230769	1.212121
GCG	1.090909	0.363636	0.615385	0.606061
GCT	0.727273	1.454545	1.230769	1.333333
GGA	2.153846	1.259259	1.428571	1.55
GGC	0.615385	1.111111	0.857143	0.95
GGG	0.615385	0.888889	1.428571	0.8
GGT	0.615385	0.740741	0.285714	0.7
GTA	0.444444	0.444444	0.421053	0.126984
GTC	0.888889	0.622222	0.842105	1.206349
GTG	1.333333	2.044444	1.894737	2.222222
GTT	1.333333	0.888889	0.842105	0.444444
TAC	0	0.8	1.333333	1.153846
TAT	0	1.2	0.666667	0.846154
TCA	1.5	1.862069	3.333333	1.222222
TCC	1.5	1.034483	0	1.555556
TCG	0	0.206897	0.666667	0.222222
TCT	3	0.413793	0	0.777778
TGC	2	0.666667	0.5	0.941176
TGT	0	1.333333	1.5	1.058824
TTA	0.4	0.130435	0	0.075
TTC	2	0.625	1.333333	1.304348
TTG	1.6	1.173913	0	1.05
TTT	0	1.375	0.666667	0.695652

Figure 3: Overall RSCU frequency of gene C, gene E, prM, and NS5 of the KFDV. Over represented codons (>1.6) are highlighted in yellow and under represented codons (<0.6) are in green.

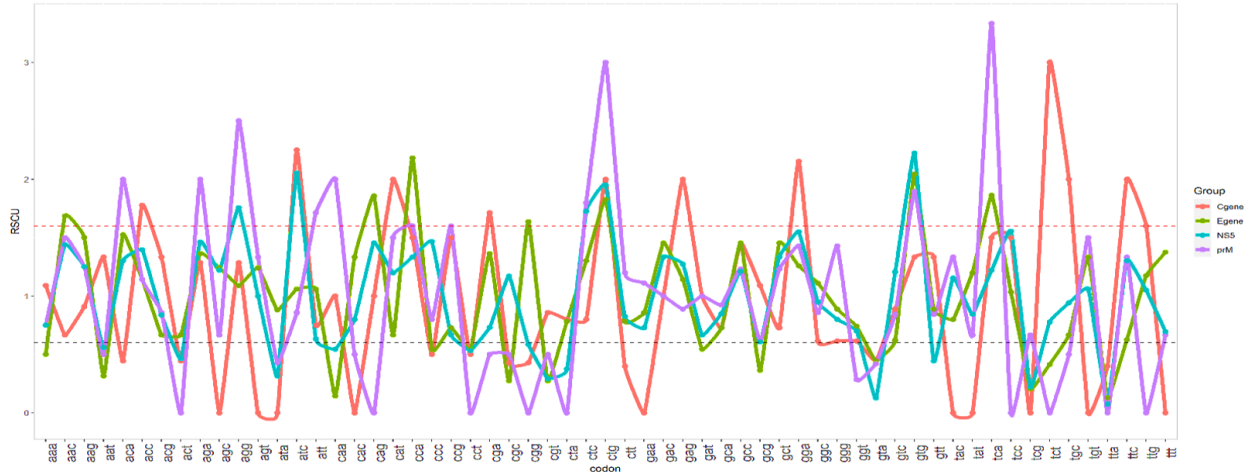

Figure 4: Representation of overall RSCU frequency of selected genes. A dotted red and black lines are indicative of over and under representation, respectively.

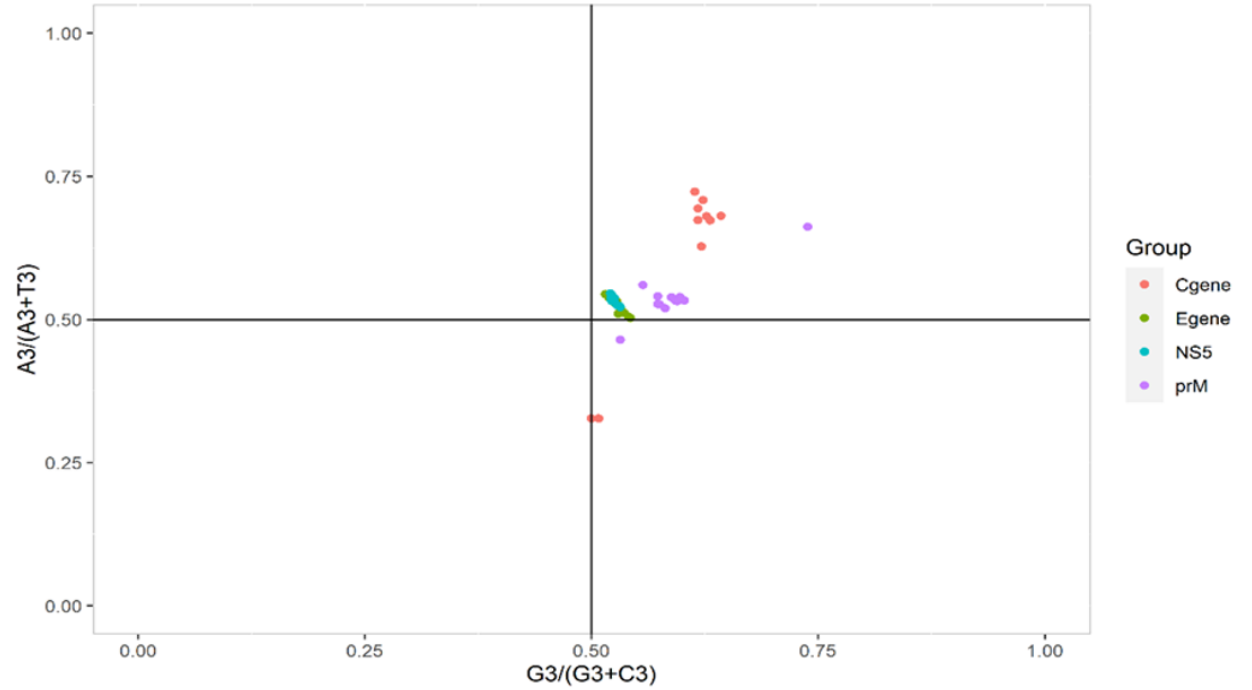

Figure 5: Illustration of PR2 bias plot. The interception of the X and Y-axis is the origin (0.50), and it is the place where nucleotide composition A=T, G=C of the DNA strand.

Where *n* represents the total of codons for particular amino acid, and nj represents the total observed number of the *j*th codon for that amino acid (Wu *et al.*, 2020).

eNC plot was illustrated to determine the relationship between an eNC and GC3 (GC nucleotide at 3^{rd} position). The plot quantifies the codon usage bias of a gene, and it is the best method to estimate the total synonymous codon usage. The following equation was used to plot the eNC plot,

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

Where GC3 contents are represented by *S*. If the eNC values pointed on the standard, it specifies that bias is influenced by the mutational pressure, whereas values that are pointed below the curve indicates the role of natural selection in shaping the codon usage (Yao *et al.*, 2020).

Neutrality plot analysis

The neutrality plot method was employed to identify the factors that affect the preference of codon usage and to examine the amount of influence of natural selection and mutational pressure in each gene, gene C, gene E, gene prM, and gene NS5 of the KFD virus. The neutrality plot demonstrates the linear

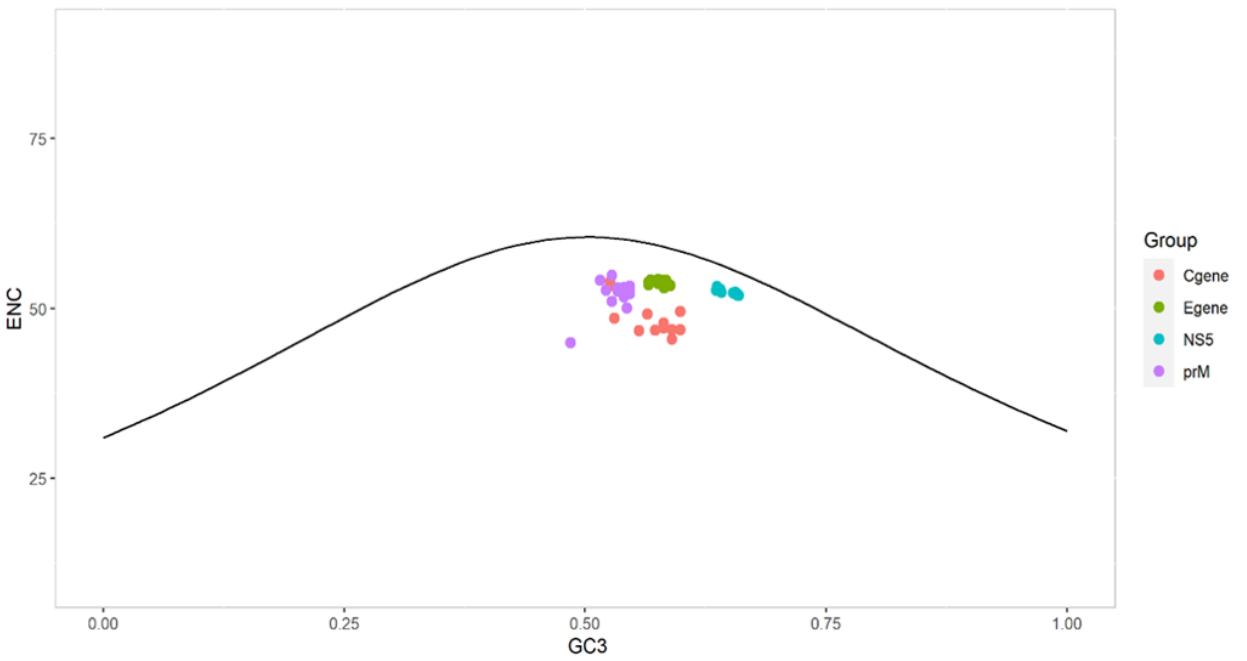

Figure 6: Each coloured points in a plot represent the different genes of the KFDV, and the illustration of a plot elucidates the relationship between the ENC value of a particular gene and its GC3 content.

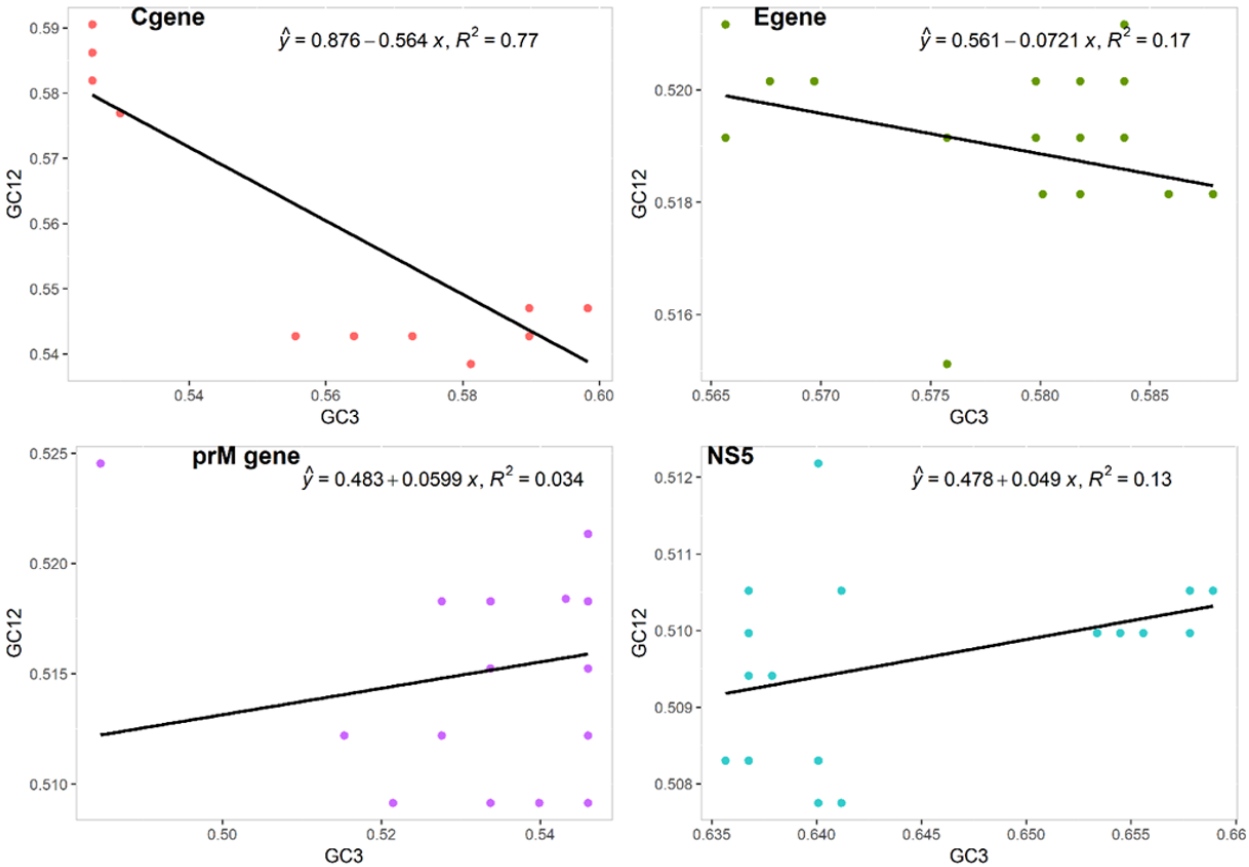

Figure 7: Graphical representation of neutrality of gene C, gene E, prM, and NS5. The Axis-X represents the frequency of GC3, and Axis-Y represents the mean value of GC12 (mean of GC at $1^{st}$and 2^{nd} position).

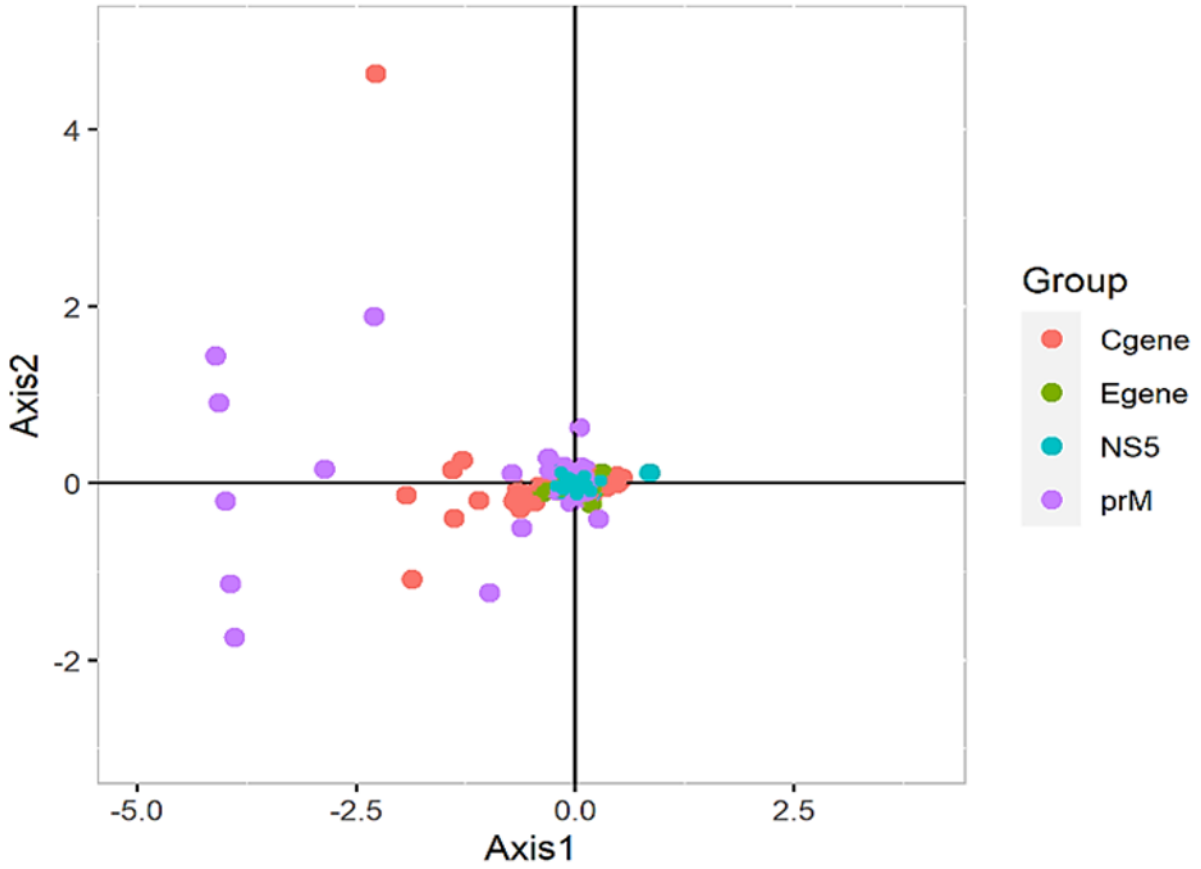

Figure 8: Graphical representation of correspondence analysis, showing a greater contribution of Axis-1 in shaping the codon usage pattern in genes- C, E, prM, and NS5 of the KFD virus.

Table 2: Relative dinucleotide abundance frequencies of genes - C, E, prM, and NS5

Dinucleotides	Gene C	Gene E	Gene prM	Gene NS5
AA	1.477	0.881	0.93	0.964
AC	0.591	1.298	1.037	0.989
AG	0.913	0.985	1.087	1.052
AT	1.035	0.831	0.923	0.974
CA	0.727	1.475	1.691	1.263
CC	1.237	1.044	0.885	1.092
CG	0.918	0.485	0.556	0.521
CT	1.203	1.09	0.898	1.331
GA	1.06	1.064	1.063	1.224
GC	1.071	0.809	0.928	0.812
GG	1.045	1.147	0.875	1.069
GT	0.774	0.929	1.19	0.817
TA	0.575	0.503	0.332	0.371
TC	1.143	0.859	1.182	1.215
TG	1.161	1.394	1.487	1.372
TT	1.065	1.207	0.909	0.947

Table 3: Selected gene's sequence of the KFDV with accession ID and year of isolation (YoI).

Gene C		Gene E		GeneprM		GeneNS5	
Accession Number	YoI	Accession Number	YoI	Accession Number	YoI	Accession Number	YoI
JQ434075.1	2007	JQ434075.1	2007	MH013227.1	2016	MG720079.1	2012
MG720081.1	2013	KP315947.1	2014	MH013226.1	2016	MG720080.1	2012
MG720082.1	2013	KY779854.1	2012	MG934430.1	2017	MG720081.1	2013
MG720085.1	2013	KY779859.1	2012	MG720122.1	2016	MG720082.1	2013
MG720086.1	2013	KY779864.1	2015	MG720121.1	2017	MG720083.1	2012
MG720087.1	2013	KY779865.1	2015	MG720120.1	2016	MG720085.1	2013
MG720091.1	2016	KY779866.1	2013	MG720119.1	2016	MG720086.1	2013
MG720092.1	2016	KY779867.1	2015	MG720118.1	2016	MG720087.1	2013
MG720096.1	2016	MF186838.1	2016	MG720117.1	2017	MG720091.1	2016
MG720098.1	2006	MF186839.1	2016	MG720116.1	2017	MG720092.1	2016
MG720101.1	2016	MF186840.1	2016	MG720115.1	2016	MG720096.1	2016
MG720102.1	2017	MF186841.1	2016	MG720114.1	2017	MG720098.1	2006
MG720104.1	2016	MF186842.1	2016	MG720113.1	2017	MG720101.1	2016
MG720105.1	2016	MF186843.1	2016	MG720111.1	2016	MG720102.1	2017
MG720106.1	2016	MF186844.1	2016	MG720110.1	2014	MG720103.1	2012
MG720108.1	2012	MG720079.1	2012	MG720108.1	2012	MG720104.1	2016
MG720111.1	2016	MG720080.1	2012	MG720106.1	2016	MG720105.1	2016
MG720113.1	2017	MG720081.1	2013	MG720105.1	2016	MG720106.1	2016
MG720114.1	2017	MG720082.1	2013	MG720104.1	2016	MG720108.1	2012
MG720115.1	2016	MG720083.1	2012	MG720103.1	2012	MG720110.1	2014
MG720117.1	2017	MG720085.1	2013	MG720102.1	2017	MG720111.1	2016
MG720118.1	2016	MG720086.1	2013	MG720101.1	2016	MG720113.1	2017
MG720119.1	2016	MG720087.1	2013	MG720098.1	2006	MG720114.1	2017
MG720120.1	2016	MG720091.1	2016	MG720096.1	2016	MG720115.1	2016
MG720121.1	2017	MG720092.1	2016	MG720092.1	2016	MG720116.1	2017
MG720122.1	2016	MG720096.1	2016	MG720091.1	2016	MG720117.1	2017
MH013226.1	2016	MG720098.1	2006	MG720087.1	2013	MG720118.1	2016
MH013227.1	2016	MG720101.1	2016	MG720086.1	2013	MG720119.1	2016
–	–	MG720102.1	2017	MG720085.1	2013	MG720120.1	2016
–	–	MG720103.1	2012	MG720083.1	2012	MG720121.1	2017
–	–	MG720104.1	2016	MG720082.1	2013	MG720122.1	2016
–	–	MG720105.1	2016	MG720081.1	2013	MH013226.1	2016
–	–	MG720106.1	2016	MG720080.1	2012	MH013227.1	2016
–	–	MG720108.1	2012	MG720079.1	2012	–	–
–	–	MG720110.1	2014	–	–	–	–
–	–	MG720111.1	2016	–	–	–	–
–	–	MG720113.1	2017	–	–	–	–
–	–	MG720114.1	2017	–	–	–	–
–	–	MG720115.1	2016	–	–	–	–
–	–	MG720116.1	2017	–	–	–	–
–	–	MG720117.1	2017	–	–	–	–
–	–	MG720118.1	2016	–	–	–	–
–	–	MG720119.1	2016	–	–	–	–
–	–	MG720120.1	2016	–	–	–	–
–	–	MG720121.1	2017	–	–	–	–
–	–	MG720122.1	2016	–	–	–	–
–	–	MH013226.1	2016	–	–	–	–
–	–	MH013227.1	2016	–	–	–	–

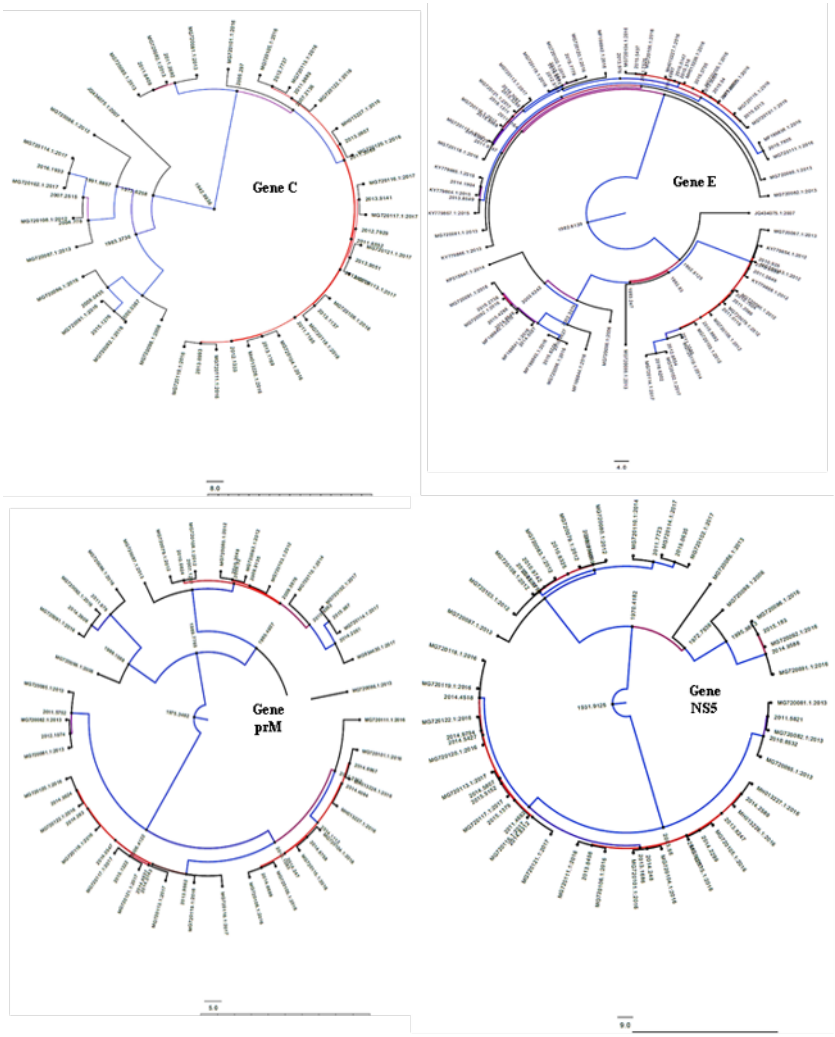

Figure 9: Phylogenetic tree of each gene - C,E,prM, and NS5 of the KFD virus indicating the tMRCA diverging age at the midpoint.

Table 4: Substitution rate and tMRCA ages of KFDV genes. The substitution rate was estimated using the Datamonkey server.

Genes	Substitution Rate (subs/site/year)			tMRCA Ages		
	Mean	95% HPD (Highest Posterior Density)		Mean	95% HPD	
		Low	High		Low	High
Gene C	5.606E-4	1.5466E-4	1.0269E-3	74.07	23.81	144.25
Gene E	8.964E-4	5.8034E-4	1.2168E-3	34.52	23.25	47.65
Gene prM	5.655E-4	2.1754E-4	9.3808E-4	41.92	18.90	74.40
Gene NS5	2.678E-4	8.1067E-5	4.4463E-4	86.31	35.14	158.61

Table 5: List of a positively selected site among the selected genes of KFD virus

KFDV Genes	Positively selected site	Overall dN/dSrate ratio
Gene C	34	0.371
Gene E	123	0.273
Gene prM	No Evidence	0.225
Gene NS5	826	0.064

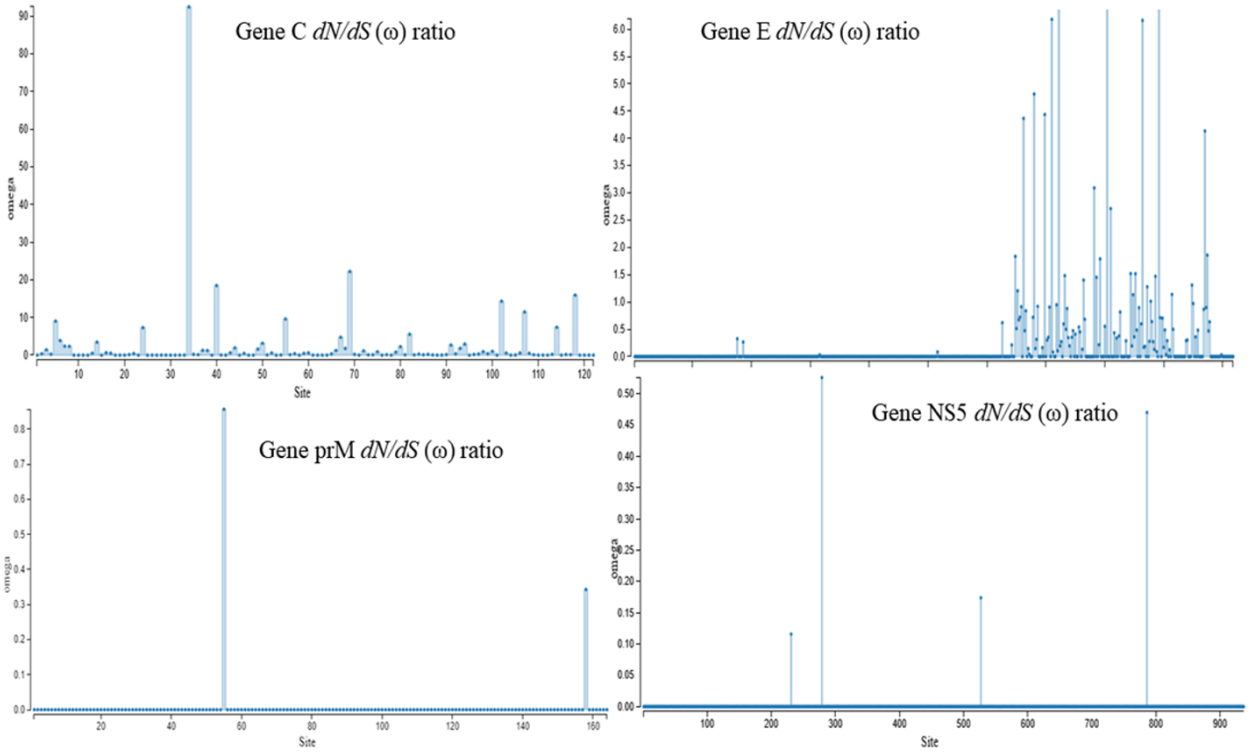

Figure 10: Overall dN/dS(ω) rate ratio of genes- C, E, prM, and NS5 of the KFD virus.

relationship between GC12 (Mean of GC1 and GC2) and GC3. To plot the scatter diagram, the GC3 values were plotted against GC12 on verticle and horizontal axes, respectively. As a selection-mutation equilibrium coefficient, a regression line was plotted on the graph. Less than 0.5 regression coefficient value indicates the influence of natural selection, and greater than 0.5 indicates the major role of mutational pressure (Deb *et al.*, 2020).

Estimation of Average Aromaticity (AROMA) and Hydrophobicity (GRAVY)

Apart from nucleotide compositions, ENC, RSCU, and CAI, GRAVY (Grand Average Hydrophobicity), and AROMO (Aromaticity) contributes to shaping the codon usage. GRAVY is well known as the sum of hydropathy values of entire amino acids in a sequence, range between -2.0 & +2.0. Hydrophobicity proteins will attain positive values, whereas hydrophilic proteins will achieve negative values (Khandia *et al.*, 2019). The AROMO value is the occurrence of aromatic amino acids, i.e. Trp, Phe, and Tyr, within an amino acid sequence (Khandia *et al.*, 2019).

Correspondence Analysis (COA)

The codon usage bias diverges from one gene to another. Therefore, the correspondence analysis was implemented based on the previous study by Greenacer (Fabra *et al.*, 2010; Yao *et al.*, 2020) to compute the relationship and variation among the genes- C, E, prM, and NS5 (Xu, 2017). The implementation of COA was to understand the variation and dissimilarities in codon usage. A plot was illustrated using the values of 59 synonymous codons of RSCU across two axes (axis-1, axis-2) (Deb *et al.*, 2020). The correspondence analysis was performed using CodonW software and visualized in R programming software.

Evolutionary characteristics Analysis

Sequence Data

Each nucleotide sequence of genes- C, E, prM and NS5 were downloaded from the NCBI website (National Center for Biotechnology Information (nih.gov)). A total of 30, 48, 34, and 33 sequences of genes- C, E, prM, and NS5 of Homo sapiens, Monkeys, and Ticks from the region of India were downloaded.

Sequence Alignment

The multiple sequence alignment and sequence editing of a nucleotide sequence that codes for genes- C, E, prM, and NS5 of the KFD virus were individually aligned using the MEGA-X software by incorporating the MUSCLE algorithm.

Homologous Recombination Detection

The Datamonkey web server, which provides an algorithm GARD (Genetic Algorithm for Recombination Detection) to estimate the homologous recombination, was used to determine the homologous

recombination genes. There was no report of genetic recombination regions found in any of the genes.

Evolutionary Rate and Coalescent Analysis

The construction of phylogenetic analysis, a choice of statistical best-fit models, is the fundamental criteria. Therefore, a phylogenetic model was selected based on the Akaike Information Criteria (AIC) obtained from the jModelTest2 tool. Whereas, the BEAUti interface of the BEAST software was used to build the input analysis. The four molecular clock models (Relaxed Clock log-normal (RCLN), Relaxed clock exponential (RCE), Strict clock, and Random Local Clock (RLC)) were considered with Coalescent: Bayesian skyGrid and Coalescent: Extended Bayesian skyline plot trees (Suresh, 2020). An algorithm Markov Chain Monte Carlo (MCMC) in the Bayesian Analysis by Sampling Tree was used to coassess the evolutionary rate and time to the most recent common ancestor. The MCMC chains was repeatedly changed until all the constraints had an effective sample extent of >200. The BEAST generated log files were analyzed using the BEAST integrated Tracer tool.

Determination of Selection Pressure or Positive Selection

The positive selection pressure was employed using the Datamonkey Adaptive Evolution (DAE) server to eliminate the superfluous sequence in a dataset of evolutionary rate analysis. Determining the ratio of non-synonymous (*dN*) to synonymous (*dS*) substitution is an approach to estimate the selection pressure. The Fixed-Effects Likelihood was employed for the evaluation of *dN*, *dS*, and *dN/dS* (ω) rate per site of a coding alignment sequence. Also, the FEL strategy expects that the determination of positive selection for each site is consistent along the whole phylogeny.

RESULTS

Codon usage analysis

Data collection and sequence editing

The CDS (coding sequence) of each gene i.e., gene C (n=30), gene E (n=48), gene prM (n=34) and gene NS5 (n=33) of the KFD virus were extracted from the NCBI (National Center for Biotechnology Information (nih.gov)) database. Removal of stop codons, sequence editing, multiple sequence alignment (using MUSCLE algorithm), and estimation of nucleotide composition was performed using MEGA-X software.

Analysis of Nucleotide Composition and Relative Dinucleotide Abundance Frequency

The nucleotide composition such as A, T, G, C, and nucleotide composition at 3^{rd} position A3, T3, G3, C3, also the G+C contents GC, GC1(GC content at 1^{st} codon position), GC2 (GC content at 2^{nd} codon position), GC3 (GC content at 3^{rd} codon position) of genes - C, E, prM, NS5 were estimated to analyze the contribution of nucleotide composition in codon usage bias. The evaluated frequency of nucleotide composition is given in **Table 1. Fig. 1** and detailed nucleotide compositions of all the genes are listed in [**supplementary table 1**].

Whereas, an estimation of relative abundance frequency of 16 dinucleotides of selected genes- C, E, prM, and NS5 of KFDV was calculated using R studio software. The dinucleotide that has a frequency value >1.23 is known as over represented, and < 0.78 are underrepresented. The overall abundance frequency f all four genes are listed in **Table 2. Fig. 2.**

1. Gene C: Among all the 16 dinucleotide bases, two dinucleotide AA (1.477) and CC (1.237) were overrepresented >1.23. And, AC (0.591), GT (0.774), and TA (0.575) were observed as underrepresented < 0.78.

2. Gene E: AC (1.298), CA (1.475), and TG (1.394) are overrepresented, and CG (0.485), TA (0.503) were underrepresented.

3. Gene prM: CA (1.691) and TG (1.487) are overrepresented, and CG (0.556), TA (0.332) seems to be underrepresented.

4. Gene NS5: The dinucleotide CA (1.263), CT (1.331), TG (1.372) were observed as overrepresented, and CG (0.521), TA (0.371) are underrepresented.

Analysis of Relative Synonymous Codon Usage (RSCU)

In the present study, the relative synonymous codon usage of each gene, gene C, gene E, prM, and NS5, were determined and plotted using the R studio programming software. Fig. 3. The frequency value of each synonymous codon is segregated based on the RSCU range of 0.6 to 1.6. The values >1.6 are known as an over represented synonymous codon, and < 0.6 are represented as underrepresented synonymous codons. The determined over and under-represented codons are highlighted yellow and green colour, respectively. **Table 3.** The codons that achieve a significant frequency value >1.0 are known as high frequency or positively biased codons. Whereas, the frequency < 1.0 is

termed as a lower frequency or negatively biased codon.

1. Gene C: Among all the 59 codons, eight codons (
2. Gene E: Seven codons (AAC, CAG, CCA
3. Gene prM: Nine codons (ACA, AGA, AGG, ATT, CAA, CTC, CTG, GTG, TCA), and eight codons (AAT, CAC, CGA, CGC, CGT, GGT, GTA, TGC) was identified as over and under-represented codons, respectively. Also, 25 high-frequency & 24 low-frequency codons were obtained. 7 codons were terminated with nucleotide A among the 25 high-frequency codons in the prM gene.
4. Gene NS5: Five (AGG, ATC, CTC, CTG¸GTG), and nine codons (AAT, ACT, ATA, CAA, CCT, CGG, CGT, CTA, TCG) were over and under-represented, respectively. Also, 28 high-frequency & 30 low-frequency codons were determined. Among the 28 high-frequency codons, 13 codons were dominantly terminated with nucleotide C.

Analysis of natural selection and mutational pressure on the codon usage bias

To analyze the influencing factors of mutational pressure and natural selection among the genes- C, E, prM, and NS5 of the KFD virus, the estimation of Parity Rule 2 (PR2), eNC, and Neutrality was examined and plotted by using R studio programming software.

Analysis of Parity rule 2 – plot

The origin of the PR2 signifies the direction and degree of bias. Parity rule 2 bias plot is comparatively informative when the PR biases are estimated at the third position of AT and GC content. The nucleotide composition of DNA is A=T, G=C, according to Chargraff's 2nd parity rule (PR2). So the origin is the place where there is no accumulation of bias. The PR2 plot is constructed by plotting the [G3/(G3+C3)] values on X-axis and [A3/(A3+T3)] values on the Y-axis. In this study, the mean value of[G3/(G3+C3)] and [A3/(A3+T3)] of each selected genes of the KFD virus were as follow,

1. Gene C: The calculated mean value of GC and AT bias was 0.59 and 0.61, respectively. The domination of AT over the GC indicates
2. Gene E: Mean value of GC and AT bias was 0.52 and 0.52, respectively. Equal contribution of both purines and pyrimidines.
3. Gene
4. Gene

In the present study, none of the genes has A=T, G=C composition, specifying the bias among selected genes. The NS5 and gene E shows a slightly less bias compared to gene C, and prM since the points of gene C and prM were situated away from the origin. Fig. 4. Therefore, the parity rule 2 plot specifies an occurrence of bias at the third position of AT and GC of all the selected genes, also suggests the major role of natural selection over the mutational pressure.

Analysis of Effective Number of Codons (eNC)

The effective number of codons was calculated for all the selected genes of KFDV, and it is to evaluate the pattern of extent codon usage in a particular gene or genome. Whereas in this study, the eNC value of genes- C, E, prM, and NS5 obtained was 45.52-53.92 (SD $\pm$2.64), 53.04-54.28 (SD $\pm$0.33), 44.96-54.85 (SD $\pm$1.53), and 51.93-53.25 (SD $\pm$0.30), respectively (**supplementary table 1**). eNC of all the selected genes was plotted in a single frame to illustrate and compare the selection and mutational pressure. Fig. 5. Each different colour point represents the different four genes. Points situating right below and close to the standard curve specifies a major role of natural selection and slight influence of GC3.

Analysis of Neutrality plot

The neutrality was analyzed and plotted by calculating the nucleotide composition of GC12 (mean value of GC1 and GC2) against GC3 to determine the prompting factors of natural selection and mutational pressure. Whereas as in the graph, the slope of the regression line acts as an indicator or expressed as the evolutionary rate of natural selection and mutational pressure. Also, the regression coefficient against GC12 and GC3 is considered as a natural-mutational equilibrium coefficient. In the present study, a negative regression line and negative significant R-value were observed in gene C with y= 0.876-0.564, R^2 = 0.77, and 56.4% neutrality that indicates mutational pressure over selectional pressure. The negative regression line and negative significant score between GC12 & GC3 of gene E were observed as y=0.561-0.0721, R^2= 0.17, and 1.7% of neutrality that shows a mutational pressure. whereas, in the prM and NS5 genes, the positive regression line and positive significant coefficient were observed with y=0.483+0.0599, R^2=0.034 & 3.4% neutrality, and y= 0.478+0.049, R^2=0.13 &1.3% neutrality, respectively. Indicating natural selection over mutational pressure. Fig. 6.

Estimation and correlation of Average Aromaticity (AROMA) and Hydrophobicity (GRAVY)

To determine the correlation of hydrophobicity and aromaticity between eNC and GC12 of the genes- C, E, prM, and NS5 were evaluated [**supplementary table 2**]. In gene C, a negative significant value was observed between eNC-GRAVY and GC12-GRAVY, -0.2429 and -0.2901, respectively, indicating the influence of hydrophobicity. Whereas, the non-significant correlation of aromaticity between eNC and GC12 signifies the absence of aromaticity effect while shaping the codon bias pattern in gene C. The non-significant correlation score of GRAVY between eNC and GC12 0.0131 and 0.0349, respectively, signifies the non-impact of hydrophobicity in gene E, but a significant correlation of AROMA between eNC and GC12 -0.0631 and -0.1144 was seemed to be a contributing factor for shaping the codon usage bias in gene E. Significant correlation -0.0154 and -0.0171 of GRAVY were observed and shows the impact of hydrophobicity, but the observed non-significant correlation of AROMA between eNC and GC12 were 0.1537 and -0.1621, respectively in gene prM. Although in non-structural gene NS5, non-significant correlation values of GRAVY between eNC& GC12 and AROMA between eNC& GC12, i.e., 0.1010&-0.1224 and 0.0631&0.0351, respectively, were noticed as a non-contribution factor while shaping the codon usage bias.

Correspondence analysis (COA)

The correspondence analysis was performed using the R program software. Fig. 7 represents the COA of each gene, in which each set of coloured points are representative of different selected genes of KFDV.In correspondence analysis (COA), 59 codons (except Trp, stop codons, and Met) were illustrated along 59 orthogonal axes in high-dimensional space. In a COA plot, the RSCU value was plotted in the high-dimensional space. Whereas, Axis-1 and Axis-2 were considered as major contributors (these axes are the two major dimensional coordinates). In the analysis, axis-1 elucidated 54.3%, 85.8%, 49.9%, and 93.3% contribution, and axis-2 elucidated 20.5%, 5.6%, 28.7%, and 4.1% contribution for genes- C,E, prM,and NS5 of the KFD virus, respectively. Hence, axis-1 shows a higher usage for NS5 (93.3%) and comparatively slightly less for the rest genes of the KFDV. Therefore, axis-1was more contemplated while shaping the codon usage pattern. Also, the contribution of Axis-1 is an indication of greater codon usage variation among the gene C, gene E, prM, and NS5 genes of the KFDV.

Evolutionary characteristics Analysis

Data collection and Sequence Editing

To analyze the evolutionary characteristics, the profound nucleotide sequence of gene C, gene E, prM, and NS5 of KFDV from the region of India were downloaded from the NCBI database (https://www.ncbi.nlm.nih.gov). **Table 4**. The sequence of gene C (n=29), gene E ((n=48), prM (n=34), and NS5 (n=33) were aligned and edited using MEGA-X software.

Homologous Recombination Detection

A Genetic Algorithm for Recombination Detection (GARD) was used to determine the homologous recombination in the dataset of genes- C, E, prM, and NS5 genes of the KFDV. However, none of the gene's found any recombination evidence. Therefore, the dataset containing the FASTA sequence of the selected genes was taken directly to analyze the evolutionary rate.

Evolutionary Rate Analysis

Evolutionary rate analysis was employed to assess the significant changes in evolutionary rate over the period. In this study, the complete gene sequence of genes- C, E, prM, and NS5 of the KFD virus were used to evaluate the time of Most Recent Common Ancestor (tMRCA) and substitution rate (s/s/y) using the Bayesian-based coalescent method. As a preliminary criterion, the DNA substitution model selection was done using the jModelTest2 tool. Based on the AIC (Akaike Information Criterion) score generated in the jModelTest2 tool, the best fit substitution models- GTR was observed as the best fit for gene C, GTR+G for both gene E and NS5, and HKY for the prM gene. The required parameters and priors, MCMC chain length, clock rate was specified using the BEAUti tool to generate XML format. The tree files (.trees) and logarithmic (.log) were obtained from the XML file after performing the execution of the BEAST. The estimated evolutionary rate and tMRCA are listed in Table 5.

To each dataset of the genes, the MCMC chain cycle 1-10 million generations were run to achieve 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 file generated from BEAST and the Tracer tool was used to visualize the achieved statistical scores.

The evolutionary rates for the genes- C,E, prM, and NS5 from the KFD virus are 5.6 x10^{-4}, 8.9 x10^{-4}, 5.6 x10^{-4}, and 2.6 x10^{-4} respectively. The recorded tMRCA ages as 74.07 years with 95% HPD (lowest 23.81, highest 144.25), 34.52 years with 95% HPD (lowest 23.25, highest 45.65), 41.92 years with

95% HDP (lowest 18.90, highest 74.40), and 86.31 years with 95% HPD (lowest 35.14, highest 158.61). **Table 5.** Analyzing the phylogenetic tree of gene -C, E, prM, and NS5 of KFDV reveals the tMRCA ages as 1942, 1982, 1975, and 1931, respectively. Fig. 8

The evolutionary rate of gene NS5 (86) years (1931-2017) was high compared to the genes- gene C (75), gene E (35), and prM (42), respectively. This indicates the gene NS5 was found to be the first synthesized virulent gene.

Positive Selection Analysis

Selection pressure was analyzed from the Datamonkey server using the FEL algorithm. The results discovered that one site in genes- C, E, NS5 has undergone positive selection (ω < 1) i.e., sites34, 123, 826, respectively with overall *dN/dS*(ω) rate ratio of 0.371, 0.273, and 0.0640. Whereas, there was no evidence of a positively selected site found in the prM gene of the KFD virus. **Table 6.**

DISCUSSION

The KFD virus that belongs to *Flaviviridae*has been one of the high zoonosis infectious diseases. That was first reported in the Kyasanuru forest at Shimoga among the monkeys (Kasabi, 2011). A tick called *Haemaphysalisspinigera*was the first observed major transmission vector when the infection was transferred from monkey to human (Author n.d.). Although there are 16 ticks were found to have the capability of carrying the virulent gene and the ability to transmit (Yadav, 2020). The KFD virus contains a positive–sense and single-strand RNA genome that is almost 11kb in length, which encodes a single polyprotein amino acid.

In the present study, the three structural genes and one non-structural gene, i.e., NS5, have been used to examine the codon usage bias and evolutionary rate among them. To determine mutational pressure and natural selection with a site-specific region among the selected genes, the integration of two different methodologies was employed. Based on the literature survey, we found that there are specific codon usage bias analyses have been employed on the KFDV virus. Although, districtwide analysis of evolutionary rate was observed in a previous study (Yadav, 2020; Mehla, 2009) by involving the PCR and NGS. The result of the previous study emits lower genetic divergence among general *flaviviruses* (Yadav, 2020). Whereas, in the present study, we have used the overall available gene sequences from the NCBI database around India.

The codon usage bias analytics was employed to determine the pattern of codon bias. The evaluated nucleotide composition of each gene revealed that nucleotide G has been used most frequently in all genes of the KFD virus. The frequent usage of nucleotide G could be the genomic feature of the KFD virus. Whereas, the dinucleotides AA-CC, AC-CA-TG, CA-TG, and CA-CT-TG were observed to be the most abundant dinucleotides in genes- C, E, prM, and NS5, respectively. Indicating that each gene has different abundant dinucleotides, and CA seems to be common in genes- E, prM, and NS5. Also, dinucleotides AC-GT-TA, CG-TA, CG-TA, CG were reported as under-represented in genes- C, E, prM, and NS5, respectively. The relative synonymous codons (RSCU)that ended with under-represented dinucleotides are do not seem to be favourably preferred while shaping the codon usage. Thus, observing the variations or substitutions of odd preferring dinucleotides illustrates that dinucleotide and mononucleotide are contributing factors in shaping the codon usage pattern of the KFD virus.

The magnitude between mutational pressure and natural selection was estimated for each gene- C, E, prM, and NS5 of the KFD virus. A result of PR2 signifies that there was a noticeable and high bias among the prM and C genes compared to E and NS5. Since the NS5 and E genes were almost closer to the origin, specifying that the base pairs A=T and G=C nucleotide pattern. According to a previous study about codon usage bias (Chen *et al.*, 2014), genes that situate on the origin of the PR2 plot defines that no bias in the sequence. So, considering this feature of the PR2 plot, we can conclude that the genes E and NS5 has very slight bias compared to gene C and prM.

The evaluated eNC values varied at the average of 49.17 (SD $\pm$2.64), 53.70 (SD $\pm$0.33), 52.85 (SD $\pm$1.53), and 52.43 (SD $\pm$0.30) among the genes- C, E, prM, and NS5, respectively. Whereas the previous research states that, lower the eNC score defines the high gene expression. Since, the eNC scores that range between 35-55 indicates a moderate level of bias among the sequences (Tao and Yao, 2020; Wang and Zhang, 2020). It is revealed that the eNC value of each gene- C, E, prM, and NS5-was found to be moderately biased.

To confirm the driving forces of the bias and the extent of evolutionary forces, the analysis of neutrality and plot was employed. A less than 0.5 regression coefficient score is an indication of natural selection whereas, greater than 0.5 is mutational pressure. The methodology considers the regression coefficient as mutational-natural coefficient equilibrium (Deb *et al.*, 2020). The obtained results spec-

ify the mutational pressure taking the major role among the genes- C and E, whereas genes- prM and NS5 genes experiencing the natural selection. The mutations in the viral genome arise from the errors in the replication process. Also, a genetic variation within the population can be considered as a rate of mutation, and a high mutational rate is proportional to the population size. This mutation evolution can lead to a greater degree of genetic diversity (Deb *et al.*, 2020).

The influencing physical properties of amino acid, such as hydrophobicity and aromaticity, were estimated, and correlation was evaluated for each gene of the KFD virus. Previous research suggests that hydrophobicity and aromaticity are indicative of the consequence of natural selection and translational (Singh and Tyagi, 2017). Therefore, in the present study, hydrophobicity is taking a positive role among genes- C and prM, but the influence of aromaticity was only observed in gene E. Thus, hydrophobicity and aromaticity are taking a positive role while shaping the codon usage pattern among the genes of the KFD virus.

The multivariate statistical method, correspondence analysis, was used to determine the synonymous codon usage among selected genes of the KFD virus. The obtained results indicate the first principal axis among all the genes has a greater contribution compared to axis 2. So, the correspondence score directs that though the contribution of the first axis high and significant variation, also the 2^{nd} principal axis has a slight appreciable influence in the synonymous codon usage of each gene of the KFD virus.

Once the codon usage pattern is observed among each gene- C, E, prM, and NS, the evolutionary analysis (tMRCA) of each gene was estimated using the BEAST software. Based on previous findings of KFDV genes (Yadav, 2020), the best fit model for each gene was determined by the jModelTest tool based on the AIC scores of each gene. We observed that the substitutional model GTR has the lowest AIC scores for the gene C, GTR+G for the gene E and NS5, whereas model HKY for the prM gene. And best fit models were selected with clock type and prior tree. A strict clock and random local clock were used with coalescent: Bayesian sky grid and coalescent: Extended Bayesian skyline plot prior trees. The sky grid model is flexible that allows multiple loci, and the changes occur at the prespecified points in real-time. Also, capture the complex population dynamics. Among all the nonparametric coalescent-based models, the sky grid has a well-suited starting point (Gill *et al.*, 2016). Whereas, the skyline method assists to extract information about past population genetics in the nonparametric method. The coalescent skyline plot splits the time between the root of the tree (tMRCA) and present into section, and evaluates a diverse, effective population for each section (Skyline plots n.d.) (Heller *et al.*, 2013).

The estimated evolutionary rate of the gene NS5 was noticed as the first evolved virulent gene among the genes- C, E, and prM, which was almost 86 years [1931-2017]. While recognization of genetic loci that undergoes adaptation is the major concept of evolutionary biology. Several statistical algorithms have been developed to quantify the selection pressure (positive selection pressures) that acts on protein-coding sites. Among these statistical approaches, *dN/dS*is the most widely used (Kryazhimskiy and Plotkin, 2008). This method enumerates the selection pressure by comparing the rate of the substitution at a silent site (*dS*) against the rate of substitution at non-silent sites (*dN*). The strong positive selection is anticipated as $dN/dS < 1$ or $\omega < 1$ (Kryazhimskiy and Plotkin, 2008). Considering these features of positive selection, we observed the single positive selection site among genes- C, E, and NS5.

CONCLUSIONS

The integrated analysis of codon usage bias and evolutionary rate analysis has proven that the genes- C, E, prM, and NS5 possess a bias where mutational and natural selection both played a major role. Whereas the eNC value indicated the moderate bias, and neutrality signifies the mutational pressure in C and E gene. The correspondence analysis concludes the major contribution of axis-1 in the synonymous codon usage pattern. The observed evolutionary and phylogenetic analysis testified that gene NS5 was the first virulent synthesized gene. Considering all these analyzed data, this study will aid the future surveillance of KFDV and researchers to understand the evolutionary characteristics of the KFD virus.

ACKNOWLEDGEMENT

We would like to thank the Spatial Epidemiology lab, Indian Council for Agriculture Research (ICAR) — National Institute of Veterinary Epidemiology and Disease Informatics, Department of Veterinary Public Health and Epidemiology, VeterinaryCollege and Outreach project on Zoonotic diseases, ICAR and JSS AHER, Mysuru, Karnataka, India for providing necessary support to carry out this research work.

Ethical statement

The authors declare that the ethical statement is not applicable as we have not collected any animal sample for the study.

Authors contribution

Mallikarjun S Beelagi and Uma Bharathi Indrabalan: Conceptualization, Data curation, Writing – original draft. Sharanagouda S Patiland Chandan Shivamallu: Writing - review & editing, Supervision.

Kuralayanapalya Puttahonnappa Suresh: Methodology & Formal analysis. Shiva Prasad Kollur, Ashwini Prasad and Chandrashekar Srinivasa: Writing – review & editing.

Conflict of Interest

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

Funding Support

The authors declare that they have no funding support for this study.

REFERENCES

Behura, S. K., Severson, D. W. 2013. Codon usage bias: causative factors, quantification methods and genome-wide patterns: with emphasis on insect genomes. *Biological Reviews*, 88(1):49–61.

Chen, H., Sun, S., Norenburg, J. L., Sundberg, P. 2014. Mutation and Selection Cause Codon Usage and Bias in Mitochondrial Genomes of Ribbon Worms (Nemertea). *PLoS ONE*, 9(1):e85631–e85631.

Comeron, J. M., Aguadé, M. 1998. An Evaluation of Measures of Synonymous Codon Usage Bias. *Journal of Molecular Evolution*, 47(3):268–274.

Deb, B., Uddin, A., Chakraborty, S. 2020. Codon usage pattern and its influencing factors in different genomes of hepadnaviruses. *Archives of Virology*, 165(3):557–570.

Dodd, K. A. 2011. Ancient Ancestry of KFDV and AHFV Revealed by Complete Genome Analyses of Viruses Isolated from Ticks and Mammalian Hosts. *PLoS Neglected Tropical Diseases*, 5(10):1–8.

Donnelly, P., Tavare, S. 1995. Coalescents and Genealogical Structure Under Neutrality. *Annual Review of Genetics*, 29(1):401–421.

Drummond, A. J., Rambaut, A. 2007. BEAST: Bayesian evolutionary analysis by sampling trees. *BMC Evolutionary Biology*, 7(1):214–214.

Fabra, U., Pompeu, R. T., Fargas 2010. Correspondence Analysis of Raw Data. *Research Gate*, 91:958–63.

Fajardo, T. 2020. The Flavivirus Polymerase NS5 Regulates Translation of Viral Genomic RNA. *Nucleic acids research*, 48(9):5081–93.

Gill, M. S., Lemey, P., Bennett, S. N., Biek, R., Suchard, M. A. 2016. Understanding Past Population Dynamics: Bayesian Coalescent-Based Modeling with Covariates. *Systematic Biology*, 65(6):1041–1056.

Gupta, N., Wilson, W., Neumayr, A., Saravu, K. 2020. Kyasanur forest disease: a state-of-the-art review. *QJM: An International Journal of Medicine*, 310.

Heller, R., Chikhi, L., Siegismund, H. R. 2013. The Confounding Effect of Population Structure on Bayesian Skyline Plot Inferences of Demographic History. *PLoS ONE*, 8(5):e62992–e62992.

Kasabi, G. S. 2011. Kyasanur Forest Disease. *Emerging infectious diseases*, 19(2):278–81.

Khandia, R., Singhal, S., Kumar, U., Ansari, A., Tiwari, R., Dhama, K., Das, J., Munjal, A., Singh, R. K. 2019. Analysis of Nipah Virus Codon Usage and Adaptation to Hosts. *Frontiers in Microbiology*, 10:1–18.

Kryazhimskiy, S., Plotkin, J. B. 2008. The Population Genetics of dN/dS. *PLoS Genetics*, 4(12):e1000304–e1000304.

Mehla, R. 2009. Recent Ancestry of Kyasanur Forest Disease Virus. *NCBI*, 15(9):1431–1438.

Pan, S., Mou, C., Wu, H., Chen, Z. 2020. Phylogenetic and codon usage analysis of atypical porcine pestivirus (APPV). *Virulence*, 11(1):916–926.

Rahman, S., Ur 2017. Analysis of Codon Usage Bias of Crimean-Congo Hemorrhagic Fever Virus and Its Adaptation to Hosts. *Genetics and Evolution*, 58:1–16.

Singh, N. K., Tyagi, A. 2017. A detailed analysis of codon usage patterns and influencing factors in Zika virus. *Archives of Virology*, 162(7):1963–1973.

Suresh, K. P. 2020. Evolutionary Analysis and Detection of Positive Selection of Hemagglutinin and Neuraminidase Genes of H5N1 Avian Influenza From Chicken, Duck and Goose Across Asia. " *Explor Anim Med Res*, 10(2):169–78.

Takaaki, K., Tatsuo, S. 2018. Analysis of factors affecting codon usage bias in human papillomavirus. *Journal of Bioinformatics and Sequence Analysis*, 9(1):1–9.

Taming the BEAST 2018. A community teaching material resource for BEAST 2. *Systematic Biology*, 67(1):170–174.

Tao, J., Yao, H. 2020. Comprehensive analysis of the codon usage patterns of polyprotein of Zika virus. *Progress in Biophysics and Molecular Biology*, 150:43–49.

Tao, P. 2009. Analysis of Synonymous Codon Usage in Classical Swine Fever Virus. *Virus Genes*,

38(1):104–116.

Venugopal, K., Gritsun, T., Lashkevich, V. A., Gould, E. A. 1994. Analysis of the structural protein gene sequence shows Kyasanur Forest disease virus as a distinct member in the tick-borne encephalitis virus serocomplex. *Journal of General Virology*, 75(1):227–232.

Wang, Y.-L., Zhang, Y.-Y. 2020. cg04448376, cg24387542, cg08548498, and cg14621323 as a Novel Signature to Predict Prognosis in Kidney Renal Papillary Cell Carcinoma.

Wu, H., Bao, Z., Mou, C., Chen, Z., Zhao, J. 2020. Comprehensive Analysis of Codon Usage on Porcine Astrovirus. *Viruses*, 12(9):991–991.

Xu, X. 2017. Comparative Characterization Analysis of Synonymous Codon Usage Bias in Classical Swine Fever Virus. *Microbial Pathogenesis*, 107:368–71.

Yadav, P. D. 2020. Phylogeography of Kyasanur Forest Disease Virus in India (1957-2017) Reveals Evolution and Spread in the Western Ghats Region. *Scientific Reports*, 10(1):1–12.

Yao, X., Fan, Q., Yao, B., Lu, P., Rahman, S. U., Chen, D., Tao, S. 2020. Codon Usage Bias Analysis of Bluetongue Virus Causing Livestock Infection. *Frontiers in Microbiology*, 11:1–12.

Zhou, J., Teo, Y.-Y. 2016. Estimating time to the most recent common ancestor (TMRCA): comparison and application of eight methods.

Zhou, Z. 2016. Codon Usage Is an Important Determinant of Gene Expression Levels Largely through Its Effects on Transcription. *Proceedings of the National Academy of Sciences of the United states of America*, 113:6117–6125.