Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID-19. The COVID-19 resource centre is hosted on Elsevier Connect, the company's public news and information website.

Gene 778 (2021) 145470

Contents lists available at ScienceDirect

Gene

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

Research paper

Mutational analysis and assessment of its impact on proteins of SARS-CoV-2 genomes from India

Rezwanuzzaman Laskar, Safdar Ali *

Clinical and Applied Genomics (CAG) Laboratory, Department of Biological Sciences, Aliah University, Kolkata, India

ARTICLE INFO

Keywords:
SARS-CoV-2
Mutation
Protein
Asymptomatic

ABSTRACT

Mutational status of SARS-CoV-2 genomes from India along with their impact on proteins was ascertained through multiple tools including MEGA, Genome Detective, SIFT, PROVEAN and ws-SNPs&GO. Excluding gaps and ambiguous sequences, 493 variable sites (152 parsimony informative and 341 singleton) were observed. NSP3 had the highest incidence of 101 sites followed by S protein (74), NSP12b (43) and ORF3a (31). Average mutations per sample for males and females was 2.56 and 2.88 respectively. Non-uniform geographical distribution of mutations suggests that sequences in some regions are mutating faster than others. There were 281 mutations (198 Neutral and 83 Disease) affecting amino acid sequence. NSP13 has a maximum of 14 Disease variants followed by S protein and ORF3a with 13 each. Disease mutations in genomes from asymptomatic people was mere 11% but those from deceased patients was at 38% indicating contribution of these mutations to the pathophysiology of the SARS-CoV-2.

1. Introduction

The ongoing COVID-19 global pandemic began from Wuhan, China and has devastated millions of lives, economies and even nations as a whole. The first reported case was in December 2019 and as of 1st September 2020, there have been 2,56,21,967 reported cases and 8,54,235 deaths worldwide (www.worldometers.info/coronavirus/). Of these, 36,21,245 cases and 64,469 deaths have happened in India making it one of the most affected countries in the world (www.mygov.in/covid-19).

The causative agent identified for COVID-19 is Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2) which belongs to family *Coronaviridae* characterized by single strand positive sense RNA genome. Though this is novel virus but the outbreak is not the first one from members of *Coronaviridae*. Previously, Severe Acute Respiratory Syndrome (SARS) coronavirus in 2003 and Middle East Respiratory Syndrome (MERS) coronavirus in 2012 created a scare as they had a relatively higher mortality rate. However, SARS-CoV-2 is by far the most contagious one (Chan et al., 2020; Lee et al., 2003; Peiris et al., 2004; Zaki et al., 2012).

The higher incidence of viral infections would imply a faster evolution process for SARS-CoV-2 (Rahman et al., 2020). This is so because more the virus replicates, higher are the chances of it accumulating mutations with the possibility of it leading to altered dynamics of its virulence, pathogenesis and interactions with host. The changes may not be necessarily favoring the virus; however, the unpredictability demands caution.

The SARS-CoV-2 genome encodes for 16 non-structural proteins in addition to the replicase polyprotein, the spike (S) glycoprotein, envelope (E), membrane (M), nucleocapsid (N) and other accessory proteins (Ren et al., 2020). The impact of mutations in all the regions of the genome needs to be assessed to understand viral evolution.

With a definitive possibility of India becoming the most affected country by SARS-CoV-2 in near future and the demographic burden involved, it is pertinent to be analyze the accumulating variations in the genome accounting for possible changes in protein and their potential to alter the virus in any manner. On 6th June 2020 we retrieved 611 FASTA sequence congregations from India along their rational meta data from GISAID EpiCoV server to construct the phylo-geo-network and analyze the haplogroups along with their geographical distribution across

Abbreviations: SARS-CoV-2, Severe Acute Respiratory Syndrome Coronavirus 2; SARS, Severe acute respiratory syndrome; MERS, Middle East respiratory syndrome; S protein, Spike glycoprotein; E protein, Envelope; M protein, Membrane; N protein, Nucleocapsid; MSA, Multiple Sequence Alignment (MSA); MCL, Maximum Composite Likelihood; PI, Parsimony informative; SNP, Single nucleotide Polymorphism; MV, Multi-variable.

* Corresponding author at: Department of Biological Sciences, Aliah University, IIA/27, Newtown, Kolkata 700160, India.
E-mail address: ali@aliah.ac.in (S. Ali).

https://doi.org/10.1016/j.gene.2021.145470
Received 22 October 2020; Received in revised form 22 January 2021; Accepted 28 January 2021
Available online 4 February 2021

different states of India (Laskar and Ali, 2020). Herein we extend our study using the same congregation of sequences to analyze the nature and composition of the observed mutations and their impact on proteins of SARS-CoV-2.

2. Methodology

2.1. Sequence congregations collection and alignment

GISAID EpiCoV is an open access repository of genomic and epidemiologic information about novel corona viruses from across the world from wherein sequences were extracted and alignment performed as previously reported (Laskar and Ali, 2020). Therein, 611 FASTA sequence congregations along their rational meta data from GISAID EpiCoV (www.epicov.org) server was used.

In order to perform mutational profile analysis with clinical correlation, we selected 15 genomes of deceased patients from existing congregation. However, there were just two genomes for asymptomatic patients in the congregations. So, on 09.12.2020, we downloaded 775 FASTA sequences with patient status from the same server and selected 30 genomes from asymptomatic patients. As the data filter for genome extraction, we used hCoV-19 as a virus name, human as a host, India as a location and complete sequence with high coverage. Details of the asymptomatic samples are given in Supplementary file 1.

The sequence congregations of 611 genomes and 30 genomes of asymptomatic patients were aligned separately against the reference sequence (NC_045512.2) using the FFT-NS-fragment method which is a light-weight algorithm of MAFFT v7 web-server (https://mafft.cbrc.jp/alignment/software/closelyrelatedviralgenomes.html) and further studied using MEGA(v.10) (Katoh et al., 2019; Kumar et al., 2018).

2.2. Nucleotide analysis

MEGA(v.10) is a multithreaded tool for molecular and evolutionary analysis. Multiple Sequence Alignment (MSA) of the extracted sequences (Laskar and Ali, 2020) was initially visualized by this software then the variable sites are exported into spreadsheets with or without missing/ambiguous and gap sites along their respective positions (Kumar et al., 2018). Using this software, we estimate the MCL (Maximum Composite Likelihood) nucleotide substitution pattern and Tajima's Neutrality test to understand transition transversion bias and nucleotide diversity (Tajima, 1989; Tamura et al., 2004). PIRO IGLSF is a MATLAB-based simulation software, we used this for the identify the location of mutated nucleotide position on specific gene (Alam et al., 2019). The nucleotide similarity percentage was validated by NCBI BLAST (blast.ncbi.nlm.nih.gov) to investigate the sequence diversity.

2.3. Protein analysis

Coronavirus Typing Tool of Genome Detective (v.1.13) and COVID-19 Genome Annotator of Coronapp are webtools for analysis of protein and nucleotide mutation (Mercatelli et al., 2020; Vilsker et al., 2019). We used these tools for annotation, identification and classification of mutated protein followed by verification and validation of the positions with the mutated nucleotide sites by the output of MEGA(v.10). The tools refer to the variable sites as SNP (affecting the protein sequence); SNP-silent (not affecting the protein sequence) and SNP-stop (introducing of a stop codon).

2.4. Protein prediction Analysis:

SIFT, PROVEAN and ws-SNPs&GO are the prediction tools which report positive or negative impact of variants on protein phenotype. The assessments are focused upon scores using several algorithms. It is expected that a SIFT score of <0.05 is diseased ("affect protein function"), and that >0.05 is neutral ("tolerated"). This is stated that a PROVEAN score of <-2.5 is diseased ("deleterious"), and >-2.5 is neutral. ws-SNPs&GO 's PHD-SNP method is estimated to be >0.5 mutation in the probability of disease, and <0.5 is neutral (Capriotti et al., 2005; Choi et al., 2012; Sim et al., 2012)

Table 1
Tajima's Neutrality Test.

m^*	S^*	Ps^*	Θ^*	π^*	D^*
612	841	0.028124268	0.004021699	0.000412328	−2.69529519

*m = number of sequences, n = total number of sites, S = Number of segregating sites, ps = S/N, Θ = ps/a1, π = nucleotide diversity, D = Tajima test statistic.

3. Results and discussion

3.1. Composition and distribution of variable sites

The observed MSA length was 29903 bp wherein the variable sites could be extracted through two different options. If we included gaps and ambiguous sequences, a total of 841 variable sites were observed with a percentage coverage of 2.81%, where percentage coverage = [(No of variable site / MSA Length) * 100]. All the sites have been shown in Supplementary file 2. The Tajima's D statistic was also analysed (Table 1) and its negative value indicated the significance of these variable sites.

However, excluding the gaps and ambiguous sequences reduced this percentage coverage to 1.65% encompassing 493 variable sites which we have used for subsequent analyses reported in this study. This included 152 parsimony informative (PI) sites and 341 singleton sites (SNP: Single Nucleotide Polymorphism). The PI sites are those whose incidence was observed in multiple samples whereas singleton sites had a restricted single sample incidence. The distribution of these sites according to various substitutions, protein localizations and impact therein has been summarized in Fig. 1, Supplementary file 2. As evident therein, C → T (181 sites) forms the most prevalent mutation in both PI and singleton sites and G → T (95 sites) comes a distant second. The common aspect of two most prevalent mutations is "T" being the substituted nucleotide. Further, there were two multi-variable (MV) sites each in PI and singleton category wherein two separate mutations were observed at the same site in different samples. The details of observed MV sites have been summarized in Table 2.

The distribution of the variable sites across proteins of SARS-CoV-2 in a non-uniform manner is reflective of the differential contributions of proteins in evolution. As per our data, NSP3 had the maximum of 101 variable sites followed by S protein (74 sites), NSP12b (43 sites) and ORF3a (31 sites) (Fig. 1; Supplementary file 2). These four proteins account for over half of the total variable sites of the genome and may be considered as drivers of genomic evolution for SARS-CoV-2. The mutations of S protein have been the focus for multiple research groups owing to its plausible impact on viral entry to the host cell but the mutations elsewhere may be equally relevant as the viral genome is known to harbor only what's essential (Hassan et al., 2020; Mercatelli and Giorgi, 2020; Sanjuán and Domingo-Calap, 2016). We believe a holistic approach is required to understand the evolution as more often than not the selection advantage being offered by any mutation is a chance event and can be from any part of the genome.

In terms of the impact of these variable sites on amino acid sequence of the viral proteins we classified them into four categories. First, the sites located in the extragenic region and hence no influence on the coding proteins. There were 10 such variable sites localized to the UTR regions (2 in 3′UTR and 8 in 5′UTR). Secondly, SNP-silent included those variable sites wherein the nucleotide change was leaving the amino acid sequenced unaltered. A total of 186 such sites were distributed across the genome. Thirdly, the variable sites which were leading to the introduction of a stop codon were referred to as SNP-stop and there were

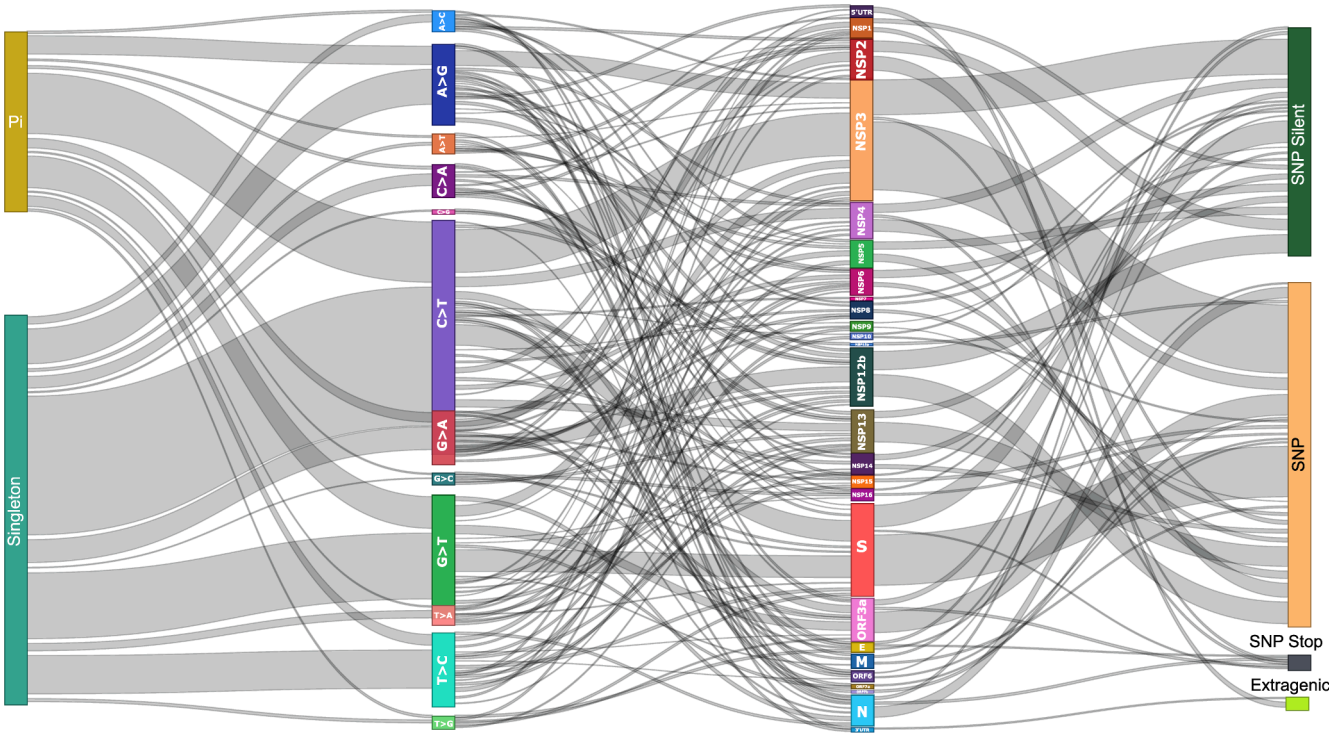

Fig. 1. Summary of variations observed in SARS-CoV-2 genomes from India. The nature of variations (Singleton/PI); type of mutation, genome localization and impact on protein (SNP, SNP-Silent, SNP-Stop, Extragenic) has been represented along with their interlinking. The width of the connecting lines represent number; broader the line more the number of that parameter.

Table 2
Localization and mutations observed at the multi-variable (MV) sites.

S No	Nucleotide Position	Category	Nucleotide Variant 1	Amino Acid Variant 1	Nucleotide Variant 2	Amino Acid Variant 2
1	4893	Pi	C > A	T725K	C > T	T725I
2	5821	SNP	A > G	L1034L	A > T	L1034F
3	23,282	SNP	G > C	D574H	G > T	D574Y
4	23,593	Pi	G > T	Q677H	G > C	Q677H

8 such sites in our study. Lastly, the variable sites which were affecting the protein sequence are referred as SNP in the study and there were 281 such sites (Supplementary file 3). The prevalence and distribution of these sites has been summarized in Fig. 1 and results of the prediction of their impact on protein has been discussed later.

3.2. Constitution of genome, variable sites and the substitutions

In order to understand the underlying dynamics of substitutions, we performed the maximum composite likelihood estimate of nucleotide substitution as shown in Table 3. The reference and substituted nucleotide have been shown in rows and columns respectively. The values therein represent the probability of substitution (r) from one base to another. For simplicity, the sum of r values is made equal to 100. The nucleotide frequencies of the MSA are 29.87% (A), 32.14% (T/U), 18.37% (C), and 19.63% (G). The transition/transversion rate ratios are k1 = 2.195 (purines) and k2 = 7.799 (pyrimidines). The overall transition/transversion bias is R = 2.356, where R = [A*G*k1 + T*C*k2]/[(A + G)*(T + C)]. This substantiates the prevalence of certain mutations (C → T and G → T) over others.

We thereon looked at these variations in combination with their prevalence across samples. The most prevalent nucleotide at the variable sites in reference sequence was C (209) followed by G (137) whereas T was by far the predominantly substituted nucleotide (293, 60%). Also, the other three nucleotides had an almost equal representation in substitutions (A-68, G-68, T-64). This biased prevalence was not restricted to the alignment but was also getting translated to population incidence. There was a total of 723 mutations with C as reference nucleotide and 1032 mutations with T as substituted nucleotide across 611 studied genomes. The composition of 493 variable sites, their substitutions and prevalence across samples has been summarized in Fig. 2 and Supplementary file 2. Evidently, any particular mutation may be incident across multiple samples and a single sample can harbor

Table 3
Maximum Composite Likelihood Estimate of Nucleotide Substitution.

	Substituted				
	Nt	A	T	C	G
Reference	**A**	–	*4.57*	*2.61*	**6.13**
	T	*4.25*	–	**20.39**	*2.79*
	C	*4.25*	**35.68**	–	*2.79*
	G	**9.33**	*4.57*	*2.61*	–

* Rates of transitional substitutions are **bold** and transversional substitutions are *italicized*

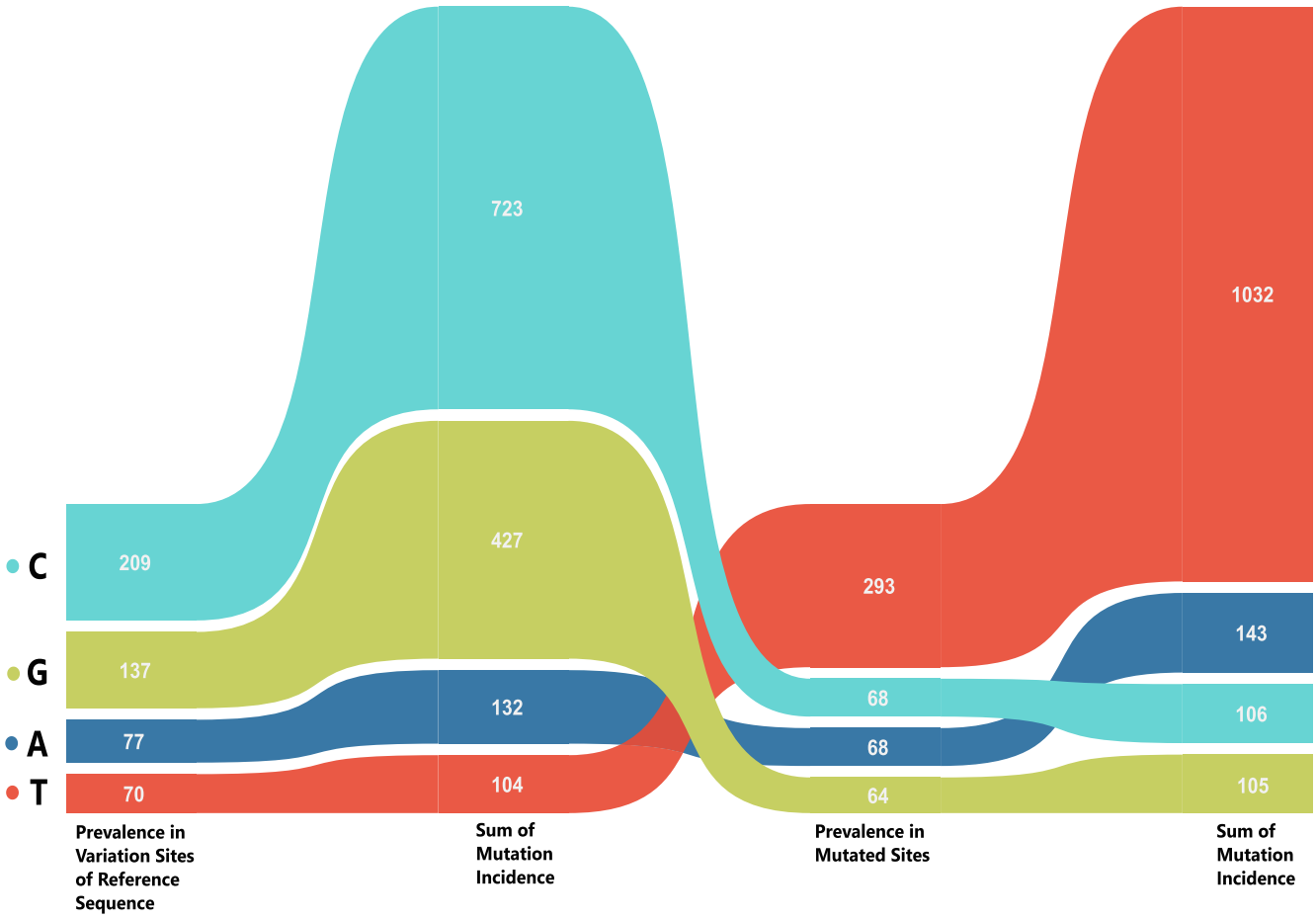

Fig. 2. Prevalence and composition of different nucleotides across reference and substituted positions in SARS-CoV-2 genomes from India.

multiple mutations. A cumulative number for the same has been referred to as "Sum of Mutation Incidence" herein and thereafter in this study.

3.3. Age and gender wise distribution of samples and mutations therein

We subsequently analyzed the patient's dataset with reference to age and gender for the incidence of mutations. However, since patients' data wasn't cumulatively available, the data for this aspect isn't exhaustive but representative for 255 samples (104 females and 151 males). The patients whose genomes were used in the study and age was known were classified into seven categories from infancy to over 75 years. The maximum number of patients for both males and females belonged to mature adulthood category of 50 to 74 years with 57 and 40 samples respectively (Fig. 3, Supplementary file 4). This adheres to the fact that the older population is at a greater risk for not clearing the infection owing to a possibly weaker immune system and other physiological conditions.

The simple question of whether or not age and gender are associated with accumulation of genome variations has a not so simple answer. The overall average number of mutations per sample was 2.69 and the corresponding values for males and females separately was 2.56 and 2.88 respectively. Thus, women were contributing more to the mutational accumulation as compared to males. The individual mutational load for different age groups in males and females has been represented in Fig. 4. Evidently, women are contributing more to the mutational load except for three age groups; 5–9 years, 10–14 years and 20–34 years. The highest difference on the basis of gender is for 15–19 years (2.67) but since there was just one female sample in that age group, it can't be emphasized much in isolation but the overall pattern does seem relevant. This is more so because, in terms of incidence, males are almost 1.5 times of the females but in terms of variations, fewer females are contributing more to the mutational load. Possibly, the virus is behaving differently depending on gender.

3.4. Geographical distribution and accumulation of variable sites

The mutational distribution across different states of India was subsequently ascertained. Generally speaking, more the virus replicates more should be the accumulated variations. The fact that the samples used in the study aren't uniformly distributed across states provides for an intriguing template for analysis. The number of samples and the mutations therein for different states has been summarized in Fig. 5 and Supplementary file 4. Evidently, Gujarat with highest number of 199 samples had maximum of 694 mutations. However, the correlation is neither uniform nor universal. For instance, Maharashtra with 94 samples had 203 variations whereas Telangana with 97 samples had 154 variations. A plausible explanation for this can be one sequence in Telangana (Genome ID 458080) to be identical to reference sequence as reported (Laskar and Ali, 2020). That means, Maharashtra samples had more variations from reference sequence to begin with. But if we look at Odisha and Tamil Nadu with 30 and 31 samples accounting for 109 and 40 variations respectively, it's evident that sequences in some regions are mutating faster than others. Another contrasting example pair is Delhi (63 samples, 54 mutations) and West Bengal (40 samples, 70 mutations). The exact mutational route can be revealed only with tracking the route of samples and spreading of infection which has not been feasible for present dataset due to paucity of information. However, we can surely say that some sequences are mutating more than the

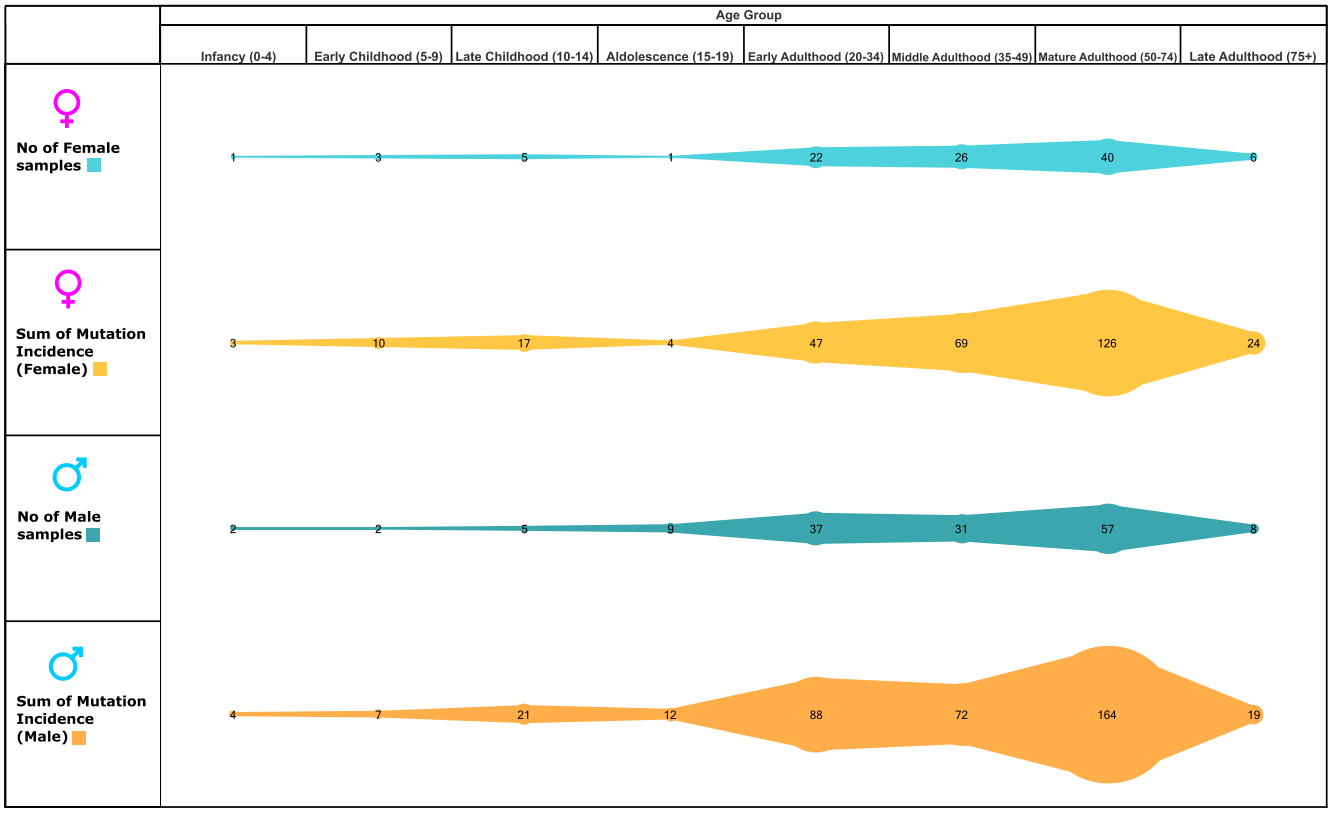

Fig. 3. Age and gender wise distribution of mutations in SARS-CoV-2 genomes from India. Number of male/female samples and sum of mutations incidence therein according to age group.

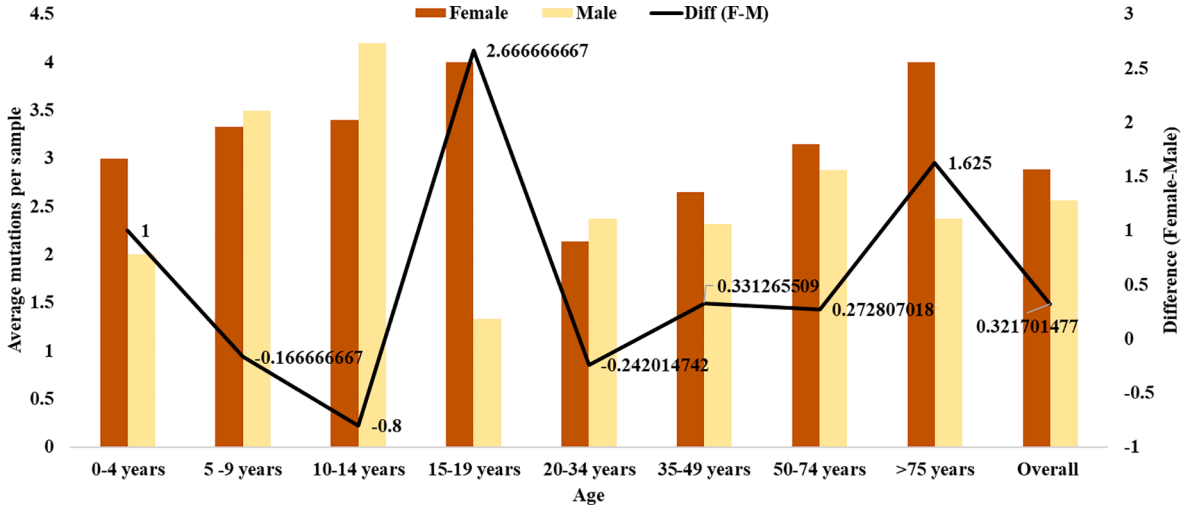

Fig. 4. Average number of mutations per sample in different age groups of males/females and the differences therein.

others but whether the geographical location is playing a role needs to be ascertained.

3.5. Impact of variables sites on viral proteins

A total of 281 SNPs which were present which were altering the amino acid sequence. Their details and positions have been summarized in Table 4 and Supplementary file 5. We also ascertained the prevalence of these variants across samples. The most incident variant Q57H localized to ORF3a was present in 127 samples followed by A994D in NSP3 present in 29 samples. Amongst the silent SNPs, Y71Y in M protein was present in 117 samples followed by D294D in S protein with 69 incidences. The overall data for variants present in 10 genomes or more has been summarized in Fig. 6a. Conversely, we also assessed the accumulation of variations in a given genome as summarized in Fig. 6b. Interestingly, one sample (Genome ID 461495) had highest incidence of

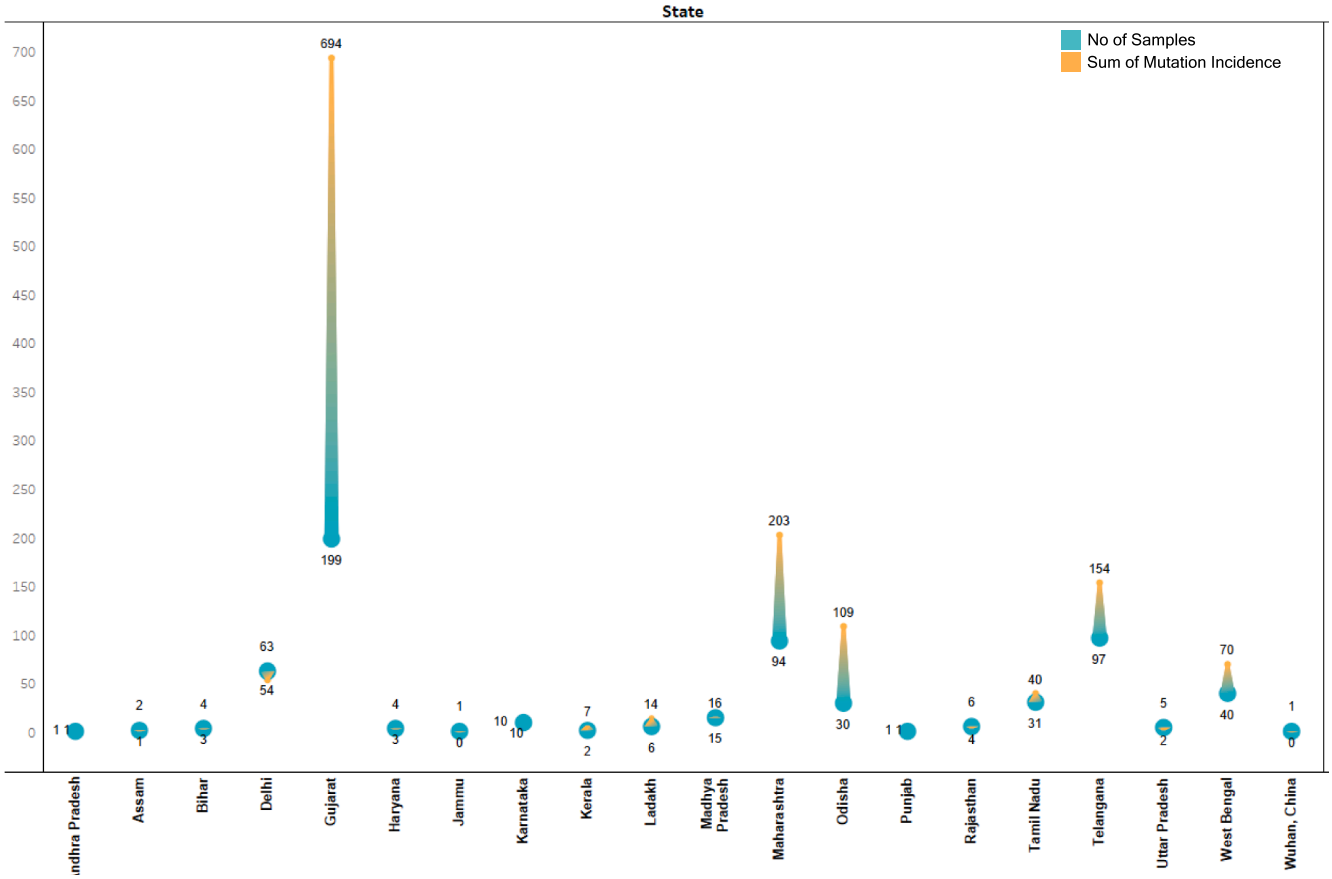

Fig. 5. Number of samples and corresponding Sum of Mutation Incidence of SARS-CoV-2 across different states of India.

Table 4
Distribution of SNPs across different proteins.

S No	Gene/ Protein	No of Variable Sites	Sum of Mutations Incidence	Mutations affecting Protein (N = N2 + N3; D = D2 + D3)	Neutral by Three (N3)	Disease by One/ Neutral by Two (N2)	Disease by Two/ Neutral by One (D2)	Disease by Three (D3)
1	**5′UTR**	**8**	**16**	–				
2	**ORF1a**	**231**	**578**	127				
	NSP1	14	19	7 (N = 6; D = 1)	H81Y, G137C, D147E, V169A	R24C; V38F	R124C	
	NSP2	29	46	15 (N = 14; D = 1)	Y196H, L204F, K338R, I616F, Y621S, E633D, V682L, T708I, I739V, P765S	G192D; P626T; I671T; V710F	T592I	
	NSP3	101	284	63 (N = 54; D = 9)	A872T, T882I, Y925C, E940D, P971S, G989V, S1029I, P1054L, M1083I, H1141Y, A1268T, S1534I, T1543K, T1567I, M1588I, V1629A, S1733G, T1761I, G1861S, T1822I, T1854A, T1854I, N1871T, K1973R, S2103F, K2029E, G2035E, P2144S, L2146F, A2249V, V2372I, T2274I, T2300I, L2323V, P2480L, H2520Y, A2593V, S2625F, S2661F	D930G, D1036G, A1298V, A1649S S1515F, A1812D S1978P, P2046L P2079L, S2015K S2015R, G2118C S2303Y, V2613F V2629A	A1306V P1472S D1625V L2146P S2242P	G1069E Y1675C C2691S A2732D
	NSP4	26	109	14 (N = 11; D = 3)	W2769L, M2796I, H2831Y, A2994V, F3031Y, D3042N A3143V, L3161I	T2777I T3058I Y3160H	L2781P F3071Y	T3223I
	NSP5	20	41	9 (N = 4; D = 5)	T3453A, Q3390R, P3395S	N3405L	S3386P	L3338F A3379V Y3472H S3517F
	NSP6	17	30	7 (N = 6; D = 1)	Q3826H Q3826R	I3731T, P3613L, D3681N, I3835T,	I3731F	
	NSP7	2	5	0				
	NSP8	11	26	6 (N = 3; D = 3)	K4081R	E3962K, K4069T	M4032S, L4033F	R3993C
	NSP9	6	11	3 (N = 2; D = 1)		V4181I, V4242I		T4249I
	NSP10	4	4	2 (N = 0; D = 2)				A4271V A4273V
	NSP12a	1	3	1 (N = 1; D = 0)	S4398L			
3	**ORF1ab**	**103**	**182**	60				
	NSP12b	43	84	22 (N = 12; D = 10)	A4487V, I4593L, S4621G, A4645G, T4685I, A4798V S5305L	K4451N, R4565H M4588I, E4670D, L4721I	K4483N, D4532G, A4577V,D4676Y, S4710T, M5148I, V5272I	D5076G G5200E T4801I
	NSP13	30	49	21 (N = 8; D = 14)	P5377S, S5490A, K5669R, M5798V, V5894L I5899V, T5923I	A5770V	K5364R, E5492Q, I5617V, A5620S, P5624L, V5820L, N5827K, A5833G. M5900I	W5830C Y5865C P5726S
	NSP14	15	20	7 (N = 7; D = 0)	A5926S, P5971L, M5974I, K6274N	L5952I, T5930A S6180I		
	NSP15	8	17	5 (N = 3; D = 2)	V6518L, A6533T	G6581D		D6491Y K6486T
	NSP16	7	12	5 (N = 4; D = 1)	P6805S, L6909F, A6914S	K6958R		G6837C
4	**S**	**74**	**227**	49 (N = 36; D = 13)	L54F, N148Y, E156D, A243S, S255F, G261S, Q271R, T299I, T323I, E471Q, A520S, T572I, E583D, T602I, V622I, Q677H, A706S, T761S, G769V, T827I, A831S, A879S, T1027I, H1101Y, V1104L, G1124V, K1181R, K1191N, G1251V, Q1201K	I434K, S494P, D574Y, A892V H1083Q, P1263L	T723I, F797C, L828P, T941K, V1068F, D1153Y, C1243F	G857C A930T A930V S1021F I1179N C1250F
5	**ORF3a**	**31**	**178**	22 (N = 9; D = 13)	V13L, G18V, S74A, V77F, T175I	L41F, S74F, S171L, T190I	I62T, L83F, T176I	I35T, L46F, L53F, Q57H, C81F, L85F, L86W, G172C, G196V, G251V
6	**E**	**5**	**8**	1 (N = 0; D = 1)				V29S
7	**M**	**10**	**136**		D3G, A68S, H125Y	A69S, V70F		

(*continued on next page*)

Table 4 (*continued*)

S No	Gene/ Protein	No of Variable Sites	Sum of Mutations Incidence	Mutations affecting Protein (N = N2 + N3; D = D2 + D3)	Neutral by Three (N3)	Disease by One/ Neutral by Two (N2)	Disease by Two/ Neutral by One (D2)	Disease by Three (D3)
				5 (N = 5; D = 0)				
8	**ORF6**	**6**	**17**	2 (N = 1; D = 1)	I60V		D61L	
9	**ORF7a**	**2**	**2**	2 (N = 0; D = 2)			P45L	G38V
10	**ORF7b**	**1**	**1**	**0**				
11	**ORF8**	**–**	**–**	**–**				
12	**N**	**20**	**38**	13 (N = 12; D = 1)	S37L, L139F, A152S	P6T, P13L, G18V, G30V, S33I, D63N, D144Y, A156S, Q160P		R92S
13	**ORF10**	**–**	**–**	**–**				
14	**3′UTR**	**2**	**3**	**0**				
				281				

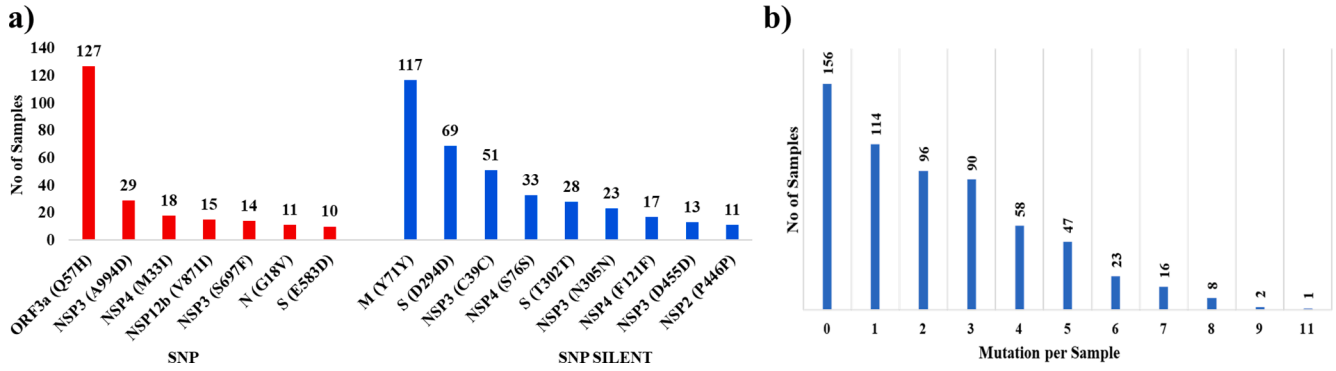

Fig. 6. Prevalence of variant sites across studied genomes. a) Most prevalent SNP and SNP_Silent across studied samples. Variants which had incidence across more than ten genomes haven been represented. b) Samples with accumulated variations in genomes.

11 mutations while 114 samples harbored just a single mutation. There were 156 samples with no mutations and 340 with more than one mutation. To account for these, the Sum of Mutation Incidence has been used in this study as explained above.

The impact of mutations on proteins was predicted through three different tools; SIFT, PROVEAN, ws-SNPs&GO; which classified the mutations as "Neutral/Tolerated" or "Disease/Affect Protein Function/ Deleterious". For the sake of simplicity, we have referred the results from all sites as Neutral and Disease. Though the prediction outcomes of the three tools were not in sync for all sites but since the classification of outcomes were on similar lines, the results can be represented in a binary manner with four categories. First two categories represent wherein the three tools have the same prediction; either all predicting a site to be "Neutral" or "Disease". The other two categories represent deviation between prediction outcomes. They are "Disease by One, Neutral by Two" and "Disease by Two and Neutral by One". For comparison between variants, any mutation predicted as Disease by Two or Three tools are considered as Disease and mutations predicted as Neutral by Two or Three tools are considered as Neutral.

The distribution of Disease and Neutral variants across the different genes of SARS-CoV-2 has been shown in Table 4 and Supplementary file 5. These could be analyzed in three aspects. First, in terms of overall incidence. The maximum variants affecting protein sequence were present in NSP3 (63) followed by Spike (S) protein with 49 variants. Secondly, if we focus only on variants with predicted outcome as "Disease" then NSP13 has a maximum of 14 such variants followed by S protein and ORF3a with 13 variants each. Thirdly, we looked at those proteins which had more Disease variants as compared to Neutral. There were five such proteins namely: NSP5, NSP10, NSP13, ORF3a, E, ORF7a. Of these NSP10 had just two variants and both of them were predicted as Disease by all three tools. Others had differential bias towards Disease variants. Thus, we can say that though some regions of the genome have more variations but mostly Neutral while others with fewer variations are more impactful in terms of their predicted impact due to more Disease variants. Conversely, mutations in some proteins can be relatively better tolerated by the viral genome

The overall protein prediction outcomes of the 611 genomes have been summarized in Fig. 7. There were total of 198 mutations (70%) and 83 mutations (30%) which are predicted to be Neutral and Disease respectively by at least two tools. These predictions suggest that even though mutations are accumulating in SARS-CoV-2, they are predominantly neutral. This is the possible reason that no major virulence or physiological deviations have been observed so far.

3.6. *Mutational profile of asymptomatic and deceased samples*

In order to further assess impact of these variations we compared their prevalence across samples which were asymptomatic with those wherein the patient died. The idea was that if predictions are true, then asymptomatic samples should have more of neutral mutations whereas deceased ones should have more of disease mutations. The present congregation of samples in the study had just 2 asymptomatic samples and 15 deceased. Thereon, we included 30 new asymptomatic samples (Supplementary file 1) and compared their amino acid mutations with those of 15 deceased samples. Their comparative data has been shown in Table 5. The p value therein represents the probability that a given variant chosen at random to be Neutral or Disease.

Taking the threshold as common prediction by at least two tools the

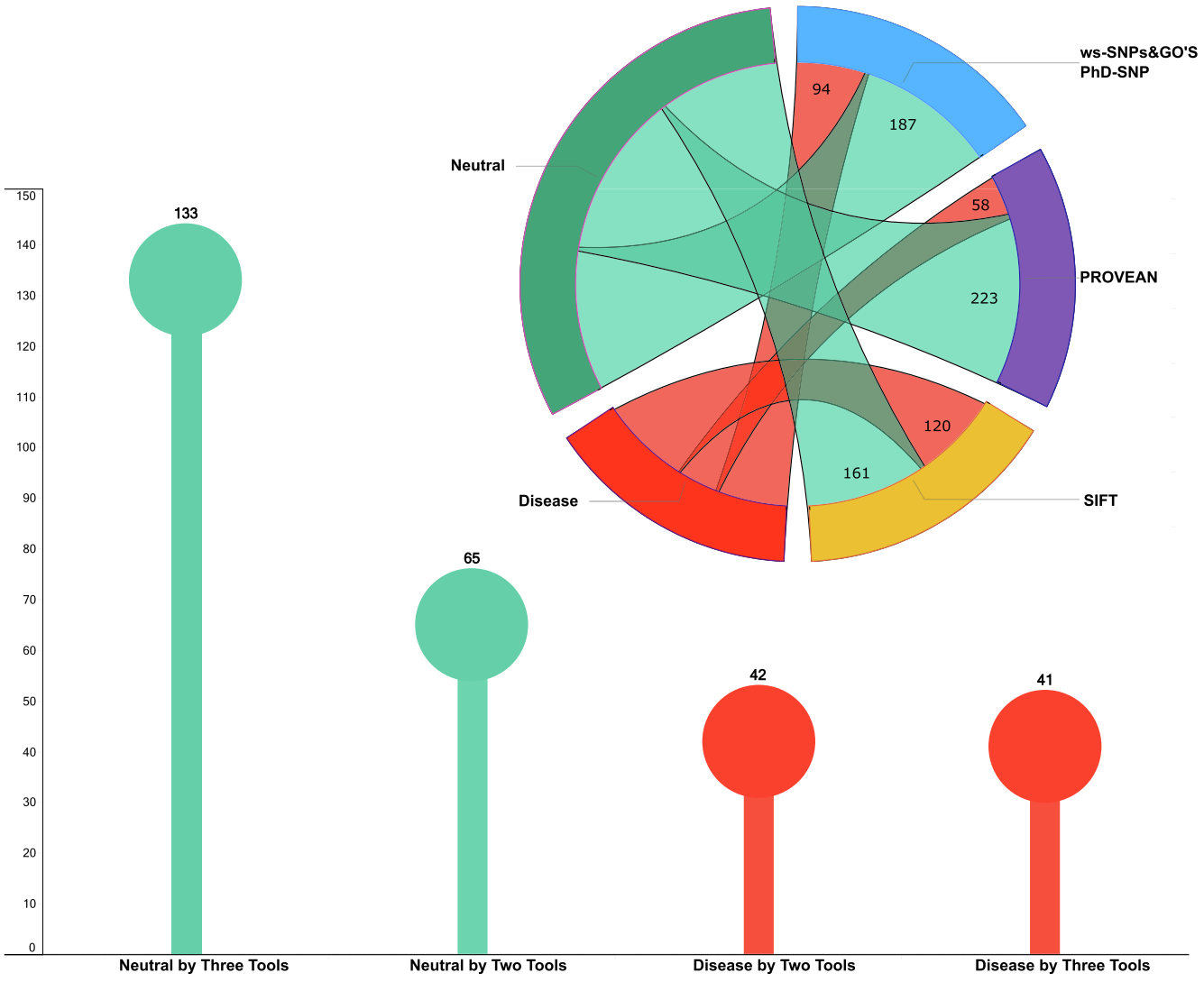

Fig. 7. Predicting the impact of mutations affecting amino acid sequence on proteins through multiple tools.

Table 5
Predicting the impact of mutations on proteins from different set of genomes.

		Original congregation			Asymptomatic			Deceased		
No of Samples		611			30			15		
Samples with no mutation		156			0			01		
Total Variants		281			55			13		
Mutation Type	Protein Prediction by tools	No of variants	p value	Sum of Mutation Incidence	No of variants	p value	Sum of Mutation Incidence	No of variants	p value	Sum of Mutation Incidence
NEUTRAL	**Three Tools (N3)**	133	0.47	254	31	0.56	76	2	0.15	4
	Two Tools (N2)	65	0.23	160	18	0.33	60	6	0.46	10
	Total = (N3 + N2)	**198**	**0.7**	**414**	**49**	**0.89**	**136**	**8**	**0.61**	**14**
DISEASE	**Three Tools (D3)**	41	0.15	187	2	0.04	4	1	0.08	2
	Two Tools (D2)	42	0.15	70	4	0.07	2	4	0.31	6
	Total = (D3 + D2)	**83**	**0.3**	**257**	**6**	**0.11**	**6**	**5**	**0.39**	**8**
	Grand Total	**281**	**1**	**671**	**55**	**1**	**142**	**13**	**1**	**22**

data gives interesting insights. As shown and previously mentioned, for the original congregation of 611 samples, 70% mutations were Neutral (p value 0.7) and 30% were Disease (p value 0.3). The asymptomatic samples with a total of 55 SNPs affecting amino acid sequence had p value of 0.89 for Neutral variants and 0.11 for Disease variants. Corresponding data for the deceased samples with a total of 13 mutations affecting amino acid sequence had p values of 0.61 and 0.39 for Neutral and Disease variants respectively. Evidently, asymptomatic samples had majorly neutral mutations (89%) but deceased samples have a reduced share of neutral mutations (61%) and enhanced share of Disease mutations (39%). The data were analyzed in terms of p value owing to the difference in number of samples in each category. We believe two

aspects need to be considered while ascertaining this differential data. First, the variation in number of asymptomatic and deceases samples. Though we have analysed in terms of percentage and p value for uniformity the fact remains that data from more samples need to be studied. Secondly, there is a relatively reduced possibility for an asymptomatic sample being sequenced and hence a chance for underestimation of mutations in asymptomatic group. Thus, a larger data set analysis for all categories with clinical correlation is essential to provide greater insight into the impact of protein variations on SARS-CoV-2.

4. Conclusions

The mutational accumulation in SARS-CoV-2 genomes is a multifactorial event with some areas of genome more prone to mutations, selective mutations being more prevalent, non-linear assimilation of mutations across various states and differential correlation between mutational impact on proteins and physiological state. Though, age and gender specific bias in incidence of mutations was observed but data has to be inferred while acknowledging the absence of an established causal relationship between the disease and gender. Further, the asymptomatic samples had higher occurrence of Neutral variants while deceased samples had relatively higher incidence of Disease variants which needs to be reaffirmed in a larger sample set. A cross-linking of mutational dynamics and patient history will provide for better correlation and understanding of the variations in SARS-CoV-2 genomes.

CRediT authorship contribution statement

Rezwanuzzaman Laskar: Methodology, Investigation, Formal analysis, Validation. **Safdar Ali:** Conceptualization, Supervision, Writing - original draft.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

The authors thank the Department of Biological Sciences, Aliah University, Kolkata, India for all the financial and infrastructural support provided. Authors acknowledge all the authors associated with originating and submitting laboratories of the sequences from GISAID's EpiFlu™ (www.gisaid.org) Database on which this research is based.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.gene.2021.145470.

References

Alam, C.M., Iqbal, A., Sharma, A., Schulman, A.H., Ali, S., 2019. Microsatellite Diversity, Complexity, and Host Range of Mycobacteriophage Genomes of the Siphoviridae Family. Front. Genet. 10, 207. https://doi.org/10.3389/fgene.2019.00207.

Capriotti, E., Fariselli, P., Casadio, R., 2005. I-Mutant2.0: predicting stability changes upon mutation from the protein sequence or structure. Nucleic Acids Res. 33, W306–W310. https://doi.org/10.1093/nar/gki375.

Chan, J.-F.-W., Yuan, S., Kok, K.-H., To, K.-K.-W., Chu, H., Yang, J., Xing, F., Liu, J., Yip, C.-C.-Y., Poon, R.-W.-S., Tsoi, H.-W., Lo, S.-K.-F., Chan, K.-H., Poon, V.-K.-M., Chan, W.-M., Ip, J.D., Cai, J.-P., Cheng, V.-C.-C., Chen, H., Hui, C.-K.-M., Yuen, K.-Y., 2020. A familial cluster of pneumonia associated with the 2019 novel coronavirus indicating person-to-person transmission: a study of a family cluster. The Lancet 395, 514–523. https://doi.org/10.1016/S0140-6736(20)30154-9.

Choi, Y., Sims, G.E., Murphy, S., Miller, J.R., Chan, A.P., 2012. Predicting the functional effect of amino acid substitutions and indels. PLoS ONE 7, e46688. https://doi.org/10.1371/journal.pone.0046688.

Hassan, Sk.S., Choudhury, P.P., Roy, B., Jana, S.S., 2020. Missense mutations in SARS-CoV2 genomes from Indian patients. Genomics 112, 4622–4627. 10.1016/j.ygeno.2020.08.021.

Katoh, K., Rozewicki, J., Yamada, K.D., 2019. MAFFT online service: multiple sequence alignment, interactive sequence choice and visualization. Briefings Bioinf. 20, 1160–1166. https://doi.org/10.1093/bib/bbx108.

Kumar, S., Stecher, G., Li, M., Knyaz, C., Tamura, K., 2018. MEGA X: Molecular Evolutionary Genetics Analysis across Computing Platforms. Mol Biol Evol 35, 1547–1549. https://doi.org/10.1093/molbev/msy096.

Laskar, R., Ali, S., 2020. Phylo-geo-network and haplogroup analysis of 611 novel Coronavirus (nCov-2019) genomes from India (preprint). 10.1101/2020.09.03.281774.

Lee, N., Hui, D., Wu, A., Chan, P., Cameron, P., Joynt, G.M., Ahuja, A., Yung, M.Y., Leung, C.B., To, K.F., Lui, S.F., Szeto, C.C., Chung, S., Sung, J.J.Y., 2003. A Major Outbreak of Severe Acute Respiratory Syndrome in Hong Kong. N. Engl. J. Med. 348, 1986–1994. https://doi.org/10.1056/NEJMoa030685.

Mercatelli, D., Giorgi, F.M., 2020. Geographic and Genomic Distribution of SARS-CoV-2 Mutations. Front. Microbiol. 11, 1800. https://doi.org/10.3389/fmicb.2020.01800.

Mercatelli, D., Triboli, L., Fornasari, E., Ray, F., Giorgi, F.M., 2020. *coronapp* : A Web Application to Annotate and Monitor SARS-CoV-2 Mutations (preprint). Bioinformatics. https://doi.org/10.1101/2020.05.31.124966.

Peiris, J.S.M., Guan, Y., Yuen, K.Y., 2004. Severe acute respiratory syndrome. Nat. Med. 10, S88–97. https://doi.org/10.1038/nm1143.

Rahman, M.S., Hoque, M.N., Islam, M.R., Akter, S., Rubayet-Ul-Alam, A., Siddique, M.A., Saha, O., Rahaman, Md.M., Sultana, M., Crandall, K.A., Hossain, M.A., 2020. Epitope-based chimeric peptide vaccine design against S, M and E proteins of SARS-CoV-2 etiologic agent of global pandemic COVID-19: an in silico approach. PeerJ 8, e9572. https://doi.org/10.7717/peerj.9572.

Ren, L.-L., Wang, Y.-M., Wu, Z.-Q., Xiang, Z.-C., Guo, L., Xu, T., Jiang, Y.-Z., Xiong, Y., Li, Y.-J., Li, X.-W., Li, H., Fan, G.-H., Gu, X.-Y., Xiao, Y., Gao, H., Xu, J.-Y., Yang, F., Wang, X.-M., Wu, C., Chen, L., Liu, Y.-W., Liu, B., Yang, J., Wang, X.-R., Dong, J., Li, L., Huang, C.-L., Zhao, J.-P., Hu, Y., Cheng, Z.-S., Liu, L.-L., Qian, Z.-H., Qin, C., Jin, Q., Cao, B., Wang, J.-W., 2020. Identification of a novel coronavirus causing severe pneumonia in human: a descriptive study. Chin. Med. J. 133, 1015–1024. https://doi.org/10.1097/CM9.0000000000000722.

Sanjuán, R., Domingo-Calap, P., 2016. Mechanisms of viral mutation. Cell. Mol. Life Sci. 73, 4433–4448. https://doi.org/10.1007/s00018-016-2299-6.

Sim, N.-L., Kumar, P., Hu, J., Henikoff, S., Schneider, G., Ng, P.C., 2012. SIFT web server: predicting effects of amino acid substitutions on proteins. Nucleic Acids Res 40, W452–457. https://doi.org/10.1093/nar/gks539.

Tajima, F., 1989. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics 123, 585–595.

Tamura, K., Nei, M., Kumar, S., 2004. Prospects for inferring very large phylogenies by using the neighbor-joining method. Proc. Natl. Acad. Sci. USA 101, 11030–11035. https://doi.org/10.1073/pnas.0404206101.

Vilsker, M., Moosa, Y., Nooij, S., Fonseca, V., Ghysens, Y., Dumon, K., Pauwels, R., Alcantara, L.C., Vanden Eynden, E., Vandamme, A.-M., Deforche, K., de Oliveira, T., 2019. Genome Detective: an automated system for virus identification from high-throughput sequencing data. Bioinformatics 35, 871–873. https://doi.org/10.1093/bioinformatics/bty695.

Zaki, A.M., van Boheemen, S., Bestebroer, T.M., Osterhaus, A.D.M.E., Fouchier, R.A.M., 2012. Isolation of a novel coronavirus from a man with pneumonia in Saudi Arabia. N. Engl. J. Med. 367, 1814–1820. https://doi.org/10.1056/NEJMoa1211721.