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 Reports 23 (2021) 101139

Contents lists available at ScienceDirect

Gene Reports

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

Whole genome comparison of Pakistani Corona virus with Chinese and US Strains along with its predictive severity of COVID-19

Rashid Saif [a,*], Tania Mahmood [a], Aniqa Ejaz [a], Saeeda Zia [b], Abdul Rasheed Qureshi [c]

[a] Decode Genomics, 323-D, Punjab University Employees Housing Scheme (II), Lahore, Pakistan
[b] Department of Sciences and Humanities, National University of Computer and Emerging Sciences, Lahore, Pakistan
[c] Out Patients Department-Pulmonology, Gulab Devi Chest Hospital, Ferozepur Road, Lahore, Pakistan

ARTICLE INFO

Keywords:
Pakistani SARS-nCoV2
Phylogenetic analysis
Variant calling pipeline
3D structural modeling

ABSTRACT

Initially submitted 784 SARS-nCoV2 whole genome sequences on NCBI Virus database were selected for phylogenetic analysis to look into their similarities with two of Pakistani sequenced coronavirus strains having accessions of MT240479 and MT262993. The MT240479 named (Gilgit1-Pak) was found in close proximity to MT184913 named (CruiseA-USA), while MT262993 named (Manga-Pak) was in neighboring to MT039887 named (WI-USA) strain, which were further chosen for variant calling analysis along with reference genome NC_045512 as out-group to construct concluding cladogram and looked for evolutionary distance with PAUP software in this article. Aforementioned Pakistani strains each of having 29,836 bases were compared with MT263429 (WI-USA) of 29,889 bases and MT259229 (Wuhan-P.R. China) of 29,864 bases. Whole genome variant calling pipeline revealed 31 variants in both Pakistani strains collectively (Manga-Pak vs USA having 2del & 7SNPs, while different from Chinese strain with 2del & 2SNPs, similarly Gilgit1-Pak vs USA having 10SNPs, while different from Chinese strains having 8SNPs). These variants harbour *ORF1ab*, *ORF1a* and *N* genes having their role is viral replication/translation, host innate immunity and viral capsid formation respectively. These novel variants may be one of the reasons for low mortality rate in Pakistan with 385 deaths as compared to USA with 63,871 and P.R. China with 4633 by May 01, 2020. However functional characterization of these variants and their integrations with other viral proteins including variability of human receptors (ACE2 & NRP1) may be the other reasons for unlikely COVID-19 statistics in Pakistan which need further confirmatory studies. Moreover, mutated N and ORF1a proteins in Pakistani strains were also analyzed by 3D structure modeling, which give another dimension of comparing these alterations at amino acid level. In a nutshell, these novel variants are correlated with reduced mortality of COVID-19 severity in Pakistan while more robust results can be obtained by wet lab experimentation. This also gives insight of genomic landscape of these indigenous strains to develop diagnostics kits, vaccines and therapeutic interventions.

1. Introduction

The SARS pandemic engendered new avenues to ponder and identify variations in this animal based SARS-nCoV2 that how human receptor ACE2 and NRP1 become ideally compatible with the spike region of this virus and as a result COVID-19 spread human population globally (Zhu et al., 2003). In the current century, the first wave of transmission started from SARS-CoV in Guangdong province, P.R. China and thereafter disseminated worldwide which resulted in 916 fatalities (Cherry and Krogstad, 2004). Next in 2012, MERS emerged in Saudi Arabia with 858 associated deaths by the end of November 2019 (*MERS monthly summary*, 2019). On December 12, 2019, first patient of COVID-19 was reported infected with SARS-nCoV2 strain in Wuhan, Hubei Province, P. R. China (Ren et al., 2020). This infection got widespread and of May 01, 2020, 2,036,770 active cases, 234,279 (7.06%) deceased and 1,048,807 (31.59%) recovered has been reported worldwide (*COVID-19*, n.d.). This strain mainly belongs to the B-beta corona viruses genus (Lefkowitz et al., 2018; Letko et al., 2020). It has been observed that mortality rate vary from country to country, as fatality rate due to corona is low in Pakistan, which pondered the scientists to look into linkage between

Abbreviations: SARS-nCoV, severe acute respiratory syndrome novel coronavirus; DNA, deoxyribonucleic acid; NCBI, National Center for Biotechnology Information; BAM, binary alignment maps; SAM, sequence alignment maps.

* Corresponding author.
E-mail address: rashid.saif37@gmail.com (R. Saif).

https://doi.org/10.1016/j.genrep.2021.101139
Received 18 July 2020; Received in revised form 7 March 2021; Accepted 8 April 2021
Available online 15 April 2021

Table 1
Accession numbers of SARS-nCoV2 strains from various geographical regions used in phylogenetic tree construction.

Sequence accession no.	Geographic region	References
MT240479	Gilgit city, Pakistan	https://www.ncbi.nlm.nih.gov/nuccore/MT240479
MT262993	Washington city, USA	https://www.ncbi.nlm.nih.gov/nuccore/MT262993
MT263429	Manga city, Pakistan	https://www.ncbi.nlm.nih.gov/nuccore/MT263429
MT259229	Wuhan, Hubei Province, P.R. China	https://www.ncbi.nlm.nih.gov/nuccore/MT259229

different variants of the SARS-nCoV2 with its severity along with other influencing factors e.g. temperature, testing facility, healthcare system, lockdown and preventive measures, aging factor and hygienic practices. Recently published genomic characterization study speculated the proximal origin of human Corona virus from Bat (*Rhinolophus affinis*) and Pangolin (*Maris javinica*) strains by natural selection in an animal host before zoonotic transfer or natural selection in humans following zoonotic transfer due to the novel observed variants in RBD and polybasic cleavage site of the Spike region (Andersen et al., 2020).

In this comparative genomics study, phylogenetic analysis was carried out with 784 whole genome sequences available from NCBI Virus database in order to trace the closest Pakistani Corona virus homologues. These closely related sequences (two Pakistani Gilgit1 MT240479 and Manga MT262993, USA MT263429 and P.R. China MT259229 strains) plus 1 reference sequence on NCBI from P.R. China NC_045512 as an outgroup was taken for tree construction and detection of evolutionary distance through PAUP software. Further variant calling analysis using Galaxy platform on these 4 strains was performed to predict the effect of novel variants of Pakistani strains on the severity of COVID-19. SWISS-MODEL 3D structural modeling analysis was also performed for comparing mutant proteins of Pakistani strains to have an insight of potential role of these mutant proteins on pathogenicity of SARS-nCoV2. In Pakistan preventive measures especially ablution practices and use of herbal products such as Senna leaves are also one of influencing factors for being asymptomatic or contributing to host immunity against COVID-19 but genetic factors are more valid to address and as a preliminary study to predict association of genetic variation with severity of virulence on the basis of which further wet lab experiments and confirmatory studies on functional effects of these variations could be conducted for developing indigenous vaccines, drugs and nucleic acid based detection kits.

2. Materials and methods

2.1. Phylogenetic analysis

Phylogenetic analysis was carried through online NCBI Virus database (NCBI, n.d.) based on GenBank sequence type and whole genome sequenced data taxid: 2697049 of 784 SARS-nCoV2 strains from all around the globe. The sequences neighboring to Pakistani strains (MT240479 and MT262993) were taken along with NCBI SARS-nCoV2 reference genome sequence (NC_045512-P.R. China) (Table 1) for constructing and inferring evolutionary tree by implementing Neighbor Joining (NJ) algorithm using PAUP software. Optimality criteria were set to parsimony. Maximum likelihood distance with accelerated transformations were choose as optimized order method to construct tree on which further molecular clock analysis was carried out to check evolutionary distance among the sequences used.

2.2. Whole genome variant calling analysis

FASTA sequences of two Pakistani (Gilgit1 MT240479 and Manga MT262993), Chinese (Wuhan) MT259229 and USA (Washington)

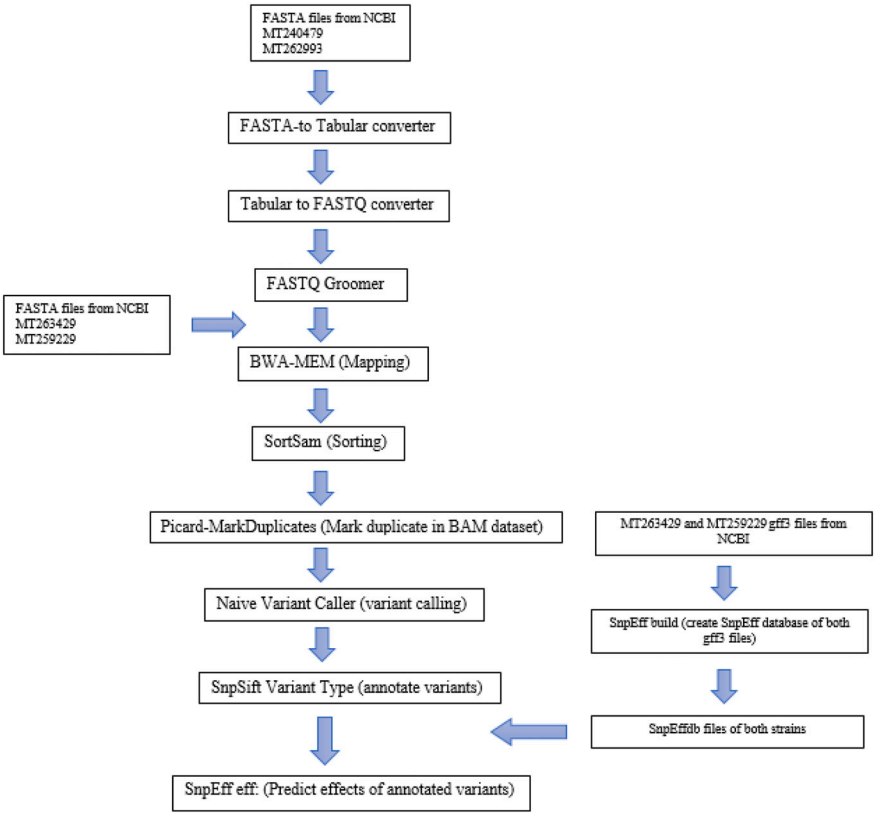

Fig. 1. Variant calling workflow.

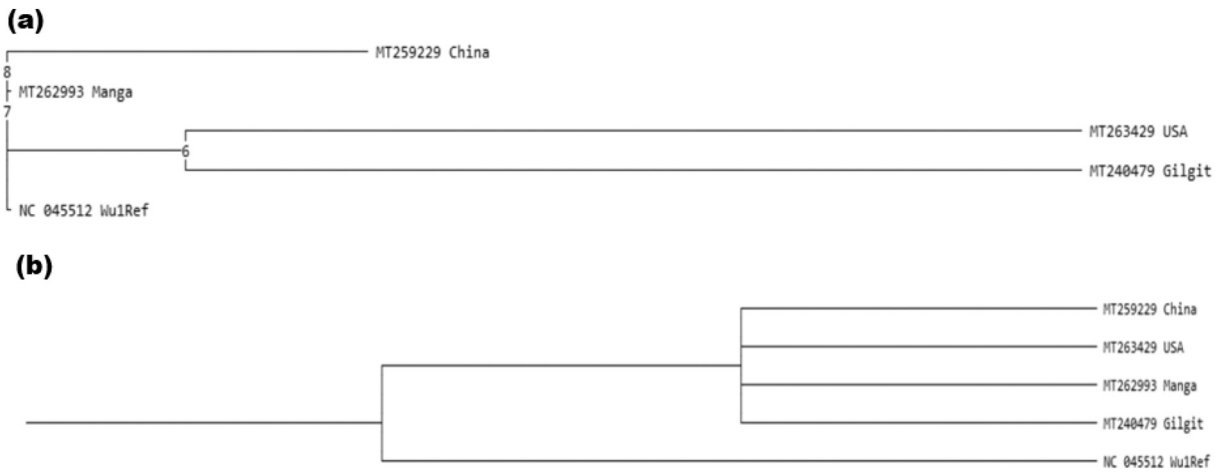

Fig. 2. Unrooted phylogram and rooted cladogram are constructed on MT240479, MT262993, MT263429, MT259229 and NC_045512 as outgroup. Phylogram is labelled at internal nodes (a) and cladogram (b), both labelled with strain identifier names.

MT263429 strains were retrieved from NCBI GenBank (NCBI National Center for Biotechnology Information, n.d.). Pakistani FASTA sequences were converted to Fastq format using FASTA-to-Tabular-to-FASTQ tools (Galaxy Version 1.1.0) (Blankenberg et al., 1783) and then mapped against reference genome using BWA MEM v 0.7.17.1 (Li, 2013). Mapped reads were coordinate sorted using SortSam feature and duplicate sequences were marked using MarkDuplicates feature of Picard tool. Aligned sequencing reads were processed for per position variant call using Naive Variant Caller (NVC) v 0.0.3 (Blankenberg et al., 2014). SnpSift Variant type (Ruden et al., 2014) and SnpEff eff was used to annotate variants by custom building of reference sequence databases using SnpEff build v 4.3+T.galaxy4 (Fig. 1) (Cingolani et al., 2012).

2.3. Protein structure modeling

Proteins bearing variation in Pakistani strains were subjected to homology based structure modeling using Promod3 on Swiss model platform (Waterhouse et al., 2018; *SWISS-MODEL*, n.d.). MT259229 (Chinese) and MT263429 (USA) were first subjected to search for homologues templates from which a prophesied 3D model was built. Next PDB files of these were used as template to build Pakistani mutant proteins (N and ORF1a) 3D models based on target-template alignment along with QSQE score complementing the GMQE for tertiary structure evaluation which is accomplished by supervised built-in SVM algorithm (Studer et al., 1765; Bertoni et al., 2017).

3. Results

3.1. Phylogenetic analysis

Both unrooted phylogram and rooted cladogram were constructed considering 5 sequences (MT240479, MT262993, MT263429, MT259229 and NC_045512 as outgroup) (Fig. 2). Cladogram is demonstrating the relatedness of all strains sharing same ancestral node at a point. However, branch lengths were deduced from the phylogram which depicts that the distance between reference sequence and internal node 7 is −1.1481e−007. MT263429 and MT240479 are very close relatives having branch lengths of 0.00017. Internal node 8 is connected to 7 with a branch length of 6.9704e−008 while node number 6 is connected to 7 consisting of 3.3446e-005 branch length. Difference in the branch length of MT259229 Chinese strain from reference sequence is −6.71e−05. MT262993 has a branch length of 9.2986e−008 which is −2.08e−07 distance away from the NC_045512. While distance of MT263429 and MT240479 from reference strain is −1.70e−04. Likelihood ratio analysis was applied in automated molecular clock test. This resulted in score of molecular clock which is =40715.2587 and non-clock model score = 40706.9349. The p-value of clock vs non-clock model is 0.00084 too lower than significance value ≤ 0.05. So this rejected the molecular clock assumption.

3.2. Whole genome variant calling analysis of SARS-nCoV2

Pakistani strain MT262993 (29,836 bp) is more closely related to the virus strain MT259229 (28,964 bp) from P.R. China with the difference of 4 variations and each variant occurs on an average distance of 7466 bp (Table 2), while MT263429 (29,889 bp) strain from USA is different with 9 variants loci at the rate of 1 variant per 2988 bp (Table 3).

Gilgit1 MT240479 strain differ from MT259229 (Chinese) strain at 8 loci (SNPs) (Table 4), variants occurring after 3733 bp while MT263429 (USA) differ from 10 variants (SNPs) each after 2988 bp (Table 5).

The effects of variations by functional class for MT262993 (Manga) vs MT263429 (USA) are 05 missense variants, for MT240479 (Gilgit1) vs MT263429 (USA) are 13 missense and 02 silent while MT240479 vs MT259229 are 08 missense and 03 silent variations and 01 silent variant of MT262993 vs MT259229. Distribution of all variants impact, type and region are demonstrated in (Table 6) and the graphical representation of variants impacts is shown in (Fig. 3).

Table 2
Details of different variants observed in Manga-Pak and Chinese strains.

Manga-Pak (MT262993) vs Chinese strains (MT259229)									
Strain	Pos	Ref	Alt	AC	AF	Eff	Vartype	GT	NC
MT259229	9706	ATTTCTATTGGTTCTT	A	1	0.5	Codon_deletion, ttctattggttcttt, FYWFF3153,	Del	0	A = 1,d15 = 1
MT259229	19501	ATTGTATCTCGATGCTTATAAC	A	1	0.5	Codon change, codon deletion gattgtatctcgatgcttataaca/gaa DCISMLIT6417E	Del	0	A = 1,d21 = 1
MT259229	20678	T	C	1	1	Synonymous coding tgT/tgC, C6809	SNP	1	C = 1
MT259229	29854	C	G	1	1	Intergenic	SNP	1	G = 1

Table 3
Detail of different variants observed in Manga-Pak vs USA.

MT262993 (Manga-Pak) vs MT263429 (USA)									
Strain	Pos	Ref	Alt	AC	AF	Eff	Vartype	GT	NC
MT263429	21	H	A	1	1.0	Intergenic	Interval, SNP, SNP	1	A = 1
MT263429	22	T	A	1	1.0	Intergenic	SNP	1	A = 1
MT263429	23	T	C	1	1.0	Intergenic	SNP	1	C = 1
MT263429	24	A	C	1	1.0	Intergenic	SNP	1	C = 1
MT263429	1427	A	G	1	1.0	Non synonymous coding gAc/gGc, D392G	SNP	1	G = 1
MT263429	2878	A	G	1	1.0	Non synonymous coding Aca/Gca, T876A	SNP	1	G = 1
MT263429	9707	ATTTCTATTGGTTCTT	A	1	0.5	Codon deletion, ttctattggttcttt, FYWFF3153	Del	0	A = 1, d15 = 1
MT263429	19502	ATTGTATCTCGATGCTTATAAC	A	1	0.5	Codon change, codon deletion gattgtatctcgatgcttataaca/ gaa. DCISMLIT6417E	Del	0	A = 1, d21 = 1
MT263429	28841	T	C	1	1.0	Non synonymous coding, tTa/tCa, L194S.	SNP	1	C = 1

Table 4
Detail of different variants observed in Gilgit1-Pak vs P.R. China.

MT240479 (Gilgit1-Pak) vs MT259229 (China)									
Strain	Pos	Ref	Alt	AC	AF	Eff	Vartype	GT	NC
MT259229	227	C	T	1	1	Intergenic	SNP	1	T = 1
MT259229	870	C	T	1	1	Non synonymous coding, Cgt/Tgt, R207C	SNP	1	T = 1
MT259229	1334	C	T	1	1	Synonymous coding, ccC/ccT, P361	SNP	1	T = 1
MT259229	1383	G	A	1	1	Non synonymous coding, Gta/Ata, V378I	SNP	1	A = 1
MT259229	9145	C	T	1	1	Non synonymous coding, cCt/cTt, P2965L	SNP	1	T = 1
MT259229	11069	G	T	1	1	Non synonymous coding, ttG/ttT, L3606F	SNP	1	T = 1
MT259229	20678	T	C	1	1	Synonymous coding, tgT/tgC|C6809	SNP	1	C = 1
MT259229	29854	C	G	1	1	Intergenic	SNP	1	G = 1

Table 5
Detail of different variants observed in Gilgit1-Pak vs USA.

MT240479 (Gilgit1-Pak) vs MT263429 (USA)									
Strain	Pos	Ref	Alt	AC	AF	Eff	Vartype	GT	NC
MT263429	24	A	C	1	1.0	Intergenic	SNP	1	C = 1
MT263429	228	C	T	1	1.0	Intergenic	SNP	1	T = 1
MT263429	871	C	T	1	1.0	Non synonymous coding, Cgt/Tgt, R207C	SNP	1	T = 1
MT263429	1335	C	T	1	1.0	Synonymous_coding, ccC/ccT, P361	SNP	1	T = 1
MT263429	1384	G	A	1	1.0	Non synonymous coding, Gta/Ata, V378I	SNP	1	A = 1
MT263429	1427	A	G	1	1.0	Non synonymous coding, gAc/gGc, D392G	SNP	1	G = 1
MT263429	2878	A	G	1	1.0	Non synonymous coding, Aca/Gca, T876A	SNP	1	G = 1
MT263429	9146	C	T	1	1.0	Non synonymous coding, cCt/cTt, P2965L	SNP	1	T = 1
MT263429	11070	G	T	1	1.0	Non synonymous coding, ttG/ttT, L3606F	SNP	1	T = 1

3.3. Alteration in Pakistani SARS-nCoV2 genes and their effects

Detailed analysis of SARS-nCoV2 genome revealed some key alterations occurring in the sequences that codes for ORF1a polyprotein (4405aa) by cds-QIS61085.1 gene and by cds-QIS30017.1 gene. ORF1ab polyprotein (7096aa) by cds-QIS30016.1 gene and *ORF1ab* gene and nucleocapsid phosphoprotein (419aa) expressed by *N* gene (Tables 7 and 8).

N protein has multifarious activities, having 3 major domains of M-M (matrix protein), M-N and N-S (S-spike protein) interaction (C-k et al., 2014). It binds RNA tightly and packages the viral genome into capsid (a ribonucleoprotein) (Lai and Cavanagh, 1997). ORF1ab (replicase polyprotein 1) encodes 7096 amino acids present at 5′ end and is involved in replication and translation of viral RNAs (Phan, 2020). It interacts with the host innate immune response and is responsible for host virulence. Results also indicated that ORF1a polyprotein (4405aa) shows a rate of nonsynonymous substitutions usually (Graham et al., 2008).

3.4. Pakistani CoV2 alteration on codon and amino acid level

Quantity and distribution of SNPs and deletions against USA and Chinese strains are shown in (Fig. 4).

Genotypes of all samples were heterozygous. We considered relationship between codon usage, base substitutions SNVs and base deletions in annotated VCFs (Fig. 5).

Amino acid changes in annotated VCF were also identified. Surprisingly, there is a multiple nucleotide deletion in MT262993 strain TTTCTATTGGTTCTT (FYWFF3153), ATTGTATCTCGATGCTTATAAC (DCISMLIT6417E) at loci 9706, 19501 with MT259229 and at positions 9707, 19502 when comparing it with MT263429 that might be causing alteration in protein effectiveness (Fig. 6).

3.5. Variants distribution across the whole genome of Pakistani Corona virus strains

Division of variants across the whole genome of SARS-nCoV2 before 2900 bases is shown in (Fig. 7).

Seven SNPs were observed in first 2900 bases while other 02 are present in the remaining genome between Pakistani Manga and USA strain, while Pakistani Gilgit1 strain has total 10 SNPs, first seven within 2900 bases are shown in (Fig. 7a, b). Total eight variants appeared in Gilgit1 vs Chinese strain, first four variants are shown in (Fig. 7c).

Table 6
Detail of Pakistani variants, their impact, type and region.

MT240479 (Gilgit1-Pak) vs MT263429 (USA)

Type	Count	Percent
Low	2	11.765
Moderate	13	76.471
Modifier	2	11.765

MT240479 (Gilgit1-Pak) vs MT263429 (USA)

Type			Region		
Type	Count	Percent	Type	Count	Percent
Intergenic	2	11.765	Exon	15	88.235
Non synonymous coding	13	76.471	Intergenic	2	11.765
Synonymous	2	11.765			

MT262993 (Manga-Pak) vs MT263429 (USA)

Type	Count	Percent
Moderate	8	61.538
Modifier	5	38.482

MT262993 (Manga-Pak) vs MT263429 (USA)

Type			Region		
Type	Count	Percent	Type	Count	Percent
Codon change, codon deletion	1	7.602	Exon	8	61.538
Codon deletion	2	15.385	Intergenic	5	38.482
Intergenic	5	38.462			
Non synonymous coding	5	38.462			

MT240479 (Gilgit1-Pak) vs MT259229 (P.R. China)

Type	Count	Percent
Low	3	23.077
Moderate	8	61.538
Modifier	2	15.385

MT240479 (Gilgit1-Pak) vs MT259229 (P.R. China)

Type			Region		
Type	Count	Percent	Type	Count	Percent
Intergenic	2	15.385	Exon	11	84.615
Non synonymous coding	8	61.538	Intergenic	2	15.385
Synonymous	3	23.077			

MT262993 (Manga-Pak) vs MT259229 (P.R. China)

Type	Count	Percent
Low	1	20
Moderate	3	60
Modifier	1	20

MT262993 (Manga-Pak) vs MT259229 (P.R. China)

Type	Count	Percentage	Type	Count	Percent
Codon change, codon deletion	1	20	Exon	4	80
Codon deletion	2	40	Intergenic	1	20
Intergenic	1	20			
Non synonymous coding	1	20			

Similarly total 04 variants (SNPs = 2, del = 2), none of the polymorphism within first 2900 bases are observed in Manga vs Chinese strain shown in (Fig. 7d), exact positions and genes are mentioned in (Tables 2–5, 7, 8).

3.6. Homology modeling of Pakistani SARS-nCoV2 mutant proteins

N protein which has mutations when Pakistani strains were compared to MT263429 USA was modeled from partial alignment and structurally analyzed. Comparison between Gilgit1 and USA has sequence similarity of 62% with 26% sequence coverage, aligning amino acids in the range of 250–360 from total 419aa, while Manga strain has sequence similarity with USA of 63%, sequence coverage 30% aligning 46–172 amino acids out of 419aa (Fig. 8).

MT240479 Gilgit1 has more loop structures than MT262993 Manga N protein and no ligand interaction domains were found in the template

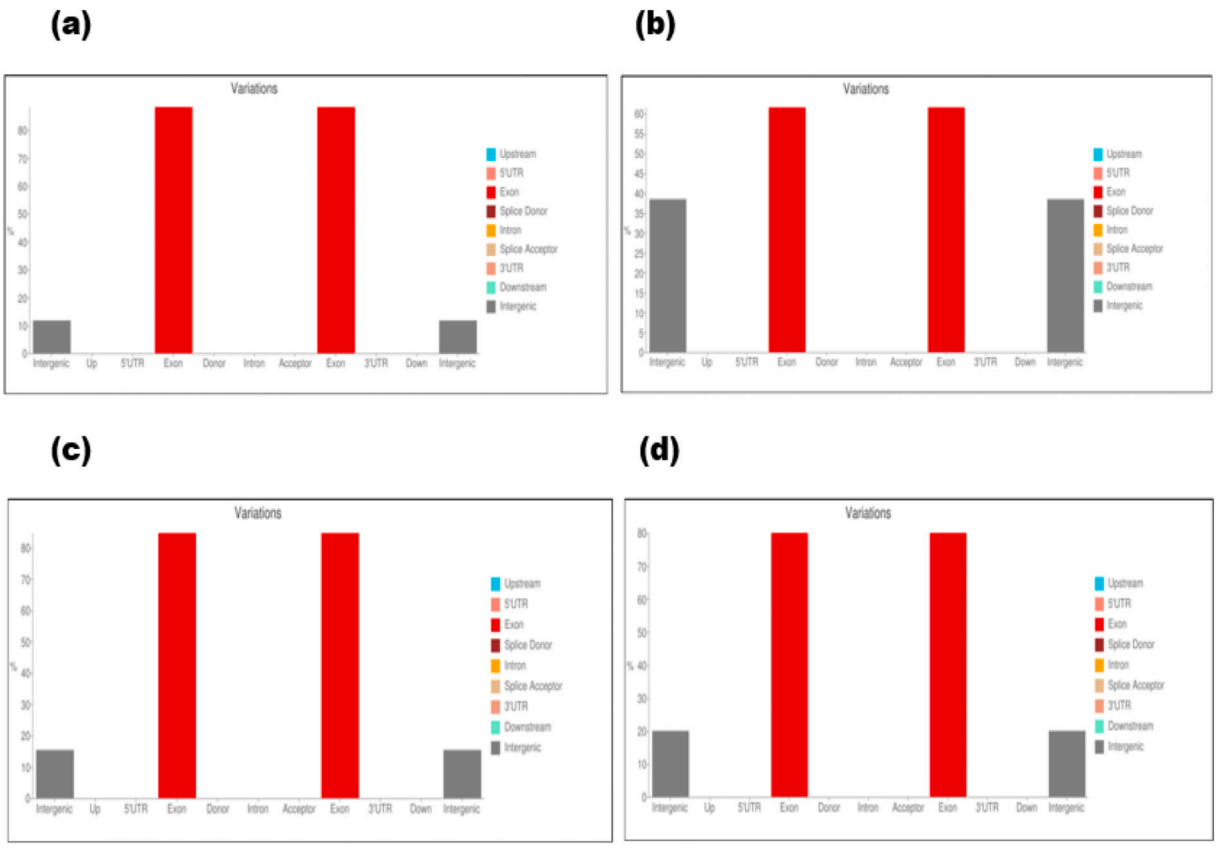

Fig. 3. Graphical illustration of variants type on x-axis and its occurrence percentage on y-axis. MT240479 (Gilgit1) vs MT263429 (USA) (a), MT262993 (Manga) vs MT263429 (USA) (b), MT240479 (Gilgit1) vs MT259229 (P.R. China) (c), MT262993 (Manga) vs MT259229 (P.R. China) (d). Grey bar showing intergenic variants and red bar of exonic variants.

Table 7
Altered genes of Gilgit1-Pak and Manga-Pak vs USA strain.

MT240479 (Gilgit1-Pak) vs MT263429 (USA)					
		Variants impact		Variants effect	
Gene	Biotype	Low	Moderate	Non synonymous coding	Synonymous coding
N	Protein coding	0	1	0	1
Orf1AB	Protein coding	1	6	1	6
cds-QIS61085.1	Protein coding	1	6	1	6

MT262993 (Manga-Pak) vs MT263429 (USA)					
		Variants impact	Variants effect		
Gene	Biotype	Moderate	Codon change, codon deletion	Codon deletion	Non synonymous coding
N	Protein coding	1	0	0	1
Orf1AB	Protein coding	4	1	1	2
cds-QIS61085.1	Protein coding	3	0	1	2

(USA) (Fig. 9).

ORF1ab protein amino acid sequence was too large to be modeled. At last, mutant ORF1a proteins were then structurally modeled by using both Chinese and USA strain sequences as template. Partial alignment of 1567–1878 amino acids (7% of total sequence coverage) among Gilgit1 and Manga strains were compared against USA and Chinese strain which resulted in 62% sequence similarity in all four alignments (Fig. 10).

Structures resulting from ORF1a protein alignment revealed three major non-covalent protein-ligand interaction domains in all ORF1a protein structures. Ligand Zn^{+2} forms interaction with their chain A (Fig. 11).

4. Discussion

Around 70% of human pathogens are of zoonotic origin including all Corona viruses specially novel SARS-CoV2, which is causing COVID-19, the ongoing pandemic (Ou et al., 2020). By today May 1, 2020 210 countries and territories are affected. In the current era Next Generation Sequencing (NGS) techniques are booming to let genomic data available seamlessly to scientists and researchers for deciphering an unseen in depth view of viruses on genomic level to overcome and combat this disease. We performed phylogenetic analysis along with variant calling and detected 31 variants in *N*, *ORF1ab* and *ORF1a* genes while comparing Pakistani strains with MT263429 and MT259229. These

Table 8
Altered genes of Gilgit1-Pak and Manga-Pak vs Chinese strain.

MT240479 (Gilgit1-Pak) vs MT259229 (P.R. China)					
		Variants Impact		Variants effect	
Gene	Biotype	Low	Moderate	Non synonymous coding	Synonymous coding
cds-QIS30016.1	Protein coding	2	4	4	2
cds-QIS30017.1	Protein coding	1	4	4	1

MT262993 (Manga-Pak) vs MT259229 (P.R. China)						
		Variant Impact		Variant effect		
Gene	Biotype	Low	Moderate	Codon change, codon deletion	Codon deletion	Synonymous coding
cds-QIS30016.1	Protein coding	1	2	1	1	1
cds-QIS30017.1	Protein coding	0	1	0	1	0

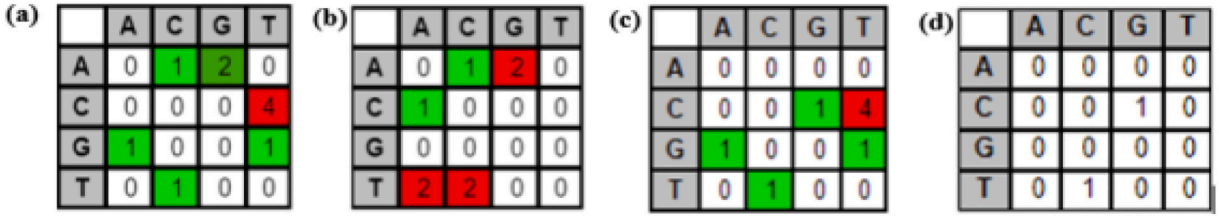

Fig. 4. Comparison between MT240479 (Gilgit1-Pak) vs MT263429 (USA) having 10 SNPs with 8 transitions (Ts) and 2 transversions (Tv) (a), MT262993 (Manga-Pak) vs MT263429 (USA), 8 SNPs 4Ts, 2Tv + deletions (b), MT240479 (Gilgit1-Pak) vs MT259229 (P.R. China) 8 SNPs, 6Ts, 2Tv (c), and MT262993 (Manga-Pak) vs MT259229 (P.R. China) strain of SARS CoV-2 displaying 2 bp changes (SNPs), 1 Ts and 1 Tv (d).

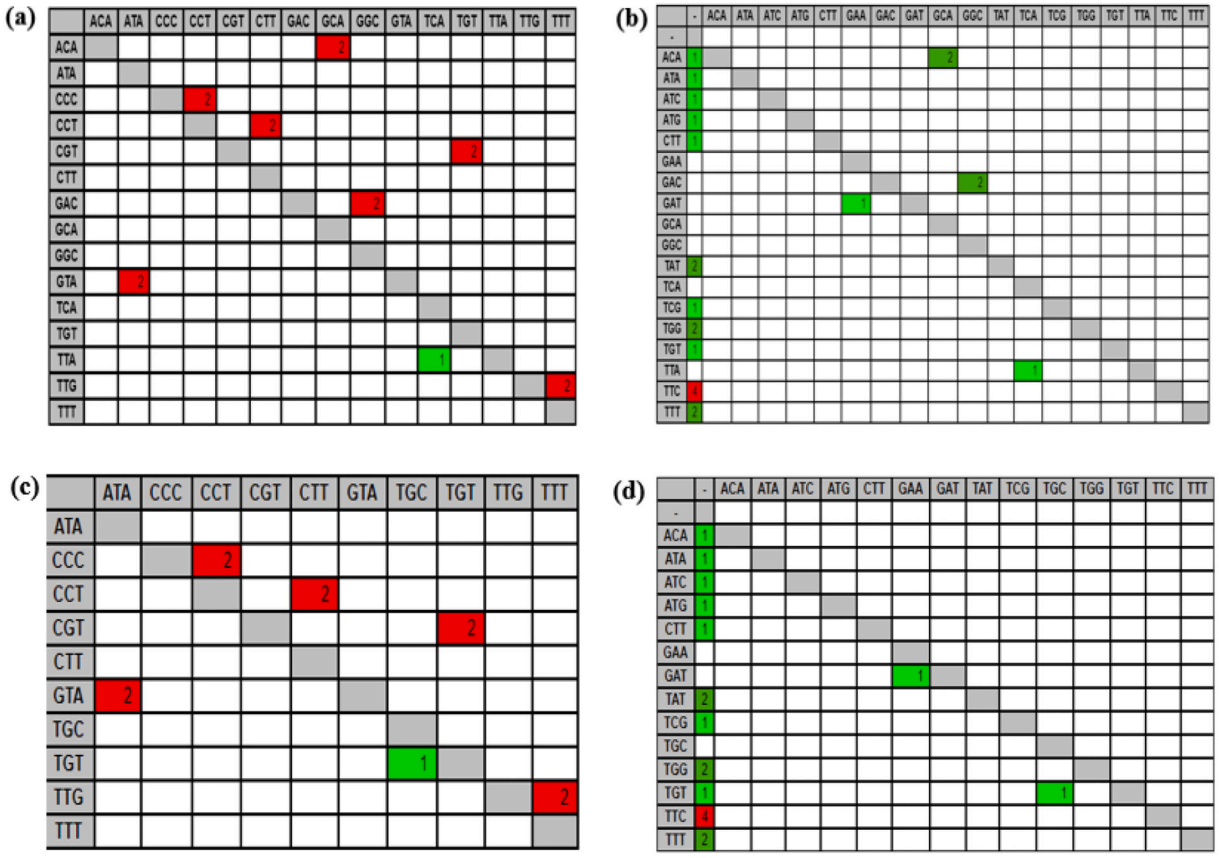

Fig. 5. Results of codon changes. MT240479 (Gilgit1-Pak) vs MT263429 (USA) (a), MT262993 (Manga-Pak) vs MT263429 (USA) (b), MT240479 (Gilgit1-Pak) vs MT259229 (P.R. China) (c), MT262993 (Manga-Pak) vs MT259229 (P.R. China) (d), strains of SARS CoV-2. Columns indicating changed codons as a result of variation. Red color is signifying heat-map i.e. higher variation.

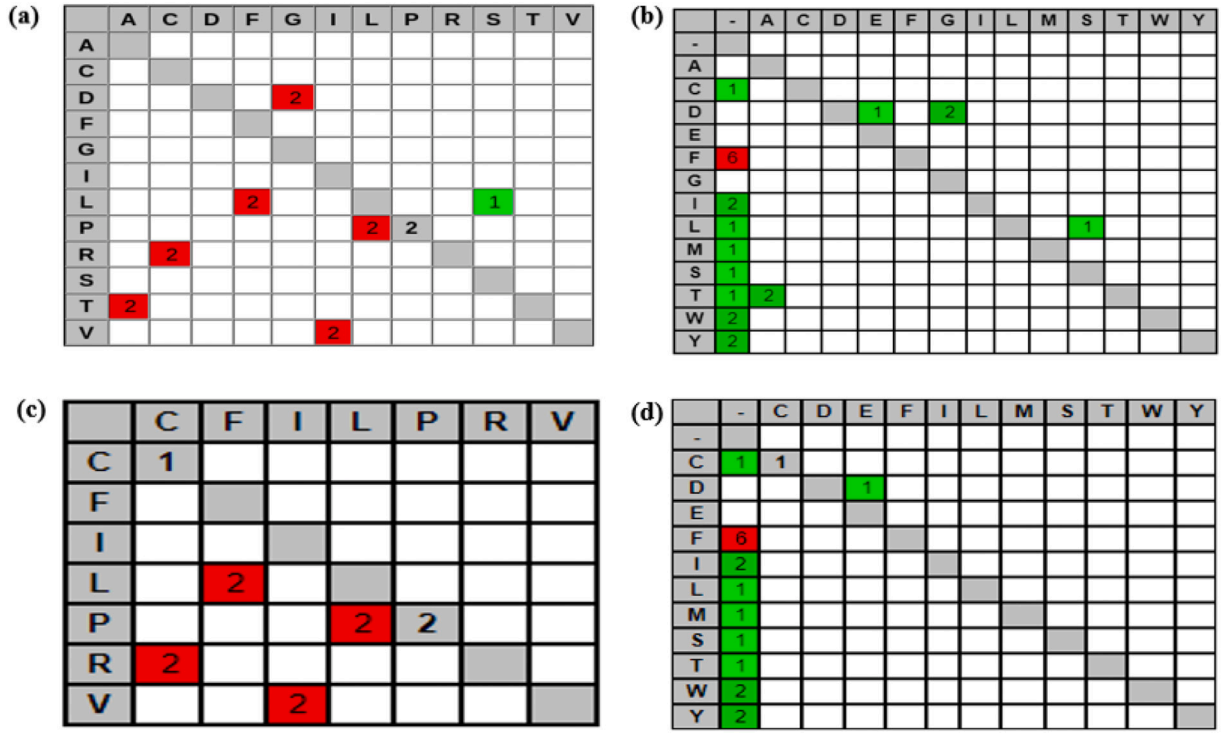

Fig. 6. Detailed explanation of amino acid change in MT240479 (Gilgit1-Pak) vs MT263429 (USA) (a), MT262993 (Manga-Pak) vs MT263429 (USA) (b), MT240479 (Gilgit1-Pak) vs MT259229 (P.R. China) (c), MT262993 (Manga-Pak) vs MT259229 (P.R. China) (d), strains of SARS CoV-2. Where rows are reference codons and columns are changed amino acids. The count of impact is showing high variation. Red background color indicates that more changes happened.

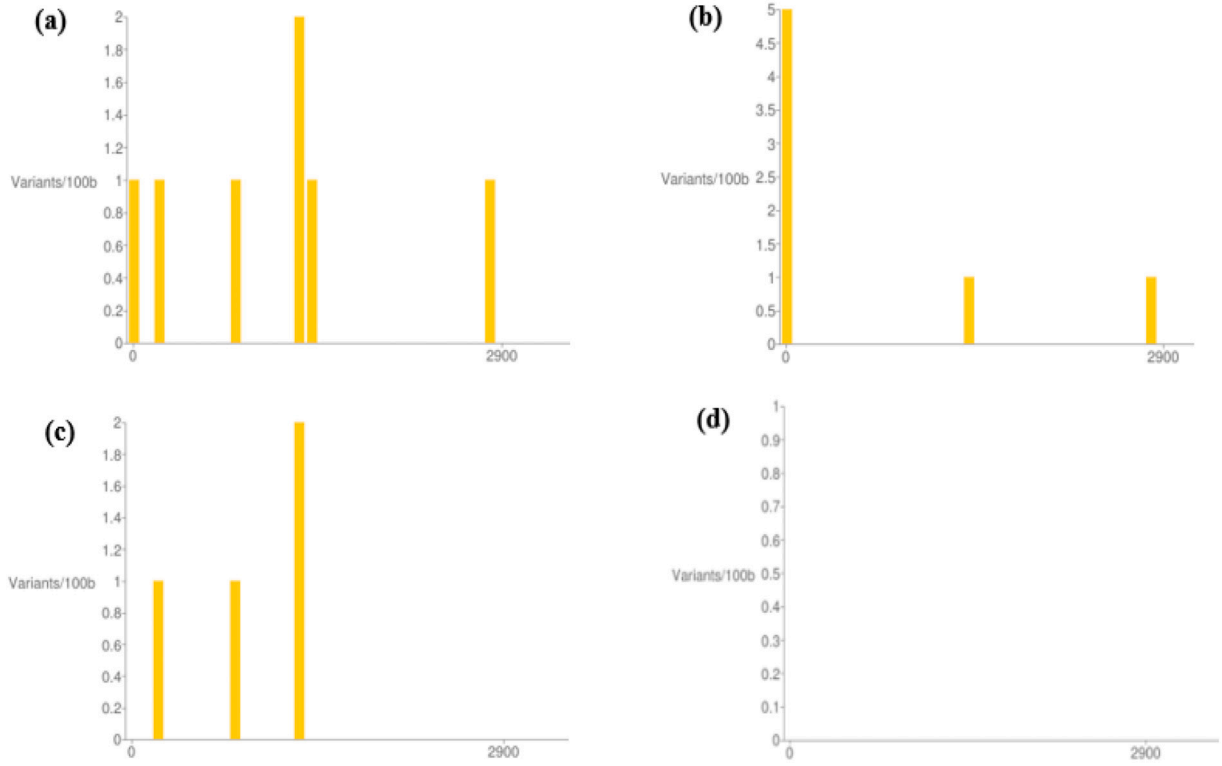

Fig. 7. Bar chart illustrating variants across the whole genome (first 2900 bases are shown here). MT240479 (Gilgit1-Pak) vs MT263429 (USA) (a), MT262993 (Manga-Pak) vs MT263429 (USA) (b), MT240479 (Gilgit1-Pak) vs MT259229 (P.R. China) (c), MT262993 (Manga-Pak) vs MT259229 (P.R. China) (d). X-axis showing bases and y-axis showing variants/100 bases.

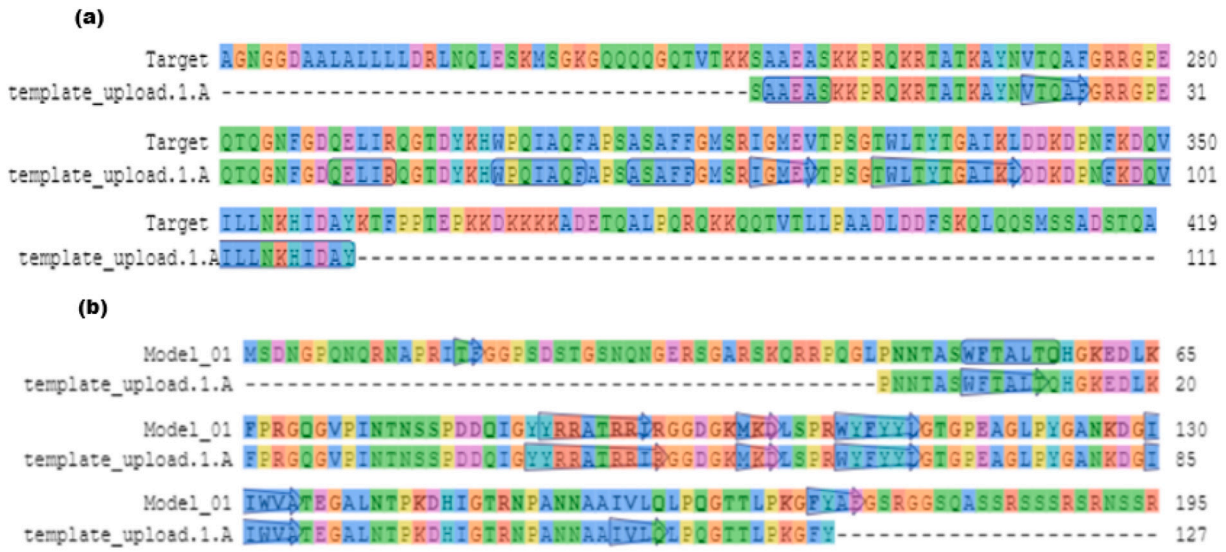

Fig. 8. Target-template alignment of N protein. 250–360 amino acids aligned between target MT240479 (Gilgit1-Pak) and template MT263429 (USA) (a), 46–172 amino acids are aligned between target MT262993 (Manga-Pak) and template MT263429 (USA) (b).

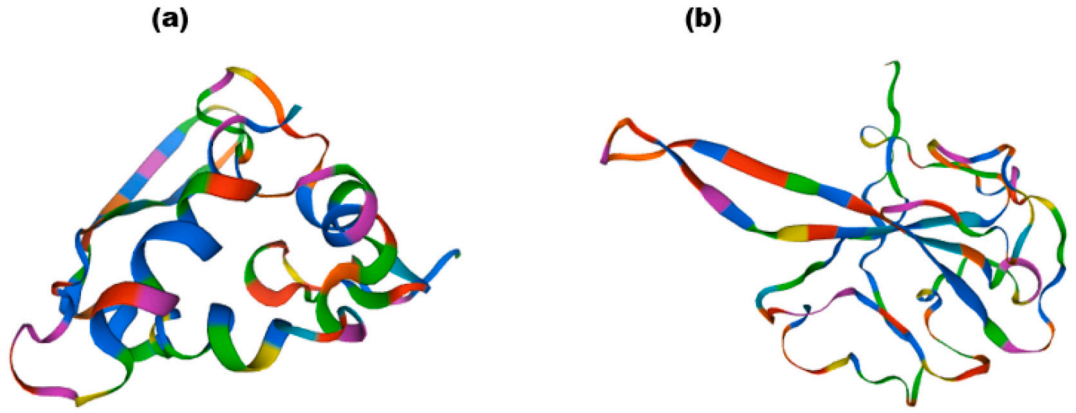

Fig. 9. Protein 3D structures modeled. MT240479 (Gilgit1-Pak) N protein (a), and MT262993 (Manga-Pak) N protein (b). For both MT263429 (USA) was used as template.

mutational events come up with some predictive evidence about the outcomes of cases. On January 19, 2020, first case of COVID-19 was reported in Washington, USA and within one and half month time, death rate was 27%, while in Pakistan, first case was reported on February 26, 2020, and death rate after the same time was 10.14% with 89.86% recovery rate, even current death rate in Pakistan by today's (July 18, 2020) statistics is 3%. Similarly, P.R. China also suffered from very high ratio of deceased people 43.18% in the same time span after its first case was being reported (*Worldometer*, n.d.). Gilgit1 and Manga strain variants are linked with low severity of this pandemic along with other factors of high temperature, less elderly people, hygienic conditions (ablution) and usage of eastern herbs as home remedy practices in Pakistan. Other factors which may also include healthcare system, prevention strategies and host responses, possibly, variations in Pakistani SARS-nCoV2 proteins might also have affected its interaction with the ACE2 receptors in humans causing low virulence and vice-a-versa, but for validation of this assumption further functional studies are required. We also performed structural analysis and modeled mutant proteins 3D structure from target-template alignments using N and ORF1a proteins of USA and Chinese strains as template with both Pakistani N and ORF1a protein as target sequences to visualize it at amino acid level to further ensure the differences in studied strains.

5. Conclusion

We conclude our discussion by making the instance that N and ORF1a genes variants in Pakistani Corona virus as reported in this paper are associated with some functional phenotypes causing low mortality rate in Pakistan vs USA and Chinese strains, however no variants were found in RBD and polybasic cleavage site of spike region which is more critical region for the virulence of this virus. This dry lab analysis still needs more research to validate and to find out association of these mutant genes and other influencing factors with the pathogenicity of this virus.

CRediT authorship contribution statement

Rashid Saif (RS): Conceive the idea, analyze the data with Galaxy, trouble shooting, contribution from initial drafting to proofreading of the manuscript, contributed as mentor for other co-authors understanding, correspondence with journal.
Tania Mahmood (TM): Helped in analyzing the data, initial drafting of the manuscript, editing.
Aniqa Ejaz (AE): Helped in analyzing the data, initial drafting of the manuscript, editing.

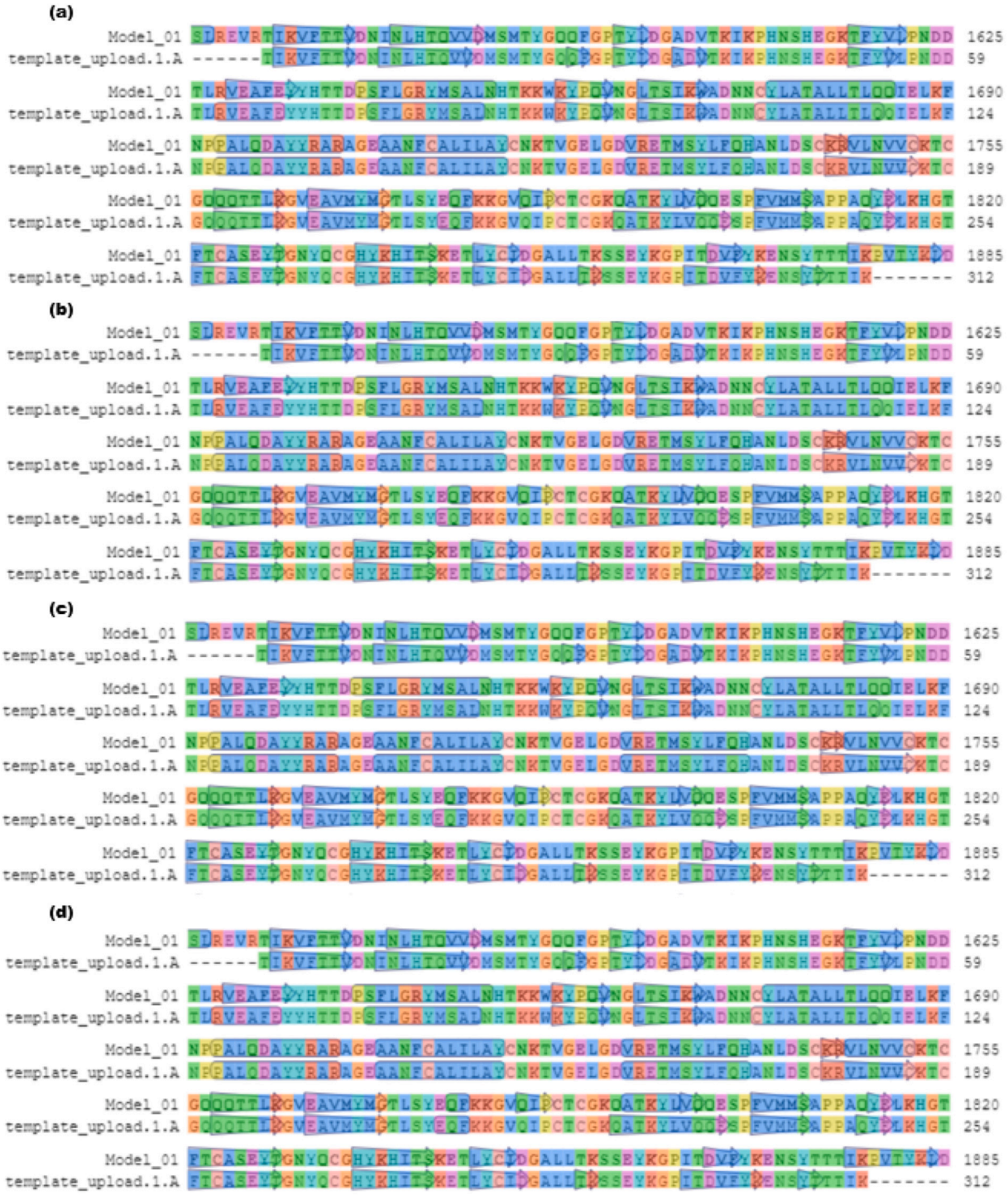

Fig. 10. Target-template partial alignment. MT240479 (Gilgit1-Pak) vs MT263429 (USA) (a), MT262993 (Manga-Pak) vs MT263429 (USA) (b), MT240479 (Gilgit1-Pak) vs MT259229 (P.R. China) (c), MT262993 (Manga-Pak) vs MT259229 (P.R. China) (d). Alignment from 1567 to 1878 amino acids out of 4405aa is shown in all four alignments.

Saeeda Zia (SZ): Helped in analyzing the data and understanding mathematical/statistical aspects of data analysis and refining the whole pipeline, helped in understanding the functioning of the bio-informatics tools and developing further computational biology tools of similar kind for bioinformatics data science specialization, proofreading.

Abdul Rasheed Qureshi (ARQ): Helped in understanding clinical aspects of COVID-19, day by day updates on the prevalence rate of COVID-19 in Pakistan, prevention strategies, proofreading and participated in refinement of this manuscript.

Declaration of competing interest

There is no conflict of interest among authors.

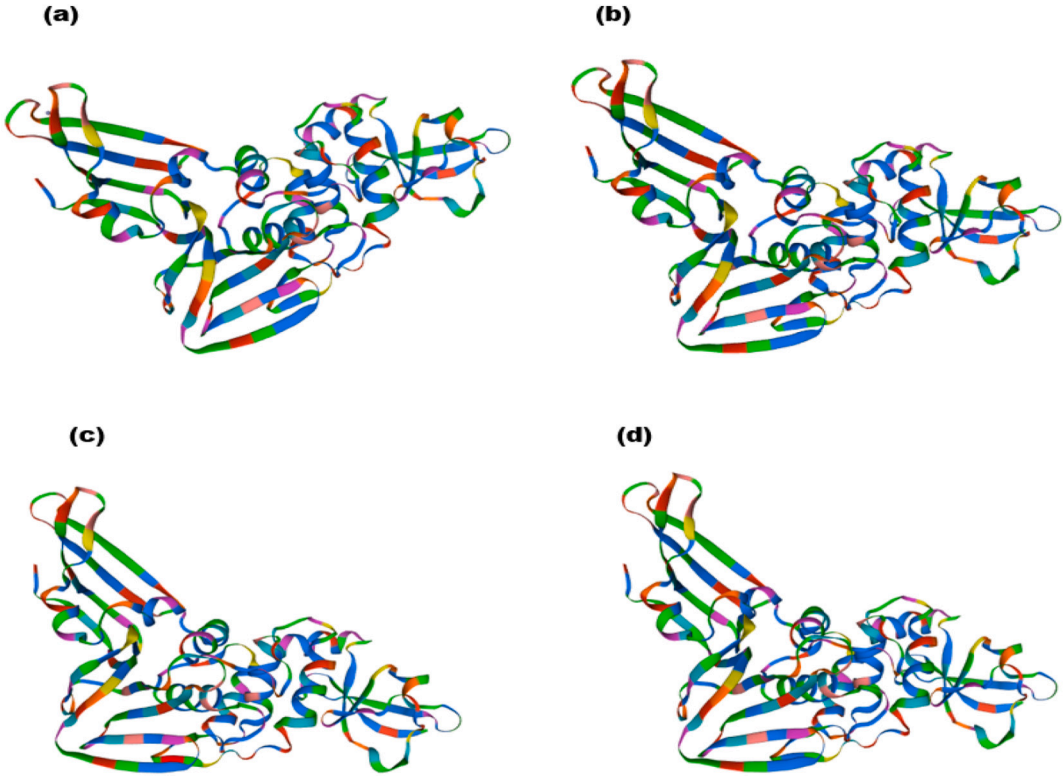

Fig. 11. ORF1a protein structurally modeled MT240479 (Gilgit1-Pak) (a), and MT262993 (Manga-Pak) (b), both modeled using MT263429 (USA) as template. MT240479 (Gilgit1-Pak) (c), while MT262993 (Manga-Pak) (d), which were modeled against MT240479 (P.R. China) ORF1a protein.

Acknowledgements

Authors are thankful to the University Institute of Biochemistry and Biotechnology, Pir Mehr Ali Shah Arid Agriculture University, Rawalpindi and Department of Healthcare Biotechnology, National University of Sciences and Technology (NUST), Islamabad, Pakistan for their generous sequencing efforts and making their data publically available on NCBI to facilitate the local and international researcher's community.

References

Andersen, K.G., Rambaut, A., Lipkin, W.I., Holmes, E.C., Garry, R.F., 2020. The proximal origin of SARS-CoV-2. Nat. Med. 1, 3. https://doi.org/10.1038/s41591-020-0820-9.

Bertoni, M., Kiefer, F., Biasini, M., Bordoli, L., Schwede, T., 2017. Modeling protein quaternary structure of homo-and hetero-oligomers beyond binary interactions by homology. Sci. Rep. 1, 15. https://doi.org/10.1038/s41598-017-09654-8.

Blankenberg, D., Gordon, A., Von Kuster, G., Coraor, N., Taylor, J., Nekrutenko, A., 1783. Manipulation of FASTQ data with Galaxy. Bioinformatics 2020, 1785. https://doi.org/10.1093/bioinformatics/btq281.

Blankenberg, D., Von Kuster, G., Bouvier, E., Baker, D., Afgan, E., Stoler, N., Taylor, J., Nekrutenko, A., 2014. Dissemination of scientific software with Galaxy ToolShed. Genome Biol. 403 https://doi.org/10.1186/gb4161.

Cherry, J.D., Krogstad, P., 2004. SARS: the first pandemic of the 21 st century. Pediatr. Res. 1, 5. https://doi.org/10.1203/01.PDR.0000129184.87042.FC.

Cingolani, P., Platts, A., Wang, L.L., Coon, M., Nguyen, T., Wang, L., Land, S.J., Lu, X., Ruden, D.M., 2012. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain w1118; iso-2; iso-3. Fly 80 (92). https://doi.org/10.4161/fly.19695.

C-k, Chang, Hou, M.-H., Chang, C.-F., Hsiao, C.-D., Huang, T.-h., 2014. The SARS coronavirus nucleocapsid protein–forms and functions. Antivir. Res. 39, 50. https://doi.org/10.1016/j.antiviral.2013.12.009.

Graham, R.L., Sparks, J.S., Eckerle, L.D., Sims, A.C., Denison, M.R., 2008. SARS coronavirus replicase proteins in pathogenesis. Virus Res. 88, 100. https://doi.org/10.1016/j.virusres.2007.02.017.

SWISS-MODEL. https://swissmodel.expasy.org/.

COVID-19. https://www.covidvisualizer.com/.

MERS monthly summary. https://www.who.int/emergencies/mers-cov/en/, 2019 (November).

Worldometer. https://www.worldometers.info/coronavirus/.

Lai, M.M., Cavanagh, D., 1997. The molecular biology of coronaviruses. In: Advances in Virus Research, vol. 48(1). Elsevier, p. 100. https://doi.org/10.1016/S0065-3527(08)60286-9.

Lefkowitz, E.J., Dempsey, D.M., Hendrickson, R.C., Orton, R.J., Siddell, S.G., Smith, D.B., 2018. Virus taxonomy: the database of the International Committee on Taxonomy of Viruses (ICTV). Nucleic Acids Res. D708, D717. https://doi.org/10.1093/nar/gkx932.

Letko, M., Marzi, A., Munster, V., 2020. Functional assessment of cell entry and receptor usage for SARS-CoV-2 and other lineage B betacoronaviruses. Nat. Microbiol. 562, 569. https://doi.org/10.1038/s41564-020-0688-y.

Li, H., 2013. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. preprint, arXiv. arXiv:1303.3997.

NCBI. Virus. https://www.ncbi.nlm.nih.gov/labs/virus/vssi/#/.

NCBI National Center for Biotechnology Information. https://www.ncbi.nlm.nih.gov/.

Ou, X., Liu, Y., Lei, X., Li, P., Mi, D., Ren, L., Guo, L., Guo, R., Chen, T., Hu, J., 2020. Characterization of spike glycoprotein of SARS-CoV-2 on virus entry and its immune cross-reactivity with SARS-CoV. Nat. Commun. 1, 12. https://doi.org/10.1038/s41467-020-15562-9.

Phan, T., 2020. Genetic diversity and evolution of SARS-CoV-2. Infect. Genet. Evol. 104260 https://doi.org/10.1016/j.meegid.2020.104260.

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, H., 2020. Identification of a novel coronavirus causing severe pneumonia in human: a descriptive study. Chin. Med. J. https://doi.org/10.1097/CM9.0000000000000722.

Ruden, D.M., Cingolani, P., Patel, V.M., Coon, M., Nguyen, T., Land, S.J., Lu, X., 2014. Using Drosophila melanogaster as a model for genotoxic chemical mutational studies with a new program, SnpSift. Front. Genet. 35 https://doi.org/10.3389/fgene.2012.00035.

Studer, G., Rempfer, C., Waterhouse, A.M., Gumienny, R., Haas, J., Schwede, T., 1765. QMEANDisCo—distance constraints applied on model quality estimation. Bioinformatics 2020, 1771. https://doi.org/10.1093/bioinformatics/btz828.

Waterhouse, A., Bertoni, M., Bienert, S., Studer, G., Tauriello, G., Gumienny, R., Heer, F. T., de Beer, T.A.P., Rempfer, C., Bordoli, L., 2018. SWISS-MODEL: homology modelling of protein structures and complexes. Nucleic Acids Res. W296, W303. https://doi.org/10.1093/nar/gky427.

Zhu, Q., Yu, M., Fan, B., Chang, G., Si, B., Peng, W., Jiang, T., Liu, B., Deng, Y., Liu, H., 2003. A complete sequence and comparative analysis of a SARS-associated virus (Isolate BJ01). Chin. Sci. Bull. 941, 948. https://doi.org/10.1007/BF03184203.