Genomic Variations in *ORF3a* and *ORF1ab* Genes of SARS-nCoV2 Strain from Southern Pakistan

Rashid Saif[1,2*], Aniqa Ejaz[1], Tania Mahmood[1], Saeeda Zia[3], Abdul Rasheed Qureshi[4]

[1] Institute of Biotechnology, Gulab Devi Educational Complex, Lahore, Pakistan
[2] Decode Genomics, 323-D, Punjab University Employees Housing Scheme (II), Lahore, Pakistan
[3] Department of Sciences and Humanities, National University of Computer and Emerging Sciences, Lahore, Pakistan
[4] Out-Patient Department-Pulmonology, Gulab Devi Chest Hospital, Ferozepur Road, Lahore, Pakistan

*Corresponding Author: rashid.saif37@gmail.com

Abstract

Emergence of COVID-19 pandemic has resulted in 8,578,283 total cases and 456,286 deaths worldwide as of June 19, 2020. We previously analysed genomic variants in two Northern Pakistani SARS-nCoV2 strains against USA and Chinese strains as reference, and hypothesized the putative role of observed variants in low severity of COVID-19 in Pakistan. Due to high variation rate in this virus, we further analysed the whole genome of Southern Pakistani SARS-nCoV2 MT500122 strain (Karachi-Pak) vs NC_045512 (Wuhan1-China) from NCBI and observed 4 variants (3=SNPs,1=del). Three of variants at g.1604 (del ND447N), SNPs at g.1912 (p.=), g.10582 (p.=) and g.26022 (p.=) in *ORF1ab* and *ORF3a* genes respectively. *ORF1ab* encodes 16 non-structural polyproteins (nsps1-16) and play role in viral replication. The codon change deletion in its sequence (observed in MT500122) might have caused conformational alterations particularly in nsp2&5 structures which may obstruct its effectiveness. *ORF3a* is unique to SARS-nCoV2 and located in-between envelope and spike genes, which assist viral entry into the host cell by interacting with *S* gene. Alteration in its sequence might have hampered the activation of *S* gene and affect its binding capacity to host ACE2 and NRP1 receptors, which may greatly weaken its pathogenicity in its different strains and hence may vary severity of COVID-19. Nevertheless, intensive data and conclusive wet lab experiments are needed for validating this postulated hypothesis. Moreover, these variants have modifier to silent impact on further 9 genes e.g. *M, N, S, E, ORFs 6, 7a, 7b, 8* and *10* as well. Advancements in understanding the role of these Pakistani SARS-nCoV2 genomic variations will be helpful in developing indigenous vaccines, diagnostic kits and drug targets.

Key words Southern Pakistani SARS-nCoV2, *ORF1ab*, *ORF3a*, whole genome variations.

Introduction

Severe Acute Respiratory Syndrome Corona Virus 2 (SARS-nCoV2), the causative agent of COVID-19 is associated with person to person transfer by droplet, contact and fomites, [1] has spread around the globe like a wildfire with ~ 9 million people affected and 456,286 fatalities as of June 19, 2020 [2]. The SARS-nCoV2 has extensive host spectrum including bats, mice, camel, avian, masked palm civets, cats, dogs etc., indeed it is speculated that SARS-nCoV2 is derived from bat (*Rhinolophus affinis*) and pangolin (*Maris javinica*) reservoirs in China market [3]. SARS-nCoV2 belongs to lineage B of betacoronavirus genus which in total has 4 lineages (A-D) while other two formerly known disease causing coronaviruses, SARS-CoV and Middle Eastern Respiratory Syndrome (MERS) which had previously caused epidemics of current century in 2003 and 2012, belongs to lineage B and C respectively [4]. SARS-nCoV2 differs from SARS-CoV and MERS by 12 nucleotide substitution in its Receptor Binding Domain (RBD) and polybasic cleavage site in its spike region [3]. This high nucleotide substitution rate in single stranded ~30K bp RNA genome of SARS-nCoV2 is the basis of intra population divergence and rapid evolution of this virus due to which the nucleic acid based detection kits lack sensitivity and specificity. Currently, wide spectrum anti-viral drugs are being used to treat

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 21 June 2020 doi:10.20944/preprints202006.0247.v1

COVID-19 patients as no vaccine is available to treat this deadly disease [5]. Genomic characterization of SARS-nCoV2 is a prerequisite in order to develop indigenous vaccines, diagnostic kits and drug development. Recent progresses in high throughput sequencing technologies and computational tools has led quick data availability to scientific community in order to have a deep insight of this unseen virus on genomic level. Previously we reported 31 variants in 2 Pakistani SARS-nCoV2 strains MT240479 (Gilgit1) and MT262993 (Manga) while comparing them with MT259229 (China) and MT263429 (USA) strains and predicted the association between these variants and reduced mortality rate due to COVID-19 in Pakistan [6].

In current study, we did genomic variation analysis on Southern Pakistani SARS-nCoV2 MT500122 (Karachi-Pak) strain by comparing it with NC_045512 (Wuhan1-China) strain. Variations in observed genes and other genes affected by these variations are elucidated and alterations in them are hypothesized to have function in modifying viral replication, division, attachment to host cell receptors and its entry into the cell causing low virulency associated with reduced mortality rate in Pakistan which is 4.85% up to today.

Materials and Methods:

Variation analysis pipeline:

Illumina single end data of Pakistani SARS-nCoV2 strain was retrieved and uploaded from NCBI Nucleotide database [7] under accession number MT500122 (Karachi-Pak) on galaxy platform. This was converted to FASTQ format using FASTA-Tabular-FASTQ tool of galaxy [8]. Fastp tool applied trimmed low quality reads while MultiQC tool generated quality check graphs [9]. Trimmed fastq files were then mapped against NCBI SARS-nCoV2 reference sequence NC_045512 (Wuhan1) using Bowtie2 tool [10]. Bam file generated was coordinate sorted and adapter sequences were marked using Picard tool MarkDuplicates feature [11]. Indel-alignment qualities for aligned reads were calculated using Add LoFreq alignment quality scores tool (Galaxy Version 2.1.4+galaxy0). SNVs and Indels were called using LoFreq (Galaxy Version 2.1.4+galaxy2) tool specifically designed for microbial sequence variant detection [12]. Annotation and prediction of resulting variants and their effects were inferred using SnpEff eff (Galaxy Version 4.3+T.galaxy1) tool [13]. “CHROM POS ID REF ALT ANN[*].EFFECT” fields were extracted from VCF file using SnpSift Extract Fields (Galaxy Version 4.3+t.galaxy0) [14]. The complete variant calling workflow run on Pakistani SARS-nCoV2 is shown in (Figure 1).

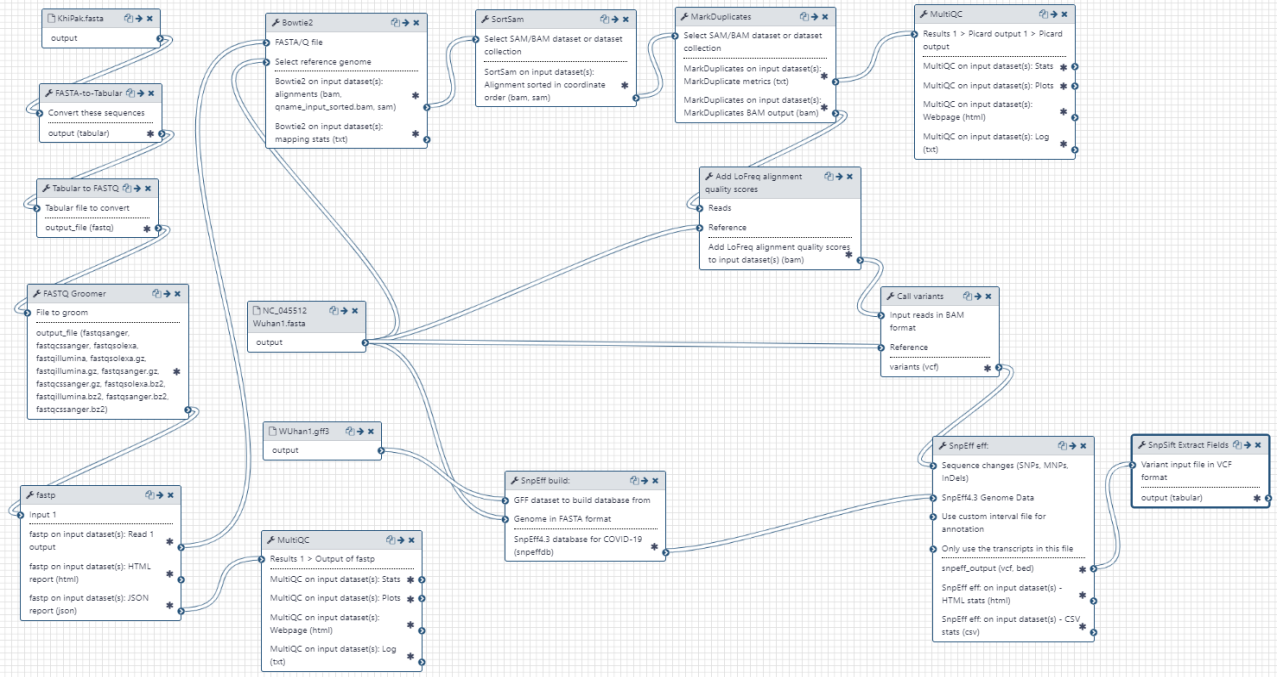

Figure 1: Variant calling workflow run on single-end Illumina read MT500122 (Karachi-Pak) vs NC_045512 (Wuhan1-China).

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 21 June 2020 doi:10.20944/preprints202006.0247.v1

Results and discussion:

Alterations in *ORF3a* and *ORF1ab* genes in Southern Pakistani SARS-nCoV2 strain:

Alignment of MT500122 (Karachi-Pak) SARS-nCoV2 strain whole genome sequence with NC_045512 (Wuhan1-China) strain revealed 4 alterations in accessory proteins, 3 in *ORF1ab* and 1 in *ORF3a* gene (Table 1).

Table 1: Alterations observed in MT500122 (Karachi-Pak) vs NC_045512 (Wuhan1-China) strain.

Gene	Position	REF	ALT	Effects	Variant type
ORF1ab	1604	AATG	A	Codon change plus codon deletion, moderate, aatgac/aac, ND447N.	Del
ORF1ab	1912	C	T	Synonymous codon change, silent mutation, tcC/tcT, S549	SNP
ORF1ab	10582	C	T	Synonymous codon change, silent mutation, gaC/gaT, D3439	SNP
ORF3a	26022	C	T	Synonymous codon change, silent mutation, gaC/gaT, D210	SNP

Reverse genetic studies on *ORF1ab* suggested its modifying role in gene expression, in virulence and virus-host cell interactions besides its prominent role in viral replication [15]. *ORF1ab* gene expresses 16 non-structural polyproteins (nsps1-16). The codon deletion and codon change AATGAC/AAC at position 1604 and SNP at 1912 C/T lies on nsp2 which binds to host prohibitin 1 and 2 proteins (PHB 1,2) and plays a significant role in disrupting host cellular environment while SNP 10582 C/T is present on nsp5 (3C–like proteinase) of *ORF1ab* gene [16]. *ORF3a* on the other hand is located between spike and envelope genes which encode membrane protein and assists viral entry into the host cell by interacting with S protein thus promoting virulence [17]. These variations in both genes of Sothern Pakistani SARS-nCoV2 might causing halt in its replication and also impeding its invasion into the host cell by interacting with the human Angiotensin-converting enzyme 2 (ACE2) receptors on type II alveolar cells and neuropilin-1 (NRP1) respiratory and olfactory epithelium cells receptor, which might be one of the factors contributing to the less pathogenicity and mortality rate due to COVID-19 in Pakistan even after relaxed lockdown measures in the country.

Impacts of variations:

Base substitutions (SNPs), amino acid and codon changes in genes and extent of their impacts are coloured and are graphically presented in (Figure 2).

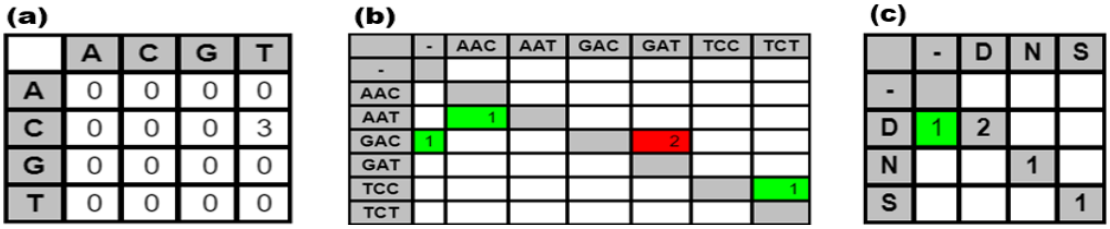

Figure 2: Results of Base substitutions (SNPs) (a), codon change (b), and amino acid alteration (c) in MT500122 (Karachi-Pak) vs NC_045122 (Wuhan1-China) SARS-nCoV2 strain. Rows represents reference while columns present altered bases, codons or amino acids respectively. Red color indicates that more changes happened.

Alterations occurring in *ORF1ab* and *ORF3a* genes in Pakistani SARS-nCoV2 strain have seven types of impacts on 11 genes. Non-coding upstream regulatory regions of Envelope (*E*), Membrane (*M*), Nucleocapsid phosphoprotein (*N*), *ORFs 6, 7a, 7b, 8* and *10* genes are affected by variants which have modifying effects on these genes. In exonic region, four effects are identified in which two variations in *ORF1ab* gene at g.1604, g.1912 and g.26022 of *ORF3a* gene have synonymous alterations with low/silent effect. Remaining exonic region variation in *ORF1ab* gene at g.10582 is disruptive inframe deletion which cause disruptive decrease in coding sequence length of *ORF1ab* gene while one impact of variations is observed in non-coding downstream regulatory region of *ORF1ab* which might have modify the expression of this gene. At last Spike (*S*) gene has also modifying impact in its downstream regulatory region (Table 2) which may have inhibited its activation by conformational changes in its structure thus debilitating its attachment to the host cell receptors ACE2 and NRP1. Still further wet lab experiments are required to corroborate this hypothesis.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 21 June 2020 doi:10.20944/preprints202006.0247.v1

Table 2: Effects of variations in MT500122 (Karachi-Pak) vs NC_045512 (Wuhan1-China) on various genes encoding proteins.

Gene	Variants impact low*	Variants impact moderate*	Variants impact modifier*	Variants effect codon change, codon deletion	Variants effect downstream	Variants effect synonymous coding	Variants effect upstream
E	0	0	1	0	0	0	1
M	0	0	1	0	0	0	1
N	0	0	1	0	0	0	1
ORF10	0	0	1	0	0	0	1
ORF1ab	2	1	1	1	1	2	0
ORF3a	1	0	0	0	0	1	0
ORF6	0	0	1	0	0	0	1
ORF7a	0	0	1	0	0	0	1
ORF7b	0	0	1	0	0	0	1
ORF8	0	0	1	0	0	0	1
S	0	0	1	0	1	0	0

*__low:__ Harmless proteins usually include synonymous variations
*__moderate:__ A missense variant or inframe deletion that might change protein effectiveness
*__modifier:__ Usually exonic variants that may affect non coding genes or downstream gene variant.

The graphical output of number of effects of variations by type and region along with their percentages are shown in (Figure 3).

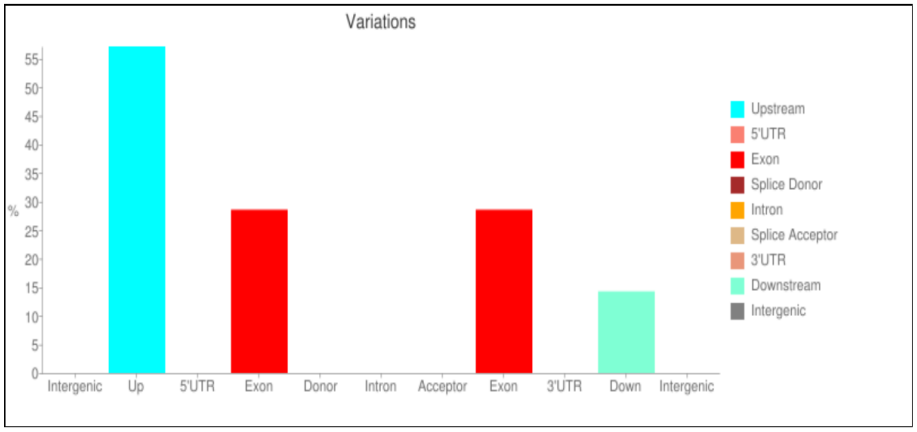

Figure 3: Graphical illustration of variant types on x-axis and its prevalence percentage on y-axis in MT500122 (Karachi-Pak) vs NC_045512 (Wuhan1-China) SARS-nCoV2 strains.

Every 4 variations in which 3 transitions C/T and 1 deletion in MT500122 (Karachi-Pak) strain occurred at an average rate of 7,475 bp. These variations affected 11 genes in total with various types of impact which is elaborated with their occurrence rate across whole genome in (Table 3).

Table 3: Comprehensive detail on MT500122 (Karachi-Pak) variants type, impact and genomic location

MT500122 (Karachi-Pak) vs NC_045512 (Wuhan1-China)					
Type		**Count**	**Percent**		
Low		3	21.429		
Moderate		1	7.143		
Modifier		10	71.429		
Type			**Region**		
Type	**Count**	**Percent**	**Type**	**Count**	**Percent**
Codon change, codon deletions	1	7.143	Downstream	2	14.286
Downstream	2	14.286	Exon	4	28.571
Synonymous coding	3	21.429	Upstream	8	57.143
Upstream	8	57.143			

These findings underscores the significance of variants in genes *ORF1ab* and *ORF3a* in Southern Pakistani SARS-nCoV2 strain that might be affecting their functions of viral replication and entry into the host cell respectively. Moreover, impacts of aforementioned variants on other genes especially the *S* gene which is critically involved in virulency of this virus by attaching to the host cell receptors suggested to be one of the reasons for causing less

pathogenicity in Pakistani population. This in silico analysis on novel genomic variation is a prerequisite for development of PCR-based assays for disease detection and to discriminate native SARS-nCoV2 strains with other circulating strains.

Acknowledgements: Authors are thankful to the Jamil-ur-Rahman Center for Genome Research, Dr. Panjwani Center for Molecular Medicine and Drug Research, International Center for Chemical and Biological Sciences, University of Karachi, Pakistan for their generous sequencing efforts and making their data publically available on NCBI to facilitate the researcher community.

References

1. Tilocca B, Soggiu A, Musella V, Britti D, Sanguinetti M, Urbani A, Roncada P (2020) Molecular basis of COVID-19 relationships in different species: a one health perspective. Microbes Infection
2. Worldometer. https://www.worldometers.info/coronavirus/.
3. Andersen KG, Rambaut A, Lipkin WI, Holmes EC, Garry RF (2020) The proximal origin of SARS-CoV-2. Nature medicine 26 (4):450-452
4. Letko M, Marzi A, Munster V (2020) Functional assessment of cell entry and receptor usage for SARS-CoV-2 and other lineage B betacoronaviruses. Nature microbiology 5 (4):562-569
5. Rothan HA, Byrareddy SN (2020) The epidemiology and pathogenesis of coronavirus disease (COVID-19) outbreak. Journal of autoimmunity:102433
6. Saif R, Mahmood T, Ejaz A. 2020. Whole Genome Comparison of Pakistani Corona Virus with Chinese and US Strains along with its Predictive Severity of COVID-19. bioRxiv. doi:10.1101/2020.05.01.072942.
7. NCBI National Center for Biotechnology Information. https://www.ncbi.nlm.nih.gov/
8. Blankenberg D, Gordon A, Von Kuster G, Coraor N, Taylor J, Nekrutenko A, Team G (2010) Manipulation of FASTQ data with Galaxy. Bioinformatics 26 (14):1783-1785
9. Chen S, Zhou Y, Chen Y, Gu J (2018) fastp: an ultra-fast all-in-one FASTQ preprocessor. Bioinformatics 34 (17):i884-i890
10. Langmead B, Salzberg SL (2012) Fast gapped-read alignment with Bowtie 2. Nature methods 9 (4):357
11. Broad Institute. Picard. http://broadinstitute.github.io/picard/ .
12. Wilm A, Aw PPK, Bertrand D, Yeo GHT, Ong SH, Wong CH, Khor CC, Petric R, Hibberd ML, Nagarajan N (2012) LoFreq: a sequence-quality aware, ultra-sensitive variant caller for uncovering cell-population heterogeneity from high-throughput sequencing datasets. Nucleic acids research 40 (22):11189-11201
13. Cingolani P, Platts A, Wang LL, Coon M, Nguyen T, Wang L, Land SJ, Lu X, Ruden DM (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 6 (2):80-92
14. Ruden DM, Cingolani P, Patel VM, Coon M, Nguyen T, Land SJ, Lu X (2012) Using Drosophila melanogaster as a model for genotoxic chemical mutational studies with a new program, SnpSift. Frontiers in genetics 3:35
15. Graham RL, Sparks JS, Eckerle LD, Sims AC, Denison MR (2008) SARS coronavirus replicase proteins in pathogenesis. Virus research 133 (1):88-100
16. Yoshimoto FK (2020) The Proteins of Severe Acute Respiratory Syndrome Coronavirus-2 (SARS CoV-2 or n-COV19), the Cause of COVID-19. The Protein Journal:1
17. Lu W, Xu K, Sun B (2010) SARS accessory proteins ORF3a and 9b and their functional analysis. In: Molecular Biology of the SARS-Coronavirus. Springer, pp 167-175

 doi:10.20944/preprints202006.0247.v1