

# Whole genome sequence comparison between Pakistani SARS-COV-2 strain MW447609 and mutated South African variant 501y.v2 to predict the severity of disease

Kifayat Ullah\*, Obaidullah Qazi\*\*, Zilwa Mumtaz\*, Zubia Rashid\*\*\*, Muhammad Zubair Yousaf\*

\* KAM School of Life Sciences, Forman Christian College University, Ferozpur road Lahore, 54660, Pakistan

\*\* Department of Microbiology, Abdul wali khan university, Mardan, Pakistan

\*\*\* Pure Health Laboratory, Mafraq Hospital, Abu Dhabi 1227788, United Arab Emirates

**Abstract-** Recently submitted whole genome sequences of the South African variant (501Y.V2) were taken for variant calling analysis to see how this strain is related and discriminated from the Pakistani strain (MW447609). After variant calling of 501Y.V2 (South African) vs MW447609 (Karachi Pak) strain, a total of 104 variants were detected in which 77 SNPs, 22 deletions and 5 insertions were included. The functional effect of these variants was spread between all classes of mutations such as missense (69.3%), nonsense (5.3%) and silent (25.3%). The highest number of changes (deletion) were observed in the upstream (51.04%) and downstream (29.143%) regions. These changes may be causing RBD of 501Y.V2 variant to become more functional and critical in binding with a host. As the objective of the study was to identify genomic variations and mutations to predict the severity of disease. We may say that the disease severity, viral load etc. would not be the same in Pakistan as of South African (501Y.V2) variant caused in south Africa. In addition to this, we also hypothesize that vaccines and other therapeutic interventions would be equally affective to MW447609 (Pakistani) strains in contrast to South African (501Y.V2) variant.

**Index terms-** SARS-COV-2 strain, Whole genome sequence, Variant, Variant calling, South African variant, Genomic variations

## I. INTRODUCTION

SARS-CoV-2 comes in the category of beta corona viruses. Their outermost layer is called envelop which is very important during throughout its life cycle because it protect the virus when it is travelling from one host to another(1). Their genome is positive sense RNA. These viruses use the host cell machinery directly to form protein from their RNA(2). Rise in COVID-19 outbreak in Pakistan within few days after 26 February 2020 was due to pilgrim influx. Before 19 February the outbreak was spread in many cities due to religious tourism(3). The cases of coronavirus rise from 3% to 10% from previous days. On 27 March 2021 more than 450 cases are reported in 24 hours in Pakistan. But the actual number is much more than this. The variant circulating in the 3rd wave is highly contagious than the previous one. It's also deadlier(4). The strain circulating in this wave is UK strain (5).

The south African variant 501Y.V2, has an important mutation at 501 position from N (asparagine) to Y (tyrosine). Basically, this virus gained 23 new mutations regarding the first Wuhan strain.

Out of all these new mutations 20 causes the change in amino acid eight are present in the spike protein(6). As we know eight mutations were present in spike protein, so these mutations were given that much importance. These mutations may help to infect more easily or to survive better. This betterment worse the human survival(7). Another hypothesis was that; this virus may have greater opportunity to escape from antibodies which were produced during previous infection to same patient. It means that if someone is infected in previous episode, he will also be at a risk by new strain(8).

The researcher from south Africa try to study the effect of plasma from previously affected patient from COVID-19, they found a plasma concentration of about 200 fold greater is required to 501Y.V2 for neutralizing and in effectivity(9). Due to virus evading capacity and more transmission rate whole genome sequence analysis is performed with our Pakistani strain (Karachi)(10).

## II. METHODOLOGY

### *Uploading of input files for GATK4 pipeline.*

Recently submitted whole genome paired end data of south African variant (501Y.V2) and Karachi Pakistan strain of corona virus were taken and analyzed by using given methodology for variant calling. Input files of SARS-CoV-2 mutant strain 501Y.V2 were downloaded from European Nucleotide Archive with the accession number as: EBI SRA SRR: 13620326 (read 1 and read 2) & also reference SARS-CoV-2 genome Karachi-Pak (MW447609) in both FASTA and GFF format(11). Full length Homo sapiens genome was also downloaded from NCBI genome database in FNA format(12). After uploading the files, paired end data of variant genome (501Y.V2) was concatenated by using concatenate dataset.

### *Trimming and quality checks of variant data*

After concatenation, FASTA and GFF file of reference genome were combined by using SnpEff build(13). Cutting, trimming, cleaning and duplicate detection of variant data were done by Fastp: (Galaxy Version 0.19.5+galaxy1). MultiQC (1.9 version of Galaxy) were used to sequence quality determination and filtering the variant data for duplication rate(14). Faster download and extract read in fastq were used to fetch data from SRA, NCBI.

### *Mapping and filtering of variant genome (501Y.V2)*

Variant genome was mapped with human genome by BWA-MEM. It aligned the variant genome sequence read against human genome(15). Aligned BAM file was filtered using Filter SAM. By using Samtool fastx unmapped BAM file is converted

to FASTQ format for aligning against reference genome(16). A high speed Bowtie2: (Galaxy Version 2.3.4.3+galaxy0) tool were applied to mapped reads with reference genome.

#### Sorting and duplicate detection in variant strain

Aligned BAM file of variant strain was sorted for adding and replacing of read groups. in this file duplicates were detected and marked by Mark duplicates software(17). It examined the aligned records in BAM datasets and the duplicate molecules were highlighted and flagged.

#### Variant calling of (501Y.V2) against reference genome (MW447609)

Mutect2 software of GATK4 pipeline was used for variant calling of 501Y.V2 against reference genome. Various short variant including insertions and deletions (indels) and single nucleotide variants (SNV) were detected(18).

#### Variation annotation and effect prediction

The variants which were detected in the previous step were annotated by SnpEff eff: (Galaxy Version 4.3+T.galaxy1) 36. It predicted the effects of variants such as change in amino acid and predicted their effect on known genes(13).

### III. RESULTS

#### Whole genome variant calling analysis of SARS-nCov2.

The south African variant genome size is 29,903 bases. After variant calling 104 variants were observed. These variants include insertions, deletions and single SNPs. There variants are spread all over the genome including upstream, downstream regions, ORF1ab, S gene etc.

TABLE 1  
VARIANTS DETAIL OBSERVED IN SRR13620326 (SOUTH AFRICAN) AND MW447609 (KARACHI PAK) STRAIN.

Chromosome	Length	Variants	Variants rate
MW447609.1	29,903	104	287
Total	29,903	104	287

#### Number variants by type

Out of all 104 variants, SNPs were found to be the highest percentage. South African strain differ from Pakistani strain at 77 loci (SNPs). In this study we also observed 5 insertions and 22 deletions.

TABLE 2  
DETAIL OF DIFFERENT VARIANTS OBSERVED BETWEEN THE TWO STRAINS.

Type	Total
SNP	77 (74%)
MNP	0
INS	5 (4.8%)
DEL	22 (21.1%)
MIXED	0
INV	0
DUP	0
BND	0
INTERVAL	0

Type	Total
Total	104

#### Number of effects by impact, functional class and region

Distribution of all variants by impact are given below. The highest impact recorded was a modifier 80.571%. The variants effect by their function were 52 missense, 04 nonsense and 19 silent. The Missense/Silent ratio: 2.7368. These variants were found in all regions and genes. But the highest percentage was upstream and downstream. There were 0.76% stop gain variant and 19.42% in exons.

TABLE 3  
NUMBER OF EFFECTS BY IMPACT, FUNCTIONAL CLASS AND REGION

SRR13620326 (South African) MW447609 (Karachi Pak)					
Type	Count	Percent			
High	26	4.952%			
Low	19	3.619%			
Moderate	57	10.857%			
Modifier	423	80.571%			
Type	Count	Percent			
Missense	52	69.333%			
Nonsense	4	5.333%			
Silent	19	25.333%			
Type		Region			
Type	Region	Percent	Type	Count	Percent
Conservative in frame deletion	2	0.38%	Downstream	153	29.1
Disruptive in frame deletion	3	0.57%			
Downstream gene variant	153	29.14%	Exon	102	19.4
Frameshift variant	22	4.19%			
Intergenic region	2	0.38%			
Missense variant	52	9.91%	Intergenic	2	0.38
Stop gained	4	0.76%			
Synonymous variant	19	3.62%	Upstream	268	51.1
Upstream gene variant	268	51.05%			

#### Altered genes of 501Y.V2 their effect by region, function and impact

Detail analysis revealed a key alteration between two strains. We have observed missense, stop gain and synonymous as well as disruptive and conservative in frame deletion in almost all genes. In this study we have found 1 disruptive in frame deletion and 18 downstream, 26 upstream and 2 missense mutations in E gene. ORF1ab shown 2 conservative inframe deletions, 1 disruptive inframe deletion, 28 downstream, 1 upstream, 14 frameshifts, 28 missenses, 3 stop gained and 9 synonymous mutations. There is a change in S gene which codes for spike protein. These changes include 1 stop gain mutation and 5 synonymous and 12 missense mutations. The details of all genes with change are given below in table 4.6.

TABLE 4: ALTERED GENES OF SRR13620326 (SOUTH AFRICAN FROM MW447609 (KARACHI PAK)

Gene	Bio Type	Variants effect by impact				Variant effect by region				Variant effect by function			
		High	Low	Moderate	Modified	Conservative inframe deletion	Disruptive inframe deletion	Downstream	Frameshift	Missense	Stop gained	Synonymous	Upstream
E	Coding	0	0	3	44	0	1	18	0	2	0	0	26
M	Coding	1	2	3	40	0	0	12	1	3	0	2	28
N	Coding	1	2	3	33	0	0	3	1	3	0	2	30
ORF10	Coding	1	0	1	28	0	1	1	1	0	0	0	27
ORF1ab	Coding	17	9	31	29	2	1	28	14	28	3	9	1
ORF3a	Coding	0	0	4	43	0	0	21	0	4	0	0	22
ORF6	Coding	0	0	0	45	0	0	12	0	0	0	0	33
ORF7a	Coding	0	0	0	44	0	0	12	0	0	0	0	32
ORF7b	Coding	0	0	0	43	0	0	12	0	0	0	0	31
ORF8	Coding	2	1	0	38	0	0	9	2	0	0	1	29
S	Coding	4	5	12	34	0	0	25	3	12	1	5	9

Base

*changes (SNPs) between two strains*

SNPs of south African SRR13620326 variant are represented in table (a) in which change of base which change of base are also counted. Rows represent the reference base and columns represent the changed base. The highest change was observed T (thymidine) to C (cytosine) and also between A (Adenine) to G (guanosine). In table (b) the transition and transversion mutation are given. We have found 55 transition mutations in which a purine base (A-G) changes with purines and pyrimidine (C-T) changes with pyrimidines, 20 transversion mutations in which a purine is changes with pyrimidine and vice versa.

TABLE 5  
COMPARISON BETWEEN SRR13620326 (SOUTH AFRICAN) VS MW447609 (KARACHI PAK). TABLE (A)

	A	C	G	T
A	0	1	18	3
C	1	0	0	12
G	5	1	0	10
T	2	21	3	0

**(b) REPRESENT THE BASE CHANGE (SNP) AND TRANSITION AND TRANVERSION COUNT AND RATIO**

<b>Transitions</b>	56
<b>Transversions</b>	21
<b>Ts/Tv ratio</b>	2.66

*Amino acid changes*

The amino acid changes along with heat map are given below. Rows represent reference amino acids and columns are changed into amino acids. The red color indicates the highest change, and the highest changed was recorded from Phenylalanine to Leucine and from Leucine to Serine. While the rest of the changed amino acid along with their heat map are given below.

	*	-	?	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y	
*																								
-			5																					
?																								
A		1		1															1	1	3			
C		1																1						
D				1					2													1		
E		2							1			3												
F		3						3					5						1		1			
G									1															
H		1																						
I		1									1									1	1			
K	1	2					2								2			1						
L		1						1					2						5					
M											1	1			1						1			
N		2				1						1			1				1					
P													1			3				1				
Q	3											1	1											
R		2																	1	1				
S		3						1			1	1			1					2				
T		1		1							2				1						2			
V		4		1				3														2		
W																				1				
Y		2																						1

Fig 1. Detail of amino acid change between SRR13620326 (South African) vs MW447609 (Karachi Pak)

#### IV. DISSCUSSION

Corona virus is one of the zoonotic origin pathogens affecting humans. These pathogens make 70% of all disease causing pathogen to humans(19). The recent pandemic overlapping the whole world is also caused by one of these pathogen Coronavirus, the pathogen behind COVID-19. In the present day of modern world Next generation sequencing procedures are providing and facilitating researcher and scientists to decrypt hidden depth of virus at genomic level to tackle the disease(20). This coronavirus S protein have a 12 nucleotide insertion in the receptor binding domain that make it possible to easily attach to human angiotensin converting enzyme 2 receptor(21). The evolution of this virus is rapid due to the presence of its RNA genome and also by the presence of RNA dependent RNA polymerase which make mistakes during bases addition(22). Therefore, it is important to sequence and analyze whole genome by that mean we can only be able to understand the common mutation, transmission of virus, spread rate, disease severity as well as viral load and evolution(23).

Another hypothesis was that; this virus may have greater opportunity to escape from antibodies which were produced during previous infection to same patient. It means that if someone is infected in previous episode, he will also be at a risk by new strain(9). Despite from these antibodies the previously

used vaccines will also not at a position to stop the virus. When the researcher from south Africa try to study the effect of plasma from previously affected patient from COVID-19 to measure the neutralizing and ineffective capacity, they found a plasma concentration of about 200 fold greater is required to 501Y.V2 for neutralizing and in effectivity(24).

After variant calling between SRR13620326 (South African) vs MW447609 (Karachi Pak) we identified 104 variants out of them 77 SNPs, 22 deletions and 5 insertions. The effect of these variants was 80% modifier. The functional effect of these variants was spread between all classes of mutations. The missense mutation was detected in higher percentage, 69.3% (52) and silent and nonsense was in 25.3% (17), 5.33% (4). The south African strain is much different from Pakistani strain because various mutation and polymorphism are detected by comparing both strains. May be the gain of stop codon (0.76%) in S protein is likely linked with truncated protein or change in spike protein which make south African strain more transmissible than Pakistani strain(25).

Various deletion is also indicated as compared to Karachi strain of coronavirus. some of this deletion is disruptive and conservative. The highest number of changes are observed in upstream (51.04%) and downstream region (29.143%) and also in exons (19.14%). Only 2 of these are detected in intergenic

regions. We have also detected various SNPs(26). These SNPs include 55 transitions in which purines are changed with purines and pyrimidine are changed with pyrimidine bases. Along with 55 transition we have also identified 20 transversion in south African strain vs Pakistani strain (Karachi). (27) also reported transition is favored over transversion mutations. In these transition there is purine to purine change or pyrimidine to pyrimidine change, while in transversion either the purines are changed with pyrimidine and vice versa(28). May be these SNPs are making south African strain worse than other strains. Highest changed was recorded at the amino acid level from Phenylalanine to Leucine and from Leucine to Serine.

In this study a mutational hotspot is detected in ORF1ab in which 2 conservatives in frame deletions, 1 disruptive in frame deletion, 28 downstream, 1 upstream, 14 frameshifts, 28 missenses, 3 stop gained and 9 synonymous mutations are included. As 14 frameshift mutations were detected in the present study these mutations would be creating a completely different translational product. The stop gained mutations would be creating a truncated protein of ORF1ab. Similar findings also observed in another study by Olabode E. Omotoso and coworker(29). The effect of these mutations would be assessed by further study on computational analysis.

We concluded our discussion by saying that various missense, synonymous, disruptive in frame deletions, stop gain might be

1. Schoeman D, Fielding BC. Coronavirus envelope protein: current knowledge. *Virology journal*. 2019;16(1):1-22.
2. Mallapaty S. Why does the coronavirus spread so easily between people? *Nature*. 2020;579(7798):183-4.
3. Badshah SL, Ullah A, Badshah SH, Ahmad I. Spread of Novel coronavirus by returning pilgrims from Iran to Pakistan. *Journal of Travel Medicine*. 2020;27(3):taaa044.
4. Wallace DJ, Ackland GJ. Abrupt increase in the UK coronavirus death-case ratio in December 2020. *medRxiv*. 2021.
5. Hashmi A. Pakistan increases COVID restrictions amid a third wave 2021 [updated March 15. Available from: <https://www.aljazeera.com/news/2021/3/15/pakistan-increases-coronavirus-restrictions-amid-third-wave>.
6. Tang JW, Toovey OT, Harvey KN, Hui DD. Introduction of the South African SARS-CoV-2 variant 501Y. V2 into the UK. *The Journal of infection*. 2021.
7. Wise J. Covid-19: The E484K mutation and the risks it poses. *British Medical Journal Publishing Group*; 2021.
8. Cele S, Gazy I, Jackson L, Hwa S-H, Tegally H, Lustig G, et al. Escape of SARS-CoV-2 501Y. V2 variants from neutralization by convalescent plasma. *medRxiv*. 2021.
9. Wibmer CK, Ayres F, Hermanus T, Madzivhandila M, Kgagudi P, Oosthuysen B, et al. SARS-CoV-2 501Y. V2 escapes neutralization by South African COVID-19 donor plasma. *Nature medicine*. 2021;27(4):622-5.
10. Cheng MH, Krieger JM, Kaynak B, Arditi MA, Bahar I. Impact of South African 501. V2 variant on SARS-CoV-2 spike infectivity and neutralization: A structure-based computational assessment. *bioRxiv*. 2021.
11. ENA. [Available from: <https://www.ebi.ac.uk/ena/browser/home>.
12. NCBI. [Available from: <https://www.ncbi.nlm.nih.gov/genome/?term=>.
13. Cingolani P, Platts A, Wang LL, Coon M, Nguyen T, Wang L, et al. 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*. 2012;6(2):80-92.
14. Ewels P, Magnusson M, Lundin S, Källner M. MultiQC: summarize analysis results for multiple tools and samples in a single report. *Bioinformatics*. 2016;32(19):3047-8.
15. Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv preprint arXiv:13033997*. 2013.
16. Blankenberg D, Gordon A, Von Kuster G, Coraor N, Taylor J, Nekrutenko A, et al. Manipulation of FASTQ data with Galaxy. *Bioinformatics*. 2010;26(14):1783-5.
17. Langmead B. Aligning short sequencing reads with Bowtie. *Current protocols in bioinformatics*. 2010;32(1):11.7. 1-.7. 4.
18. Benjamin D, Sato T, Cibulskis K, Getz G, Stewart C, Lichtenstein L. Calling somatic SNVs and indels with Mutect2. *Biorxiv*. 2019:861054.
19. Cui J, Li F, Shi Z-L. Origin and evolution of pathogenic coronaviruses. *Nature Reviews Microbiology*. 2019;17(3):181-92.
20. Harari YN. The world after coronavirus. *Financial Times*. 2020;20(03):2020.
21. Hoffmann M, Kleine-Weber H, Krüger N, Müller M, Drosten C, Pöhlmann S. The novel coronavirus 2019 (2019-nCoV) uses the SARS-coronavirus receptor ACE2 and the cellular protease TMPRSS2 for entry into target cells. *BioRxiv*. 2020.

## References

22. Forni D, Cagliani R, Clerici M, Sironi M. Molecular evolution of human coronavirus genomes. *Trends in microbiology*. 2017;25(1):35-48.
23. Lucey M, Macori G, Mullane N, Sutton-Fitzpatrick U, Gonzalez G, Coughlan S, et al. Whole-genome sequencing to track severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) transmission in nosocomial outbreaks. *Clinical Infectious Diseases*. 2021;72(11):e727-e35.
24. Andreano E, Piccini G, Licastro D, Casalino L, Johnson NV, Paciello I, et al. SARS-CoV-2 escape in vitro from a highly neutralizing COVID-19 convalescent plasma. *BioRxiv*. 2020.
25. Hassan SS, Kodakandla V, Redwan EM, Lundstrom K, Choudhury PP, Abd El-Aziz TM, et al. An Issue of Concern: Unique Truncated ORF8 Protein Variants of SARS-CoV-2. *bioRxiv*. 2021.
26. Vankadari N. Overwhelming mutations or SNPs of SARS-CoV-2: A point of caution. *Gene*. 2020;752:144792.
27. Zhao Z, Li H, Wu X, Zhong Y, Zhang K, Zhang Y-P, et al. Moderate mutation rate in the SARS coronavirus genome and its implications. *BMC evolutionary biology*. 2004;4(1):1-9.
28. Alam I, Radovanovic A, Incitti R, Kamau A, Alarawi M, Azhar EI, et al. An interactive COVID-19 virus Mutation Tracker (CovMT) with a particular focus on critical mutations in the Receptor Binding Domain (RBD) region of the Spike protein. 2021.
29. Omotoso OE, Babalola AD, Matareek A. Mutational hotspots and conserved domains of SARS-CoV-2 genome in African population. *Beni-Suef University journal of basic and applied sciences*. 2021;10(1):1-7.
30. Wang WB, Liang Y, Jin YQ, Zhang J, Su JG, Li QM. E484K mutation in SARS-CoV-2 RBD enhances binding affinity with hACE2 but reduces interactions with neutralizing antibodies and nanobodies: Binding free energy calculation studies. *bioRxiv*. 2021.
31. Hotez PJ, Nuzhath T, Callaghan T, Colwell B. COVID-19 Vaccine Decisions: Considering the Choices and Opportunities. *Microbes and Infection*. 2021:104811.
32. Villoutreix BO, Calvez V, Marcelin A-G, Khatib A-M. In silico investigation of the new UK (B. 1.1. 7) and South African (501y. v2) SARS-CoV-2 variants with a focus at the ace2-spike rbd interface. *International journal of molecular sciences*. 2021;22(4):1695.

## AUTHORS

**First Author** – Kafayatullah, M.Phil, KAM School of Life Sciences, Forman Christian College University, Ferozpur road Lahore, 54660, Pakistan.

**Second Author** – Obaidullah Qazi, Department of Microbiology, Abdul wali khan university, Mardan, Pakistan.

**Third Author** – Zilwa Mumtaz, PhD scholar, KAM School of Life Sciences, Forman Christian College University, Ferozpur road Lahore, 54660, Pakistan

**Correspondence Author** – Muhammad Zubair Yousaf, KAM School of Life Sciences, Forman Christian College University, Ferozpur road Lahore, 54660, Pakistan.