Volume 66, Issue 5, 2022
Journal of Scientific Research
Institute of Science,
Banaras Hindu University, Varanasi, India.

Analysis of mutational spectra in *SARS-CoV-2* spike protein and its course of evolution to predict the 'Alarming Variants' for the upcoming wave of pandemic

Shuvam Banerjee*[1], Sreejita Dutta[1,2], Shrinjana Dhar[1], Pritha Bhattacharjee[1]

[1]Environmental Epigenomics Laboratory, Department of Environmental Science, University of Calcutta, 35 Ballygunge Circular Road, Kolkata-700019, shuvam16190@gmail.com, shrinjana.dhar.100@gmail.com, 777.pritha@gmail.com
[2]Institute of Environment and Sustainable Development, Banaras Hindu University, duttasreejita2018@gmail.com
*Environmental Epigenomics Laboratory, Department of Environmental Science, University of Calcutta, 35 Ballygunge Circular Road, Kolkata-700019, shuvam16190@gmail.com

Abstract:**COVID-19 disease extended to different parts of the globe developing variations in the *SARS-CoV-2* viral sequence since its emergence. Some of those variations had functional implications with major concerns, which the WHO termed as 'Variant of Concern' during the second wave of infection. The signatures of VOCs are specific to countries where they are observed and are popularly recognized with the country names. It is noteworthy to mention that some new mutations are always evolving and it is expected that some of these new mutations could play a havoc role in hitting the upcoming wave of the pandemic. This work aimed to list all the mutations of the Spike protein that evolved with time and how they have increased their presence in a given course of time (three-time point analysis) and stabilized in the *SARS-COV-2* infected population. The mutations that will play predominant role in the upcoming pandemic wave are mentioned here:T19I, V213G, D405N, R408S,G142D, G339D, K417N, N440K, D614G, Q954H, N969K mutations in India; G142D, G339D, S373P, S375F, K417N, N440K, S477N, T478K, Q493R, Q498R, N501Y, Y505H, D614G, H655Y, N679K, P681H, N764K, D796Y, Q954H, N969K in South Africa; T95I, G339D, S373P, S375F, N501Y, D614G, H655Y in Brazil; G339D, S373P, S375F, K417N, N440K, N501Y, D614G, P681H in the UK.**

Index Terms:**COVID-19, Pandemic, *SARS-CoV-2*, Spike protein.**

I. INTRODUCTION

The COVID-19 pandemic started spreading from Wuhan, China in late December 2019 and *SARS-CoV-2* is the causative pathogen. To date, the extent of molecular divergence in *SARS-CoV-2* is well perceived (Tang et al. 2020). In general, genomes of Coronaviruses are mostly around 29 to 31 kb long, of which about 2/3rd contains the ORF1ab gene expressing the largest and most complex polyproteins of RNA viruses (Banerjee et al. 2020) and the rest of the part expresses the structural protein spike glycoprotein. *SARS-CoV-2* genome contains a positive-sense single-stranded RNA and is covered by the capsid formed of the nucleocapsid protein (N) which is further packed by three structural proteins viz. membrane protein (M), spike protein (S), and envelope protein (E) (Brian and Baric, 2005).

The spike-like surface glycoprotein or S protein plays a pivotal role in viral transmission into the host by binding with the Angiotensin-converting enzyme 2 (ACE-2) receptor (Du et al. 2009). Classification of newer variants, especially the Variant of Concerns is based on new mutations at the S protein because any mutation or combination of multiple mutations at the S protein can change the conformation of the protein (Stobart et al. 2013). This structural alteration might lead to altered interaction with the receptor or with the antibody in post-vaccination cases. Hence, it is imperative to identify and accumulate the mutations of S protein that evolved gradually since the first wave and stabilized in the population.

Since the onset of COVID-19, the disease extended to different parts of the globe developing variations in its genetic sequence. The chance of mutation increases while a virus spreads widely and causes rapid infections in populations (WHO, 2021). Some of the mutations commonly found in the

first wave have increased their presence in the population during the second wave whereas a few were either abolished or decreased. Among all mutations, synonymous mutations do not make any change to *SARS-CoV-2* proteins, but the missense ones can significantly alter the structure and function of the protein and thus our study focuses on the missense mutations (Wang et al. 2021).

II. OBJECTIVE

The ancestral sequence of Wuhan with the sequence ID NC_045512.2 is considered the reference sequence to identify new mutations in the *SARS-CoV-2* viral sequences. Since the emergence of COVID-19, new viral strains have evolved during the tenure of the first wave, second wave and third wave. The World Health Organization (WHO) coined some of those variants like the 'Variant of Concern' (VOC) and those became the focus of the mutation study in present days. Each VOC are known to carry a combination of newer mutations that were not found in the early days of the onset of the pandemic. With this background, the present work aims to accumulate all the mutations of the S protein that evolved with time and became stable in the infected population. Along with this, it tries to recognize the mutations that might be playing the most crucial and dominant role in the upcoming wave of COVID-19 infection.

III. METHODOLOGY

A. Listing of known mutations at S protein from Literature review

Articles on *SARS-CoV-2* mutation and its variants published till 31st May 2021 were filtered and extracted from "PubMed", "Scopus", "Web of Science" and "Google Scholar" and were systematically reviewed. The keywords include Medical Subject Headings (MeSH) terms like "COVID-19, "Mutation", "Spike glycoprotein", "*SARS-CoV-2* variants", and "Variant of Concerns", "Evolution", and "Spike protein mutations" (Abdullahi et al. 2020). Articles describing mutations, genetic diversity and variations at the spike protein of *SARS-CoV-2* were consideredandevaluated to list out the mutations already reported at the S protein.

B. Retrieval of S protein sequences from the database

Simultaneously, sequences were retrieved from the "NCBI Virus" or "GISAID" database in three phases, specific input was "*SARS-CoV-2*" and "Spike Glycoprotein". In the case of the NCBI virus database, the output was refined with sequence lengths 1250 to 1280, as the length of the wild type sequence of the targeted S protein is 1273.

C. Multiple sequence alignment (MSA) and analysis of mutational spectra

Clustal Omega was employed to align multiple sequences retrieved from the database, against the ancestral sequence of Wuhan (NC_045512.2 as reference sequence) and screened the alignment result to bridge the missing information of the viral S protein evolution and to include those mutations (if any) that have not been reported yet in any published articles. Three phases were analyzed; one phase spans between the 1st and 15th March 2021 (considering the sequences submitted till February), the second phase was between 11th and 25th June (considering sequences submitted between 1st April and 10th June 2021) and the third phase spans between 16th and 31st March (considering sequences submitted between 25th and 28th February 2022). During Multiple Sequence Alignment (MSA), the Gap opening penalty and gap extension penalty were set at 12 and 2 respectively to ensure that unnecessary gaps are not created and alterations are observed easily. MSA results were thoroughly screened to find out the exact location of the mutations and alterations linked to that position.

D. Identification of mutations to play a decisive role in the upcoming wave of infection

The number of occurrences of each mutated variant in phase-1, phase-2 and phase-3 was calculated (suppose a, b and c) and then was divided by the total number of sequences analyzed in that phase (suppose x in phase-1,y in phase-2 and z in phase-3). The proportion of each mutated variant would be a/x, b/y and c/z in phases-1,phase-2 and phase-3 respectively. Those mutant variants with a significant increase in proportion in this tenure are expected to play a crucial role in the upcoming wave of COVID-19.

IV. THE VARIANT OF CONCERNS (VOC) AND ALLIED MUTATIONS AT THE SPIKE PROTEIN

The VOC of the United Kingdom (UK), South Africa, Brazil and India – all carry some signature mutations at the spike protein and the viral strains were being classified based on the presence of those mutations.

The UK was hit by the second wave of the pandemic in the last week of September 2020 and the infection curve depicting new cases/day took a sharp increase from the first week of December and reached its top by 8th January 2021 (68053 new cases/day). The UK lineage (also known as B.1.1.7) was reported in December 2020 and it has been estimated as a quickly spreading strain dominantly circulating in England and its surroundings. Signature mutations of this variant include N501Y in spike protein that causes a conformational change in the receptor-binding domain (RBD). This variant also possesses

a deletion mutation at positions 69 and 70 (del69-70) which is resulted in increased transmissibility (Public Health England, 2020; Volz et al. 2021). Along with these mutations, D614G and E484 are also commonly reported from the UK and 614G has now become a worldwide dominant variant of *SARS-CoV-2*.

New cases/day in South Africa took a sharp rise since the first week of December 2020 and reached the maximum on 8th January 2021 (21980 new cases/day) during the second wave. Although the South African strain B.1.351 is an independent strain found in late December 2020, it shares some common mutations with other VOCs. E484K and N501Y at the spike protein are the major characteristic mutations of this strain, but it doesn't contain del69-70 like B.1.1.7. The E484K mutation may be responsible for the neutralization of monoclonal antibodies (Weisblum et al. 2020; Resende et al. 2021) though it is imprecise whether this variant has any role to increase disease severity.

On 2nd January 2021, the number of the new cases reported in Brazil was 15827 and since then the infection rate started to increase again, which may be considered the second wave and the number of new cases detected in a single day crossed the bar of 1 lakh on 25th March and it is still the highest number in Brazil till date. The new *SARS-CoV-2* variant of Brazil, denoted by P.1 was first detected in Japan on 6th January from the samples collected from four travellers arriving from Brazil at the Haneda airport in Tokyo. The main characteristics mutations for this strain are K417T, E484K, and N501Y in the spike protein however a few more are included under the P1 lineage.

The number of cases of coronavirus infection had started to decline in India since September 2020 and dropped to less than 10000 new cases/day during January (on January 25, the number of new infections reported in the whole country was only 9102). The number of new infections in India began to increase again exponentially from the first week of March as the second wave struck. The number of new infections on March 1 was 12,286, rising to 1, 15, 736 by April 6 and reaching a maximum of 4,14,188 on May 6. That 4, 14, 188 new cases in a single day are the highest reported number till now and the rate of infection is decreasing in India since then.

The Delta variant (also known as B.1.617.2) of India which is part of the B.1.617 lineage, is the fourth one to be declared a variant of global concern by WHO, along with the other VOCs discussed above (Ralph, 2021). It carries L452R, T478K and P681R mutations (Centres for Disease Control and Prevention, 2021), but unlike B.1.617.1 it does not carry E484Q. Along with these mutations mentioned, D614G is also dominant in the Indian population and part of the same lineage. Indian delta variant has also been referred to as a "double mutant", due to the presence of both E484Q and L452R mutation.

These two mutations have individually been found in several other coronavirus variants, but the presence of both these mutations together was found in some coronavirus genomes from India and thus "double mutant" term was coined.

V. Results and Discussion of the Mutational Spectra Analysis

The present study attempted to identify the potential VOCs of *SARS-CoV-2* in the upcoming wave of the COVID-19 pandemic. The characteristics of the virus (in terms of infectivity rate and mortality rate) have been altered with continuous evolution and possess some unique genetic variations.

In the Indian population, a total of 31 mutations were found by analysing three consecutive phases (Table I). An N>K change at position 440of the S protein (which belongs totheReceptor Binding Motif region of the spike protein) was found in 7.5% of the samples in phase-1 and 100% of the samples in phase-3, but not found in our study population during phase-2 analysis. Unlikely, in the N-terminal Domain of spike protein, two mutations, i.e, 154 (E>K) and 95 (T>I) were observed in 14% and 11.5% of the samples respectively during phase-1; these increased their presence by 1.5 fold during phase-2 but diminished in phase-3. Similarly, A570D, P681H and D1118H were found in phase-1 and 2 but absent in phase-3. The only variation which is consistently present (100% of the samples) in the three phases is D614G. During phase-3, 11 new mutations were found among which 7 (G142D, G339D, K417N, N440K, D614G, Q954H and N969K) were the signature mutation of the Omicron variant. Apart from signature variations, some novel mutations like T19I, V213G, D405N and R408S were found in 100% of the studied samples indicating a strong chance to be present in the upcoming *SARS-CoV-2* strain.

Table I: Mutations at Spike Protein in Indian Population and proportional presence of those mutations during phase-1, phase-2 and phase-3 of the analysis

Mutation Position (nucleotide)	Mutation Position (Amino Acid)	% of the population carrying Mutated Variant in phase-1 (N=200)	% of the population carrying Mutated Variant in phase-2 (N=83)	% of the population carrying Mutated Variant in phase-3 (N=140)
C21618T	T19I	0	0	100
21633-641	L24del-P26del	0	0	97.14
21802-846	P80del-T95del	0	0	38.57
C21846T	T95I	11.5	15.7	0
G21987A	G142D	0	0	100
T22200G	V213G	0	0	100
G22022A	E154K	14	20.5	0
G22578A	G339D	0	0	100
22581-22623	V341del-N354del	0	0	72.86
G22775A	D405N	0	0	100
A22786C	R408S	0	0	100
G22813T	K417N	0	0	100
T22882G	N440K	7.5	0	100
T22917G	L452R	11	50.6	0
G23012C	E484Q	15.5	42.2	0
A23063T	N501Y	5.5	32.5	0
C23271A	A570D	6.5	27.7	0
A23403G	D614G	100	100	100
C23525T	H655Y	0	0	0
G23593T	Q677H	8	0	0
T23599G	N679K	0	0	0
C23604G	P681R	13	47	0
C23604A	P681H	12.5	28.9	0
C23854A	N764K	0	0	0
G23948T	D796Y	0	0	0
A24424T	Q954H	0	0	100
T24469A	N969K	0	0	100
G24914C	D1118H	6	27.7	0
21767-69	H69-70V	1	26.5	0
21992-94	Y144	0	12.1	0
21995-97	Y145	0	16.7	0

69-70HV deletion, a signature mutation for the UK variant B.1.1.7, was observed in only 1% population in India during phase-1, 26.5% during phase-2 and absent in phase-3. Similarly, 144-145YY deletion was a newly evolved mutation found at around 12.05% and 16.7% of cases during phase-2 but it was not present in phase-3. Hence, it can be predicted that these mutations will not carry forward to the upcoming new strain.

D80A, D215G, K417N, E484K, N501Y, D614G and A701V were the signature missense mutations of the South African variant B.1.351. After analysis of three consecutive phases of infection, a total of 41 mutations were retrieved among which 20 (G142D, G339D, S373P, S375F, K417N, N440K, S477N, T478K, Q493R, Q498R, N501Y, Y505H, D614G, H655Y, N679K, P681H, N764K, D796Y, Q954H and N969K) were the signature mutations of Omicron (Table II). It has been seen that

some mutations like D80A, D215G,E484K, and A701V were found in a comparatively higher proportion in earlier phases, but they were completely absent in the post Omicron phase. Two mutations, i.e, N501Y and D614G have been consistently present in three phases indicating their strong presence in the future strains.

Table II: Mutations in Spike Protein in South African Populationand proportional presence of those mutations during phase-1, phase-2 and phase-3 of the analysis

Mutation Position (nucleotide)	Mutation Position (Amino Acid variation)	% of the population carrying Mutated Variant in phase-1 (N=139)	% of the population carrying Mutated Variant in phase-2 (N=59)	% of the population carrying Mutated Variant in phase-3 (N=114)
C21614T	L18F	28.78	28.8	0.00
C21618T	T19I	0	0	89.47
C21618G	T19R	0	10.2	0.00
A21792C	K77T	0	10.2	0.00
A21801C	D80A	96.4	76.3	0.00
G21974T	D138Y	0	6.8	0.00
G21987A	G142D	0	10.2	100.00
T22200G	V213G	0	0	89.47
A22206G	D215G	97.1	76.3	0.00
G22578A	G339D	0	0	100.00
T22679C	S373P	0	0	100.00
C22686T	S375F	0	0	100.00
G22775A	D405N	0	0	73.68
A22786C	R408S	0	0	42.11
G22813T	K417N	0	0	70.18
T22882G	N440K	0	0	64.91
G22813T	K417N	97.8	72.9	0.00
G22992A	S477N	0	0	82.46
C22995A	T478K	0	0	82.46
G23012A	E484K	97.1	71.2	0.00
A23040G	Q493R	0	0	75.44
A23055G	Q498R	0	0	82.46
A23063T	N501Y	97.8	79.7	82.46
T23075C	Y505H	0	0	82.46
C23271A	A570D	0	10.2	0.00

A23403G	D614G	71.9	100	96.49
C23525T	H655Y	0	0	96.49
T23599G	N679K	0	0	96.49
C23604A	P681H	0	0	96.49
C23664T	A701V	97.8	74.6	0.00
C23709T	T716I	0	10.2	0.00
C23854A	N764K	0	0	71.93
G23948T	D796Y	0	0	91.23
G24410A	D950N	0	8.5	0.00
A24424T	Q954H	0	0	92.98
T24469A	N969K	0	0	92.98
T24506G	S982A	0	10.2	0.00
G24914C	D1118H	0	10.2	0.00
21767-69	H69-70V Del	1	6	0.00
21992-94	Y144 Del	1	6	0.00
21995-97	Y145 Del	0	0	0.00

L18F, T20N, P26S, D138Y, R190S, K417T, E484K, N501Y, D614G, H655Y, T1027I are the already known mutations that are part of the P.1 lineage of Brazil. But analysis of three consecutive phases revealed 37 mutations, among them, 24 mutations were the part of Omicron (Table III). In this population, three mutations, i.e, N501Y, D614G, and H655Y were present in all three phases with good percentage. Therefore, it can be confirmed that these mutations have become stabilized in the population and will be present in the upcoming *SARS-CoV-2* strains. Besides these, the signature mutations of Omicron were found in more than 70% of the studied samples, indicating their chance of presence in the next phases too. However, mutations like L18F, T20N, P26S, D138Y, etc. were found in 100% of the studied cases during phase-2 but abolished in phase-3 which means that these mutations will not pass through the next strains.

Table III: Mutations at Spike Protein in Brazil Population and proportional presence of those mutations during phase-1, phase-2 and phase-3 of the analysis

Mutation Position (nucleotide)	**Mutation Position (Amino acid)**	**% of the population carrying Mutated Variant in phase-1 (N=120)**	**% of the population carrying Mutated Variant in phase-2 (N=104)**	**% of the population carrying Mutated Variant in phase-3 (N=134)**
C21614T	L18F	35.8	100	0.00
C21621A	T20N	33.3	98	0.00
C21638T	P26S	35.8	100	0.00
C21762T	A67V	0	0	91.04
C21846T	T95I	0	0	91.04

G21974T	D138Y	36.7	100	0.00
G22132T	R190S	35.8	99	0.00
G22578A	G339D	0	0	97.01
T22679C	S373P	0	0	95.52
C22686T	S375F	0	0	95.52
A22812C	K417T	35.8	99	0.00
G22813T	K417N	0	0	89.55
T22882G	N440K	0	0	83.58
G22992A	S477N	0	0	70.15
C22995A	T478K	0	0	70.15
G23012A	E484K	84.1	99	0.00
A23040G	Q493R	0	0	74.63
G23048A	G496S	0	0	74.63
A23055G	Q498R	0	0	74.63
A23063T	N501Y	36.7	100	74.63
T23075C	Y505H	0	0	74.63
C23202A	T547K	0	0	91.04
A23403G	D614G	100	100	100.00
C23525T	H655Y	36.7	100	100.00
C23604A	P681H	0	0	100.00
C23854A	N764K	0	0	100.00
G23948T	D796Y	0	0	100.00
C24130A	N856K	0	0	94.03
A24424T	Q954H	0	0	100.00
T24469A	N969K	0	0	100.00
C24503T	L981F	0	0	94.03
C24642T	T1027I	36.7	99	0.00
G25088T	V1176F	90.8	100	0.00
21767-69	H69-70V Del	3	0	0.00
21992-94	Y144 Del	1	1	0.00
21995-97	Y145 Del	0	1	0.00
21992-97	Y144-145 Del	0	1	0.00

While studying the UK population, a total of 30 mutations were retrieved by accumulating three pandemic phases that include 19 signature mutations of the Omicron variant. Table IV showed that three mutations (N501Y, D614G and P681H) were consistently present in all three phases and these were found in 100% of the studied population during the 3rd phase. Along with these, the signature mutations of the Omicron variant were found in more than 98% of the cases. Hence, it can be inferred that these mutations will be present in the upcoming strains of SARS-CoV-2.

Table IV: Mutations at Spike Protein in UK Population and proportional presence of those mutations during phase-1, phase-2 and phase-3 of the analysis

Mutation Position (nucleotide)	Mutation Position (Amino Acid)	% of the population carrying Mutated Variant in phase-1 (N=334)	% of the population carrying Mutated Variant in phase-2 (N=101)	% of the population carrying MutatedVariant in phase-3 (N=112)
C21614T	L18F	21.56	0.99	0
C21618T	T19I	0	0	69.64
21633-641	L24del-P26del	0	0	69.64
T22200G	V213G	0	0	69.64
C22227T	A222V	20.36	0	0
G22578A	G339D	0	0	100
T22679C	S373P	0	0	100
C22686T	S375F	0	0	100
G22775A	D405N	0	0	69.64
A22786C	R408S	0	0	42.86
G22813T	K417N	0	0	100
T22882G	N440K	0	0	98.21
G22992A	S477N	0	0	100
C22995A	T478K	0	0	100
A23040G	Q493R	0	0	100
A23055G	Q498R	0	0	100
A23063T	N501Y	69.5	87.1	100
T23075C	Y505H	0	0	100
C23271A	A570D	69.5	94.1	0
A23403G	D614G	84.13	100	100
C23525T	H655Y	0	0	100
T23599G	N679K	0	0	100
C23604A	P681H	70.36	93.1	100
C23709T	T716I	70.36	93.1	0
C23854A	N764K	0	0	100
G23948T	D796Y	0	0	100
A24424T	Q954H	0	0	100
T24469A	N969K	0	0	100
T24506G	S982A	70.06	94.1	0
G24914C	D1118H	70.06	93.1	0

CONCLUSION

D405N, D614G, Q954H and N969Kare the mutations that have increased their presence in India at a significant level (Fig. 1) and all these mutated variants are expected to dominate in the coming days and shall be playing an important role in the upcoming wave of the *SARS-CoV-2*infection.In Brazil, the mutant variant of N501Y, D614G, and H655Y along with a few novel mutations have occupied the majority of the genomes of the infected population and are crucial for the upcoming wave of infection. On the other hand, G142D, G339D, S373P, S375F, K417N, N440K, S477N, T478K, Q493R, Q498R, N501Y, Y505H, D614G, H655Y, N679K, P681H, N764K, D796Y, Q954H and N969K are the mutations that are increasing their presence in the South African population at a faster rate. In the United Kingdom, the mutant variant of N501Y, D614G and P681H mutations were the dominant variants in the third wave

and now all these variants are found in more than 85% of the population and are expected to dominate during the next wave also.

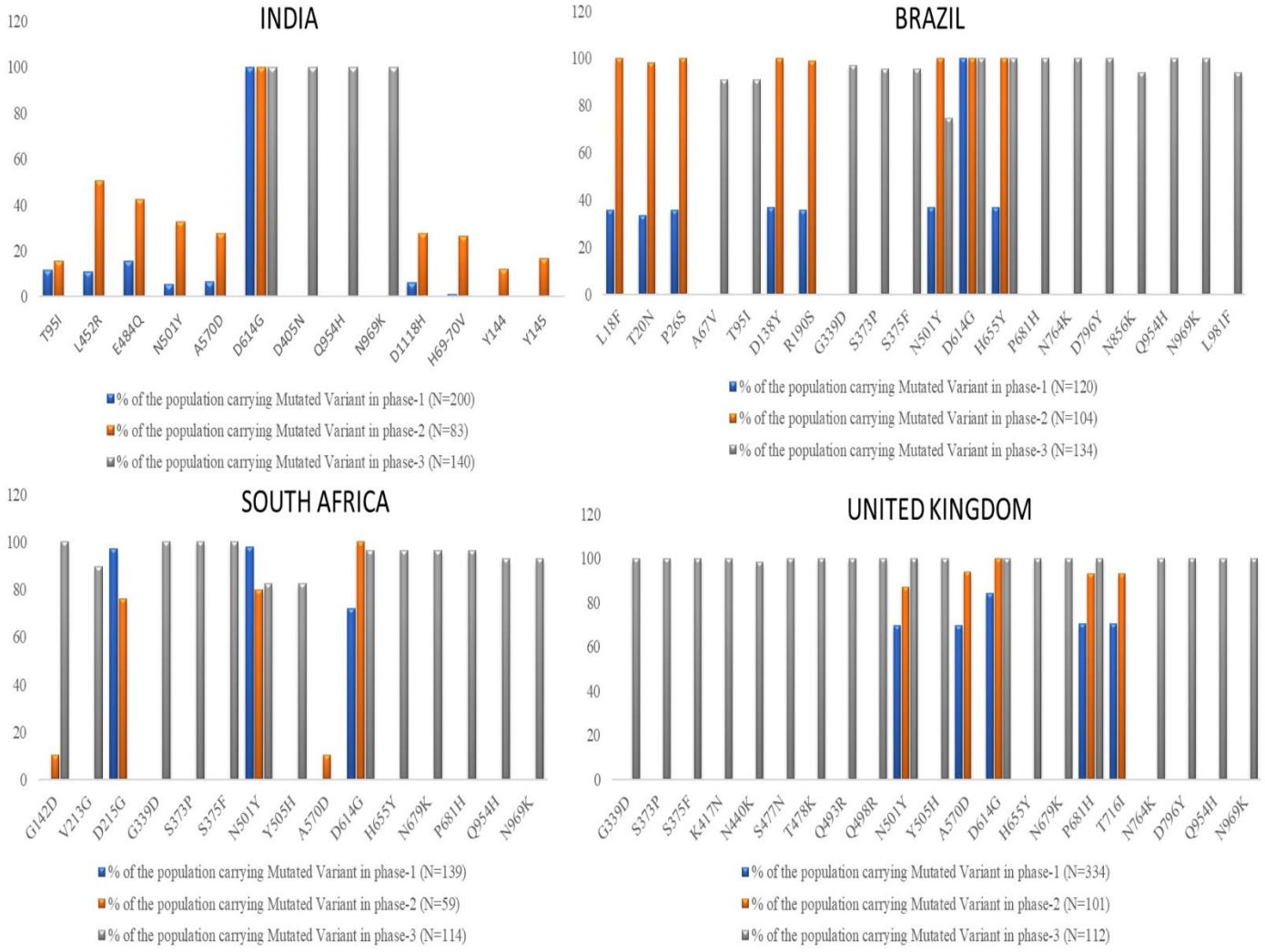

Fig. 1. Graphical representation of the signature mutations of South Africa, India, Brazil and UK that may play a crucial role in the upcoming Wave of COVID-19 infection (Gray bar indicating the highly stabilized mutations which have a strong possibility to be present in the upcoming *SARS-CoV-2* strains).

References

Abdullahi I, Emeribe A, Ajayi O et al. (2020) Implications of *SARS-CoV-2* genetic diversity and mutations on pathogenicity of the COVID-19 and biomedical interventions. J Taibah Univ. Medical Sci. 15(4): 258-264.

Banerjee S, Seal S, Dey R et al. (2020) Mutational spectra of *SARS-CoV-2* orf1ab polyprotein and signature mutations in the United States of America. J. Med. Virol.

Brian D. A., Baric R. S. (2005). Coronavirus genome structure and replication. Curr. Topics Microbiol. Immunol. 287, 1–30. doi: 10.1007/3-540-26765-4_1.

Du L, He Y, Zhou Y et al. (2009) The spike protein of SARS-CoV – a target for vaccine and therapeutic development. Nat. Rev. Microbiol.

Investigation of novel *SARS-CoV-2* variant: variant of concern 202012/01, technical briefing 3. London, United Kingdom: Public Health England; 2020. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/950823/Variant_of_Concern_VOC_202012_01_Technical_Briefing_3_-_England.pdf

Ralph Ellis. WHO Says Indian COVID Strain ‘a Variant of Concern’. May11, 2021, Webmd.com. (https://www.webmd.com/lung/news/20210510/who-says-indian-covid-strain-a-variant-of-concern).

Resende PC, Bezerra JF, de Vasconcelos RHT, at al. Spike E484K mutation in the first *SARS-CoV-2* reinfection case confirmed in Brazil, 2020external icon. [Posted on virological.orgexternal icon on January 10, 2021]

SARS-CoV-2 Variant Classifications and Definitions. CDC.gov. Centres for Disease Control and Prevention. 12 May 2021. Retrieved 16 May 2021

Stobart CC, Sexton NR, Munjal H, et al. (2013) Chimeric exchange of coronavirus nsp5 proteases (3CLpro) identifies common and divergent regulatory determinants of protease activity. J Virol. 87(23): 12611-12618. https://doi.org/10.1128/JVI.02050-13

Tang X, Wu C, Li X et al. (2020) On the origin and continuing evolution of *SARS-CoV-2*. National Science Review. 7(6):1012–1023

The effects of virus variants on COVID-19 vaccines; March, 2021; World Health Organisation

Volz E, Mishra S, Chand M, et al. (2021) Transmission of *SARS-CoV-2* lineage B.1.1.7 in England: insights from linking epidemiological and genetic data. medRxiv [Preprint posted online January 4, 2021]. https://www.medrxiv.org/content/10.1101/2020.12.30.20249034v2external icon

Wang R, Chen J, GaoKet al. (2021) Analysis of *SARS-CoV-2* mutations in the United States suggests presence of four substrains and novel variants. Commun Biol. 4(1):311. doi: 10.1038/s42003-021-01867-y.

Weisblum Y, Schmidt F, Zhang F, et al. (2020) Escape from neutralizing antibodies by *SARS-CoV-2* spike protein variantsexternaliconexternal icon. eLife 9:e61312.

***