Genetic epidemiology of variants associated with immune escape from global SARS-CoV-2 genomes

Bani Jolly[1,2,$], Mercy Rophina[1,2,$], Afra Shamnath[1], Mohammed Imran[1,2], Rahul C. Bhoyar[1], Mohit Kumar Divakar[1,2], Pallavali Roja Rani[3], Gyan Ranjan[1,2], Paras Sehgal[1,2], Pulala Chandrasekhar[3], S. Afsar[3], J. Vijaya Lakshmi[3], A. Surekha[3], Sridhar Sivasubbu[1,2], Vinod Scaria[1,2,*]

[1]*CSIR Institute of Genomics and Integrative Biology (CSIR-IGIB), New Delhi, India*
[2]*Academy of Scientific and Innovative Research (AcSIR), CSIR-HRDC Ghaziabad, Uttar Pradesh, India Ranjan, P. Sehgal, S. Sivasubbu, V. Scaria)*
[3]*Kurnool Medical College, Kurnool, Andhra Pradesh, India*

[$]Authors contributed equally and would like to be known as joint first authors
*Address for correspondence: Vinod Scaria, vinods@igib.in

Abstract
Many antibody and immune escape variants in SARS-CoV-2 are now documented in literature. The availability of SARS-CoV-2 genome sequences enabled us to investigate the occurrence and genetic epidemiology of the variants globally. Our analysis suggests that a number of genetic variants associated with immune escape have emerged in global populations.

Keywords: COVID-19, SARS-CoV-2, Antibody, Mutations, Epidemiology

Text
Antibodies are one of the emerging therapeutic approaches being explored in COVID-19. These antibodies typically target the receptor-binding motif or structural domains of the Spike protein of SARS-CoV-2, in an attempt to inhibit binding of Spike protein with the host receptors. Cocktails of antibodies which target distinct structural and functional domains of spike proteins are also being currently developed considering redundant mechanisms of targeting the virus and therefore minimising escape mechanisms. Genomic documentation of the spread of SARS-CoV-2 across the globe has provided unique insights into the genetic variability and variants of functional consequence. In-depth studies in recent months have unravelled a wealth of information on the immune response in COVID-19 and offered insights into the development of therapeutics.

Recent investigations suggest a number of genetic variants in SARS-CoV-2 are associated with immune escape and/or resistance to antibodies. Their structural and functional features and mechanisms of immune evasion are also being extensively studied (*1*) . The natural occurrence and genetic epidemiology of these variants across the global populations are poorly understood. We were motivated by the wide availability of SARS-CoV-2 genomes from across the world and the increasing numbers of genetic variants suggested to contribute to escape from antibody inhibition.

We analysed a comprehensive compendium of genetic variants associated with immune escape and curated by our group from literature and preprint servers (*2*). This compendium included 120 unique variants reported in literature. To understand the genetic epidemiology of these variants in the global compendium of genomes, we compiled the dataset of

265,079 SARS-CoV-2 from GISAID (as of 17 December 2020) (*3*) apart from 1,154 genomes sequenced in-house (BioProject ID: PRJNA655577). Genome sequences with more than 5% Ns, more than 10 ambiguous nucleotides, higher than expected divergence and mutation clusters were excluded from the analysis. After quality control, the final dataset encompassed 240,133 genomes from 133 countries. Only countries with at least 100 good quality genome submissions were considered for the analysis.

86 of the 120 genetic variants associated with immune escapes were found in a total of 26,917 genomes from 63 countries (**Figure 1A**), out of which 9 variants had >1% frequency in the respective countries. Phylogenetic analysis was performed following the Nextstrain protocol for a total of 3,679 genomes, including 1,501 randomly selected genomes having these variants (**Figure 1B**) (*4*). Homoplasies were identified in the phylogeny using HomoplasyFinder (*5*). Out of 86, 43 variant sites were found to be homoplasic, suggesting they could emerge independently in different genetic lineages, out of which 9 were found to be at >1% frequency in at least one of the countries analysed.

Out of 14,222 genomes analysed from Australia, 24 immune escape associated variants mapped to 9,895 genomes (70%). Of significant frequency was the S:S477N variant which was present in 9,541 genomes (67%) from Australia. High frequency of this variant was also found in a number of other countries particularly in Europe. S:N439K was also found at high frequencies in genomes from a number of countries in Europe (*6*).

S:N501Y, one of the variants in the recently reported emergent SARS-CoV-2 lineage from the United Kingdom, was present in a total of 290 genomes, including genomes from the United Kingdom, Australia, South Africa, USA, Denmark and Brazil (*7,8)*. All 7 genomes from South Africa having S:N501Y also had the S:E484K variant and S:K417N was present in 2 of these genomes (*9*).

The ORF3a:G251V variant was also found to be prevalent across global genomes, with the highest frequencies in Hong Kong and South Korea. This variant is also one of the defining variants for the Nextstrain clade A1a (GISAID Clade V) (**Figure 1B**).

19 of the 86 genetic variants were found in genomes from India (**Supplementary Figure**). The S:N440K variant was found to have a frequency of 2.1% in India and a high prevalence in the state of Andhra Pradesh (33.8% of 272 genomes). The variant site was homplasic and the variant was found in genomes belonging to different clades and haplotypes. Time-scale analysis suggested the variant emerged in recent months (**Figure 1C**). The S:N440K variant was also reported in a case of COVID-19 reinfection from North India (*10*).

Put together, our analysis suggests that a number of genetic variants which are associated with immune escape have emerged in global populations, some of them have been found to be polymorphic in many global datasets and a subset of variants have emerged to be highly frequent in some countries. Homoplasy of the variant sites suggests that there could be a potential selective advantage to these variants. Further data and analysis would be needed to investigate the potential impact of such variants on the efficacy of different vaccines in these regions.

Acknowledgements
Authors acknowledge Disha Sharma and Abhinav Jain for the analysis of in-house genomes and the researchers, originating and submitting laboratories of the sequences retrieved from GISAID (https://doi.org/10.6084/m9.figshare.13365503.v2). BJ and MKD acknowledge a research fellowship from the Council of Scientific and Industrial Research (CSIR India). The funders had no role in the study design or the decision to publish.

References

1. Weisblum Y, Schmidt F, Zhang F, DaSilva J, Poston D, Lorenzi JCC, et al. Escape from neutralizing antibodies by SARS-CoV-2 spike protein variants. 2020 Oct 28 [cited 2020 Dec 22]; https://elifesciences.org/articles/613121

2. Rophina M, Pandhare K, Mangla M, Shamnath A, Jolly B, Sethi M, et al. FaviCoV - a comprehensive manually curated resource for functional genetic variants in SARS-CoV-2. 2020 Nov 17 https://doi.org/10.31219/osf.io/wp5tx

3. Yuelong Shu JM. GISAID: Global initiative on sharing all influenza data – from vision to reality. Eurosurveillance [Internet]. 2017 Mar 30 [cited 2020 Dec 9];22(13). https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5388101/

4. Nextstrain [Internet]. [cited 2020 Dec 9]. https://nextstrain.org/sars-cov-2/

5. Crispell J, Balaz D, Gordon SV. HomoplasyFinder: a simple tool to identify homoplasies on a phylogeny. Microbial Genomics [Internet]. 2019 Jan [cited 2020 Dec 9];5(1). https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6412054/

6. Hodcroft EB, Zuber M, Nadeau S, Crawford KHD, Bloom JD, Veesler D, et al. Emergence and spread of a SARS-CoV-2 variant through Europe in the summer of 2020. medRxiv : the preprint server for health sciences [Internet]. 2020 Nov 27 [cited 2020 Dec 9]; https://pubmed.ncbi.nlm.nih.gov/33269368/

7. Rambaut A, Loman N, Pybus O, Barclay W, Barrett J, Carabelli A, et al. Preliminary genomic characterisation of an emergent SARS-CoV-2 lineage in the UK defined by a novel set of spike mutations [Internet]. 2020 [cited 2020 Dec 22]. https://virological.org/t/preliminary-genomic-characterisation-of-an-emergent-sars-cov-2-lineage-in-the-uk-defined-by-a-novel-set-of-spike-mutations/563

8. Shang E, Axelsen PH. The Potential for SARS-CoV-2 to Evade Both Natural and Vaccine-induced Immunity [Internet]. Cold Spring Harbor Laboratory. 2020 [cited 2020 Dec 24]. p. 2020.12.13.422567. https://www.biorxiv.org/content/10.1101/2020.12.13.422567v1.abstract

9. Emergence and rapid spread of a new severe acute respiratory syndrome-related coronavirus 2 (SARS-CoV-2) lineage with multiple spike mutations in South Africa [Internet]. [cited 2020 Dec 22]. https://www.krisp.org.za/publications.php?pubid=315

10. Gupta V, Bhoyar RC, Jain A, Srivastava S, Upadhayay R, Imran M, et al. Asymptomatic reinfection in two healthcare workers from India with genetically distinct SARS-CoV-2. Clin Infect Dis [Internet]. [cited 2020 Dec 9]; https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7543380/

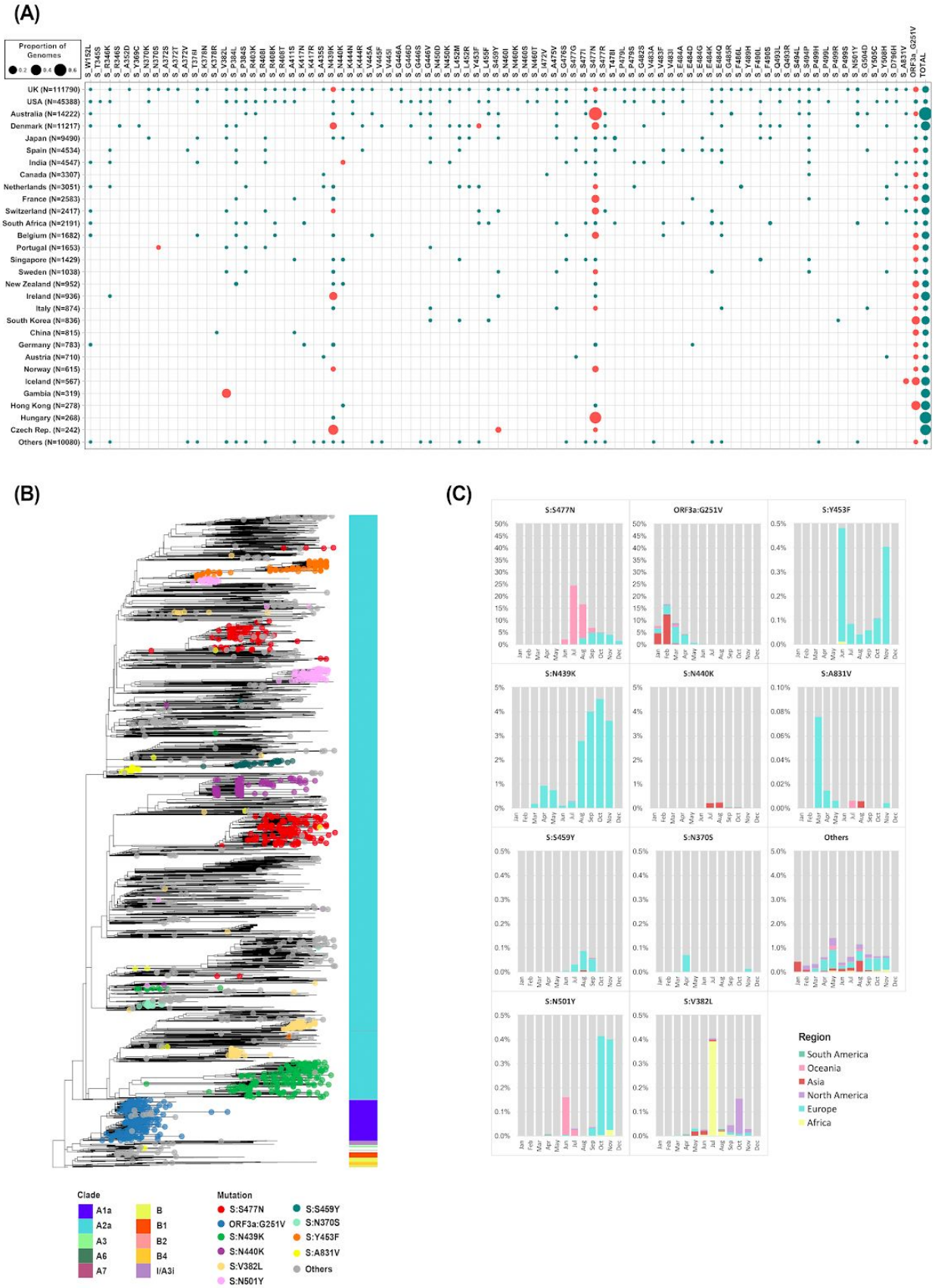
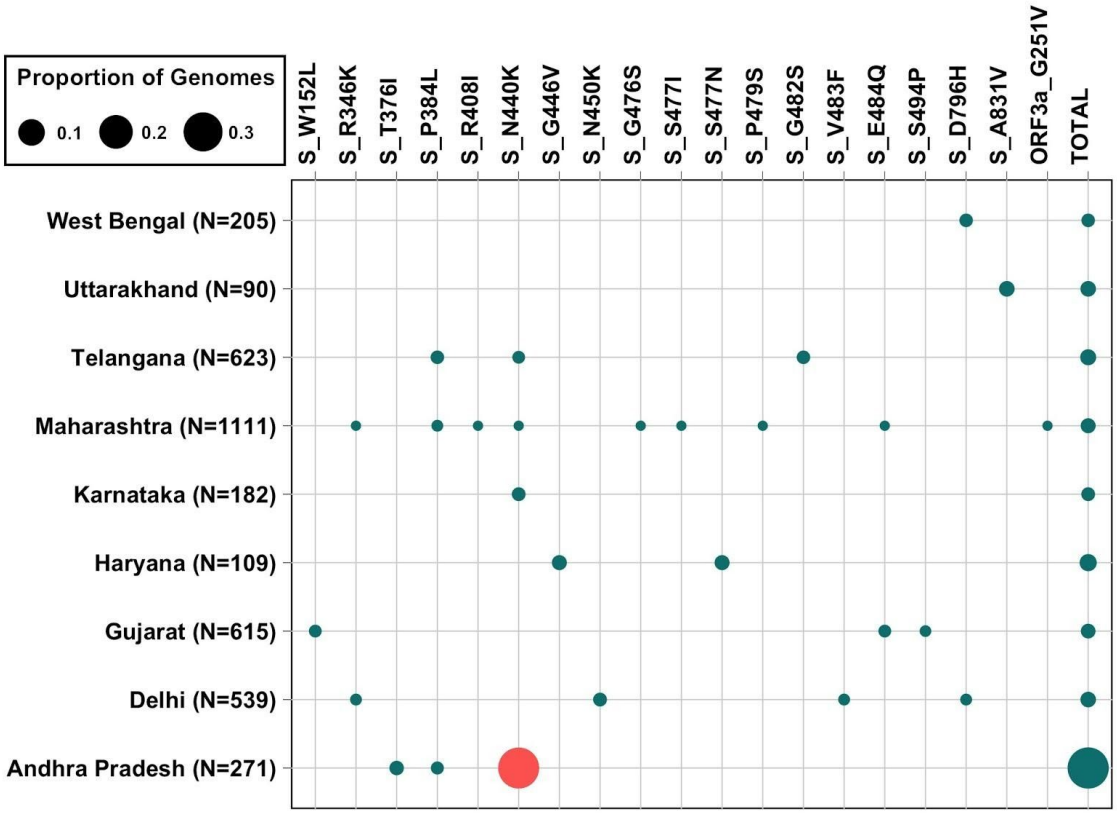

(A)
Proportion of Genomes
0.2
0.4
0.6
S_W152L
S_T345S
S_R346K
S_R346S
S_A352D
S_Y369C
S_N370K
S_N370S
S_A372S
S_A372T
S_A372V
S_T376I
S_K378N
S_K378R
S_V382L
S_P384L
S_P384S
S_R403K
S_R408I
S_R408K
S_R408T
S_A411S
S_K417N
S_K417R
S_A435S
S_N439K
S_N440K
S_K444N
S_K444R
S_V445A
S_V445F
S_G446A
S_G446D
S_G446S
S_G446V
S_N450D
S_N450K
S_L452M
S_L452R
S_Y453F
S_L455F
S_S459Y
S_N460D
S_N460K
S_N460S
S_N460T
S_I472V
S_A475V
S_G476S
S_S477G
S_S477I
S_S477N
S_S477R
S_T478I
S_P479L
S_P479S
S_G482S
S_V483A
S_V483F
S_V483I
S_E484A
S_E484D
S_E484G
S_E484K
S_E484Q
S_G485R
S_F486L
S_Y489H
S_F490L
S_F490S
S_Q493L
S_Q493R
S_S494L
S_S494P
S_P499H
S_P499L
S_P499R
S_P499S
S_S501Y
S_G504D
S_Y505C
S_Y508H
S_D796H
S_A831V
ORF3a_G251V
TOTAL
UK (N=111790)
USA (N=45388)
Australia (N=14222)
Denmark (N=11217)
Japan (N=9490)
Spain (N=4534)
India (N=4547)
Canada (N=3307)
Netherlands (N=3051)
France (N=2583)
Switzerland (N=2417)
South Africa (N=2191)
Belgium (N=1682)
Portugal (N=1653)
Singapore (N=1429)
Sweden (N=1038)
New Zealand (N=952)
Ireland (N=936)
Italy (N=874)
South Korea (N=836)
China (N=815)
Germany (N=783)
Austria (N=710)
Norway (N=615)
Iceland (N=567)
Gambia (N=319)
Hong Kong (N=278)
Hungary (N=268)
Czech Rep. (N=242)
Others (N=10080)
(B)
Clade
A1a
A2a
A3
A6
A7
B
B1
B2
B4
I/A3i
Mutation
S:S477N
ORF3a:G251V
S:N439K
S:N440K
S:V382L
S:N501Y
S:S459Y
S:N370S
S:Y453F
S:A831V
Others
(C)
S:S477N
ORF3a:G251V
S:Y453F
S:N439K
S:N440K
S:A831V
S:S459Y
S:N370S
Others
S:N501Y
S:V382L
Jan
Feb
Mar
Apr
May
Jun
Jul
Aug
Sep
Oct
Nov
Dec
Region
South America
Oceania
Asia
North America
Europe
Africa

Figure 1. (A) Variant frequencies of the immune escape variants in genomes of SARS-CoV-2. The total number of genomes analyzed from each country is specified. Variants with frequency >1% in the respective countries are highlighted in red. (B) Global phylogenetic context of the variants. The vertical bar indicates the clade assigned according to the Nextstrain nomenclature (C) Time-series data on prevalence for the genetic variants showing the region-wise proportion of genomes per month for the variants



Supplementary Figure. Variant frequencies of the immune escape variants in genomes isolated from different states in India.