Integrated genomic view of SARS-CoV-2 in India

Pramod Kumar[1,#], Rajesh Pandey[2,#], Pooja Sharma[2,#], Mahesh S Dhar[1,#], Vivekanand A[2], Bharathram Uppili[2], Himanshu Vashisht[1], Saruchi Wadhwa[2], Nishu Tyagi[2], Saman Fatihi[2], Uma Sharma[1], Priyanka Singh[1], Hemlata Lall[1], Meena Datta[1], Poonam Gupta[1], Nidhi Saini[1], Aarti Tewari[1], Bibhash Nandi[1], Dhirendra Kumar[1], Satyabrata Bag[1], Deepanshi[2], Surabhi Rathore[2], Nidhi Jatana[2], Varun Jaiswal[1], Hema Gogia[1], Preeti Madan[1], Simrita Singh[1], Prateek Singh[2], Debasis Dash[2], Manju Bala[1], Sandhya Kabra[1], Sujeet Singh[1], Mitali Mukerji[2], Lipi Thukral[2], Mohammed Faruq[2,#], Anurag Agrawal[2,*], Partha Rakshit[1,*]

[1]National Centre for Disease Control (NCDC), 22- Shamnath Marg, Delhi-110054, India.

[2]CSIR-Institute of Genomics and Integrative Biology (CSIR-IGIB), Mall Road, Delhi-110007, India.

Running Head: prevalent SARS-CoV-2 genomes in India

[#] Equal contribution

[*] Joint Senior Authors

Keywords: COVID-19, SARS-CoV-2, Whole Genome Sequencing (WGS), MinION and Phylogenetics.

Address of Lead contact:
Biotechnology Division
National Centre for Disease Control,
22-Shamnath Marg, Delhi-110054, INDIA
E mail: partho_rakshit@yahoo.com
Phone: +91-**11-23970593**

SUMMARY

India first detected SARS-CoV-2, causal agent of COVID-19 in late January-2020, imported from Wuhan, China. March-2020 onwards; importation of cases from rest of the countries followed by seeding of local transmission triggered further outbreaks in India. We used ARTIC protocol based tiling amplicon sequencing of SARS-CoV-2 (n=104) from different states of India using a combination of MinION and MinIT from Oxford Nanopore Technology to understand introduction and local transmission. The analyses revealed multiple introductions of SARS-CoV-2 from Europe and Asia following local transmission. The most prevalent genomes with patterns of variance (confined in a cluster) remain unclassified, here, proposed as A4-clade based on its divergence within A-cluster. The viral haplotypes may link their persistence to geo-climatic conditions and host response. Despite the effectiveness of non-therapeutic interventions in India, multipronged strategies including molecular surveillance based on real-time viral genomic data is of paramount importance for a timely management of the pandemic.

INTRODUCTION

The rapid worldwide spread of a novel coronavirus following its first appearance in China has pressed the global community to take measures to flatten its transmission (Chan et al., 2020; Zhu et al., 2020). The virus was named as "SARS-CoV-2" and the disease it causes has been defined as "coronavirus disease 2019" (COVID-19) (CSG, 2020). WHO declared COVID-19 a Pandemic on 11th of March, 2020 based on the speed and scale of transmission with almost 4.7 million cases reported across the globe, so far (WHO, 18th May 2020). The common signs of infection by SARS-CoV-2 include cough, fever, sore throat, respiratory symptoms inclusive of shortness/difficulties in breathing. More severe symptoms can include pneumonia, severe acute respiratory syndrome, kidney failure and even death with coalescence of factors (Zhu et al., 2020, Youg et al., 2020). Many COVID-19 cases have been reported to be asymptomatic and may serve as carrier of SARS-CoV-2 (Xu et al., 2020; He et al., 2020). Whole Genome Sequences (WGS) of SARS-CoV-2 suggest bat-CoV to be its closest progenitor with 96% homology but, the RNA binding domain (RBD) of its spike protein with an efficient binding to ACE-2, the receptor for SARS-CoV-2 in human cell seems to have been derived from Pangolin-CoVs (Wan et al., 2020, Andersen et al., 2020). Next generation sequencing (NGS) aided understanding of evolution of SARS-CoV-2 genomes and its transmission patterns after it enters a new population is proving to be an important step towards formulating strategies for management of this pandemic (Chen and Li et al., 2020).

The first three cases in India were reported from the state of Kerala in late January and early February, with a travel history of Wuhan, China. India took drastic steps to contain the further spread of the virus including imposition of travel restrictions to-and-from the affected countries. There were no new cases of COVID-19 for almost a month. All three cases subsequently tested negative making India free of the disease at that point of time (Press Information Bureau, PIB, India, 2020). However, while the global focus was on China and other eastern countries like South Korea and Japan; European countries, middle-east and the USA reported a surge in cases of COVID-19, pressing the WHO to declare it as a pandemic. March 2020 onwards, India also witnessed a surge of imported cases from countries other than China which has been further assisted with local transmission. In March, imposition of nationwide lockdown checked the epidemic curve. Despite these measurements, India is at the verge of a large outbreak as the transmission is rapidly increasing with more than 100,000 cases of COVID-19 having been reported in the third week of May, 2020.

We carried out WGS of SARS-CoV-2 (n=104) from Pan-India through the network of Integrated Disease Surveillance Program (IDSP) of National Centre of Disease Control (NCDC), Delhi. We report here a comprehensive and integrative genomic view of SARS-CoV-2 in the Indian subcontinent. In this study, we combine genetic and epidemiological data to understand the genetic diversity, evolution, and epidemiology of SARS-CoV-2 across India. The spectrum of variations would be an important tool towards contact tracing, effective diagnostics and backbone for drug and vaccine development.

METHODS

Subject recruitment

The study was conducted jointly by the NCDC and CSIR-Institute of Genomics and Integrative Biology (CSIR-IGIB). Institutional ethical clearance was obtained at both the places prior to initiation of research. A total of 127 laboratory confirmed cases of COVID-19 from a targeted testing representing different locations (as described in **Table-1** and **supplementary figure-1**) were included in the study for genomic analyses. Targeted testing involved suspected cases; having symptoms (fever, cough and breathlessness) with recent travel history to high-risk countries or positive contacts of COVID-19 cases.

Sample collection and Viral RNA isolation

The nasopharyngeal and oropharyngeal swabs (in viral transport medium) received at NCDC, Delhi through IDSP were subjected to viral inactivation followed by RNA extraction using QIAamp Viral RNA Mini Kit (Qiagen). To ensure that sub-optimal RNA samples are also included in the study, we made use of SuperScript IV (Cat. No. 18091050, Thermo Fisher Scientific, Waltham, MA, USA), for superior first strand cDNA synthesis and included them for sequencing.

Molecular diagnosis of COVID-19

Reverse transcribed (RT)-PCR assay was used on purified RNA for detection of SARS-CoV-2 in the samples. Quantitative RT-PCR was carried out using Taqman assay chemistry on ABI7500 platform. The primer/probe concentrations and reaction conditions for diagnostics were as per the WHO protocols (Corman et al., 2020). Two target genes were used for diagnosis of SARS-CoV-2, envelope (E) gene for screening and RNA dependent RNA polymerase (RdRp gene) for confirmation. The positive samples were analyzed based on the country of origin (traveler), contact with positive case, geographical location (community), gender and age. Samples from each group were selected and further processed for WGS of the SARS-CoV-2.

WGS of SARS-CoV-2

cDNA synthesis, ONT library preparation and sequencing: cDNA synthesis, ONT library preparation and sequencing has been mentioned in detail in supplementary section. Briefly, 50 ng of total RNA was used for cDNA (first and second strand) synthesis using random hexamer.

A 100ng of double stranded cDNA was used for NGS using a highly multiplexed PCR amplicon approach on the Oxford Nanopore Technologies (ONT, Oxford, United Kingdom) MinION using V3 primer pools (ARTIC Protocol). The amplicons were end repaired, ligated with native barcodes (EXP-NBD104 and EXP-NBD114, ONT) and purified using Ampure beads. The purified product was used for adaptor ligation using Adapter Mix II and purified using a combination of short fragment buffer and Ampure beads. The library was quantified using the Qubit dsDNA HS assay kit and 70ng of the library was used for sequencing. Barcoding, adaptor ligation, and sequencing were performed on samples with Ct values between 16-31. Sequencing flowcell was primed and used for sequencing using MinION Mk1B.

Illumina library preparation, sequencing and data analysis

The details of methodology for illumina library preparation, sequencing and data analysis are mentioned in supplementary material. In brief, the cDNA pool used for Nanopore sequencing libraries was also used for Illumina library preparation and subsequent sequencing employing Nextera XT protocol. Illumina's MiSeq platform was used for sequencing. The raw reads were quality checked, trimmed and the mapped human transcripts were removed. The unmapped reads from the human were used to align to the SARS-CoV-2.

Phylogeny and network analysis

The fasta sequences were aligned using MAFFT considering the MN90894.3 version as the reference sequence. Phylogenetic tree has been constructed using the Neighbour joining algorithm as statistical method and Maximum Composite Likelihood as model in MEGA10 software. FIGTREE was used for the graphical visualisation of Phylogenetic analysis (http://tree.bio.ed.ac.uk/software/figtree). Pheatmap and complex heatmap packages from R were used to plot the heatmaps. The Haplotype Network analysis has been done using PopART [Leigh and Bryant, 2015].

Protein based annotation and 3-Dimensional protein models

The protein-based annotation was performed according to the reference genome of SARS-CoV-2 (NC_045512 in the NCBI database) to categorize the specific amino acid change. Computational protein structure models of SARS-CoV-2 was created to map high frequency mutations on Spike, Nucleocapsid and nsp3 and RdRp (nsp12) proteins. The detailed methodology for the protein based annotation and 3-dimensional protein models have been mentioned in the supplementary materials.

RESULTS

Demographic details of the subjects

The mean (standard deviation) age of the total 127 subjects was 41.4±17.5 years with age range 0.5-76 years and median of 39 years. The gender ratio of male: females in the age group <39 years was 35:28 while the remaining 46 subjects >40 years had the ratio of 58:6.

Geographical location and travel history

The majority of the SAS-CoV-2 positive samples were obtained from New Delhi covering the national capital region of Delhi, India and few clusters identified as per surveillance team (covering various states of Delhi, Tamil Nadu, Maharashtra, Uttar Pradesh, Andhra Pradesh, West Bengal, Bihar, Orissa, Rajasthan, Haryana, Punjab, Assam and Union territory of Ladakh). The exposure to the COVID-19 was suggestive of travel history of subjects to Europe, West Asia and East Asia. Few subjects were from foreign countries i.e. Indonesia (n=14), Thailand (n=2) and Kyrgyz Republic (n=2). The identified localities of the subjects will further help in molecular surveillance of SARS-CoV-2 in respective geographical regions.

Profile of SARS-CoV-2 Genome Sequences

Amplicon coverage

The average amplicon coverage for the V3 ARTIC primers used in the study was more than 1000X coverage across the majority of the samples (**Figure 1A**). We observed that 73 out of 98 amplicons had more than 1000X in 90% of the samples. Conversely, there were only 25 amplicons which had less than 1000X coverage in 10% of the samples used in the study.

Genome coverage vs Ct value

We also looked into whether lower Ct values are a good indicator of genome coverage using a minimal set of virus mapping reads. We plotted genome coverage and average sequencing depth across Ct value of both the genes (E and RdRp). It was observed that higher Ct values (27 onwards) have increased possibility of lower genome coverage (**Figure 1B**) although some lower Ct value samples also had incomplete genome coverage. This may be due to the integrity of the RNA used for sequencing. It was observed that the average sequencing depth was higher (1000X-3000X) for samples with lower Ct values, whereas higher Ct value samples have relative lower average sequencing depth (100X-1000X). This would be important for informed decision making for reads required for ONT based sequencing.

Sequencing platform comparison

We sequenced a subset of samples on orthogonal platforms and sequencing methods (shotgun and amplicon) using ONT and Illumina platform. Significantly, we observed that the genetic variants were common between both the platforms (**Figure 1C**).

NGS analysis for SARS-CoV-2 sequence

A total of 104 samples passed the quality threshold for mapping full genome coverage threshold for SARS-CoV-2 genome (accession ID:) <0.05 N content with median coverage-~1500X.

Twenty-three samples that did not qualify the threshold criteria were excluded from strain identification. However, few samples were retained for variant analysis wherever the sequencing coverage was sufficiently of high quality for variant calling.

Phylogeny and variant analysis

The phylogenetic analysis of 104 high quality sequences reveal all the strains to be grouped into two major clades, a sub-clade and other clades (**Figure 2** and **Supplementary Figure 2**). In total we observed 163 variants representing singletons (107 variants), rare: 2-5% (45 variants), and common variants: >5% (11 variants). The common variants observed were 241 (Leader sequence), 3037 (NSP3), 6310 (NSP3), 6312 (NSP3), 11083 (NSP6), 13730 (NSP12/RdRp), 14408 (NSP12/RdRp), 23403 (S-Protein), 23929 (S-Protein), 25563 (ORF3a), 28311 (N-capsid).

Cluster-1: This cluster of these sequences (n=26) have shown G-clade [(Global Initiative on Sharing All Influenza Data, GISAID based nomenclature for variant at 23403 (Spike protein: D614G) position]. Based on Nextstrain classification 26 sequence belonged to A2a clade [(denoted by positions C241T; C3037T, A23403G (S: D614G); C14408T (ORF1b/RdRp: P314L)] (**Figure 2A**). The additional frequent variants in this cluster observed were 25563 (ORF3a, n=9),18877 (NSP14/Exonuclease, n=7), 26735 (M-protein, n=6), 22444 (S-protein, n=5), and 28854 (N-capsid, n=5). One novel variant *1947 T>C (NSP2)* was observed in two strains in this cluster.

Cluster-2: In our study large numbers of strains (n=65) belong to this unclassified cluster (as per GISAID and Nextstrain). The strains in this cluster had the predominant variants, G11083T variant (NSP6) (n=65), C13730T in RdRp (n=65), C28311T (N-capsid); n=65, C6312A (NSP3 variant); n=64 (one sequence being called N), C23929T (S-protein), n=50 (other being low depth/N bases). The variant C6310A (NSP3); n=22 being observed as another frequent alteration (**Figure 2B**). We also observed few novel variants, *G12685T (NSP8)*; n=5 and *TC1706T (NSP2)*; n=4, *T7621C (ORF1a/b)*; n=3, *A21792T (S protein)*; n=3, *G13920A (NSP12/RdRp)*; n=2, *A16355G (NSP13/Helicase)*; n=2 and *G18803T (NSP14/Exonuclease)*; n=2 in this cluster.

The majority of the key cluster variants 11083, 13730, 28311, 6312, 23929 are also shared in sequences submitted from Singapore region and Brunei (GISAID) and also similar clade sequences were observed in India submitted by National Institute of Mental Health and Neuro-Sciences (NIMHANS) and Gujarat Biotechnology Research Centre (GBRC) cohort. From history, based on the geographical location of the subjects of this cluster, a significant number of natives of Indonesia (n=7) and two each from Thailand and Kyrgyzstan were part of this cluster from our study site. This probably suggests introduction of this particularly from East Asian countries into India.

Cluster-3: This subclass of strains (n=7) harboring a common variant G11083T (NSP6), G1397A (NSP2) and T28668C (N-capsid) are described for the A3 clade (Nextstrain) in additions to G29742T (**Figure 2A**). Other mutated positions i.e. C884T (NSP2), G8653T (NSP4) were observed in 5 samples, whereas T16993C (NSP13), n=4; T25461C (ORF3a), n=4 and A27191G (M-protein), n=2 are putative novel sites.

The phylogeny analysis of these clusters segregated with the other Indian SARS-CoV-2 genome sequences as recently reported (GISAID) (**Figure 2B**).

Other SARS-CoV-2 genomes: Two SARS-CoV-2 belonging to A1a clade had a SNP profile of 11083(NSP6)/14805 (NsP12/RdRp)/2480 (NSP2)/2558 (NSP2)/26144 (ORF3a). In addition, we observed three B clade sequences having position 8782 C>T (NSP4) and 28144 T>C (ORF8; S clade GISAID) mutated and with one sequence with an additional C18060T B1 variant. One genome from Maharashtra had no variants and probably represented the first genome sequenced from Wuhan, China.

Re-defining Cluster-2 with neighbourhood re-joining

With over represented variants in cluster-2 for variants 11083/13730/28311/6312/23929, we defined this cluster with A4 clade. This has similarity with sequences submitted from Singapore, Brunei and other Indian sequences submitted. The haplotype network analysis suggests that these sequences are having a common origin from East Asia/South-East Asia (**Figure 2C** and **Figure 3**). This A4 clade has multiple variants in important region of viral genome, RdRp (A97V), N-capsid (P12L), NSP3 (T2016K), NSP6 (L37F) and NSP3 (S1197R) variants. In our cohort of samples, the majority of subjects were from Tamil Nadu, Delhi and Indonesia and others were from various other states (**Figure 3**).

Protein-wise analysis of SARS-CoV-2 variants

To provide quantitative insights into the mutant proteins, we characterized amino acid substitutions across the 104 viral genomes. Of the 53 point mutations identified, 29 were missense that resulted in amino acid substitutions. **Figure 4** plots the occurrence of these mutations as a function of each viral protein. The frequency of amino acid variations were maximum in nsp6 (L37F) present in 68 genomes, followed by nsp12 (A97V) in 65, nsp3 (T1198K) in 62 and nucleocapsid (P13L) in 53 genomes. Interestingly, D614G mutation in spike protein which is considered as a prevalent global mutation [Korber et al., 2020; Chandrika et al., 2020], was present in only 26 of the 104 sequenced genomes.

Next, we correlated the occurrence of each mutation with the type of amino acid change (**Figure 5A**). Interestingly, ~45% of these mutations showed no change in the nature of the amino acids, suggesting that the viral proteins harbors subtle changes in the protein shape or function. Within high-occurring mutations, P13L, L37F, A97V also showed no major residue alterations. However, there are other loci that involve complete change in amino acid type once mutated such as interchange between polar, hydrophobic or charged. For instance, high frequency positions, including T1198K in nsp3 involve acquisition of a charged group. In addition, the key mutation in spike protein (D614G) also involves loss of the charged group. These mutations that lead to positively charged groups may cause more severe structural and functional effects.

We also compared SARS-CoV-2 mutation sites with other six coronavirus sequences (**Figure 5B**). Most of the mutations were present mostly in variable locations. Out of 29 mutations, 10 are present on highly conserved residue locations. Interestingly, higher frequency mutations are at positions that evolve faster/are variable across the coronaviruses except A97L and L37F, which are present on conserved locations.

To understand how these mutations are changing the local environment, we mapped higher frequency mutations onto SARS-CoV-2 protein structures, including nucleocapsid, nsp3, nsp12, and spike protein (**Figure 6**). In nucleocapsid, there are two structural domains, the N- and C-terminal connected via a long linker region that facilitates RNA packing [Kang et al., 2020]. We found that all the mutations, including P13L found in 53 genomes, are present outside these domains, specifically, linker regions (S194L, G204R, R203K), and the N-arm of the protein

(P13L). The nsp12 is a highly conserved protein with multiple domains. The observed mutations are overlaying onto the interface (P323L) and NiRAN (A97L) region. The latter is critical as it contains a Zn+ binding site, however, little is known about the exact functional output. In contrast, the P323L mutation is present on protein interaction junctions where a hydrophobic cleft is known to bind to inhibitors (**Figure 6B**). This variant will result in loss of kink due to amino acid substitution from proline to leucine, i.e., 5-membered amino acid which resides in the buried area of the protein from to an acyclic amino acid. The nsp3 protein has a similar higher order domain arrangement, and the mutation is present on the NAB domain, which is a nucleic acid binding domain and also interacts with nsp12. This mutation may impact RNA synthesis machinery; however, little is known about its exact mechanism of action. Lastly, the D614G mutation in spike protein is an interesting substitution and has been reported with increased tally [Korber et al., 2020; Chnadrika et al., 2020]. Structurally, this mutation is located in the S1 subunit that also contains the RBD domain. Although present outside the functional region, the proximity of D614G around S1 cleavage site implicates an important change in the local environment.

Geographical distribution of SARS-CoV-2 in India

SARS-CoV-2 classified for their preponderance of distribution across geographical locations of India vis-a-vis states of India. Enrichment and depletion of specific clades/classes was observed in certain states of India. We explored the possibility of whether travel history based viral strain showed signs of adding new variants once it diversified in India.

DISCUSSION

This is the first comprehensive genomic picture of the SARS-CoV-2 prevalent in the Indian population during the early phase of outbreaks. The understanding is important keeping in view the vast geographical expanse and population density of India. There were three major waves of viral entry in India associated with multiple outbreaks (**Figure 7**). First wave includes importation of SARS-CoV-2 (A2a cluster) through travelers from Europe (Italy, UK, France etc) and the USA. Second wave of SARS-CoV-2 (A3 cluster) was linked with the Middle East (Iran and Iraq). Third wave comprises combined viral (haplotype redefined as A4) entries from

Southeast Asia (Indonesia, Thailand and Malaysia) and Central Asia (Kyrgyzstan). The study taken together with other reported genomes (Potdar et al., 2020) revealed A4 cluster (previously unclassified) is the most prevalent in Indian population. Many novel mutations identified may be specific to Indian conditions but more genomic data is needed to strengthen the assumption to rule out sampling bias and other factors (Lu et al., 2020). During the early phase of the epidemic in India, imported cases followed by local transmission were restricted to urban regions where effective IDSP network, diagnostic support, health care services and timely placed interventions (nationwide lockdown, establishment of containment zones and practicing social distance) interrupted community transfer. The mounting burden of epidemic in urban regions associated with job crunch forced migrant workers to return to their home mostly situated in rural areas which is an impending threat of the viral entries in rural regions (comprising major portion of the Indian population) and expected community transmission. To tackle domestic transmission at larger scale and higher risk of community transfer, preventive measures shall be strengthened in rural regions in addition to safe transportation arrangements for domestic migrants.

The national lockdown may have led to evolution/gain of certain variants with potential role in adapting to Indian conditions. This may have given rise to a distinct lineage awaiting its inclusion in Nextstrain (an open-source project to harness the scientific and public health potential of pathogen genome data). A more detailed analysis of these genomes might provide information whether these variations need to be considered during design of diagnostic primers as the need for testing shoots up. It may allow for creation of cost effective panels to trace the movement of lineage specific strains across geographical regions more rapidly and effectively. Lots of efforts are ongoing to identify suitable vaccine candidates through docking studies. These observations are important to consider the variants especially that map to the Indian genomes during such prioritization studies since these strains would now form a major fraction of the genomes that are likely to become more prevalent in India after lockdown. Mapping of these variant genomes in conjunction with the clinical history in terms of recovery, hospitalization and co-morbidity might allow identification of variants that should be actionable and would also have relevance for prognosis. It is imperative that robust genomic data based on large sample size including rural populations with even distribution can bring out the real scenario once correlated with epidemiological data eventually helping in drafting of further management policies.

SUPPLEMETARY INFORMATION

Supplementary Material, Supplementary figures1: location map; supplementary figure 2: heat map representation of SARS-CoV-2 variants per sample in respective clusters.

ACKNOWLEDGEMENTS

NCDC greatly acknowledge the support of Prof. Dr. Christian Drosten, Charite – Universitatsmedizin, Berlin for promptly providing the positive controls for qPCR. The authors do acknowledge GISAID for sharing the genomic sequences in public domain and other contributors SARS-CoV-2 genomic data. We would like to gratefully acknowledge the efforts of officials involved in IDSP network and associated hospitals in sample collection and timely data sharing. We would like to thank the financial aid provided by Ministry of Health and Family Welfare, Government of India. CSIR-IGIB would like to acknowledge Genotypic Technologies Pvt. Ltd., Bangalore, India for its role in facilitating ONT sequencing. Authors also acknowledge the role of all the technical and support staff of NCDC involved in COVID-19 testing and Subhash Gurjar (CSIR-IGIB) for facilitating reagents procurement amidst lockdown and for providing other assistance in the laboratory work. We thank CSIR, India for funding support towards SARS-CoV-2 genome sequencing.

AUTHOR CONTRIBUTIONS

A.A., P.K., P.R. and S.K. conceived the study; P.K., M.S.D., M.B. and P.R. planned and optimized diagnosis; M.S.D., H.V., U.S., Priyanka S., M.D., P.G., D.K., S.B. performed viral inactivation and diagnosis; S.S., H.L., A.T., B.N., V.J., H.G., P.M., N.S. and Sujeet S., did data compilation, field work and epidemiological analyses; R.P. and M.F. planned the sequencing experiment; P.S., S.W., N.T., R.P. and M.F. conducted the sequencing experiments; V.A., B.U. and M.F. analysed the sequencing data, Prateek S. and D.D. did independent sequencing analysis validation, S.F., D, S.R., N.J. and L.T. performed the protein analysis; P.K., M.F. and R.P. interpreted final data and wrote the manuscript; M.M., M.S.D. and M.B. edited the manuscript. All the authors read and approved the final manuscript.

DECLARATION OF INTERESTS

Authors declare no conflict of interests with the present study.

REFERENCES

Andersen., K., Rambaut, A., Lipkin, I., Holmes, E.C., Garry, R. (2020). The proximal origin of SARS-CoV-2. Nat. Med. *226*, 450–452.

Chan, J.F., Yuan, S., Kok, K.H., To, K.K., Chu, H., Yang, J., Xing, F., Liu, J., Yip, C.C., Poon, R.W. *et al* (2020). A familial cluster of pneumonia associated with the 2019 novel coronavirus indicating person-to-person transmission, a study of a family cluster. Lancet *395*, 514-523.

Bhattacharyya, C., Das, C., Ghosh, A., Singh, A.K., Mukherjee, S., Majumder, P.P., Basu, A., Biswas, N.K. (2020). Global Spread of SARS-CoV-2 Subtype with Spike Protein Mutation D614G is Shaped by Human Genomic Variations that Regulate Expression of TMPRSS2 and MX1 Genes. bioRxiv doi: https://doi.org/10.1101/2020.05.04.075911

Chen, Y., and Li, L. (2020). SARS-CoV-2, virus dynamics and host response. Lancet Infect. Dis. *20*, 515-516.

CSG (Coronaviridae Study Group) of the International Committee on Taxonomy of Viruses. (2020). The species severe acute respiratory syndrome-related coronavirus, classifying 2019-nCoV and naming it SARS-CoV-2. Nat. Microbiol. *5*, 536–544.

De Coster, W., D'Hert, S., Schultz, D. T., Cruts, M., Van Broeckhoven, C. (2018). NanoPack, Visualizing and processing long-read sequencing data. *Bioinformatics 32*, 2666–2669.

Hadfield, J., Megill, C., Bell, S.M., Huddleston, J., Potter, B., Callender, C., Sagulenko, P. Bedford, T., Neher, R.A. (2018). Nextstrain, Real-Time tracking of pathogen evolution. Bioinformatics *34*, 4121-4123.

Jian, L., Kusov, Y., Hilgenfeld, R., (2018). Nsp3 of coronaviruses: Structures and functions of a large multi-domain protein. Antiviral research *149*, 58-74.

Kang, S., Yang, M., Hong, Z., Zhang, L., Huang, Z., Chen, X., He, S., Zhou, Z., Zhou, Z., Chen, Q. et al. (2020). Crystal structure of SARS-CoV-2 nucleocapsid protein RNA binding domain reveals potential unique drug targeting sites. Acta Pharmaceutica Sinica doi:10.1016/j.apsb.2020.04.009

Korber, B., Fischer, W.M., Gnanakaran, S., Yoon, H., Theiler, J., Abfalterer, W., Foley, B., Giorgi, E.E., Bhattacharya, T., Parker, M.D. et al. (2020). Spike mutation pipeline reveals the emergence of a more transmissible form of SARS-CoV-2. bioRxiv doi: https://doi.org/10.1101/2020.04.29.069054.

Leigh, JW, Bryant D (2015). PopART: Full-feature software for haplotype network construction. Methods Ecol Evol 6(9):1110–1116.

Li. H. (2018). Minimap2, pairwise alignment for nucleotide sequences. *Bioinformatics 34*, 3094-3100.

Loman, N., Quick, J., Simpson, J. (2015). A complete bacterial genome assembled de novo using only nanopore sequencing data. Nat. Methods *12*, 733–735.

Lu, J., du Plessis, L., Liu, Z., Hill, V., Kang, M., Lin, H., Sun, J., François, S., Kraemer, M.U.G., Faria, N.R., et al. (2020). Genomic epidemiology of SARS-CoV-2 in Guangdong Province, China. Cell doi, 10.1016/j.cell.2020.04.023.

Lu, R., Zhao, X., Li, J., Niu, P., Yang, B., Wu, H., Wang, W., Song, H., Huang, B., Zhu, N., et al. (2020). Genomic characterization and epidemiology of 2019 novel coronavirus, implications for virus origins and receptor binding. Lancet *395*, 565-574.

Potdar, V., Cherian, S.S., Deshpande, G.R., Ullas, P.T., Yadav, P.D., Choudhary, M.L., Gughe, R., Vipat, V., Jadhav, S., Patil, S. (2020). Genomic analysis of SARS-CoV-2 strains among Indians returning from Italy, Iran & China, & Italian tourists in India. Indian J. Med. Res. *151*, 255-260.

Shu, Y., McCauley, J. (2017). GISAID, Global initiative on sharing all influenza data – from vision to reality. Euro Surveill. *22*(13), 30494.

WHO Mission briefing on COVID-19 - 12 March 2020., accessed on 13 March 2020 athttps,//www.who.int/dg/speeches/detail/who-director-general-s-opening-remarks-at-the-mission-briefing-on-covid-19---12-march-2020.

Xu, X.W., Wu, X.X., Jiang, X.G., Xu, K.J., Ying, L.J, Ma, C.L., Li, S.B., Wang, H.Y., Zhang, S., Gao, H.N., et al. (2020). Clinical findings in a group of patients infected with the 2019 novel coronavirus (SARS-Cov-2) outside of Wuhan., China, retrospective case series. BMJ. *368*, m606.

Song, Y., DiMaio, F., Wang, R. Y., Kim, D., Miles, C., Brunette, T., Thompson, J., Baker, D. (2013). High-resolution comparative modeling with RosettaCM. Structure. *21*, 1735–1742.

Young, B.E., Ong, S.W.X., Kalimuddin, S., Low, J.G., Tan, S.Y., Loh, J., Ng, O.T., Marimuthu, K., Ang, L.W., Mak, T.M., et al. (2020). Epidemiologic features and clinical course of patients infected with SARS-CoV-2 in Singapore. JAMA *23*(15), 1488-1494.

Zhou, F., Yu, T., Du, R., Fan, G., Liu, Y., Liu, Z., Xiang, J., Wang, Y., Song, B., Gu, X., et al. (2020). Clinical course and risk factors for mortality of adult inpatients with COVID-19 in Wuhan., China, a retrospective cohort study. Lancet *395*, 1054-1062.

Zhu, N., Zhang, D., Wang, W., Li, X., Yang, B., Song, J., Zhao, J., Huang, B., Shi, W., Lu, R., et al. (2019). A novel coronavirus from patients with pneumonia in China. N. Engl. J. Med. *382*, 727-733.

Figure Legends:

Figure 1: Sequencing data quality parameters and orthogonal platform validation. A) Arctic Amplicon coverage plot. It represents the amplicon wise genome coverage of the SARS-CoV-2 genome. 73 out of 98 (~71.5%) amplicons have more than 100X coverage in 90% of the sequenced samples. **B) Genome coverage and sequencing depth of the SARS-CoV-2 genome.** We plotted the Ct value of the two genes (E-gene and RdRp gene) used for RT-PCR vis-a-vis genome %age coverage and sequencing depth of the samples (Blue - missing N%

below 5 and Orange - N% above 5). In general, it was observed that samples with lower Ct value have higher average sequencing depth and genome coverage compared to higher Ct value samples. **C) Genetic variants across ONT and Illumina platform**. A subset of ONT sequenced samples were sequenced on MiSeq to benchmark the genetic variants observed in the positive samples. The genetic variants marked in yellow are the same between the two platforms.

Figure 2: Haplotype network and Phylogenetic analysis of SARS-CoV-2 sequences. A) The network analysis of SARS-CoV-2 sequences from this study showing distinct clades with their geographical locations. A4 clade described in this study for the first time has widespread geographical affiliations. A3 being more confined to Ladakh and mostly these cluster represents introduction of the infection through travel history. **B)** phylogenetic tree of SARS-CoV-2 genomes from sequences submitted from all across India depicting major clade based distribution of SARS-CoV-2 in India. **C)** Network analysis of A4 clade sequences submitted from Indian and neighbouring East Asians countries.

Figure 3: Distribution of A4 clade within India and globally. A) Distribution of A4 clade variants across different geographical regions from the cohort. **B)** Distribution of A4 clade variants across different geographical regions across the globe. **C) C**omparison of A4 cluster sequences across the South-East Asian region showing sharing of variants and haplotype across South-East Asian region.

Figure 4: Protein annotation of the amino acid substitutions. 29 Amino acid substitutions obtained from 104 sequenced genomes were mapped on various viral proteins as a function of frequency of these missense mutations. The majority of mutations are in Orf1ab region, which is decomposed into multiple non-structural proteins. The color mark at the end of occurrence bar indicates the protein name shown in the above legend.

Figure 5: Amino acid properties marked as a function of mutations. The first row shows amino acid change, with polar, hydrophobic, charged, and unchanged residues shown in violet, blue, green, and yellow, respectively. The second row shows the occurrence of protein mutations in 104 genomes, with the larger circle representing higher occurrence (69) and the smallest circle shows the presence in 2 genomes. The last row depicts the conservation score of that particular

mutation of SARS-CoV-2 with 6 other coronavirus. The conservation ranges between 1 to 9 with 9 depicting the highest score.

Figure 6: Mapping of higher frequency amino acid mutations on Nucelocapsid and Spike proteins. The mutations are marked in red color on the surface representation of each protein. In Spike protein, all the domains are highlighted in different colors, including NTD, RBD, HR1, Fusion peptide region, HR2, TM, and CT domains. In addition, cleavage sites are also marked onto the structure.

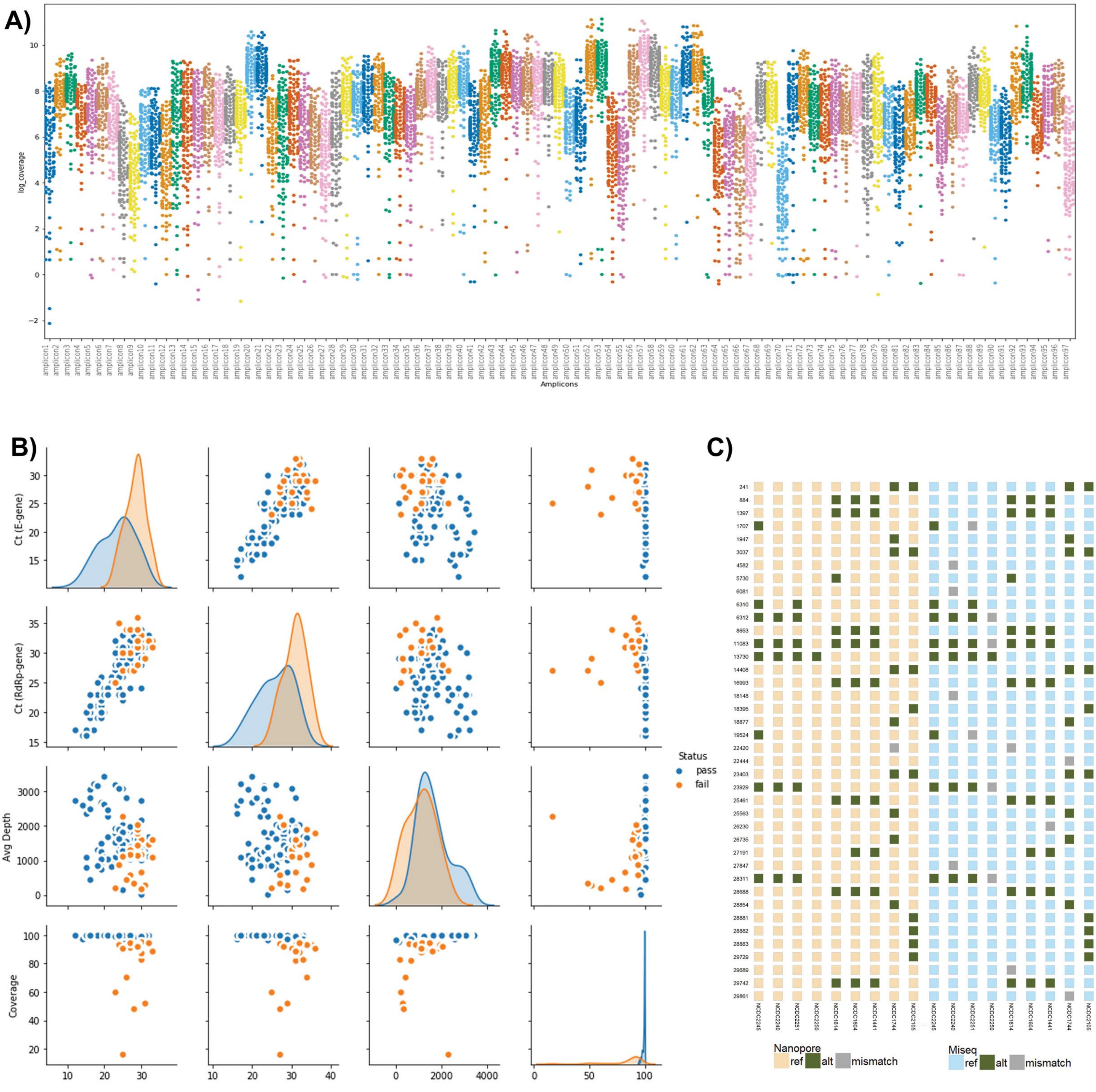
Figure 7: Scheme showing importation of prevalent SARS-CoV-2 genomes (3 major waves) in India.

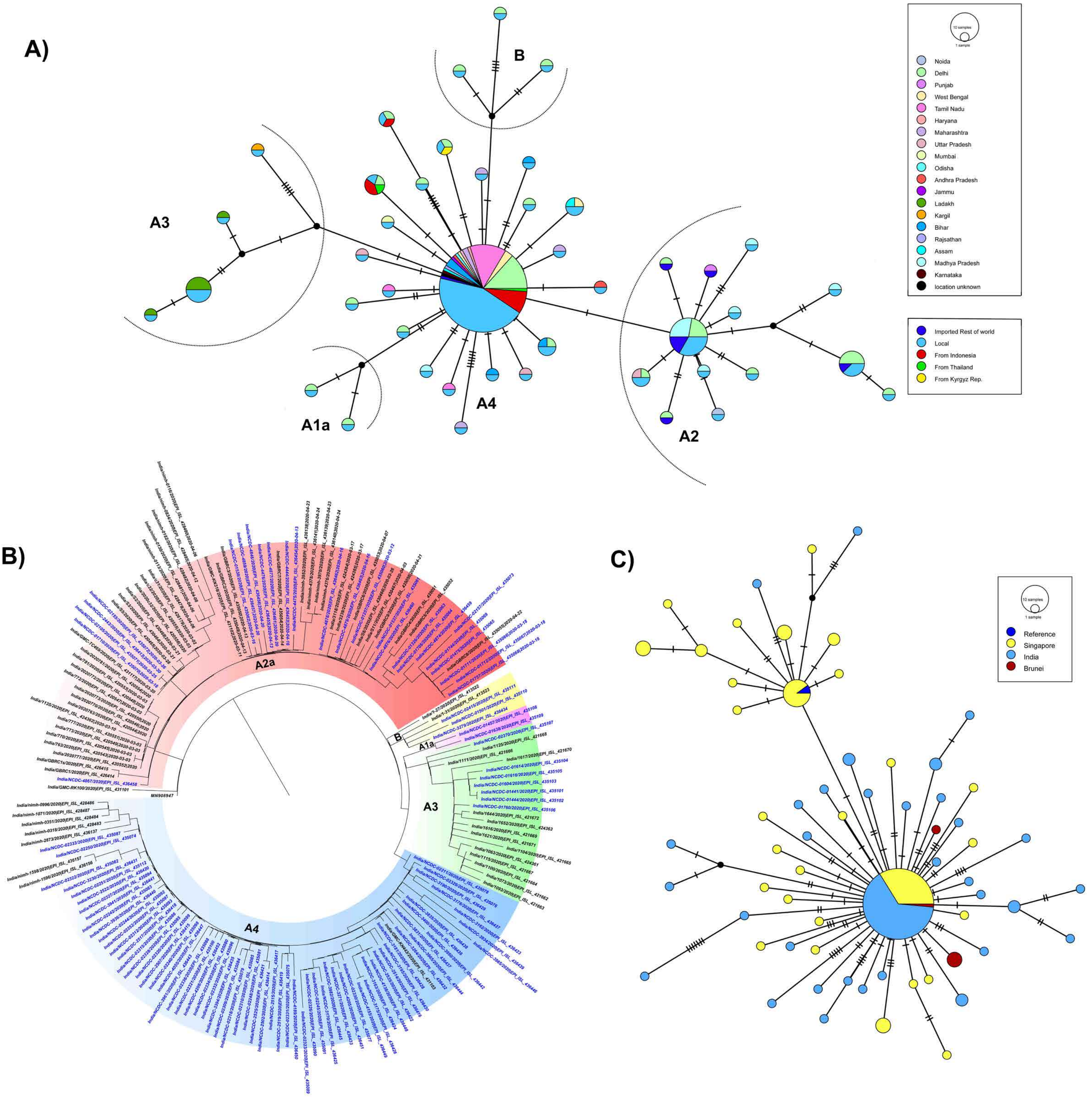
Table-1: Frequency and description the variations obtained in Cov2 genome from major identified cluster from 104 sequences

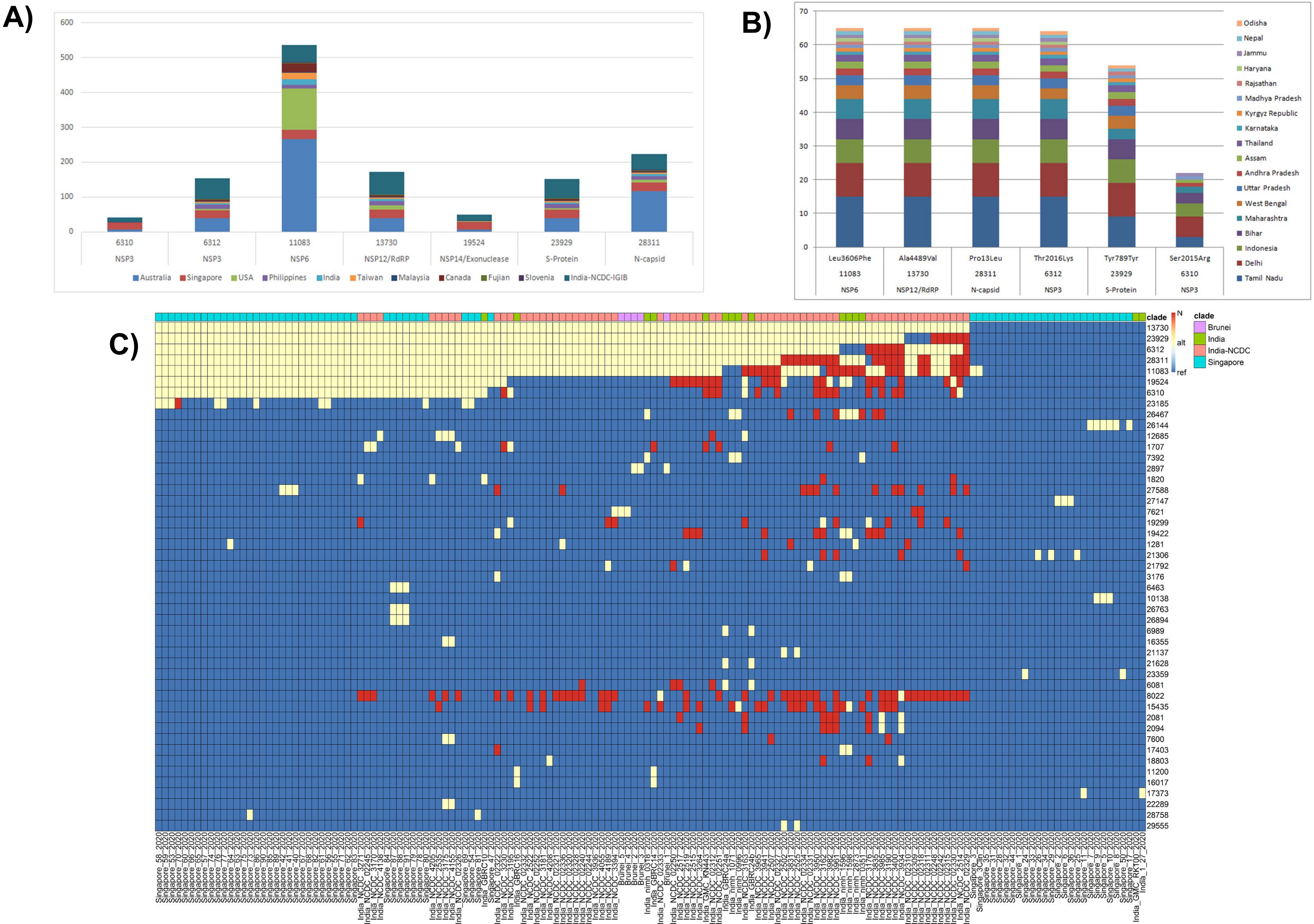
							Global Frequency					
POS	Gene	Gene	clade	Aminoacid	Amino Acid changes	effect	A	T	G	C	N	Variant count
11083	NSP6	orf1ab	A3	Leu3606Phe	QHD43415.1:p.3606L; QHD43415.1:p.3606L>F; QHD43415.1:p.3606-	LOW; MODERATE;	0	1137	7544	0	41	68
13730	NSP12/RdRP	orf1ab	-	Ala4489Val	QHD43415.1:p.4489A>V	MODERATE	0	99	0	8622	1	65
6312	NSP3	orf1ab	-	Thr2016Lys	QHD43415.1:p.2016T>K; QHD43415.1:p.2016T>I; QHD43415.1:p.2016T>R	MODERATE	93	3	0	8625	1	62
28311	N-capsid	N	-	Pro13Leu	QHD43423.2:p.13P>L; QHD43423.2:p.13P>R; QHD43423.2:p.13-	MODERATE; MODIFIER	0	160	1	8560	1	53
23929	S-Protein	S	-	Tyr789Tyr	QHD43416.1:p.789Y; QHD43416.1:p.789-	LOW; MODIFIER	0	88	0	8627	7	46
14408	NSP12/RdRP	orf1ab	A2a	Pro4715Leu	QHD43415.1:p.4715P>L; QHD43415.1:p.4715-	MODERATE; MODIFIER	0	5325	0	3381	16	26
23403	S-Protein	S	A2	Asp614Gly	QHD43416.1:p.614-; QHD43416.1:p.614D>G	MODIFIER; MODERATE	3346	0	5356	0	20	26
241	5'UTR	5'UTR	A2	0	QHD43415.1	MODIFIER,DISTANCE=25	0	5194	0	3149	379	24
3037	NSP3	orf1ab	A2	Phe924Phe	QHD43415.1:p.924F; QHD43415.1:p.924-	LOW; MODIFIER	1	5348	0	3352	21	23
6310	NSP3	orf1ab	-	Ser2015Arg	QHD43415.1:p.2015S>R; QHD43415.1:p.2015-; QHD43415.1:p.2015S	MODERATE; MODIFIER;	28	17	0	8672	5	22
25563	ORF3a	ORF3a	A2a2	Gln57His	QHD43417.1:p.57Q>H; QHD43417.1:p.57-	MODERATE; MODIFIER	0	2300	6415	1	6	9
1397	NSP2	orf1ab	A3	Val378Ile	QHD43415.1:p.378V>I	MODERATE	137	0	8584	0	1	7
18877	NSP14/Exonuclea	orf1ab	-	Leu6205Leu	QHD43415.1:p.6205L	LOW	0	208	0	8512	2	7
884	NSP2	orf1ab	-	Arg207Cys	QHD43415.1:p.207R>C	MODERATE	0	39	0	8682	1	6
8653	NSP4	orf1ab	-	Met2796Ile	QHD43415.1:p.2796M>I	MODERATE	0	37	8680	0	5	6
26735	M-Protein	M	-	Tyr71Tyr	QHD43419.1:p.71Y; QHD43419.1:p.71-	LOW; MODIFIER	0	41	0	8677	4	6
28688	N-capsid	N	A3	Leu139Leu	QHD43423.2:p.139L; QHD43423.2:p.139-	LOW; MODIFIER	0	8586	0	130	6	6
29742	3'UTR	3'UTR	A3	0	QHI42199.1	MODIFIER,DISTANCE=68	61	120	8268	0	273	6
12685	NSP8	orf1ab	-	Gln4140His	QHD43415.1:p.4140Q>H; QHD43415.1:p.4140-	MODERATE; MODIFIER	0	0	8720	0	2	5
16993	NSP13/Helicase	orf1ab	-	Tyr5577His	QHD43415.1:p.5577Y>H	MODERATE	0	8715	0	6	1	5
22444	S-Protein	S	-	Asp294Asp	QHD43416.1:p.294D	LOW	0	13	0	8438	271	5
25461	ORF3a	ORF3a	-	Ala23Ala	QHD43417.1:p.23A	LOW	0	8714	0	6	2	5
28854	N-capsid	N	-	Ser194Leu	QHD43423.2:p.194S>L; QHD43423.2:p.194-	MODERATE; MODIFIER	0	64	0	8654	4	5
1706	NSP2	#N/A	-	Ser481fs	#N/A	#N/A	0	8721	0	0	1	4
1707	NSP2	orf1ab	-	Ser481Phe	QHD43415.1:p.481S>F; QHD43415.1:p.481-	MODERATE; MODIFIER	0	3	0	8716	3	4
658	NSP1	orf1ab	-	Ala131Ala	QHD43415.1:p.131A; QHD43415.1:p.131-	LOW; MODIFIER	0	8706	0	14	2	3
1820	NSP2	orf1ab	-	Gly519Ser	QHD43415.1:p.519G>S; QHD43415.1:p.519-	MODERATE; MODIFIER	13	0	8708	0	1	3
7621	ORF1a/b	orf1ab	-	Cys2452Cys	QHD43415.1:p.2452C	LOW	0	8720	0	0	2	3
8782	NSP4	orf1ab	-	Ser2839Ser	QHD43415.1:p.2839S; QHD43415.1:p.2839-	LOW; MODIFIER	0	1370	0	7344	8	3
14805	NSP12/RdRP	orf1ab	A1a1	Tyr4847Tyr	QHD43415.1:p.4847Y; QHD43415.1:p.4847-	LOW; MODIFIER	0	892	0	7813	17	3
18486	NSP14/Exonuclea	orf1ab	-	Leu6074Leu	QHD43415.1:p.6074L	LOW	0	3	0	8717	2	3
19524	NSP14/Exonuclea	orf1ab	-	Leu6420Leu	QHD43415.1:p.6420L; QHD43415.1:p.6420-	LOW; MODIFIER	0	33	0	8501	188	3
21792	S-Protein	S	-	Lys77Met	QHD43416.1:p.77K>M	MODERATE	8717	0	0	0	5	3
27191	M-Protein	M	-	Ter223Ter	QHD43419.1:p.223-; QHD43419.1:p.223*	MODIFIER; LOW	8717	0	2	0	3	3
28881	N-capsid	N	A2a1	Arg203Lys	QHD43423.2:p.203R>K; QHD43423.2:p.203R>M; QHD43423.2:p.203-	MODERATE; MODIFIER	1357	1	7347	0	17	3
28882	N-capsid	N	A2a1	Arg203Arg	QHD43423.2:p.203R; QHD43423.2:p.203R>S; QHD43423.2:p.203-	LOW; MODERATE;	1353	0	7353	0	16	3
28883	N-capsid	N	A2a1	Gly204Arg	QHD43423.2:p.204-; QHD43423.2:p.204G>R	MODIFIER; MODERATE	0	0	7353	1353	16	3
1059	NSP2	orf1ab	A2a2a	-	QHD43415.1:p.265T>I; QHD43415.1:p.265-	MODERATE; MODIFIER	0	1900	0	6816	6	2
1281	NSP2	orf1ab	-	Ala339Val	QHD43415.1:p.339A>V; QHD43415.1:p.339-	MODERATE; MODIFIER	0	3	0	8717	2	2
1947	NSP2	#N/A	-	-	#N/A	#N/A	0	8720	0	0	2	2
2480	NSP2	orf1ab	-	Ile739Val	QHD43415.1:p.739-; QHD43415.1:p.739I>V	MODIFIER; MODERATE	8419	0	292	0	11	2
2558	NSP2	orf1ab	A1a1b	Pro765Ser	QHD43415.1:p.765P>S; QHD43415.1:p.765-	MODERATE; MODIFIER	0	321	0	8390	11	2
7600	ORF1a/b	orf1ab	-	Cys2445Cys	QHD43415.1:p.2445C; QHD43415.1:p.2445-	LOW; MODIFIER	0	0	0	8721	1	2
13920	NSP12/RdRP	#N/A	-	Lys4552Lys	#N/A	#N/A	0	0	8721	0	1	2
16355	NSP13/Helicase	orf1ab	-	Lys5364Arg	QHD43415.1:p.5364K>R	MODERATE	8721	0	0	0	1	2
18395	NSP14/Exonuclea	orf1ab	-	-	QHD43415.1:p.6044A>V	MODERATE	0	2	0	8719	1	2
18803	NSP14/Exonuclea	orf1ab	-	Ser6180Ile	QHD43415.1:p.6180S>I	MODERATE	0	0	8721	0	1	2
19684	EndoRNAse	orf1ab	-	Val6474Leu	QHD43415.1:p.6474V>L	MODERATE	0	43	8663	0	16	2
21137	O-ribose	orf1ab	-	Lys6958Arg	QHD43415.1:p.6958-; QHD43415.1:p.6958K>R	MODIFIER; MODERATE	8705	0	14	0	3	2
22289	S-Protein	S	-	Ala243Ser	QHD43416.1:p.243A>S	MODERATE	0	2	8716	0	4	2
26144	ORF3a	ORF3a	A1a	Gly251Val	QHD43417.1:p.251G>V; QHD43417.1:p.251-	MODERATE; MODIFIER	0	820	7890	0	12	2
28144	ORF8	ORF8	B	Leu84Ser	QHD43422.1:p.84L>S; QHD43422.1:p.84-	MODERATE; MODIFIER	0	7346	0	1353	23	2
28144	ORF8	ORF8	B	Leu84Ser	QHD43422.1:p.84L>S; QHD43422.1:p.84-	MODERATE; MODIFIER	0	7346	0	1353	23	2
29555	orf	Intergenic	-	0	QHI42199.1	MODIFIER,DISTANCE=3	0	1	0	8710	11	2
29555	orf	Intergenic	-	0	QHI42199.1	MODIFIER,DISTANCE=3	0	1	0	8710	11	2
29729	3'UTR	3'UTR	-	-	QHI42199.1	MODIFIER,DISTANCE=55	0	8457	1	0	264	2
84	5'UTR	5'UTR	-	-	0	MODIFIER	1	3	0	8137	581	1
167	5'UTR	5'UTR	-	-	QHD43415.1	MODIFIER,DISTANCE=99	0	1	8337	1	383	1

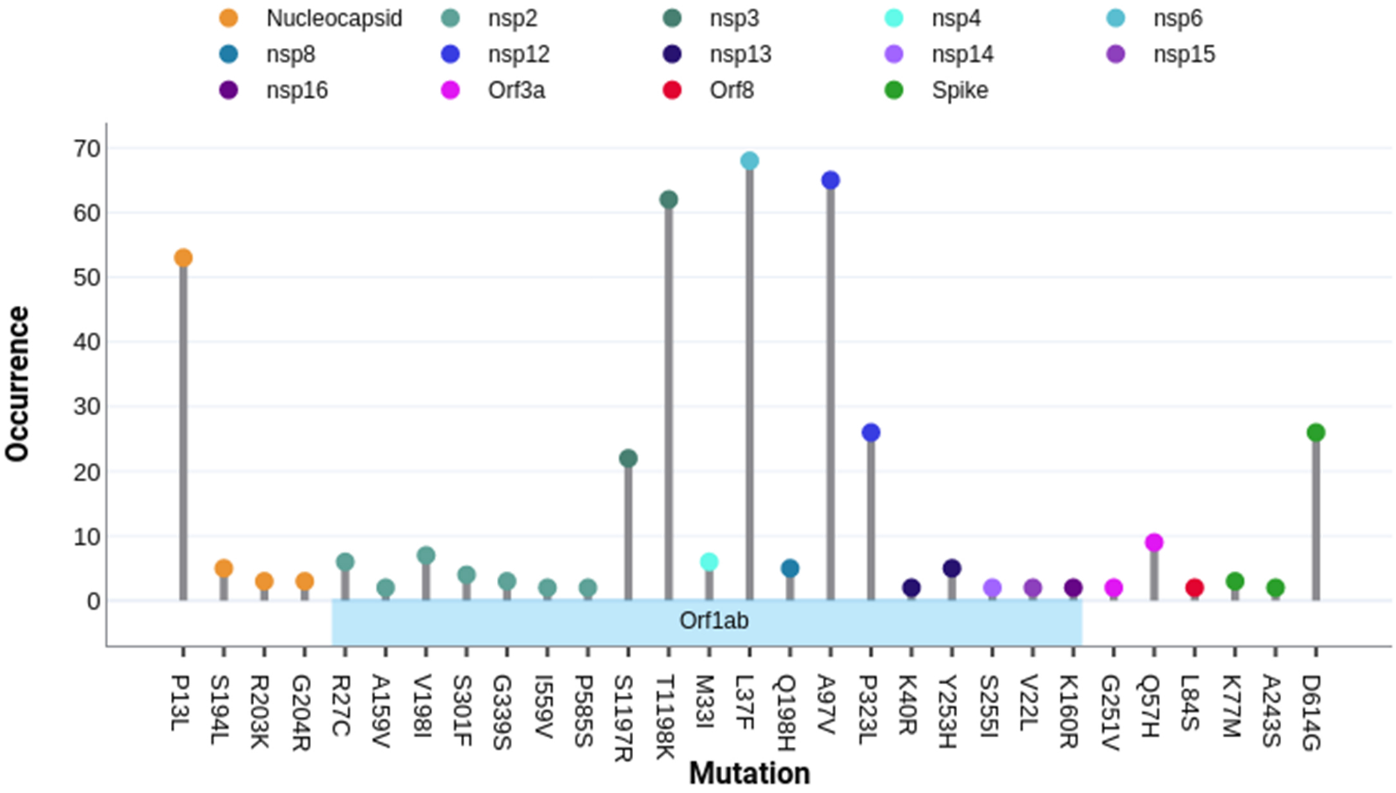
203	5'UTR	5'UTR	-	-	QHD43415.1	MODIFIER,DISTANCE=63	0	3	1	8355	363	1
337	NSP1	orf1ab	-	-	QHD43415.1:p.24R	LOW	0	2	0	8686	34	1
507	NSP1	orf1ab	-	-	QHD43415.1:p.81-86HGHVMV>H	MODERATE	8711	0	0	0	11	1
509	NSP1	orf1ab	-	-	QHD43415.1:p.82-85GHVM>V	MODERATE	0	0	8710	0	12	1
561	NSP1	#N/A	-	-	#N/A	#N/A	0	0	8716	0	6	1
683	NSP1	orf1ab	-	-	QHD43415.1:p.140-143LKSF>L; QHD43415.1:p.140L	MODERATE; LOW	0	3	0	8716	3	1
851	NSP2	orf1ab	-	-	QHD43415.1:p.196Y>H	MODERATE	0	8721	0	0	1	1
1281	NSP2	orf1ab	-	Ala339Val	QHD43415.1:p.339A>V; QHD43415.1:p.339-	MODERATE; MODIFIER	0	3	0	8717	2	1
1601	NSP2	orf1ab	-	-	QHD43415.1:p.446L>I	MODERATE	1	0	0	8719	2	1
1601	NSP2	orf1ab	-	-	QHD43415.1:p.446L>I	MODERATE	1	0	0	8719	2	1
1887	NSP2	orf1ab	-	-	QHD43415.1:p.541A>V; QHD43415.1:p.541-	MODERATE; MODIFIER	0	2	0	8717	3	1
1912	NSP2	orf1ab	-	-	QHD43415.1:p.549S	LOW	0	17	0	8703	2	1
1912	NSP2	orf1ab	-	-	QHD43415.1:p.549S	LOW	0	17	0	8703	2	1
2040	NSP2	orf1ab	-	-	QHD43415.1:p.592T>I	MODERATE	0	1	0	8719	2	1
2485	NSP2	orf1ab	-	-	QHD43415.1:p.740I	LOW	0	2	0	8719	1	1
2558	NSP2	orf1ab	A1a1b	Pro765Ser	QHD43415.1:p.765P>S; QHD43415.1:p.765-	MODERATE; MODIFIER	0	321	0	8390	11	1
3145	NSP3	orf1ab	-	-	QHD43415.1:p.960L>F	MODERATE	0	9	8710	0	3	1
3176	NSP3	orf1ab	-	-	QHD43415.1:p.971P>S	MODERATE	0	4	0	8712	6	1
3429	NSP3	orf1ab	-	-	QHD43415.1:p.1055T>I	MODERATE	0	1	0	8720	1	1
3604	NSP3	orf1ab	-	-	QHD43415.1:p.1113H	LOW	0	0	0	8721	1	1
4144	NSP3	#N/A	-	-	#N/A	#N/A	0	0	8718	0	4	1
4680	NSP3	#N/A	-	-	#N/A	#N/A	0	0	0	8721	1	1
4859	NSP3	#N/A	-	-	#N/A	#N/A	0	0	8720	0	2	1
5008	NSP3	orf1ab	-	-	QHD43415.1:p.1581T	LOW	0	0	8720	0	2	1
5151	NSP3	orf1ab	-	-	QHD43415.1:p.1629V>A	MODERATE	0	8720	0	0	2	1
5657	NSP3	orf1ab	-	-	QHD43415.1:p.1798V>I; QHD43415.1:p.1798V>L	MODERATE	1	0	8720	0	1	1
5730	NSP3	orf1ab	-	-	QHD43415.1:p.1822T>I	MODERATE	0	10	0	8711	1	1
6395	NSP3	orf1ab	-	-	QHD43415.1:p.2044L	LOW	0	3	0	8709	10	1
7071	NSP3	#N/A	-	-	#N/A	#N/A	1	0	8718	0	3	1
7350	ORF1a/b	#N/A	-	-	#N/A	#N/A	0	0	0	8720	2	1
7734	ORF1a/b	#N/A	-	-	#N/A	#N/A	0	8717	0	0	5	1
8595	NSP4	orf1ab	-	-	QHD43415.1:p.2777T>I	MODERATE	0	0	0	8721	1	1
9477	NSP4	orf1ab	-	-	QHD43415.1:p.3071F>Y	MODERATE	150	8567	0	0	5	1
9634	NSP4	orf1ab	-	-	QHD43415.1:p.3123L>F	MODERATE	8710	4	0	0	8	1
10039	NSP4	orf1ab	-	-	QHD43415.1:p.3258T; QHD43415.1:p.3258-	LOW; MODIFIER	0	1	0	8717	4	1
10449	Mpro	orf1ab	-	-	QHD43415.1:p.3395P>L; QHD43415.1:p.3395-	MODERATE; MODIFIER	0	3	0	8717	2	1
11074	NSP6	orf1ab	-	-	QHD43415.1:p.3604F>X; QHD43415.1:p.3603F; QHD43415.1:p.3603F>FX;	HIGH; LOW; MODERATE;	0	22	0	8692	8	1
11109	NSP6	orf1ab	-	-	QHD43415.1:p.3615A>V	MODERATE	0	20	0	8699	3	1
12929	NSP9	orf1ab	-	-	QHD43415.1:p.4222-	MODIFIER	0	0	8720	0	2	1
13038	NSP10	#N/A	-	-	#N/A	#N/A	0	0	0	8720	2	1
13862	NSP12/RdRP	orf1ab	-	-	QHD43415.1:p.4533T>I; QHD43415.1:p.4533-	MODERATE; MODIFIER	0	13	0	8707	2	1
14220	NSP12/RdRP	orf1ab	-	-	QHD43415.1:p.4652D	LOW	0	1	0	8720	1	1
15290	NSP12/RdRP	#N/A	-	-	#N/A	#N/A	0	0	8721	0	1	1
15344	NSP12/RdRP	#N/A	-	-	#N/A	#N/A	0	0	0	8721	1	1
15435	NSP12/RdRP	orf1ab	-	-	QHD43415.1:p.5057-	MODIFIER	8720	0	0	0	2	1
15669	NSP12/RdRP	#N/A	-	-	#N/A	#N/A	0	8720	0	1	1	1
16221	NSP12/RdRP	orf1ab	-	-	QHD43415.1:p.5319P	LOW	0	0	8721	0	1	1
16377	NSP13/Helicase	orf1ab	-	-	QHD43415.1:p.5371P	LOW	0	5	8716	0	1	1
17122	NSP13/Helicase	orf1ab	-	-	QHD43415.1:p.5620A>T; QHD43415.1:p.5620A>S	MODERATE	1	0	8719	0	2	1
17122	NSP13/Helicase	orf1ab	-	-	QHD43415.1:p.5620A>T; QHD43415.1:p.5620A>S	MODERATE	1	0	8719	0	2	1
17403	NSP13/Helicase	orf1ab	-	-	QHD43415.1:p.5713A	LOW	0	2	0	8718	2	1
17415	NSP13/Helicase	orf1ab	-	-	QHD43415.1:p.5717A	LOW	0	8721	0	0	1	1
17415	NSP13/Helicase	orf1ab	-	-	QHD43415.1:p.5717A	LOW	0	8721	0	0	1	1
17656	NSP13/Helicase	orf1ab	-	-	QHD43415.1:p.5798M>V	MODERATE	8700	0	3	0	19	1
17747	NSP13/Helicase	orf1ab	B1a1	-	QHD43415.1:p.5828P>L; QHD43415.1:p.5828-	MODERATE; MODIFIER	0	856	0	7842	24	1
17858	NSP13/Helicase	orf1ab	B1a	-	QHD43415.1:p.5865Y>C	MODERATE	7845	0	876	0	1	1
18060	NSP14/Exonuclea	orf1ab	B1	-	QHD43415.1:p.5932L; QHD43415.1:p.5932-	LOW; MODIFIER	0	886	0	7832	4	1
18078	NSP14/Exonuclea	orf1ab	-	-	QHD43415.1:p.5938K	LOW	1	0	8720	0	1	1
18312	NSP14/Exonuclea	orf1ab	-	-	QHD43415.1:p.6016V	LOW	1	2	0	8718	1	1
18312	NSP14/Exonuclea	orf1ab	-	-	QHD43415.1:p.6016V	LOW	1	2	0	8718	1	1
18573	NSP14/Exonuclea	orf1ab	-	-	QHD43415.1:p.6103S	LOW	0	8720	0	1	1	1
19644	EndoRNAse	#N/A	-	-	#N/A	#N/A	0	8697	0	0	25	1
19861	EndoRNAse	orf1ab	-	-	QHD43415.1:p.6533A>T	MODERATE	1	0	8681	0	40	1
19862	EndoRNAse	orf1ab	-	-	QHD43415.1:p.6533A>V; QHD43415.1:p.6533-	MODERATE; MODIFIER	0	5	0	8676	41	1

19875	EndoRNAse	#N/A	-	-	#N/A	#N/A	0	0	0	8682	40	1
20255	EndoRNAse	orf1ab	-	-	QHD43415.1:p.6664-; QHD43415.1:p.6664D>G	MODIFIER; MODERATE	8720	0	0	0	2	1
20429	EndoRNAse	orf1ab	-	-	QHD43415.1:p.6722P>L	MODERATE	0	1	0	8720	1	1
20580	EndoRNAse	orf1ab	-	-	QHD43415.1:p.6772V; QHD43415.1:p.6772-	LOW; MODIFIER	1	3	8715	0	3	1
20749	O-ribose	#N/A	-	-	#N/A	#N/A	0	0	8721	0	1	1
21707	S-Protein	S	-	-	QHD43416.1:p.49H>Y	MODERATE	0	29	0	8687	6	1
21989	S-Protein	S	-	-	QHD43416.1:p.143-144VY>D; QHD43416.1:p.143V>F	MODERATE	0	1	8714	0	7	1
21990	S-Protein	S	-	-	QHD43416.1:p.143-144VY>V	MODERATE	0	8712	0	0	10	1
22430	S-Protein	#N/A	-	-	#N/A	#N/A	0	0	8518	0	204	1
22458	S-Protein	S	-	-	QHD43416.1:p.299T>I	MODERATE	0	1	0	8469	252	1
23042	S-Protein	S	-	-	QHD43416.1:p.494-	MODIFIER	0	8686	0	1	35	1
23108	S-Protein	#N/A	-	-	#N/A	#N/A	0	0	8678	1	43	1
23660	S-Protein	#N/A	-	-	#N/A	#N/A	0	0	8721	0	1	1
23677	S-Protein	S	-	-	QHD43416.1:p.705V	LOW	0	8721	0	0	1	1
24166	S-Protein	#N/A	-	-	#N/A	#N/A	8719	0	0	0	3	1
24622	S-Protein	S	-	-	QHD43416.1:p.1020-	MODIFIER	0	8719	0	0	3	1
24694	S-Protein	S	B1a1a	-	QHD43416.1:p.1044G	LOW	8634	87	0	0	1	1
24904	S-Protein	S	-	-	QHD43416.1:p.1114-; QHD43416.1:p.1114I	MODIFIER; LOW	0	5	0	8715	2	1
25318	S-Protein	S	-	-	QHD43416.1:p.1252S	LOW	0	0	0	8719	3	1
25318	S-Protein	S	-	-	QHD43416.1:p.1252S	LOW	0	0	0	8719	3	1
25350	S-Protein	S	A2a10	-	QHD43416.1:p.1263P>L; QHD43416.1:p.1263-	MODERATE; MODIFIER	0	51	0	8661	10	1
25642	ORF3a	ORF3a	-	-	QHD43417.1:p.84L	LOW	0	1	0	8720	1	1
25793	ORF3a	ORF3a	-	-	QHD43417.1:p.134R>H; QHD43417.1:p.134R>L	MODERATE	0	3	8706	0	13	1
25826	ORF3a	#N/A	-	-	#N/A	#N/A	8706	1	0	0	15	1
25919	ORF3a	ORF3a	-	-	QHD43417.1:p.176T>I	MODERATE	0	1	0	8711	10	1
25979	ORF3a	ORF3a	-	-	QHD43417.1:p.196G>V; QHD43417.1:p.196-	MODERATE; MODIFIER	0	149	8570	0	3	1
27191	M-Protein	M	-	Ter223Ter	QHD43419.1:p.223-; QHD43419.1:p.223*	MODIFIER; LOW	8717	0	2	0	3	1
27195	orf	#N/A	-	-	#N/A	#N/A	8720	0	0	0	2	1
27208	ORF6	ORF6	-	-	QHD43420.1:p.3H>Y; QHD43420.1:p.3-	MODERATE; MODIFIER	0	3	0	8716	3	1
27364	ORF6	ORF6	-	-	QHD43420.1:p.55E>*; QHD43420.1:p.55E>Q	HIGH; MODERATE	0	1	8720	0	1	1
27384	ORF6	ORF6	-	-	QHD43420.1:p.61D; QHD43420.1:p.61-	LOW; MODIFIER	0	8682	0	38	2	1
27874	ORF7b	Intergenic	-	-	QHD43422.1	MODIFIER,DISTANCE=20	0	0	0	8705	17	1
28115	ORF8	ORF8	-	-	QHD43422.1:p.74I	LOW	0	1	0	8703	18	1
28115	ORF8	ORF8	-	-	QHD43422.1:p.74I	LOW	0	1	0	8703	18	1
28253	ORF8	ORF8	-	-	QHD43422.1:p.120F>L; QHD43422.1:p.120F; QHD43422.1:p.120-;	MODERATE; LOW;	3	5	2	8703	9	1
28657	N-capsid	N	-	-	QHD43423.2:p.128D; QHD43423.2:p.128-	LOW; MODIFIER	0	151	0	8569	2	1
28863	N-capsid	N	-	-	QHD43423.2:p.197S>L; QHD43423.2:p.197-	MODERATE; MODIFIER	0	149	0	8570	3	1
29574	ORF10	#N/A	-	-	#N/A	#N/A	0	8711	0	0	11	1
29614	ORF10	ORF10	-	-	QHI42199.1:p.19C	LOW	0	8	0	8703	11	1
29614	ORF10	ORF10	-	-	QHI42199.1:p.19C	LOW	0	8	0	8703	11	1
29807	3'UTR	#N/A	-	-	#N/A	#N/A	1	8035	0	1	685	1

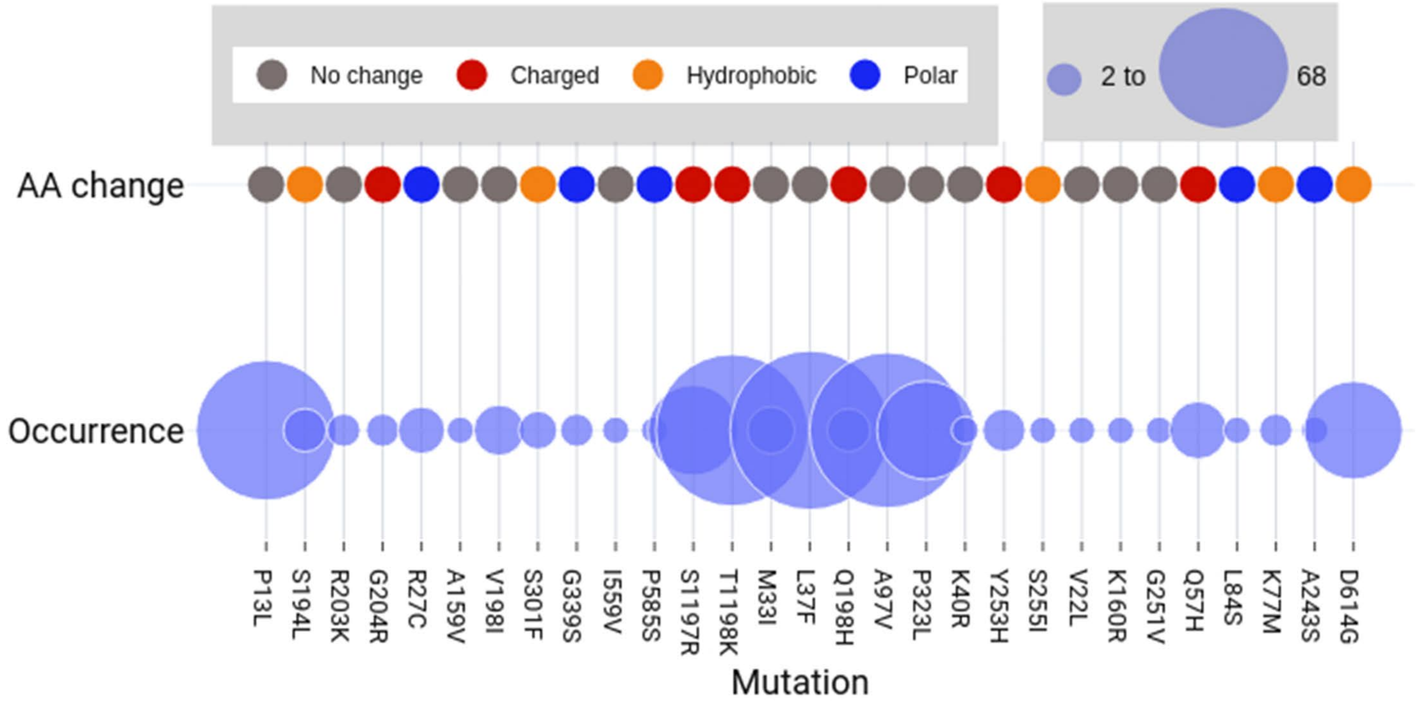

A)
log_coverage
10
8
6
4
2
0
−2
amplicon1 amplicon2 amplicon3 amplicon4 amplicon5 amplicon6 amplicon7 amplicon8 amplicon9 amplicon10 amplicon11 amplicon12 amplicon13 amplicon14 amplicon15 amplicon16 amplicon17 amplicon18 amplicon19 amplicon20 amplicon21 amplicon22 amplicon23 amplicon24 amplicon25 amplicon26 amplicon27 amplicon28 amplicon29 amplicon30 amplicon31 amplicon32 amplicon33 amplicon34 amplicon35 amplicon36 amplicon37 amplicon38 amplicon39 amplicon40 amplicon41 amplicon42 amplicon43 amplicon44 amplicon45 amplicon46 amplicon47 amplicon48 amplicon49 amplicon50 amplicon51 amplicon52 amplicon53 amplicon54 amplicon55 amplicon56 amplicon57 amplicon58 amplicon59 amplicon60 amplicon61 amplicon62 amplicon63 amplicon64 amplicon65 amplicon66 amplicon67 amplicon68 amplicon69 amplicon70 amplicon71 amplicon72 amplicon73 amplicon74 amplicon75 amplicon76 amplicon77 amplicon78 amplicon79 amplicon80 amplicon81 amplicon82 amplicon83 amplicon84 amplicon85 amplicon86 amplicon87 amplicon88 amplicon89 amplicon90 amplicon91 amplicon92 amplicon93 amplicon94 amplicon95 amplicon96 amplicon97
Amplicons
B)
Ct (E-gene)
30
25
20
15
Ct (RdRp-gene)
35
30
25
20
15
Avg Depth
3000
2000
1000
0
Coverage
100
80
60
40
20
10 20 30
10 20 30 40
0 2000 4000
0 50 100
Status
pass
fail
C)
241
884
1397
1707
1947
3037
4582
5730
6081
6310
6312
8653
11083
13730
14408
16993
18148
18395
18877
19524
22420
22444
23403
23929
25461
25563
26230
26735
27191
27847
28311
28688
28854
28881
28882
28883
29729
29689
29742
29861
NCDC2245 NCDC2240 NCDC2251 NCDC2250 NCDC1614 NCDC1604 NCDC1441 NCDC1744 NCDC2105 NCDC2245 NCDC2240 NCDC2251 NCDC2250 NCDC1614 NCDC1604 NCDC1441 NCDC1744 NCDC2105
Nanopore
ref alt mismatch
Miseq
ref alt mismatch


A)
B
A3
A4
A1a
A2
10 samples
1 sample
Noida
Delhi
Punjab
West Bengal
Tamil Nadu
Haryana
Maharashtra
Uttar Pradesh
Mumbai
Odisha
Andhra Pradesh
Jammu
Ladakh
Kargil
Bihar
Rajsathan
Assam
Madhya Pradesh
Karnataka
location unknown
Imported Rest of world
Local
From Indonesia
From Thailand
From Kyrgyz Rep.
B)
A2a
B
A1a
A3
A4
MN908947
C)
10 samples
1 sample
Reference
Singapore
India
Brunei

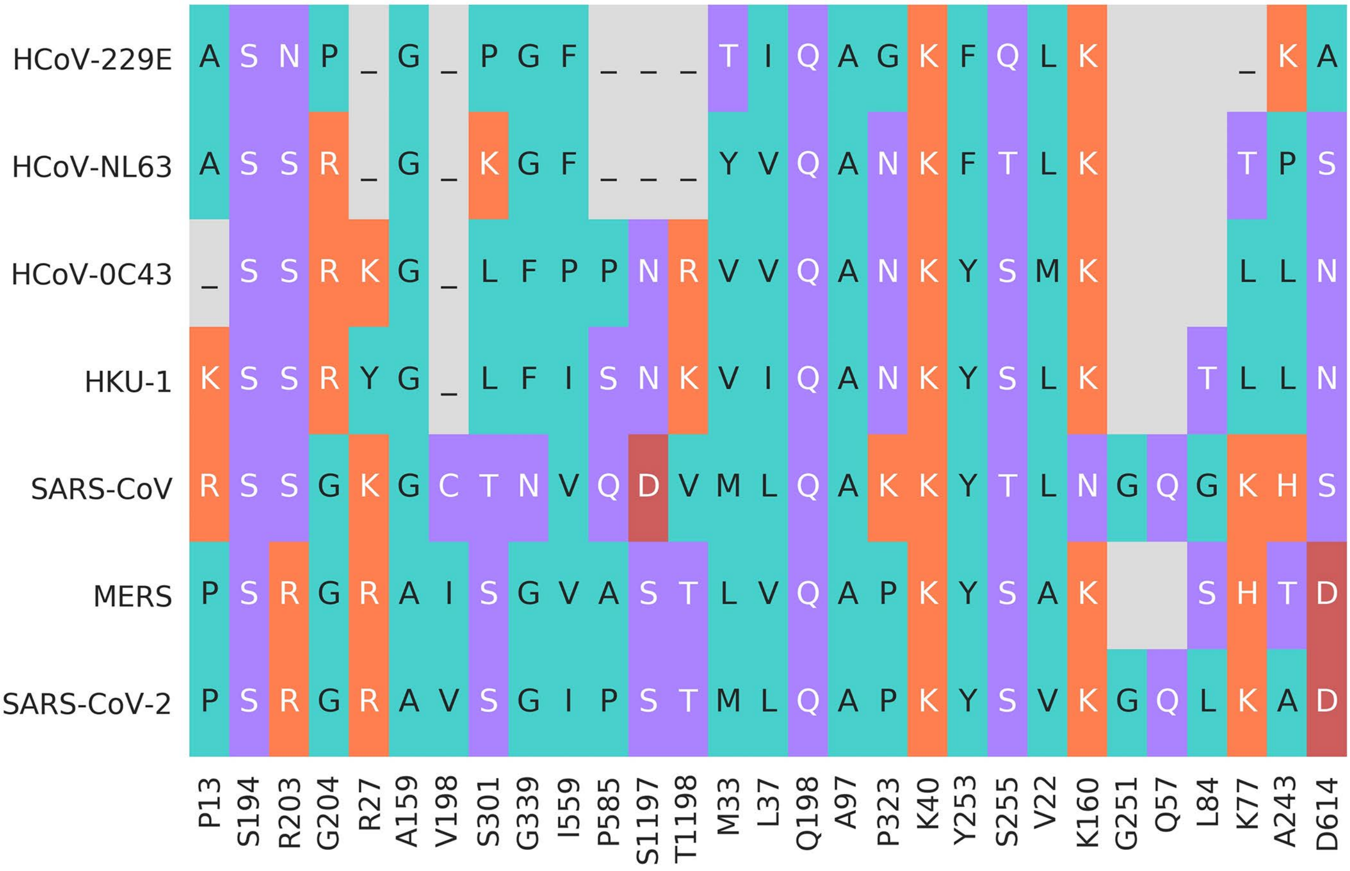

A)
600
500
400
300
200
100
0
6310
NSP3
6312
NSP3
11083
NSP6
13730
NSP12/RdRP
19524
NSP14/Exonuclease
23929
S-Protein
28311
N-capsid
Australia
Singapore
USA
Philippines
India
Taiwan
Malaysia
Canada
Fujian
Slovenia
India-NCDC-IGIB
B)
70
60
50
40
30
20
10
0
Leu3606Phe
11083
NSP6
Ala4489Val
13730
NSP12/RdRP
Pro13Leu
28311
N-capsid
Thr2016Lys
6312
NSP3
Tyr789Tyr
23929
S-Protein
Ser2015Arg
6310
NSP3
Odisha
Nepal
Jammu
Haryana
Rajsathan
Madhya Pradesh
Kyrgyz Republic
Karnataka
Thailand
Assam
Andhra Pradesh
Uttar Pradesh
West Bengal
Maharashtra
Bihar
Indonesia
Delhi
Tamil Nadu
C)
clade
13730
23929
6312
28311
11083
19524
6310
23185
26467
26144
12685
1707
7392
2897
1820
27588
27147
7621
19299
19422
1281
21306
21792
3176
6463
10138
26763
26894
6989
16355
21137
21628
23359
6081
8022
15435
2081
2094
7600
17403
18803
11200
16017
17373
22289
28758
29555
N
alt
ref
clade
Brunei
India
India-NCDC
Singapore


Nucleocapsid
nsp2
nsp3
nsp4
nsp6
nsp8
nsp12
nsp13
nsp14
nsp15
nsp16
Orf3a
Orf8
Spike
Occurrence
70
60
50
40
30
20
10
0
Orf1ab
P13L
S194L
R203K
G204R
R27C
A159V
V198I
S301F
G339S
I559V
P585S
S1197R
T1198K
M33I
L37F
Q198H
A97V
P323L
K40R
Y253H
S255I
V22L
K160R
G251V
Q57H
L84S
K77M
A243S
D614G
Mutation

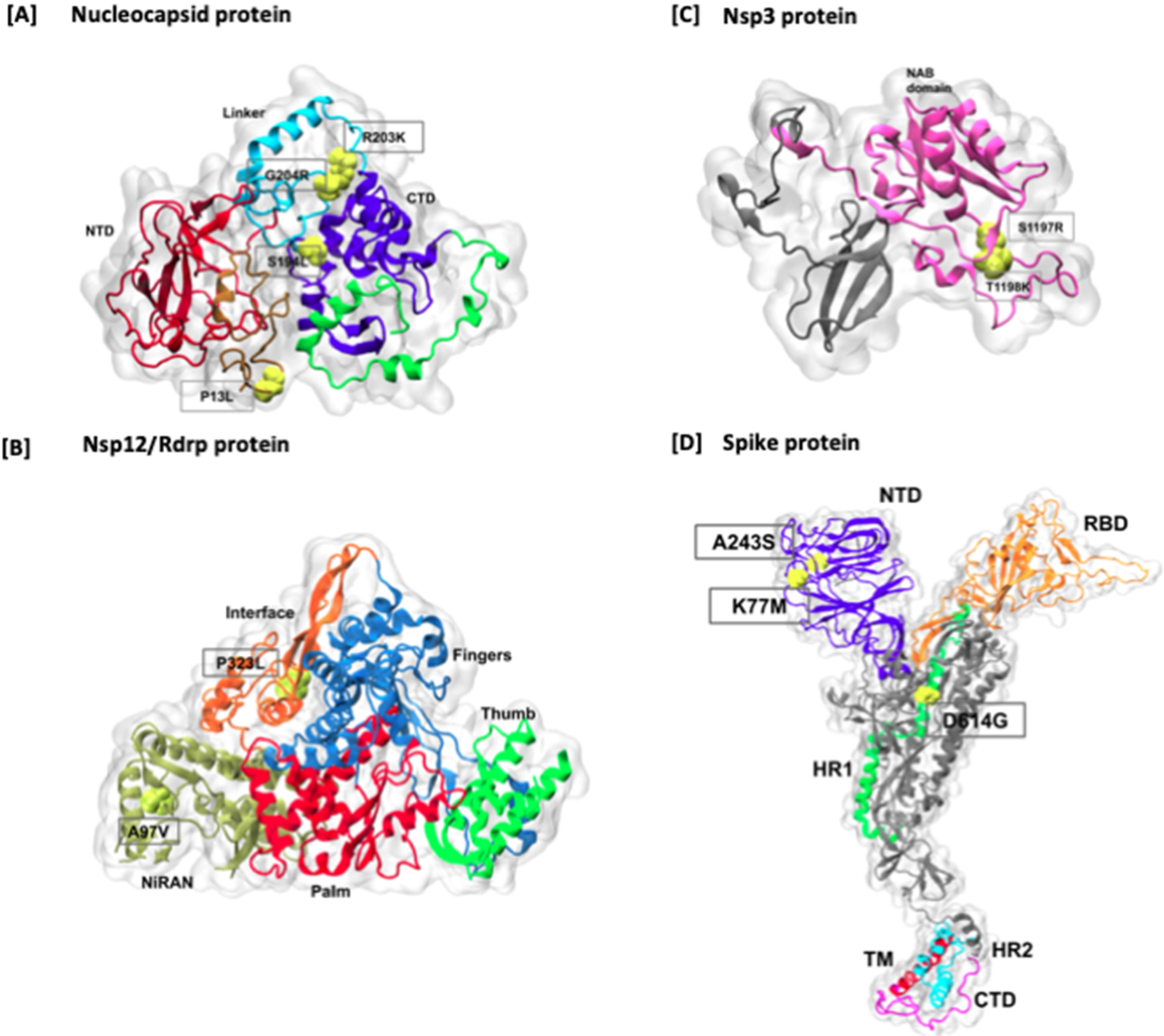
A)



B)



	P13	S194	R203	G204	R27	A159	V198	S301	G339	I559	P585	S1197	T1198	M33	L37	Q198	A97	P323	K40	Y253	S255	V22	K160	G251	Q57	L84	K77	A243	D614
HCoV-229E	A	S	N	P	_	G	_	P	G	F	_	_	_	T	I	Q	A	G	K	F	Q	L	K				_	K	A
HCoV-NL63	A	S	S	R	_	G	_	K	G	F	_	_	_	Y	V	Q	A	N	K	F	T	L	K				T	P	S
HCoV-0C43	_	S	S	R	K	G	_	L	F	P	P	N	R	V	V	Q	A	N	K	Y	S	M	K				L	L	N
HKU-1	K	S	S	R	Y	G	_	L	F	I	S	N	K	V	I	Q	A	N	K	Y	S	L	K			T	L	L	N
SARS-CoV	R	S	S	G	K	G	C	T	N	V	Q	D	V	M	L	Q	A	K	K	Y	T	L	N	G	Q	G	K	H	S
MERS	P	S	R	G	R	A	I	S	G	V	A	S	T	L	V	Q	A	P	K	Y	S	A	K			S	H	T	D
SARS-CoV-2	P	S	R	G	R	A	V	S	G	I	P	S	T	M	L	Q	A	P	K	Y	S	V	K	G	Q	L	K	A	D

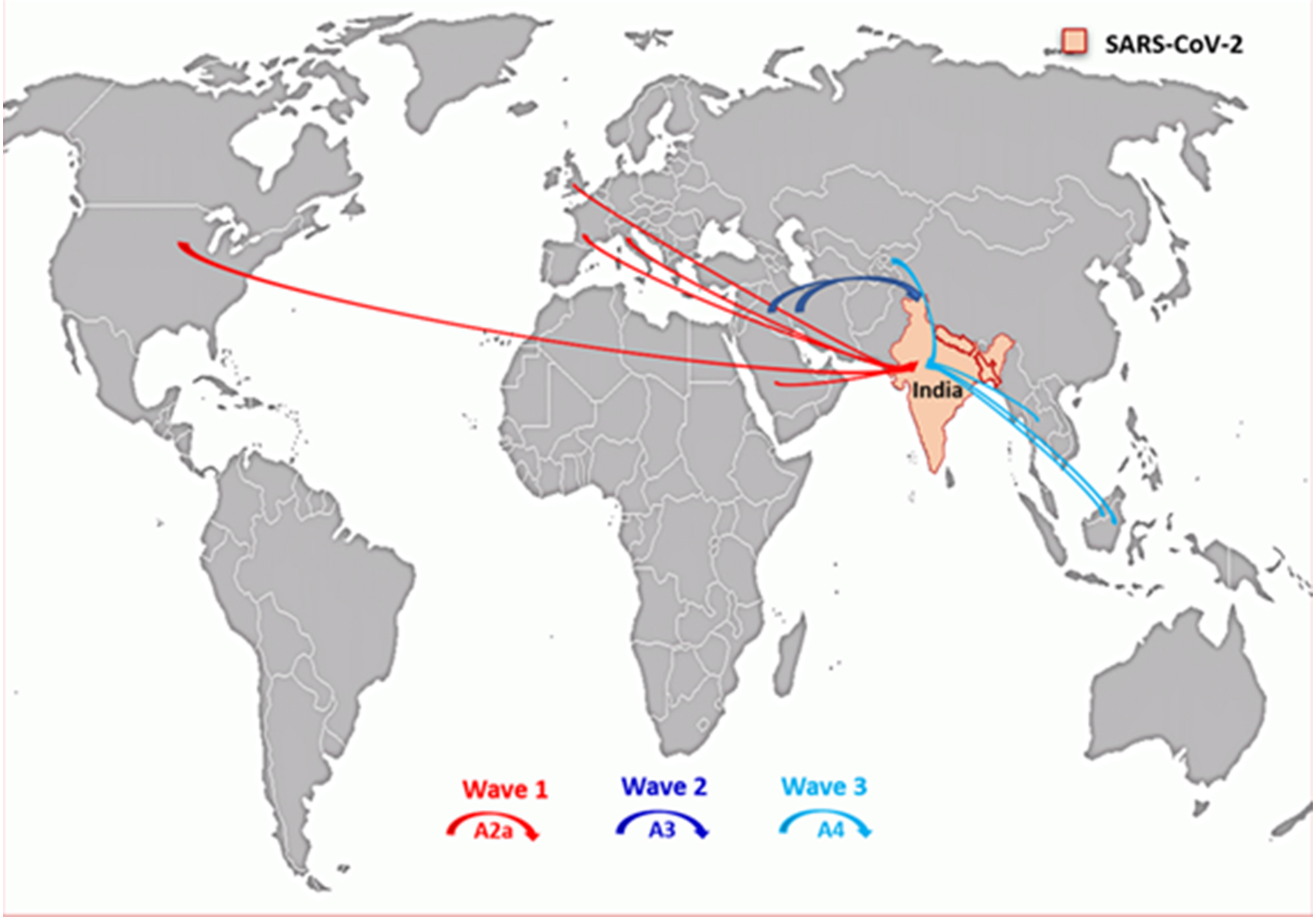


[A] Nucleocapsid protein

[C] Nsp3 protein

[B] Nsp12/Rdrp protein

[D] Spike protein


SARS-CoV-2
India
Wave 1
A2a
Wave 2
A3
Wave 3
A4