scientific reports

OPEN

Genomic epidemiology of SARS-CoV-2 from Uttar Pradesh, India

Gauri Misra[1]✉, Ashrat Manzoor[1], Meenu Chopra[2], Archana Upadhyay[1], Amit Katiyar[3], Brij Bhushan[1] & Anup Anvikar[1]

The various strains and mutations of SARS-CoV-2 have been tracked using several forms of genomic classification systems. The present study reports high-throughput sequencing and analysis of 99 SARS-CoV-2 specimens from Western Uttar Pradesh using sequences obtained from the GISAID database, followed by phylogeny and clade classification. Phylogenetic analysis revealed that Omicron lineages BA-2-like (55.55%) followed by Delta lineage-B.1.617.2 (45.5%) were predominantly circulating in this area Signature substitution at positions S: N501Y, S: D614G, S: T478K, S: K417N, S: E484A, S: P681H, and S: S477N were commonly detected in the Omicron variant-BA-2-like, however S: D614G, S: L452R, S: P681R and S: D950N were confined to Delta variant-B.1.617.2. We have also identified three escape variants in the S gene at codon position 19 (T19I/R), 484 (E484A/Q), and 681 (P681R/H) during the fourth and fifth waves in India. Based on the phylogenetic diversification studies and similar changes in other lineages, our analysis revealed indications of convergent evolution as the virus adjusts to the shifting immunological profile of its human host. To the best of our knowledge, this study is an approach to comprehensively map the circulating SARS-CoV-2 strains from Western Uttar Pradesh using an integrated approach of whole genome sequencing and phylogenetic analysis. These findings will be extremely valuable in developing a structured approach toward pandemic preparedness and evidence-based intervention plans in the future.

The severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) is extremely contagious and has spread throughout the world, leading to over 593 million confirmed cases and 6.4 million deaths till August 21, 2022, globally, since its appearance in Hubei province, China, in December 2019[1]. India reported its first case on January 30, 2020, from Kerala[2]. As on August 21, 2022, India reported over 40 million SARS-CoV-2 cases with a 5.3 lakh death toll (WHO Coronavirus Disease (COVID-19) dashboard available online at: https://covid19.who.int/). The world has witnessed three outbreaks due to *coronaviruses* with significantly high morbidity rates which were: SARS-CoV in 2002, MERS-CoV (Middle East Respiratory Syndrome coronavirus) in 2012, and COVID-19 in 2019[3,4]. SARS-CoV-2 displays higher pathogenicity and transmissibility as compared to MERS-CoV and SARS-CoV[5,6].

Coronaviruses are enveloped positive sense RNA viruses, ranging from 60 to 140 nm having an approximate genome size of 26–32 kb[7]. Belonging to the *Coronaviridae* family and genus *Betacoronavirus*, they are positive-sense single-stranded RNA genomes and encode for sixteen nonstructural (NSP1-NSP16) and four structural proteins (nucleocapsid, envelop, membrane, and spike glycoprotein). Spike (S) glycoprotein has a crown-like appearance that facilitates its attachment to the surface angiotensin-converting enzyme 2 (ACE2) receptor[8,9]. The receptor binding domain (RBD) which is 223 residues in length primarily binds to the ACE2 receptor[9,10].

Beginning in the 1960s, patients with the common cold were the first group of people infected with human *coronaviruses* (HCoVs). Following this, seven HCoVs have been reported that infect humans: 229E, OC43, SARS-CoV, NL63, HKU1, MERS-CoV, and SARS-CoV-2[11–15]. SARS-CoV-2 is subjected to genetic evolution and subsequent antigenic variations through mutations, enabling them to synchronize in human hosts. This may impact the structure and functional activity of the virus[16]. Salient mutations specifically D614G were reported to be associated with viral transmissibility, high virulence, and less protection with current vaccines[17]. One of these variants was discovered in humans, and it was this version that spread the disease from an infected farmed mink

[1]Molecular Diagnostics and COVID-19 Kit Testing Laboratory, National Institute of Biologicals (Ministry of Health and Family Welfare), A-32, Sector-62, Institutional Area, Noida, UP 201309, India. [2]National Dairy Research Institute, Karnal, Haryana, India. [3]Bioinformatics Facility, Centralized Core Research Facility, All India Institute of Medical Sciences, Ansari Nagar, New Delhi 110029, India. ✉email: gauri.misra@nib.gov.in; kamgauri@gmail.com

in Denmark. However, it did not relate to the increased transmissibility[18,19]. The World Health Organization's (WHO) Technical Advisory Group on Virus Evolution classifies variations that pose an enhanced danger to global public health as "Variants of Concern" (VOC)[20]. In January 2021, WHO classified B.1.617.2 (Delta variant) as a VOC which had a Spike (S) double mutation namely E484Q and L452R. According to WHO guidelines, every reported variant is categorized as either a variant of concern (VOC) or a variant of interest (VOI)[21]. Thereafter, in April 2021, Delta was sub-classified to Delta sub-variants namely AY.1, AY.2, and AY.3 comprising of an additional Spike protein substitution K417N. The predominant variant reported in India in October 2020 was B.1.617 which was later observed in more than 20 countries. It contained two significant mutations (E484Q and L452R) in the sequence of protein S, and therefore was also known as a "double mutant". The name was not rational due to the presence of 11 other mutations. Of the 11 mutations, the P681R mutation had the potential to induce greater pathogenicity with higher affinity towards the ACE2 receptor resulting in immune system evasion[20]. The second wave of the COVID-19 pandemic struck India adversely in 2021, along with numerous post-vaccination breakthrough infections brought on by novel strains[22].

India has witnessed an upsurge in the incidence of distinct SARS-CoV-2 strains in different states since the beginning of the pandemic. The clade I/A3i is a distinctive phylogenetic cluster identified from Indian genomes, dominated the early months of the pandemic (March and April), but by late April, a shift in clade frequency was seen, with most states revealing a higher presence of the Nextstrain clade A2a[23]. Uttar Pradesh is one of the states with the highest population densities in the nation (Census 2011, India) with an international trans-border with Nepal (Fig. 1). Cases proliferated more due to the travel from metro cities from March to April 2021 which gave rise to the deadly variants in the Western Uttar Pradesh region[24]. Basti, a district from Uttar Pradesh had reported the first instance of SARS-CoV-2 infection, which further lead to the spread of infection and transmission in the nearby areas of Uttar Pradesh[23]. Uttar Pradesh has been hit badly due to the second spike of the pandemic in 2021, in addition to many post-vaccination breakthrough infections due to new variants[25]. Our primary focus was on VOCs due to their major impact on public health.

Deployment of important functional bioinformatics tools like whole genome analysis of viral materials aids in the discovery of novel SARS-CoV-2 genetic variants spreading in populations, especially in the case of a global pandemic[26,27]. Genomic surveillance is an essential approach for investigating any outbreak and mapping the virus evolution and spread in case of emerging and re-emerging viruses and has resulted in the identification of variations with spike protein mutations that may confer increased transmissibility and immune evasion (such as alterations D614G, E484K/Q, K417T, N501Y, and P681H)[28–31]. Lineages have been used to define SARS-CoV-2 genetic variation, and an extensive nomenclature scheme has been devised[32]. During the SARS-CoV-2 pandemic, molecular and genetic characterization along with phylogenetic studies were exploited to investigate and monitor the dimensions of virus transmission[33–41]. In particular, large-scale or whole genome sequencing has culminated in the identification of genomic variants correlated with increased transmissibility, virulence, or evasion of host immune response across the world[42,43]. Tracking the emergence, dissemination, and genetic features of lineages, particularly variations of concern (VOCs), has become crucial due to their greater transmissibility, potential increased clinical severity, and ability to evade host immune responses[44].

While an array of studies on the genetic epidemiology of SARS-CoV-2 from various Indian states have been published[23,41,45–47], there has been a paucity of genomic data for SARS-CoV-2, particularly from Western Uttar Pradesh needed to assess the genetic epidemiology of COVID-19 and the prevalence of different lineages of the virus in the state. As a result, identifying and characterizing circulating lineages and sub-lineages is a cornerstone of worldwide epidemiological surveillance and policy[44]. All of these findings comprehensively underscore the

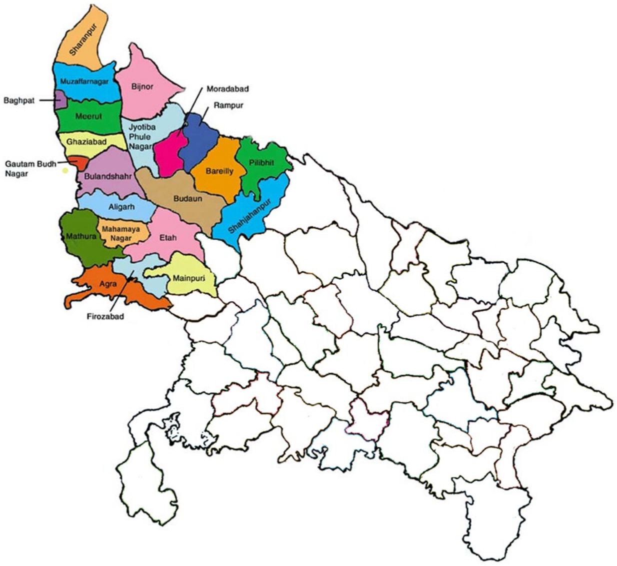


Figure 1. Highlighted regions of the map of Uttar Pradesh displaying the sample collection points.

critical significance of genomic monitoring programs in enabling countries to deploy evidence-based measures to combat the emergence and spread of novel SARS-CoV-2 variants[48]. The present study focuses on the molecular surveillance of SARS-CoV-2 strains from Western Uttar Pradesh to produce a robust database. This is the first comprehensive study from the Western region of Uttar Pradesh that recapitulates the SARS-CoV-2 variants prevalent in this region during the pandemic using molecular-based approaches.

Results

Whole genome sequence analysis

A total of 3,485 samples tested positive in real-time (RT-PCR) for SARS-CoV-2 in this study and 99 samples were processed for whole genome sequencing. The Ct values for positive tests ranged from 19 to 21 Ct which was in accordance with the WHO guidelines for clinical samples. Library preparation and quantification yielded 150 bp paired-end reads that were further diluted to final optimal loading concentrations for cluster amplification and sequencing.

Clade distribution

From March 2021–Jan 2022, 99 SARS-CoV-2 sequences from Uttar Pradesh were analyzed using Pangolin (V4.1.3) and Next Clade (V2.9.1). The phylogenetic analysis confirmed that SARS-CoV-2 sequences were grouped into two major lineages, Delta (B.1.617.2-like) and Omicron (BA.2-like) (Supplementary File 1). As per the Next Clade classification, the outbreak showed six major clades of SARS-COV-2: 21L-Omicron (75.7%) followed by 21A (Delta) 11/99 (11.1%), 21J (Delta) 10/99 (10%), 20A 2/99 (2.02%) and 21B (Kappa) 1/99 (1%) of SARS-CoV-2 (Fig. 2). Further, according to the Pangolin lineage, VOCs were detected with a majority of the Omicron variant (BA.2-like: 75.7%) followed by the Delta variant (B.1.617.2: 19.19%, AY.122: 1%, AY.88: 1.01%, B.1.36: 1.01%, B.1.633: 1.01%) and kappa variant (B.1.617: 1.01%) as represented in Fig. 3. Using the Pangolin software's web version, the Pangolin lineage of the aforementioned sequences was discovered (https://Pangolin.cog-uk.io/).

Distribution of mutations in the spike protein sequence

The presence of particular mutations is directly associated with the lineage assignment during the early pandemic period when SARS-CoV-2 genetic diversity is low. Genome analysis and study of the 99 SARS-CoV-2 genomic sequences demonstrated a total of 50 mutations in BA.2 (Omicron) and 33 in B1.167.2 (Delta) lineage as compared to a reference genome, across the genome (Supplementary File 2). Out of 50 mutations, 27 mutations were located in the spike protein of Omicron. Likewise, out of 34, seven were located in the S gene of the Delta variant (Figs. 4, 5). B.1.617.2 was initially identified in India in December 2020 and carries the spike mutation T19R, G142D, L452R, T478K, D614G, P681R, and D950N including two mutations in the NTD (T19R, G142D), two in the RBD (L452R and T478K), two mutations close to the furin-cleavage site (D614G, P681R), and one in the S2 region (D950N). Signature substitutions according to the total number of occurrences (B1.617.2 and BA.2) are D614G (266), S: T478K (261), S: G142D (165), S: N679K (150), S: P681H/R (152), S: H655Y (149), S: E484A (147), S: Q493R (147), S N501Y (145) while S: T478K (187), P681R (113) and L452R (112) are specifically present in B1.617.2 lineage (Delta variant). It is also noted that Kappa (B.1.617) carried additional mutations, namely S: E154K, and S: E484Q. Some additional mutations are also present in the Delta, Omicron, and Kappa variants, distinguishing them from each other (Fig. 6). These mutations are correlated to enhanced transmissibility, infectivity, receptor binding, and immune escape of the virus.

Phylogenetic analysis

Among the studied cases (99), one sequence (OP787530.1– Western Uttar Pradesh) demonstrated genetic relatedness to the Wuhan isolate having some unique variations that are not present in other isolates. A total of 270 genomes identified from different parts of India by the GISAID are categorized into four clades, viz. GRA, GK, G, and GH. Based on a genetic analysis of the Indian sequences, strains in the clade GRA comprised the largest

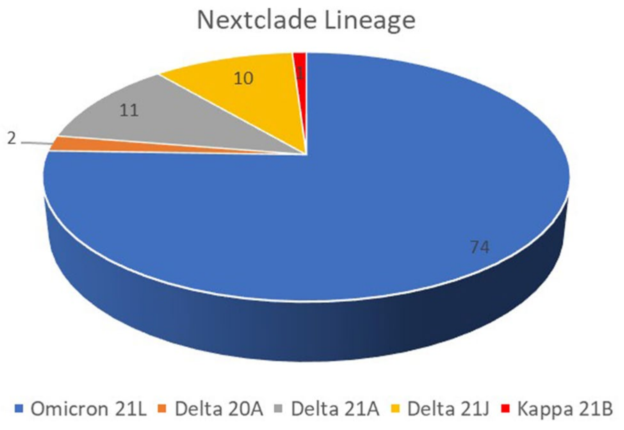


Figure 2. Next Clade classification among 99 genomic sequences of Delta and Omicron Variants.

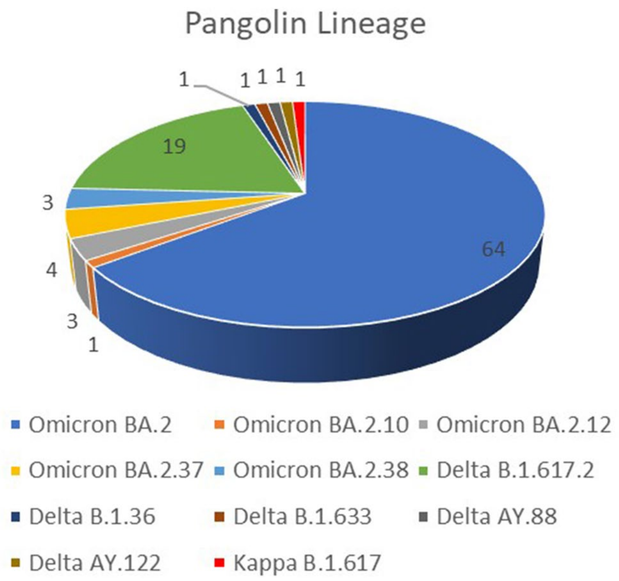


Figure 3. Pangolin classification among 99 genomic sequences of Delta and Omicron Variants.

Who Lable	Other Name	Spike Protein Mutation	ORF Mutation	Other Mutation
Omicron	BA.2	T19I,G142D,G339D,S371F,S373 P,S375F,T376A,D405N,R408S K417N,N440K,S477N,T478K, E484A,Q493R,Q498R,N501Y, Y505H,D614G,H655Y,N679K, P681H,S704L,N764K,D796Y, Q954H, N969K	ORF1a:S135R, ORF1a:T842I,ORF1a:G1307S, ORF1a:R1421K,ORF1a:L3027F,ORF1a:T3090I ORF1a:L3201F,ORF1a:T3255I,ORF1a:P3395H, ORF1a:G3676S,ORF1a:F3677L, ORF1b:P314L ORF1b:R1315C,ORF1b:I1566V,ORF1b:T2163I, ORF3a:T223I, ORF6:D61L	E:T9I M:A63T N:P13L N:R203K N:G204R N:S413R

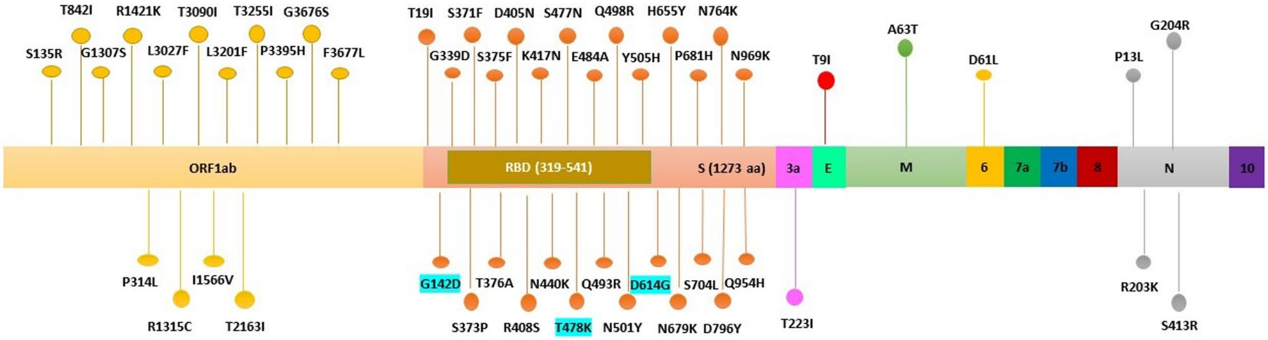
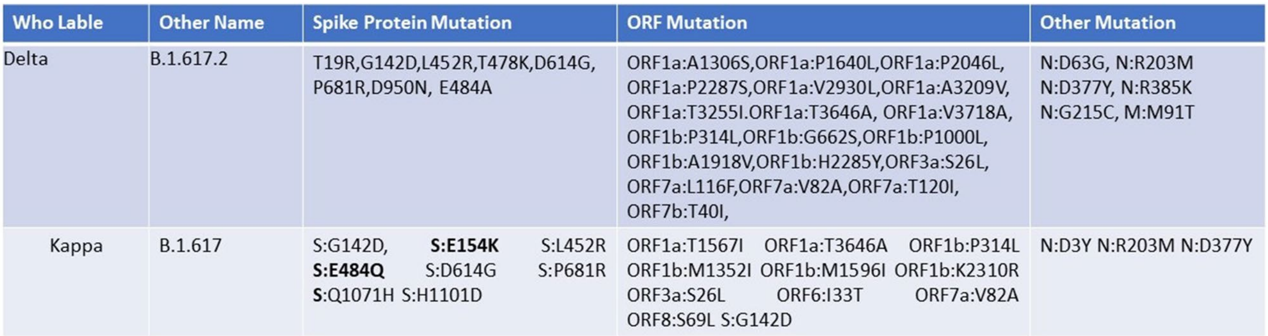


Figure 4. Representation of mutations observed within 75 samples of the Omicron Variants with different colors corresponding to different genes. Mutation common (G142D, T478K, and D614G) in Omicron and Delta variants are highlighted with blue color.

proportion (151) followed by GK (113), GH (1), and G (1). It was noted that omicron variants from all the states (Uttar Pradesh, Maharashtra, and Kerala) have belonged to clade GRA while clade GK is majorly circulated in Delta variants from all the states. During the evolutionary development of SARS-CoV-2, the group clade designating the site of diversification (marked by a red dot) emerged as the most noteworthy (Fig. 7).

Discussion

India was alerted about the SARS-CoV-2 transmission since the report of the first case of SARS-CoV-2 in the Kerala state[2]. The country-wide lockdown (March 2021 to January 2022) was announced by the Government of India to restrict viral transmission. A remarkable and distinct genetic diversity exists in India as evidenced by a few research studies[25,49,50]. Western Uttar Pradesh has been hit by two major COVID-19 spikes, leading to



Who Lable	Other Name	Spike Protein Mutation	ORF Mutation	Other Mutation
Delta	B.1.617.2	T19R,G142D,L452R,T478K,D614G, P681R,D950N, E484A	ORF1a:A1306S,ORF1a:P1640L,ORF1a:P2046L, ORF1a:P2287S,ORF1a:V2930L,ORF1a:A3209V, ORF1a:T3255I.ORF1a:T3646A, ORF1a:V3718A, ORF1b:P314L,ORF1b:G662S,ORF1b:P1000L, ORF1b:A1918V,ORF1b:H2285Y,ORF3a:S26L, ORF7a:L116F,ORF7a:V82A,ORF7a:T120I, ORF7b:T40I,	N:D63G, N:R203M N:D377Y, N:R385K N:G215C, M:M91T
Kappa	B.1.617	S:G142D, **S:E154K** S:L452R **S:E484Q** S:D614G S:P681R **S**:Q1071H S:H1101D	ORF1a:T1567I ORF1a:T3646A ORF1b:P314L ORF1b:M1352I ORF1b:M1596I ORF1b:K2310R ORF3a:S26L ORF6:I33T ORF7a:V82A ORF8:S69L S:G142D	N:D3Y N:R203M N:D377Y

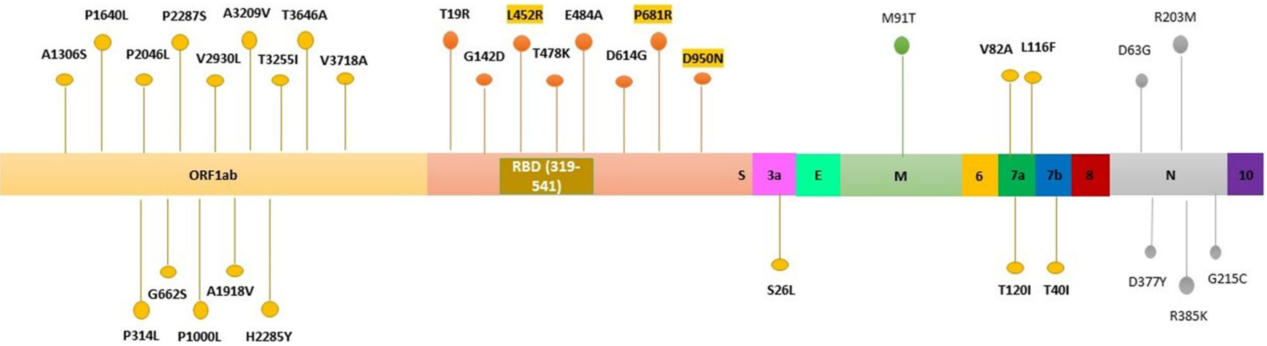


Figure 5. Representation of mutations observed within 23 samples of Delta Variant with different colors corresponding to different genes. Specific Mutations (L452R, P681R, and D950N) in the Spike protein of the Delta variants are highlighted with yellow color.

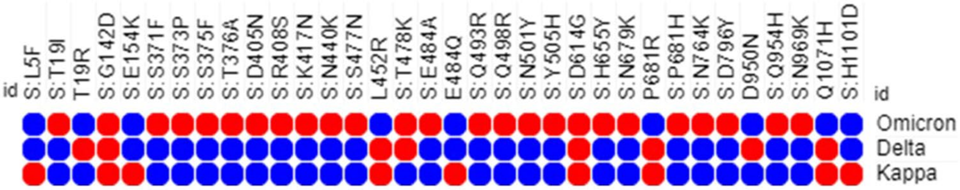


Figure 6. Heatmap of presence, and absence of spike protein mutation among 99 genome sequences. Red color represents the presence and blue represents the absence.

severe mortality in the state. Despite the higher number of cases during the ongoing pandemic period, there were several national movements such as the elections and many other related activities in public places that facilitated the spread of the SARS-CoV-2 infections from the cities to the rural areas[25].

This study unravels the prevalence of VOCs such as Omicron and Delta in the Western Uttar Pradesh region. Recent studies have shown that 56 peculiar single nucleotide polymorphism (SNP) variations among SARS-CoV-2 were found in central Uttar Pradesh which was distinguished in two major clusters that showed rigorous and detrimental effects on the genome[50,51]. In the early stages of the SARS-CoV-2 pandemic, Southeast Asia (B.6.6), Europe (B.1), and other regions of India (B.1.210 and B.1.247) reported the majority of cases[32,44]. During the second spike of the pandemic in Uttar Pradesh, VOCs such as Delta and Delta AY.1 were majorly introduced[25,50,51]. The introduction of the new SARS-CoV-2 lineage B.1.617 and sub-lineage B.1.617.2 (Delta variant), a "variant of concern" in India during the second major spike of SARS-CoV-2, has led to an upsurge in consistent and breakthrough infections[51]. According to INSACOG, the diversification and grouping of seven Pangolin lineages were discovered by the SARS-CoV-2 genomic data from Uttar Pradesh. (http://clingen.igib.res.in/ covid19genomes/).

According to our study, the 99 SARS-CoV-2 samples sequenced herein, distinctively clustered to BA.2-like and B.1.617.2 Pangolin lineages respectively. The present study showed a change in the dominant clade starting from GK to GRA. Late in 2021, the Omicron variant (clade GRA) overtook the Delta variant as the dominant variant in India as well as globally[52]. Over the different epidemic waves in India, the tree branched out according to the clades and clearly demonstrated the replacement of clades over time. SARS-CoV-2 sequences generated in the lab represented some of the genetic lineages that were spreading country-wide at that time. Sequences belonging to the same Pango branch from different states were found to be more closely related. It was noted how the predominance of the various clades varied by state. Among four clades, GRA and GK showed the mixing of

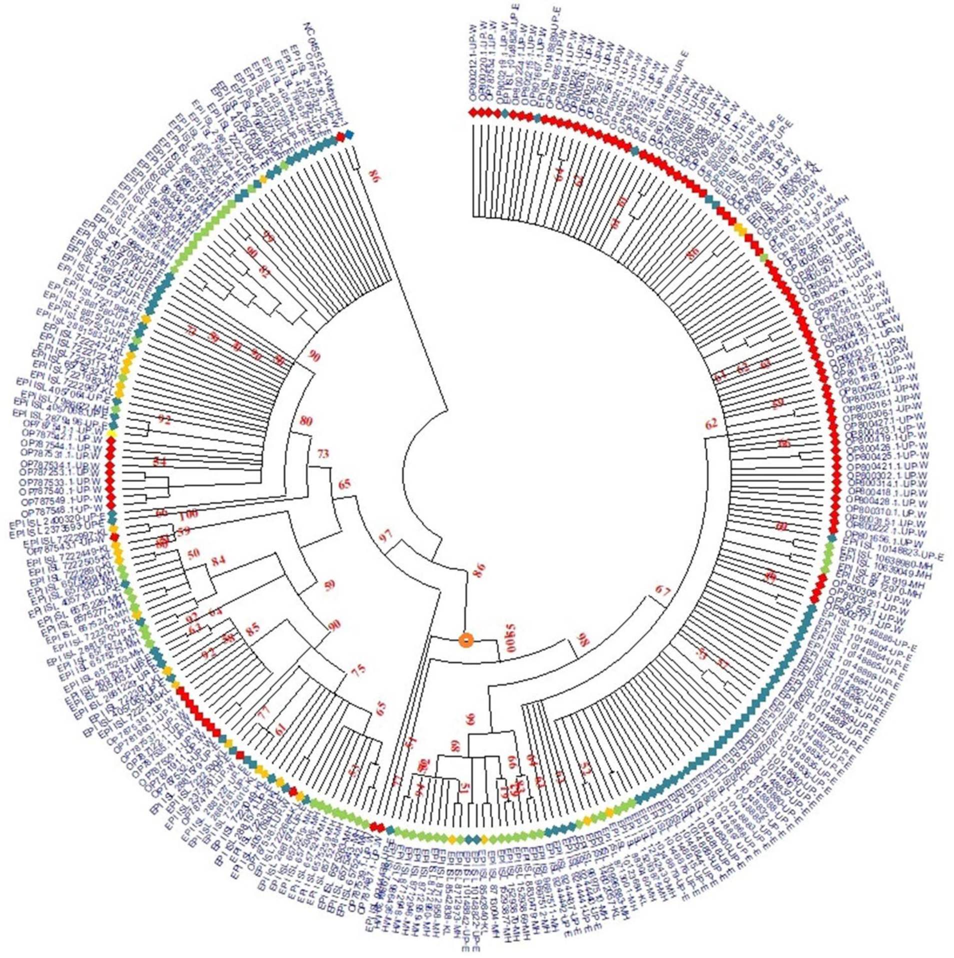


Figure 7. Maximum likelihood tree of the SARS-CoV-2 sequences. The maximum likelihood phylogenetic tree was inferred from the sequences retrieved from Western Uttar Pradesh (Red diamond) along with other GISAID sequences from Eastern Uttar Pradesh (light blue diamond), Maharashtra (green diamond), and Kerala (yellow diamond). The reference sequence of Wuhan is represented by (blue diamond).

strains (Omicron and Delta) from Uttar Pradesh, Maharashtra, and Kerala. These genomes are clustered into two clusters of Delta and Omicron strains.

One reference sequence (OP787530.1– Western Uttar Pradesh) belongs to the 20A clade and B.1.36 Pango lineage showed very close similarity with the reference sequence of Wuhan (NC_045512.2). Phylogenetic analysis showed that SARS-CoV-2 strains (OP787530.1– Western Uttar Pradesh) could be descending from the original Wuhan strain with four unique mutations (N: S194L, ORF1a: V561A, ORF1a: P971L, ORF3a: Q57H). A previous study from China reported that Q57H mutation is responsible for the fourth wave of COVID-19 in Hong Kong China[53]. Another study from India showed that G25563T/Q57H in ORF3a (n = 190/837) was the most frequent mutation found in India, principally in Western India[54]. This reinforces the fact that travel and migration were the major contributors to the pandemic spread. In fact, the psychological fear of social isolation even prevented people from disclosing their travel history. Thus, there is a high probability that OP787530.1– Western Uttar Pradesh sequence similarity could be attributed to such potential travel/ migration history.

A characteristic mutation in the spike protein of clades GRA/GH/GK is D614G, which escalates binding to the angiotensin-converting enzyme 2 (ACE2) receptor and eventually surges the viral entry into host cells. Aside from D614G, GRA (BA.2 lineage) variants with E484A mutations also exhibit substantial antibody neutralization resistance, contributing to an improved vaccine-breakthrough capability[55]. In contrast, the B.1.617.2 lineage, which carries the mutations L452R and P681R in the spike protein, may explain the increase in cases in Western Uttar Pradesh in March 2021.

Genomic epidemiology and whole genome sequencing have been widely used to monitor the transmission and evolution of the SARS-CoV-2 virus globally[29–31]. Mutations in SARS-CoV-2 have recently emerged as a major concern around the globe, possibly affecting its transmission, infectivity, virulence, and immune escape. As of now, monoclonal antibodies and vaccines primarily target the spike protein, which plays a prime role in viral attachment and entry into host cells. The receptor-binding domain (RBD) has several genomic variations and diversity in SARS-CoV-2 spike glycoprotein[56–58]. Patient samples were collected from March 2021 to January 2022 in which most of the patients had received booster doses of the vaccine and were still infected by the virus. This indicated that there was an immunity escape observed in these patients. Three escape mutations in the S gene at codon position 19 (T19I/R), 484 (E484A/Q), and 681 (P681R/H) during the fourth and fifth waves in India have a critical role in immune escape in SARS-CoV-2 infections, were dominantly found in our study subjects. Substitution at positions P681 and E484 has become increasingly common among clinical isolates. Previous studies showed that the virulence and pathogenesis of the Delta variant could be impacted by D614G and P681R mutations. Although the D614G mutation occurs in the Omicron variant, the additional presence of the P681H mutation may result in slow cleavage. Moreover, this may limit the Omicron virus replication to the upper respiratory tract resulting in less fusion and infectivity as compared to the Delta and D614G + P681R double mutants[59]. A previous study from Brazil reported that D614G mutation was detected in 90.5% of their samples, and was recently associated with higher viral loads and increased replication on human lung epithelial cells[60]. The earlier study reported that T19R and T19I mutations in the NTD spike region were significantly associated with mortality in patients by Delta and Omicron variants, respectively[61].

We found that L452R, T478K, E484A, N501Y, D614G, P681R, and D950N were the key mutations found in the spike protein including within the receptor-binding domain (RBD). The most common mutations (L452R, P681R, and D950N) were observed in the Delta variant (B.1.617.2 spike protein) however these mutations were absent in the Omicron variant, which is believed to be responsible for more adverse effects by the delta variant infections. In 2022, a study from china[62] reported that the mutated Omicron-L452R was significantly more effective at infecting humanized ACE2 mice's lung tissues. A previous study also reported that RBD mutations L452R, T478K, and E484Q may possibly result in increased ACE2 binding, whereas P681R at the furin cleavage site may improve transmissibility through an increase in S1-S2 cleavage[63]. The Omicron and Delta variants were found to have a higher transmissibility rate as compared to the original strain of SARS-CoV-2 with a capability to escape the host immune response[64] resulting in breakthrough infections. The ability of various anti-RBD-specific antibodies to bind only to the open spike protein is well recognized. Mutations that cause changes in spike glycoprotein conformation are more likely to make the RBD less susceptible to neutralizing antibodies[65–67]. A recent epidemiological and serology-based study in New Jersey revealed the presence of various mutations in the spike protein that are indicative of convergent evolution. It showed mainly L452R and T478K mapped to the Delta strain, whereas S371L, N440K, and Q493R were to the Omicron strains[68] in line with our observations except for the presence of S371F in place of S371L.

We also found one variant OP787487.1 (B.1.633– which was globally detected in February and March 2021) harbors some specific mutation (S: L5F, S: T76I, S: D253N, S: T572N, S: A575S, S: D796H, and S: T859N) which are not present in any other virus samples. In 2022, a study[69] also reported that in addition to Beta, Gamma, Delta, and Omicron VOCs, variant B.1.633 can induce vaccine breakthrough infections. The spike substitution mutant D796H showed decreased susceptibility to neutralizing antibodies, however, it also resulted in an infectivity defect[70]. These findings led us to believe that variation at this position might lead to a fitness cost for viral replication. These mutations which play a critical role in immune escape in SARS-CoV-2 infection, were found to be dominant in our study participants. In summary, our present study concluded that the newly emerged variants contributed to the second wave of COVID-19 in Uttar Pradesh. High-throughput sequencing makes it easier for researchers to identify and locate genetic variants of public health concerns that are useful for vaccine development allowing it to identify potential biomarkers and drug targets of COVID-19[71]. The present study is an attempt to derive a comprehensive study that highlights the pattern of circulating SARS-CoV-2 strains in Western Uttar Pradesh which comprised of the significant mutations G142D, N440K, E484A, N501Y, T478K, P681R, and D950N. These mutations played a critical role in immune escape in breakthrough infections. Additionally, the mutation D614G was coherent in most of the Pangolin which is specifically reported to be associated with viral transmissibility and high virulence. To our knowledge, this is the first study from Western Uttar Pradesh highlighting the molecular surveillance-based phylogenetic trends of whole genome sequences of SARS CoV-2. However, continuous and sustained monitoring of the identified global viral strains identified is required for an in-depth and detailed understanding of the evolution patterns of SARS-CoV-2 to explore the different evolutionary mechanisms adopted by the virus. The outcome will be immensely useful in designing a streamlined healthcare policy for Uttar Pradesh to contain any future spread of more evolved SARS-CoV-2 strains.

Methods

Clinical specimen collection

The National Institute of Biologicals (NIB), Noida, India is an autonomous institute under the Ministry of Health and family welfare, Government of India. It is the major testing center for SARS-CoV-2 samples as designated by the Indian Council of Medical Research (ICMR). SARS-CoV-2 suspected samples were received at NIB, Noida from various quarantine camps and hospitals located in the Western Uttar Pradesh region of India (Fig. 1) which were further processed for diagnostic testing according to the WHO guidelines[22]. During the period (March 2021 to January 2022), nasopharyngeal/oropharyngeal swabs (NPS/OPS) (n = 20,381) were collected for routine SARS-CoV-2 diagnosis. A total of 3,485 samples tested positive by real-time PCR (RT-PCR). Among 3,485 positive samples, 99 were randomly selected for sequencing. The primary inclusion criteria for the sample to be eligible for sequencing was determined by the SARS-CoV-2 positive samples that displayed a cycle threshold (Ct) of less than 30 so as to ensure maximum sequence coverage.

Nucleic acid extraction and Real-time (RT-PCR)

Extraction of viral RNA from the suspected clinical samples was performed using QIAamp Viral RNA Mini Kit using the manufacturer's instructions (Cat no. 52906; Qiagen, GmbH, Germany). The RT-PCR for diagnostic testing of SARS-CoV-2 nucleic acid was done using an NIV Multiplex Single Tube Real-Time PCR kit (Lot no. 11) on CFX96 Deep Well Real-time system (Bio-Rad) as per the manufacturer's specifications.

Next-generation sequencing

The whole genome sequencing of the selected samples was outsourced. Extracted RNA was used for the synthesis of first-strand cDNA. Commercially available PCR primers for the amplification of the complete SARS-CoV-2 genome were used for targeted enrichment. The library preparation was done and the final library distribution was evaluated on Tape Station followed by sequencing. Library preparation was done using the QIAseq DIRECT SARS-CoV2 Library Kit (Cat no.333891; Qiagen) with QIAseq DIRECT SARS-CoV-2 Enhancer (Cat no. 333884; Qiagen). Library quantification was conducted using Qubit High Sensitivity Assay. Cluster amplification on an Illumina flow cell was then achieved, followed by pooling and dilution to final optimal loading concentrations, and sequencing to produce 150 bp paired-end reads using an Illumina HiSeqX instrument (Illumina, San Diego, US).

Phylogeny construction and analysis

Genomic sequences from Eastern UP, Maharashtra, and Kerala were obtained from the Global Initiative on Sharing All Influenza Data (GISAID)[72] database, along with the reference sequences from this study (Western Uttar Pradesh) were used in the evolutionary analysis. In total, there were 270 sequences used to generate a cladogram (Supplementary File 1). The sequences were aligned using the MUSCLE program in MEGA software. We used the Model program in MEGA and Model Selection in IQ-TREE (http://iqtree.cibiv.univie.ac.at/) for finding the best-fit model with the lowest BIC (Bayesian Information Criterion) score. A maximum likelihood phylogenetic tree using the GTR + G + I model, was built with 1,000 bootstrap replications to assess the statistical robustness using MEGA11[73]. The tree was visualized using an online tool iToL (https://itol.embl.de/tree/115117108166307691660119911). Next clade and Pangolin COVID-19 Lineage Assigner were used for lineage/clade assignment.

Variant identification

Using bcl2fastq v2.20 software, raw HiSeqX data was demultiplexed from binary base call (BCL) format to FASTQ format. The paired-end FASTQ reads were then preprocessed by removing low-quality bases ($Q20 < 10$), adapter sequences, and reads with length < 30 bp using Cutadapt (version 1.18)[74]. High-quality reads were mapped to the reference genome of SARS-CoV-2 (GenBank accession number: NC_045512.2) using the Burrows-Wheeler Aligner MEM algorithm (BWA-MEM) (version 0.7.12) using default settings for paired-end mode[75]. The SAM tools package[76] was used to retain reads with high quality ($MQ > 25$), and the Mark Duplicates package was used to identify duplicate reads in the Genome Analysis Toolkit (GATK v4.1.0.0)[77]. Further, the genomic variants were predicted using uniquely-mapped reads by the GATK Haplotype Caller package. Genotypes were assigned to mutants (mutant allele frequency ≥ 0.7), degenerate nucleotides (mutant allele frequency < 0.7 and ≥ 0.3), and reference alleles (mutant allele frequency < 0.3). Further variations were also identified using the Nextclade web server (https://clades.nextstrain.org/). Ensembl Variant Effect Predictor (VEP) was used to annotate the impact of variants on genes and protein sequences[78].

Ethical statement

The National Institute of Biologicals (NIB), Noida is an apex autonomous institute under the administrative control of the Ministry of Health and Family Welfare (MoHFW), Government of India. NIB is not a hospital-based institution, however, it has been entrusted to perform COVID-19 testing since March 2020 during the pandemic. The present study has been performed using the leftover, anonymized COVID-19 samples, wherein all the methods and protocols were in concordance with the standard guidelines and regulations. All the experimental protocols carried out in the current study were approved by the Director of NIB, Noida.

Data availability

The methodology and original data of the present study are included in the article. Complete genome sequences generated from this study have been submitted to NCBI. The Bio Project accession ID is PRJNA976493 and the respective BioSample accession numbers are SAMN35370730-SAMN35370828. Additionally, the SARS-CoV-2 whole genome sequences are available in the GenBank repository in released form, and the subsequent accession IDs have been included in Supplementary File 3. Additional information has been provided in Supplementary files 4, 5.

Received: 12 May 2023; Accepted: 5 September 2023
Published online: 08 September 2023

References

1. Weekly update on COVID-19-21 August 2020. http://www.who.int https://www.who.int/publications/m/item/weekly-update-on-covid-19---16-october-2020.
2. Andrews, M. *et al.* First confirmed case of COVID-19 infection in India: A case report. *Indian J. Med. Res.* **0**, 0 (2020).
3. Peiris, J. *et al.* Coronavirus as a possible cause of severe acute respiratory syndrome. *Lancet* **361**, 1319–1325 (2003).
4. Raj, V. S., Osterhaus, A. D., Fouchier, R. A. & Haagmans, B. L. MERS: Emergence of a novel human coronavirus. *Curr. Opin. Virol.* **5**, 58–62 (2014).
5. Chan, J.F.-W. *et al.* 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 (2020).
6. Lillie, P. J. *et al.* Novel coronavirus disease (Covid-19): The first two patients in the UK with person to person transmission. *J. Infect.* **80**, 578–606 (2020).
7. Weiss, S. R. & Navas-Martin, S. Coronavirus pathogenesis and the emerging pathogen severe acute respiratory syndrome coronavirus. *Microbiol. Mol. Biol. Rev.* **69**, 635–664 (2005).
8. Richman, D. D., Whitley, R. J. & Hayden, F. G. *Clinical Virology* (ASM Press, 2017).

9. Roelle, S. M., Shukla, N., Pham, A. T., Bruchez, A. M. & Matreyek, K. A. Expanded ACE2 dependencies of diverse SARS-like coronavirus receptor binding domains. *PLoS Biol.* **20**, e3001738 (2022).
10. Tai, W. *et al.* Characterization of the receptor-binding domain (RBD) of 2019 novel coronavirus: Implication for development of RBD protein as a viral attachment inhibitor and vaccine. *Cell. Mol. Immunol.* **17**, 1–8 (2020).
11. Kin, N. *et al.* Genomic analysis of 15 human coronaviruses OC43 (HCoV-OC43s) circulating in France from 2001 to 2013 reveals a high intra-specific diversity with new recombinant genotypes. *Viruses* **7**, 2358–2377 (2015).
12. Lu, H., Stratton, C. W. & Tang, Y. Outbreak of pneumonia of unknown etiology in Wuhan, China: The mystery and the miracle. *J. Med. Virol.* **92**, 401–402 (2020).
13. Chen, N. *et al.* Epidemiological and clinical characteristics of 99 cases of 2019 novel coronavirus pneumonia in Wuhan, China: A descriptive study. *Lancet* **395**, 507–513 (2020).
14. Corman, V. M., Muth, D., Niemeyer, D. & Drosten, C. Hosts and sources of endemic human coronaviruses. *Adv. Virus Res.* **100**, 163–188 (2018).
15. Brook, C. E. & Dobson, A. P. Bats as 'special' reservoirs for emerging zoonotic pathogens. *Trends Microbiol.* **23**, 172–180 (2015).
16. Zhu, Y. *et al.* Ancestral SARS-CoV-2, but not Omicron, replicates less efficiently in primary pediatric nasal epithelial cells. *PLoS Biol.* **20**, e3001728 (2022).
17. Korber, B. *et al.* Tracking changes in SARS-CoV-2 Spike: Evidence that D614G increases infectivity of the COVID-19 virus. *Cell* **182**, 812-827.e19 (2020).
18. Oreshkova, N. *et al.* SARS-CoV-2 infection in farmed minks, the Netherlands, April and May 2020. *Eurosurveillance* https://doi.org/10.2807/1560-7917.ES.2020.25.23.2001005 (2020).
19. Rice, B. L., Lessler, J., McKee, C. & Metcalf, C. J. E. Why do some coronaviruses become pandemic threats when others do not?. *PLoS Biol.* **20**, e3001652 (2022).
20. World Health Organization. Tracking SARS-CoV-2 variants. http://www.who.int https://www.who.int/activities/tracking-SARS-CoV-2-variants (2022).
21. Holshue, M. L. *et al.* First case of 2019 novel coronavirus in the United States. *N. Engl. J. Med.* **382**, 929–936 (2020).
22. Fraser, C. *et al.* Pandemic potential of a strain of influenza A (H1N1): Early findings. *Science* **324**, 1557–1561 (2009).
23. Banu, S. *et al.* A distinct phylogenetic cluster of Indian severe acute respiratory syndrome coronavirus 2 isolates. *Open Forum Infect. Dis.* **7**, ofaa434 (2020).
24. Gardy, J. L. & Loman, N. J. Towards a genomics-informed, real-time, global pathogen surveillance system. *Nat. Rev. Genet.* **19**, 9–20 (2018).
25. Zaman, K. *et al.* Molecular epidemiology of a familial cluster of SARS-CoV-2 infection during lockdown period in Sant Kabir Nagar, Uttar Pradesh, India. *Epidemiol. Infect.* https://doi.org/10.1017/S0950268821001989 (2021).
26. Gupta, S., Misra, G. & Khurana, S. M. P. Bioinformatics: Promises and progress. *Int. J. Bioinform. Res. Appl.* **11**, 462 (2015).
27. Misra, G., Hora, S., Ginwal, S., Singh, N. & Anvikar, A. SARS-CoV-2 variants impact on key signaling pathways metamorphoses into severity. *Braz. Arch. Biol. Technol.* **66**, e23220261 (2023).
28. Janik, E., Niemcewicz, M., Podogrocki, M., Majsterek, I. & Bijak, M. The emerging concern and interest SARS-CoV-2 variants. *Pathogens* **10**, 633 (2021).
29. Fauver, J. R. *et al.* Coast-to-coast spread of SARS-CoV-2 during the early epidemic in the United States. *Cell* **181**, 990-996.e5 (2020).
30. Lu, J. *et al.* Genomic epidemiology of SARS-CoV-2 in Guangdong Province, China. *Cell* **181**, 997-1003.e9 (2020).
31. Deng, X. *et al.* Genomic surveillance reveals multiple introductions of SARS-CoV-2 into Northern California. *Science* https://doi.org/10.1126/science.abb9263 (2020).
32. Rambaut, A. *et al.* A dynamic nomenclature proposal for SARS-CoV-2 lineages to assist genomic epidemiology. *Nat. Microbiol.* **5**, 1403–1407 (2020).
33. Eden, J.-S. *et al.* An emergent clade of SARS-CoV-2 linked to returned travellers from Iran. *Virus Evol.* https://doi.org/10.1093/ve/veaa027 (2020).
34. OudeMunnink, B. B. *et al.* Rapid SARS-CoV-2 whole-genome sequencing and analysis for informed public health decision-making in the Netherlands. *Nat. Med.* **26**, 1405–1410 (2020).
35. Gudbjartsson, D. F. *et al.* Spread of SARS-CoV-2 in the Icelandic population. *N. Engl. J. Med.* **382**, 2302–2315 (2020).
36. de Jesus, J. G. *et al.* Importation and early local transmission of COVID-19 in Brazil, 2020. *Rev. Inst. Med. Trop. Sao Paulo* **62**, e30 (2020).
37. Gámbaro, F. *et al.* Introductions and early spread of SARS-CoV-2 in France, 24 January to 23 March 2020. *Eurosurveillance* **25**, 2001200 (2020).
38. Yadav, P. D. *et al.* An epidemiological analysis of SARS-CoV-2 genomic sequences from different regions of India. *Viruses* **13**, 925 (2021).
39. Starr, T. N., Greaney, A. J., Dingens, A. S. & Bloom, J. D. Complete map of SARS-CoV-2 RBD mutations that escape the monoclonal antibody LY-CoV555 and its cocktail with LY-CoV016. *bioRxiv* https://doi.org/10.1101/2021.02.17.431683 (2021).
40. Gupta, N. *et al.* Clinical characterization and genomic analysis of samples from COVID-19 breakthrough infections during the second wave among the various States of India. *Viruses* **13**, 1782 (2021).
41. Kar, S. K., Ransing, R., Arafat, S. M. Y. & Menon, V. Second wave of COVID-19 pandemic in India: Barriers to effective governmental response. *EClinicalMedicine* **36**, 100915 (2021).
42. Lopez Bernal, J. *et al.* Effectiveness of Covid-19 vaccines against the B.1.617.2 (Delta) variant. *N. Engl. J. Med.* **385**, 1532–1546 (2021).
43. Li, L. *et al.* Transmission and containment of the SARS-CoV-2 Delta variant of concern in Guangzhou, China: A population-based study. *PLoS Negl. Trop. Dis.* **16**, e0010048–e0010048 (2022).
44. McLaughlin, A. *et al.* Genomic epidemiology of the first two waves of SARS-CoV-2 in Canada. *Elife* **11**, e73896 (2022).
45. Abraham, P., Cherian, S. & Potdar, V. Genetic characterization of SARS-CoV-2 & implications for epidemiology, diagnostics & vaccines in India. *Indian J. Med. Res.* **152**, 12 (2020).
46. Radhakrishnan, C. *et al.* Initial insights into the genetic epidemiology of SARS-CoV-2 isolates from Kerala suggest local spread from limited introductions. *Front. Genet.* **12**, 630542 (2021).
47. Mondal, M., Lawarde, A. & Somasundaram, K. Genomics of Indian SARS-CoV-2: Implications in genetic diversity, possible origin and spread of virus. *Medrxiv* (2020).
48. Jimenez-Silva, C. *et al.* Genomic epidemiology of SARS-CoV-2 variants during the first two years of the pandemic in Colombia. *Commun. Med.* **3**, 1–12 (2023).
49. Muttineni, R. *et al.* Clinical and whole genome characterization of SARS-CoV-2 in India. *PLoS ONE* **16**, e0246173 (2021).
50. Parsad, P. *et al.* Unique mutational changes in SARS-CoV-2 genome: A case study for the largest state of India. Sep 2020.bioRxiv.
51. Thangaraj, J. W. V. *et al.* Predominance of Delta variant among the COVID-19 vaccinated and unvaccinated individuals, India, May 2021. *J. Infect.* **84**, 94–118 (2021).
52. Kannan, S. R. *et al.* Evolutionary analysis of the Delta and Delta Plus variants of the SARS-CoV-2 viruses. *J. Autoimmun.* **124**, 102715 (2021).
53. Chu, D. K. W. *et al.* Introduction of ORF3a-Q57H SARS-CoV-2 variant causing fourth epidemic wave of COVID-19, Hong Kong, China. *Emerg. Infect. Dis.* **27**, 1492–1495 (2021).

54. Sarkar, R. *et al.* Comprehensive analysis of genomic diversity of SARS-CoV-2 in different geographic regions of India: An endeavour to classify Indian SARS-CoV-2 strains on the basis of co-existing mutations. *Arch. Virol.* **166**(3), 801-812 (2021).
55. Groves, D. C., Rowland-Jones, S. L. & Angyal, A. The D614G mutations in the SARS-CoV-2 spike protein: Implications for viral infectivity, disease severity and vaccine design. *Biochem. Biophys. Res. Commun.* **538**, 104–107 (2020).
56. Du, Y. *et al.* Clinical features of 85 fatal cases of COVID-19 from Wuhan: A retrospective observational study. *SSRN Electron. J.* https://doi.org/10.2139/ssrn.3546088 (2020).
57. Huang, C. *et al.* Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. *Lancet* **395**, 497–506 (2020).
58. Cho, S. J. & Stout-Delgado, H. W. Aging and lung disease. *Annu. Rev. Physiol.* **82**, 433–459 (2020).
59. Khatri, R. *et al.* Intrinsic D614G and P681R/H mutations in SARS-CoV-2 VoCs Alpha, Delta, Omicron and viruses with D614G plus key signature mutations in spike protein alters fusogenicity and infectivity. *Med. Microbiol. Immunol.* **212**, 103–122 (2022).
60. Franceschi, V. B. *et al.* Genomic epidemiology of SARS-CoV-2 in Esteio, Rio Grande do Sul, Brazil. *BMC Genomics* **22**, 371 (2021).
61. Saifi, S. *et al.* SARS-CoV-2 VOCs, mutational diversity and clinical outcome: Are they modulating drug efficacy by altered binding strength?. *Genomics* **114**, 110466 (2022).
62. Zhang, Y. *et al.* SARS-CoV-2 spike L452R mutation increases Omicron variant fusogenicity and infectivity as well as host glycolysis. *Signal Transduct. Target. Ther.* **7**, 1–3 (2022).
63. Cherian, S. *et al.* SARS-CoV-2 Spike mutations, L452R, T478K, E484Q and P681R, in the second wave of COVID-19 in Maharashtra, India. *Microorganisms* **9**, 1542 (2021).
64. Yao, X. H. *et al.* A pathological report of three COVID-19 cases by minimal invasive autopsies. *Zhonghua Bing Li Xue Za Zhi Chin. J. Pathol.* **49**, 411–417 (2020).
65. Barnes, B. J. *et al.* Targeting potential drivers of COVID-19: Neutrophil extracellular traps. *J. Exp. Med.* **217**, e20200652 (2020).
66. Brinkmann, V. *et al.* Neutrophil extracellular traps kill bacteria. *Science* **303**, 1532–1535 (2004).
67. Schönrich, G. & Raftery, M. J. Neutrophil extracellular traps go viral. *Front. Immunol.* **7**, 366 (2016).
68. Mathema, B. *et al.* Genomic epidemiology and serology associated with a SARS-CoV-2 R.1 variant outbreak in New Jersey. *MBio* **13**, e0214122 (2022).
69. Hu, Y.-F. *et al.* Computation of antigenicity predicts SARS-CoV-2 vaccine breakthrough variants. *Front. Immunol.* **13**, 861050 (2022).
70. Kemp, S. A. *et al.* SARS-CoV-2 evolution during treatment of chronic infection. *Nature* **592**, 277–282 (2021).
71. Pradhan, D., Kumar, A., Singh, H. & Agrawal, U. Chapter 4—High-throughput sequencing. *ScienceDirect* 39–52 (2019).
72. Shu, Y. & McCauley, J. GISAID: Global initiative on sharing all influenza data—from vision to reality. *Eurosurveillance* **22**, 30494 (2017).
73. Tamura, K., Stecher, G. & Kumar, S. MEGA11: Molecular evolutionary genetics analysis version 11. *Mol. Biol. Evol.* **38**, 3022–3027 (2021).
74. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal* **17**, 10 (2011).
75. Wu, F. *et al.* A new coronavirus associated with human respiratory disease in China. *Nature* **579**, 265–269 (2020).
76. Li, H. *et al.* The sequence alignment/map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
77. McKenna, A. *et al.* The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20**, 1297–1303 (2010).
78. McLaren, W. *et al.* The ensembl variant effect predictor. *Genome Biol.* **17**, 122 (2016).

Acknowledgements

We thank the National Institute of Biologicals for the infrastructural support in the completion of this research. We acknowledge the contribution of Sakshi Chaudhary towards the construction of a phylogenetic tree.

Author contributions

G.M.: conceptualization, formal analysis, administration, funding, writing review, editing, and supervision. A.M.: execution, methodology, A.U.: analysis and manuscript preparation, M.C.: data analysis, B.B.: resources and data curation, and A.A.: resources and administration. All authors have read, revised, and approved the final version of the manuscript.

Funding

GM would like to acknowledge funding support from the Indian Council of Medical Research (ICMR) (Project Id: 2021-6466).

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-023-42065-6.

Correspondence and requests for materials should be addressed to G.M.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.