JOURNAL OF INTEGRATED OMICS
A METHODOLOGICAL JOURNAL
HTTP://WWW.JIOMICS.COM

ORIGINAL ARTICLE | DOI: 10.5584/jiomics.v15i2.239

A study of SARS-CoV-2 genomic profiles, evolutionary changes, and transmission dynamics in Southeastern India during three pandemic waves

Vanathy Kandhasamy[1,*], Agieshkumar Balakrishna Pillai[2], Vignesh Mariappan[2], Malarvizhi Ramalingam[1], Pajanivel Raganadin[3], Ramya Ramadoss[2], Balasubramanian Moovarkumudalvan[2], Joshy M Easow[1], Madavan Vasudevan[4], S.R. Rao[5,*]

[1]Department of Microbiology, Mahatma Gandhi Medical College & Research Institute (MGMCRI), Sri Balaji Vidyapeeth, (Deemed to be University), Puducherry, 607402, India; [2]Mahatma Gandhi Medical Advanced Research Institute (MGMARI), Sri Balaji Vidyapeeth (Deemed to be University), Puducherry, 607402, India; [3]Department of Pulmonary Medicine, Mahatma Gandhi Medical College and Research Institute (MGMCRI), Sri Balaji Vidyapeeth (Deemed to be University), Puducherry, 607402, India; [4]Theomics International Private Limited, Bangalore – 560 038, Karnataka, India; [5]Former Vice-President (Research, Innovation & Development), Sri Balaji Vidyapeeth (Deemed to be University), Puducherry – 607 403, India

Available Online: June 2025

Abstract

SARS-CoV-2 bio-surveillance at all levels is crucial for understanding its genetic evolution and vaccine effectiveness. This study investigated the emergence and evolution of new SARS-CoV-2 variants in the city of Puducherry, India throughout the three peaks of infection. A total of 128 samples were subjected to Illumina deep RNA sequencing. The results indicate that the first wave was dominated by uncommon variants, the second by Delta, and the third by Omicron. Lineages B.1.560 and B.1.617.2 were most prevalent. Analysis of 3133 common and 11 new mutations revealed Spike_D614G as the most common mutation and a novel set of mutations was observed in NS16, a key immune evasion factor. These NS16 mutations raise concerns about increased virulence, reduced vaccine efficacy, and potential antiviral resistance, warranting further investigation. Our findings contribute to SARS-CoV-2 evolutionary and genetic epidemiology research and highlight the need for ongoing surveillance to anticipate future variant threats.

Keywords: SARS-CoV-2; Genome Surveillance; Non-Structural Proteins; SARS-CoV-2 Variants; SARS-CoV-2 Lineages.

Introduction

Following its emergence in Wuhan, China, in late 2019, the Coronavirus disease-19 (COVID-19), caused by severe acute respiratory syndrome-coronavirus-2 (SARS-CoV-2) was declared a pandemic by the (World Health Organisation) WHO on March 11, 2020 [1]. India's first case was reported in January 2020, and Puducherry's on March 17, 2020. Three waves of infection, caused by different variants and ranging in severity, have swept across India [2], resulting in 44.5 million cases and 533,662 deaths as of February 3, 2025 (Ministry of Health and Family Welfare). In the Union territory (UT) of Puducherry, total stands at 166,000 cases and 1,962 deaths.

The virus has evolved through mutation, resulting in variants that exhibit increased severity of infection, more rapid transmission, and the ability to evade immune responses. These characteristics are of particular concern [3]. Bio-surveillance, which employs continuous genome sequencing, is essential for the early detection of pathogens and their variants, enabling effective mitigation strategies and limiting further spread [4]. Genomic surveillance has identified numerous SARS-CoV-2 lineages circulating in India and internationally [5]. According to the Phylogenetic Assignment of Named Global Outbreak (PANGO) classification system, the ancestral lineages are designated as A and B. The B.1 variant, carrying the D614G mutation in the spike protein, was responsible for the initial COVID-19 outbreak in Italy. Subsequent mutations within the B.1 lineage, involving the replacement of one or two amino acids, resulted in the emergence of several variants of concern (VOC). These include Alpha (B.1.1.7, with the N501Y mutation), Beta (B.1.351, with N501Y, E484K, and K417T), Gamma (P.1, derived from lineage B.1.1.28, also with N501Y, E484K, and K417T), and Delta (B.1.617.2). The Alpha variant originated in the United Kingdom, Beta in South Africa, Gamma in Brazil, and Delta in India [6-8].

***Corresponding author:** S.R. Rao, raghudbt@gmail.com | Vanathy Kandhasamy, vanathyk1@gmail.com

These variants had significantly impacted public health, influencing disease severity, hospitalization rates, mortality, and the effectiveness of vaccines and antibody treatments [3]. Genomic surveillance is crucial for quickly identifying new VOCs and implementing appropriate public health measures. Several variants of interest (VOIs), including Epsilon (B.1.427), Theta (P.3), Iota (B.1.526), and Kappa (B.1.617.1), have also been circulating [9]. The currently dominant strain worldwide is the Omicron variant (B.1.1.529), which was first reported in South Africa in November 2021. Its high transmissibility, ability to evade immune protection, and significant transmission among vaccinated populations have fueled its rapid global spread [10].

The World Health Organization has designated the Omicron sub-lineage BA.2 as the dominant strain. While BA.1 and BA.2 differ by only a few amino acids, studies suggest that BA.2 is more contagious and associated with higher reinfection rates than BA.1 [11,12]. Therefore, this lineage is being closely monitored to better understand its severity and how it causes disease. Continuous bio-surveillance and sequence analysis from cases across affected regions are essential for tracking the virus's genetic evolution and mutation rate.

In India, initial sequencing of SARS-CoV-2 suggested two separate introductions of the virus. The Delta variant, first identified in Maharashtra, subsequently became dominant, outcompeting earlier strains like Alpha and Beta [13]. With nationwide vaccine deployment, the virus's rapid mutation rate, and the potential impact of even minor genomic changes on vaccine effectiveness, continuous monitoring of circulating viral lineages and variants is crucial for effective surveillance. Therefore, this study investigated the variants and the emergence of new lineages during three infection peaks in Puducherry, a coastal region in southeastern India. Using RNA samples from swabs collected at a National Accreditation Board for Testing and Calibration Laboratories (NABL)-accredited COVID-19 testing facility within a tertiary care hospital [14], the study analyzed known and novel mutations, and the resulting data are discussed herein.

Materials and Methods

Study Approval

The study was approved by the Institute Research Committee and Institute Human Ethics Committee - MGMCRI/IRC/52/2021/04/IHEC/42. All the methods were performed in accordance to the Helsinki guidelines.

Nucleic acid extraction, COVID-19 screening and sample selection

Oropharyngeal/Nasopharyngeal samples received in viral transport medium at the diagnostic facility for molecular detection of RNA virus were used for the study. The host institute has started its Real-Time Reverse Transcriptase PCR (RT-PCR) testing service in August 2020 after obtaining NABL accreditation as per Indian Council of Medical Research (ICMR) recommendation for COVID testing. A total of 10,561 samples were processed from August 2020 to February 2022. Among them 1699 were positive. The whole genome sequencing (WGS) was done in representative positive samples (n=128). RNA was extracted from the nasopharyngeal/oropharyngeal samples using QIAamp RNA extraction kit and stored at -80°C. RT-PCR was done using various ICMR approved COVID-19 RT-PCR kits. For WGS, samples that are positive for SARS-CoV-2 by RT-PCR along with the CT (Cycle Threshold) value ≥18 & ≤25; State of Residence – Puducherry were used as the criteria for sample selection. A total of 128 samples fulfilling the above criteria were selected which included 17 from first wave (September 2020 to November 2020), 80 from the second wave (March 2021 to August 2021) and 31 from third wave (September 2021 to February 2022).

Retrieval of demographic and clinical details of patients

The demographic details were collected from ICMR specimen reference form. Since most of the forms were partially filled due to case load; the necessary details such as vaccination status, breakthrough infection (individuals turning positive for COVID-19 infection after 14 days of vaccination), co-morbidities, home quarantined or hospitalised during the course of infection were collected with an informed consent.

Library Preparation

RNA samples where in CT values ≥18 & ≤25 was considered for the WGS. The RNA was converted into cDNA using a balanced panel sets of random hexamers, Oligo DT's and COVID-19 specific oligos. SARS-CoV-2 genome is selectively amplified with oligo sets custom designed based on the ARTIC protocols. These samples are purified and barcode is added through a barcoding PCR cycle. The final samples are purified, quality checked, quantified and equimolar pooled for sequencing using Illumina Miseq deep sequencing platform with 300x2 bp Paired End chemistry. 1000x of the genome size raw data was generated to ensure maximum coverage and depth for each sample. Illumina MiSEQ platform with 300x2 was preferred for COVID-19 genome sequencing for reasons of throughput, base quality and the supported bioinformatics pipelines.

Deep sequencing data quality control and consensus genome assembly

Raw data quality control followed by adapter trimming was done using FASTQC tool kit, Trim Galore and Cut adapt. The processed reads are used to align to the reference strain NC_045512.2 using (Burrows-Wheeler Aligner) BWA and SamTools/Pileup to calculate coverage and depth. The consensus sequences are derived from the alignment files. This consensus genome sequence was used as input for identification and classification of Pangolin and Nextclade lineages.

Identification and Characterization of SARS-CoV-2 isolates

Consensus draft genomes of COVID-19 isolates were subjected to various analysis using CoVsurver (https://gisaid.org/database-

features/covsurver-mutations-app/), NextClade (https://clades.nextstrain.org/), Vigor (https://www.viprbrc.org/brc/vigorAnnotator.spg) and Genome Detective (https://www.genomedetective.com/app/typingtool/cov/). The results from the analysis tools were analysed using Microsoft Excel and GraphPad Prism tool for visualization purposes. The resulting consensus genomes (OP599771-OP599898) have been deposited to GenBank. The phylogenetic relationship between sample consensus genome sequences was analyzed using IQ-Tree tool (version 2.2.2.3) [15] by constructing a phylogenetic tree using 1000 bootstrap iterations and was visualized and annotated using iTOL online tool [16].

Results

Demographic Profile of the study participants

A total of 128 samples were sequenced, comprising 17 from the first wave, 80 from the second, and 31 from the third. The samples represented individuals aged 20 to 87, with an average age of 46.3 years. The age distribution was as follows: 20-30 years (22 cases), 31-40 years (35 cases), 41-50 years (23 cases), and 51 years and older (48 cases). Of the total, 79 (61.7%) were male and 49 (38.2%) were female. Eight patients experienced breakthrough infections after vaccination, of whom three were hospitalized, including two with diabetes mellitus. Seventeen patients presented with comorbidities such as diabetes mellitus, hypertension, hyperlipidemia, and/or heart disease. Of those quarantined at home, all but three patients (who had diabetes mellitus and/or hypertension) were without comorbidities. Fatalities occurred in the first wave (4 cases) and second wave (5 cases), but none were observed in the third wave. The deceased ranged in age from 45 to 82 years. Approximately 81% of the cases were individuals seeking testing, while the remaining 19% comprised of high-risk individuals, those with severe acute respiratory infection (SARI), patients exhibiting influenza-like illness, and asymptomatic contacts. Clinical and demographic details of the study subjects are summarized in **Table 1**.

Sequence analysis identified SARS-CoV-2 strains were classified as VOCs, VOIs, and uncommon variants. Of the 128 samples, VOCs comprised the majority, followed by a substantial proportion of uncommon strains, according to WHO classification. **Figure S1** presents the percentage distribution of variants within each category, and **Table S1** details the total number of cases for each variant. Genome coverage analysis demonstrated uniform coverage across all sequenced strains **(Figure S2)**.

Distribution of SARS-CoV-2 lineages

Table 2 provides a comprehensive overview of all sequenced lineages and variants, with **Figure 1** illustrating the distribution of individual lineages. The most prevalent lineage was B.1.617.2 (n=64, 50%), followed by BA.2 (n=15, 12%) and B.1.560 (n=14, 11%). Other lineages also circulated but decreased in prevalence over time. In the first wave, B.1.560 (n=14) was the dominant lineage in our sequenced samples. The second wave saw the emergence of six different lineages, with B.1.617.2 (n=61) being the most common, followed by B.1.1.7 (n=5). During the third wave, BA.2 (n=15) became predominant, outnumbering other lineages, and three samples also showed the presence of B.1.617.2 from the second wave. Among symptomatic individuals with influenza-like illness (ILI) and severe acute respiratory infection (SARI) (n=12/19, 63.1%), B.1.560 was the most common lineage. The most frequent lineages among deceased individuals were B.1.560 (n=3), B.1.617.2 (n=4), and B.1.1.7 (n=2).

Table 1 | Clinical and demographic details in all three waves. (*Diabets mellitus, Hypertension, Heart disease, Hyperlipidaemia, Cancer, Lung disease)

Parameters	Groups	1st wave	2nd wave	3rd wave	p-value
Age	20-30	2	8	12	≤0.01
	31-40	0	26	9	
	41-50	3	19	1	
	≥51	12	27	9	
Sex	Male	10	52	16	0.344
	Female	7	28	15	
Clinical Symptoms	Symptomatic	13	6	0	≤0.0001
	Asymptomatic	4	74	31	
Influenza like illness	Yes	9	6	0	≤0.0001
	No	8	74	31	
Severe acute Respiratory illness	Yes	4	1	0	≤0.0001
	No	13	78	31	
***Co-morbidities**	Yes	4	12	2	0.227
	No	13	68	29	
Outcome	Survived	13	75	31	0.015
	Deceased	4	5	0	

Table 2 | Genome Classification.

Lineage	Number of positives	Description	Wave	Variant
B.1.560	13	Tokyo, Japan on 20200109	First	UNCOMMON
B.1.560	1	Victoria, Australia on 20200230	First	UNCOMMON
B.1.1	2	Tokyo, Japan 20200109	First	UNCOMMON
B.1.1.354	1	Tokyo, Japan 20200109	First	UNCOMMON
AY.127	1	North Rhine-Westphalia, Germany on 20200511	Second	UNCOMMON
AY.127	1	Jammu and Kashmir, India on 20210321	Second	UNCOMMON
B.1.1.354	2	Tokyo, Japan 20200109	Second	UNCOMMON
B.1.1.7	2	California, USA on 20200708	Second	VOC_Alpha
B.1.1.7	5	North Dakota, USA on 20201023	Second	VOC_Alpha
B.1.351	4	Gauteng, South Africa on 20200801	Second	VOC_Beta
B.1.617.2	2	California, USA on 20210106	Second	VOC_Delta
B.1.617.2	1	Dakar, Senegal on 20200512	Second	VOC_Delta
B.1.617.2	1	England, United Kingdom on 20210406	Second	VOC_Delta
B.1.617.2	2	Gujarat, India on 20210407	Second	VOC_Delta
B.1.617.2	1	Indiana, USA on 20210418	Second	VOC_Delta
B.1.617.2	23	Jammu and Kashmir, India on 20210321	Second	VOC_Delta
B.1.617.2	12	North Rhine-Westphalia, Germany on 20200511	Second	VOC_Delta
B.1.617.2	2	Oromia, Ethiopia on 20210331	Second	VOC_Delta
B.1.617.2	4	Podravska, Slovenia on 20201204	Second	VOC_Delta
B.1.617.2	1	Sicily, Italy on 20210316	Second	VOC_Delta
B.1.617.2	2	Texas, USA on 20200716	Second	VOC_Delta
B.1.617.2	2	Not Associated	Second	VOC_Delta
B.1.617.2	1	Diourbel, Senegal on 20200327	Second	VOC_Delta
B.1.617.2	1	Gujarat, India on 20210401	Second	VOC_Delta
B.1.617.2	2	Minnesota, USA on 20200527	Second	VOC_Delta
B.1.617.2	1	Minnesota, USA on 20201009	Second	VOC_Delta
B.1.617.2	1	Mizoram, India on 20210213	Second	VOC_Delta
B.1.617.2	2	Red Sea Governorate, Egypt on 20201213	Second	VOC_Delta
B.1.617	2	Delhi, India on 20200303	Second	VOI_Kappa
B.1.617	1	Jharkhand, India on 20210106	Second	VOI_Kappa
B.1.617.1	1	Jharkhand, India on 20210106	Second	VOI_Kappa
AY.59	1	Not Associated	Third	VOC_Delta
B.1.617.2	2	Mecklenburg-Western Pomerania, Germany on 20210503	Third	VOC_Delta
B.1.617.2	1	Not Associated	Third	VOC_Delta
AY.127	1	North Rhine-Westphalia, Germany on 20200511	Third	VOC_Delta
AY.127	1	Not Associated	Third	VOC_Delta
BA.1.1	2	Not Associated	Third	Omicron
BA.1.1.7	4	Not Associated	Third	Omicron
BA.1.1.7	1	Lambayeque, Peru on 20220117	Third	Omicron
BA.1.1.7	1	Tennessee, USA on 20220120	Third	Omicron
BA.2	1	Karnataka, India on 20220117	Third	Omicron
BA.2	10	Karnataka, India on 20220119	Third	Omicron
BA.2	1	Santa Fe, Argentina on 20220405	Third	Omicron
BA.2	3	Not Associated	Third	Omicron
BA.2.17	1	Karnataka, India on 20220117	Third	Omicron

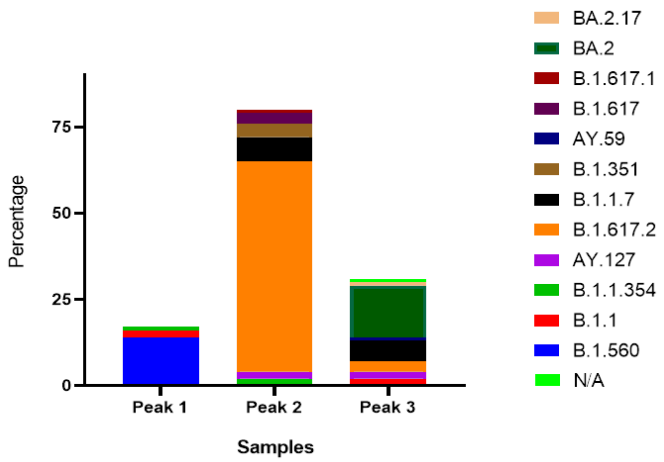

Figure 1 | Individual lineage distribution during three major peaks.

Breakthrough Infections

Vaccination status was determined through telephone interviews. Of the 128 samples, 105 individuals provided information regarding their vaccination status. Only 6% of these respondents were vaccinated at the time of sampling. Vaccinated individuals predominantly harboured uncommon variants (n=3), followed by Alpha, Beta, Kappa, and Omicron (n=1 each), as illustrated in **Figure S3**. **Figure S4** provides an overview of the sequenced variants and their corresponding outcomes.

Detection of lineages and their evolutionary relationship

Fourteen lineages were identified using PANGO lineage designation. Nextclade analysis grouped these lineages into 10 distinct clades. CoVsurver and WHO analyses identified 7 clades. The classifications based on Pangolin, Nextclade, CoVsurver, and WHO analyses are presented in **Figure 2 (A-D)**. In these figures, the x-axis represents the pandemic wave by year, and the y-axis indicates the various clades, lineages, strains, or types. The size of the shapes (circle, square, triangle, and diamond) corresponds to the total number of infected patients.

Phylogenetic relationships between the samples were determined using FastTree-2 maximum-likelihood phylogeny. FastTree-2 employs maximum-likelihood nearest-neighbour interchanges (MLIs), minimum-evolution subtree-pruning-regrafting (SPRs), and nearest-neighbour interchanges (NNIs). The samples were aligned against the Wuhan reference genome (NC_045512.2). **Figures S5 A-C** depict the branches corresponding to Pangolin, Nextstrain, and WHO lineages.

Mutational divisions of SARS-CoV-2 detected during three peaks

We analysed the frequency of both known and novel mutations, identifying a total of 3144 mutations (3133 common and 11 new), with a median of 24 mutations per sample. The highest concentration of mutations was observed in the S gene (32% of total mutations), followed by the N gene (12%) and the NSP3 gene (7%). The distribution of mutations across the sequenced strains is shown in **Figure 3**. The most frequently observed mutation was Spike_D614G, present across all three peaks. The average number of mutations per sample increased over time, from 8 in the first peak to 23 in the second and 40 in the third. Phylogenetic relationships between sample consensus genome sequences are illustrated in the phylogenetic tree (**Figure 4**), which also summarizes sample attributes, predicted Pangolin lineage,

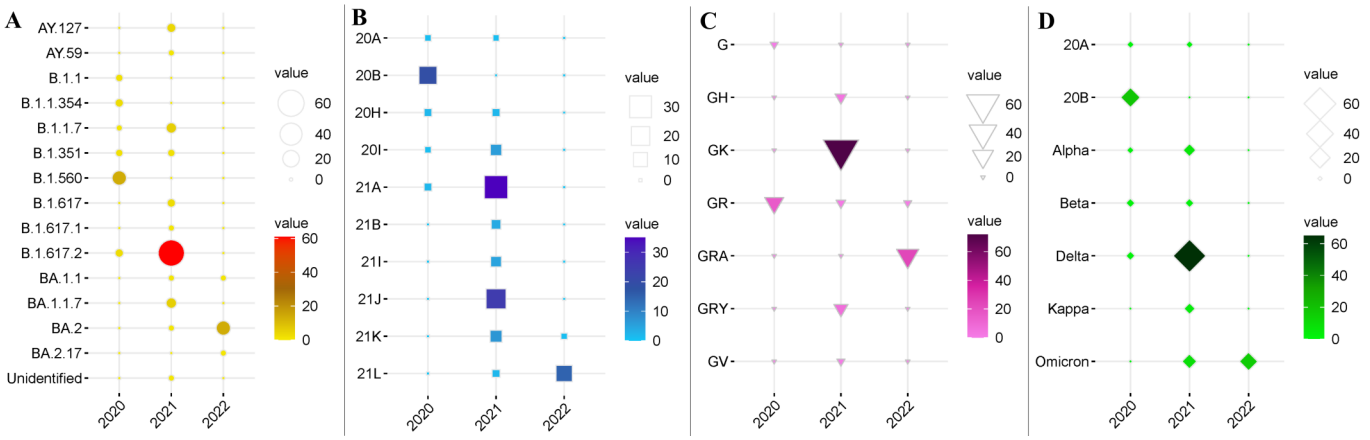

Figure 2 | Classification of lineages and number of strains sequenced (A) Classification of lineages based on Pangolin; **(B)** Classification of lineages based on Nextclade; **(C)** Classification of lineages based on CoVsuver; **(D)** Classification of lineages based on WHO.

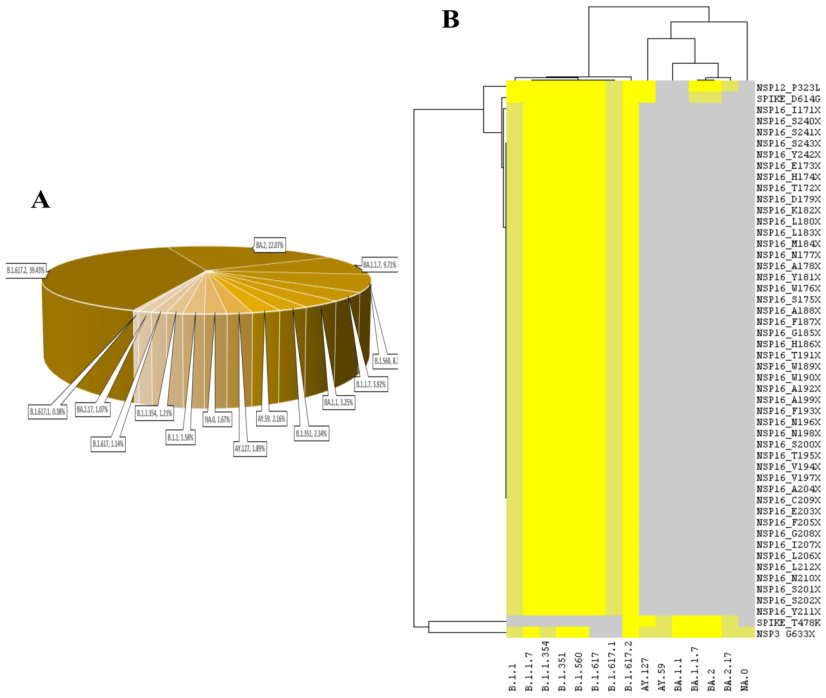

Figure 3 | Distribution of Mutations (A) Mutation percentage in the sequenced lineages **(B)** Dendrogram revealing the association between different mutation and sequenced strains.

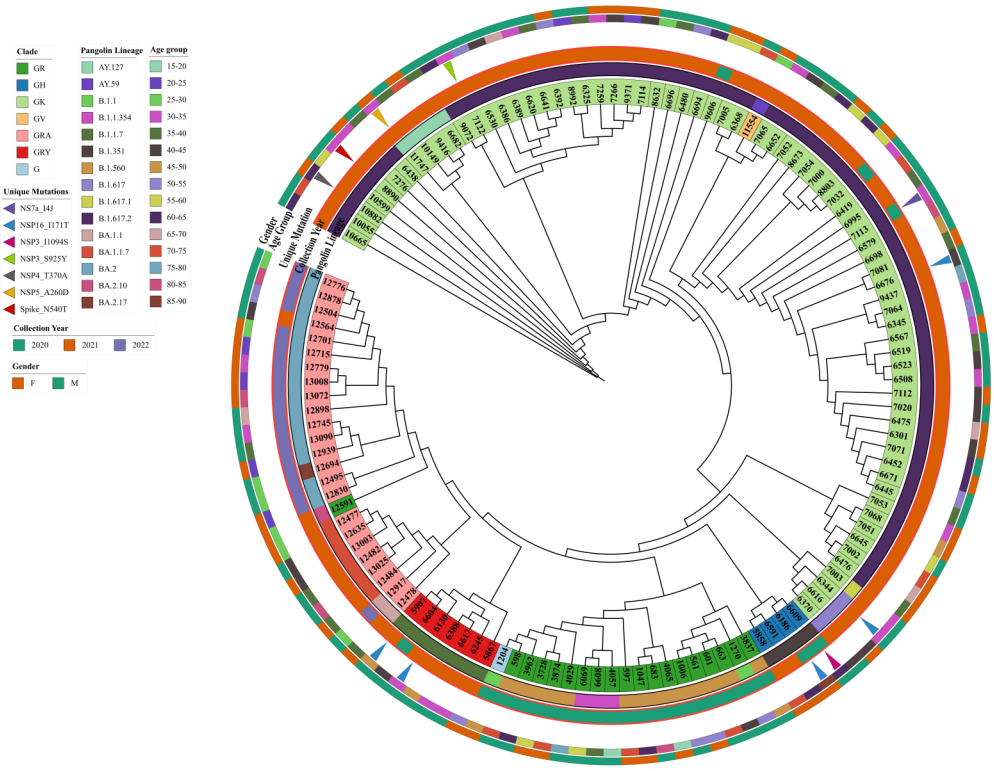

Figure 4 | Phylogenetic tree of the consensus genome sequences along with sample attributes. Tree labels are highlighted based on clade classification by CoVsuver as described in the legend. Lineages based on Pangolin are represented by the color strip 1. The corresponding sample collection year is represented by color strip 2. Samples with unique mutations are indicated by colored triangles. The patient age group and gender details are represented by color strips 3 & 4 respectively.

CoVsurver clade classification, and unique mutations detected: NS7a_I4J, NSP16_I171T, NSP3_I1094S, NSP3_S925Y, NSP4_T370A, NSP5_A260D, and Spike_N540T. Most unique mutations were found in samples from the second wave. The unique mutation NSP16_I171T was detected in five samples, representing the largest group of samples sharing a unique mutation. With one exception, these samples were primarily from male patients aged 30-50. Notably, the detected unique mutation is significant due to the high coverage (>90%) and quality control scores of the consensus genome sequences harbouring this mutation. The overall distribution of mutants, aligned with the three major peaks reported by the Government of Puducherry (https://covid19dashboard.py.gov.in/), is shown in **Figure S6**.

Discussion

The WHO declared COVID-19 a pandemic on March 11, 2020. India reported its first case on January 30, 2020, in a patient with travel history from Wuhan, China [1]. Subsequently, cases were reported across various Indian states. Puducherry's first case was recorded in March 2020, with a surge in cases observed in August 2020 [17]. Whole-genome sequencing (WGS) is a valuable tool for understanding geographical distribution, viral adaptation over time, transmission patterns, disease mechanisms, and for vaccine and drug design, dosage strategies, and hospitalization needs. Therefore, we conducted WGS on 128 representative samples collected from all three waves in Puducherry.

As a popular tourist destination with a history of French settlements, Puducherry is susceptible to virus transmission both from international travelers and through local transmission, such as mass gatherings. This study compared the SARS-CoV-2 variants circulating in Puducherry during all three waves, as determined by WGS, with those circulating globally and within India during the same periods. We identified 14 sub-lineages (B.1.560, B.1.1, B.1.1.354, B.1.351, B.1.1.7, B.1.617.2, B.1.617, AY.127, B.1.617.1, AY.59, BA.2, BA.2.17, BA.1.1, BA.1.1.7) belonging to lineages A and B. These lineages are ancestral to those that subsequently circulated globally. Lineages A and B were first sequenced in China in January 2020 [18].

The lineages were similar to those circulating in Europe, Australia, the USA, the UK, South Africa, Singapore, Malaysia, and Italy. During the first wave, the circulating lineages (B.1.560, B.1.1, and B.1.1.354) were classified as uncommon variants by the WHO. The second wave saw a mix of both A and B lineages, with the majority being B.1.617.2 (the Delta variant). A key finding of this study is the presence of numerous variants classified as "uncommon" by the WHO, meaning they did not meet the criteria for either VOC or VOI. Given their circulation during the outbreaks, the potential roles of these uncommon strains in disease severity, pathogenesis, and mutation patterns warrant further investigation for effective disease surveillance and pandemic preparedness.

An ICMR multicentric study [19], which included 26 samples from Puducherry, identified Delta (B.1.617.2) as the most prevalent variant, followed by Alpha and Kappa. That study also detected two Delta sub-lineages, AY.1 and AY.2, both carrying the K417N spike protein mutation, known to contribute to immune evasion and increased infectivity. Across India, Delta was the dominant variant, while Alpha was more prevalent in the northern region. Breakthrough infections were common with the Delta variant, likely due to reduced vaccine neutralization capacity and the variant's high transmissibility during that period [19].

An Indian Gujarat-based study of 502 sequenced samples from deceased and recovered patients, compared nationally and globally, found a missense mutation, C28854T (Ser194Leu), in the Nucleocapsid (N) gene. This mutation, considered deleterious, had an allele frequency of 47% in Gujarat's deceased patients compared to 7% globally, suggesting a distinct mutation potentially contributing to disease severity in Gujarat [20]. Another study from Delhi, involving 612 samples sequenced across the first three pandemic peaks, reported 26 lineages with novel mutations across these peaks [21]. Similar to our single-center study, the Delhi research also included samples from January to March 2022, a period dominated by Omicron (25 samples). Like the Delhi study, we also categorized our samples based on median cycle threshold (CT) values. While the minimum CT remained relatively consistent across all three peaks, the median CT in the third peak (Omicron-dominant) was less than 20. In contrast, the median CT values in the first two peaks were above 20. This differs from the findings of Gautam et al. 2022, who reported a median CT of 15 in the first two peaks [21].

The number of mutations per sample increased from the second to the third peak. This is likely related to the units of mutation per site per year and the number of substitutions per site per replication cycle. Studies have estimated the mutation rate to be 10-3 mutations per site per year, indicating continuous viral replication and subsequent mutation accumulation over time [22,23]. The lineages identified in our study also varied over time. B.1.560 was most common during the first peak, while B.1.617.2 dominated the second, mirroring findings from Delhi [21]. The third peak initially showed the presence of B.1.617.2, which gradually decreased as BA.2 and BA.1.1.7 became more prevalent. This temporal shift in dominant lineages has been observed globally [24].

The most common mutation observed was Spike_D614G. Mutations in the spike protein have been extensively documented [3,25,26]. The SARS-CoV-2 spike (S) protein is crucial for target recognition, binding to ACE2, and host cell entry. Frequent S protein mutations can affect binding affinity to host ACE2 and the receptor-binding domain (RBD) [27,28]. For example, the D614G mutation, observed pervasively in our sequenced COVID-19 cases, has been shown to alter RBD conformation and enhance the affinity between the S protein and ACE2 [25,29]. Previous studies have identified NSP2, 3, 4, 6, and 12 as recurrent mutation hotspots across various geographic regions, including Asia, Oceania, Europe, and North America, while NSP13 helicase mutants have been reported primarily in North America [30]. The new mutations detected in this study include NSP16_I171T, NS7a_I4J, NSP3_I1094S, NSP3_S925Y, NSP4_T370A, NSP5_A260D, and Spike_N540T. NSP16_I171T was found in five patients (Alpha-2, Beta-1, Kappa-1, Delta-1), while the remaining mutations were each found in a single patient. These unique mutations were confirmed by aligning fastq files generated by the Local Run Manager (Illumina) to the SARS-CoV-2 reference genome

(NC_045512.2) using Burrows-Wheeler Aligner (v. 0.7.17) [31]. NCoV-Tools was used to verify the consensus sequence for each genome. Single nucleotide variants and short insertions/deletions were detected using the GATK pipeline with an average coverage of 1000x and a minimum depth of 100x. SnpEff (v. 5.0c) [32] was used to annotate the filtered variants (VAF > 0.05). For mutation frequencies below 0.05, the existence of called reads was confirmed using the Integrative Genomics Viewer (IGV), CoVsurver, and the COVID-19 genome annotator.

The known mutations identified in this study were primarily located within NS1 to NS5, with NS3 exhibiting the highest number. NS3 is a crucial non-structural protein (NSP) involved in viral replication and host protein synthesis regulation. Among all NSPs (1-16), NSP3 has been reported to have the most mutations, although *in vitro* studies validating the effects of these mutations on host protein synthesis or viral replication efficiency are still lacking. Further investigation of the NS3 mutations observed in this study is warranted. Another important NSP utilized by the virus for replication and transcription is NSP16. NSP16 encodes methyl transferase activity, essential for viral immune evasion. The S-Adenosylmethionine (SAM)-dependent 2'-O-Methyltransferase enzyme, with its conserved catalytic tetrad ([K46-D130-K170-E203]) from SARS-CoV-2, comprises two subunits: NSP16 (catalytic) and NSP10 (stimulatory), with NSP16 being activated by heterodimerization with NSP10. Thus, NSP16's enzymatic activity depends on NSP10; otherwise, it remains in a nascent form.

In this study, unique or novel mutations in NSP16 were observed in five samples. NSP16, which, unlike NSP3, has not been reported to have many mutations, could be a potential therapeutic target for SARS-CoV-2 control. However, the identified mutations raise important questions: How will these mutations in NSP16 impact the NS10/NSP16 complex? How will the transferase activity of NSP16 be affected? And how will this modify the virus's immune evasion mechanisms? To explore these questions, we performed computational analysis, which revealed that the NSP16_I171T mutation is significant due to its proximity to the conserved catalytic tetrad. Multiple sequence alignment of NSP16 protein sequences from different species showed that this mutation occurs within a highly conserved pan-coronavirus motif (**Figure S7**). Using DynaMut2 [33], we analyzed the effect of the mutation on protein stability and dynamics by introducing the NSP16_I171T mutation into the nsp16-nsp10-SAM complex (PDB ID: 6W61). A decrease in folding free energy (ΔΔG) of -3.41 kcal/mol was observed, indicating increased flexibility in the mutated protein compared to the native structure. Furthermore, the non-polar Isoleucine is replaced by the polar Threonine, which has the potential to form hydrogen bonds. Given the mutation's proximity to the catalytic tetrad, the increased flexibility, and the hydrogen bond potential of Threonine, the NSP16_I171T mutation likely significantly influences NSP16's catalytic activity [34].

This study utilized deep sequencing to investigate the genetic diversity of the SARS-CoV-2 genome throughout the pandemic. We acknowledge several limitations. The uneven sample sizes across the different pandemic peaks limit our ability to analyze variant distribution by age and sex. We did not assess the significance of the novel mutations identified or their clinical correlations. Furthermore, because most patients were unvaccinated at the time of sampling, we could not evaluate the impact of the identified mutants on the immunology of currently available vaccines.

Concluding Remarks

The prominent mutations observed in this study suggest that the increased mutation rates during the second and third peaks may be attributed to impaired proofreading and altered polymerase activity in the virus. Further research is necessary to: (i) elucidate the molecular mechanisms driving increased mutation rates in key viral NSPs; (ii) determine if mutations in key NSPs lead to enhanced viral replication and immune evasion; (iii) investigate whether these mutations confer resistance to current antivirals; and (iv) develop broad-spectrum antivirals. Given SARS-CoV-2's propensity for frequent mutation, studying its evolutionary and genomic epidemiology is crucial. Continuous biosurveillance, including pathogen and novel mutant detection via genome sequencing, coupled with regional and global information sharing, is essential for mitigating future outbreaks and pandemics.

Acknowledgements

The authors greatly acknowledge the funding support provided by “Sri Balaji Educational and Charitable Public Trust-SBECPT”, Sri Balaji Vidyapeeth to carry out the study. We acknowledge Dr. Pravin Charles, Dr. Namrata K Bhosale, Dr. Ramyapriyadarshini for their constant support during the study, Mrs. Humera Begum and Mrs. Tamizhmani for their technical support.

References

[1] Cucinotta D, Vanelli M. WHO Declares COVID-19 a Pandemic. Acta Biomed 2020;91:157–60. https://doi.org/10.23750/abm.v91i1.9397.

[2] Mandal S, Arinaminpathy N, Bhargava B, Panda S. Plausibility of a third wave of COVID-19 in India: A mathematical modelling based analysis. Indian J Med Res 2021;153:522–32. https://doi.org/10.4103/ijmr.ijmr_1627_21.

[3] Harvey WT, Carabelli AM, Jackson B, Gupta RK, Thomson EC, Harrison EM, et al. SARS-CoV-2 variants, spike mutations and immune escape. Nat Rev Microbiol 2021;19:409–24. https://doi.org/10.1038/s41579-021-00573-0.

[4] Bartlow AW, Middlebrook EA, Romero AT, Fair JM. How Cooperative Engagement Programs Strengthen Sequencing Capabilities for Biosurveillance and Outbreak Response. Front Public Health 2021;9:648424. https://doi.org/10.3389/fpubh.2021.648424.

[5] Sarkar R, Mitra S, Chandra P, Saha P, Banerjee A, Dutta S, 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 2021;166:801–12. https://doi.org/10.1007/s00705-020-04911-0.

[6] Frampton D, Rampling T, Cross A, Bailey H, Heaney J, Byott M, et al. Genomic characteristics and clinical effect of the emergent SARS-CoV-2 B.1.1.7 lineage in London, UK: a whole-genome sequencing and hospital-based cohort study. Lancet Infect Dis 2021;21:1246–56. https://doi.org/10.1016/S1473-3099(21)00170-5.

[7] Tegally H, Wilkinson E, Giovanetti M, Iranzadeh A, Fonseca V, Giandhari J, et al. Detection of a SARS-CoV-2 variant of concern in South Africa. Nature 2021;592:438–43. https://doi.org/10.1038/s41586-021-03402-9.

[8] Faria NR, Mellan TA, Whittaker C, Claro IM, Candido D da S, Mishra S, et al. Genomics and epidemiology of the P.1 SARS-CoV-2 lineage in Manaus, Brazil. Science 2021;372:815–21. https://doi.org/10.1126/science.abh2644.

[9] Thakur V, Bhola S, Thakur P, Patel SKS, Kulshrestha S, Ratho RK, et al. Waves and variants of SARS-CoV-2: understanding the causes and effect of the COVID-19 catastrophe. Infection 2022;50:309–25. https://doi.org/10.1007/s15010-021-01734-2.

[10] Khandia R, Singhal S, Alqahtani T, Kamal MA, El-Shall NA, Nainu F, et al. Emergence of SARS-CoV-2 Omicron (B.1.1.529) variant, salient features, high global health concerns and strategies to counter it amid ongoing COVID-19 pandemic. Environ Res 2022;209:112816. https://doi.org/10.1016/j.envres.2022.112816.

[11] Lyngse FP, Kirkeby CT, Denwood M, Christiansen LE, Mølbak K, Møller CH, et al. Household transmission of SARS-CoV-2 Omicron variant of concern subvariants BA.1 and BA.2 in Denmark. Nat Commun 2022;13:5760. https://doi.org/10.1038/s41467-022-33498-0.

[12] Sahebi S, Keikha M. Clinical features of SARS-CoV-2 Omicron BA.2; Lessons from previous observations – Correspondence. Int J Surg 2022;104:106754. https://doi.org/10.1016/j.ijsu.2022.106754.

[13] Yadav PD, Potdar VA, Choudhary ML, Nyayanit DA, Agrawal M, Jadhav SM, et al. Full-genome sequences of the first two SARS-CoV-2 viruses from India. Indian J Med Res 2020;151:200–9. https://doi.org/10.4103/ijmr.IJMR_663_20.

[14] Coumare VN, Pawar SJ, Manoharan PS, Pajanivel R, Shanmugam L, Kumar H, et al. COVID-19 Pandemic-Frontline Experiences and Lessons Learned From a Tertiary Care Teaching Hospital at a Suburban Location of Southeastern India. Front Public Health 2021;9:673536. https://doi.org/10.3389/fpubh.2021.673536.

[15] Minh BQ, Schmidt HA, Chernomor O, Schrempf D, Woodhams MD, von Haeseler A, et al. IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. Mol Biol Evol 2020;37:1530–4. https://doi.org/10.1093/molbev/msaa015.

[16] Letunic I, Bork P. Interactive Tree Of Life (iTOL) v5: an online tool for phylogenetic tree display and annotation. Nucleic Acids Res 2021;49:W293–6. https://doi.org/10.1093/nar/gkab301.

[17] Laxminarayan R, Wahl B, Dudala SR, Gopal K, Mohan B C, Neelima S, et al. Epidemiology and transmission dynamics of COVID-19 in two Indian states. Science 2020;370:691–7. https://doi.org/10.1126/science.abd7672.

[18] Tan W, Zhao X, Ma X, Wang W, Niu P, Xu W, et al. A Novel Coronavirus Genome Identified in a Cluster of Pneumonia Cases — Wuhan, China 2019−2020. China CDC Weekly 2020;2:61–2. https://doi.org/10.46234/ccdcw2020.017.

[19] Gupta N, Kaur H, Yadav PD, Mukhopadhyay L, Sahay RR, Kumar A, 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 2021;13:1782. https://doi.org/10.3390/v13091782.

[20] Joshi M, Puvar A, Kumar D, Ansari A, Pandya M, Raval J, et al. Genomic Variations in SARS-CoV-2 Genomes From Gujarat: Underlying Role of Variants in Disease Epidemiology. Front Genet 2021;12:586569. https://doi.org/10.3389/fgene.2021.586569.

[21] Gautam P, Paul D, Suroliya V, Garg R, Agarwal R, Das S, et al. SARS-CoV-2 Lineage Tracking, and Evolving Trends Seen during Three Consecutive Peaks of Infection in Delhi, India: a Clinico-Genomic Study. Microbiol Spectr 2022;10:e02729-21. https://doi.org/10.1128/spectrum.02729-21.

[22] Urhan A, Abeel T. Emergence of novel SARS-CoV-2 variants in the Netherlands. Sci Rep 2021;11:6625. https://doi.org/10.1038/s41598-021-85363-7.

[23] Vilar S, Isom DG. One Year of SARS-CoV-2: How Much Has the Virus Changed? Biology (Basel) 2021;10:91. https://doi.org/10.3390/biology10020091.

[24] Volz E, Mishra S, Chand M, Barrett JC, Johnson R, Geidelberg L, et al. Assessing transmissibility of SARS-CoV-2 lineage B.1.1.7 in England. Nature 2021;593:266–9. https://doi.org/10.1038/s41586-021-03470-x.

[25] Magazine N, Zhang T, Wu Y, McGee MC, Veggiani G, Huang W. Mutations and Evolution of the SARS-CoV-2 Spike Protein. Viruses 2022;14:640. https://doi.org/10.3390/v14030640.

[26] Guruprasad L. Human SARS CoV-2 spike protein mutations. Proteins 2021;89:569–76. https://doi.org/10.1002/prot.26042.

[27] Mariappan V, S R R, Balakrishna Pillai A. Angiotensin-converting enzyme 2: A protective factor in regulating disease virulence of SARS-COV-2. IUBMB Life 2020;72:2533–45. https://doi.org/10.1002/iub.2391.

[28] Mariappan V, Ranganadin P, Shanmugam L, Rao SR, Balakrishna Pillai A. Early shedding of membrane-bounded ACE2 could be an indicator for disease severity in SARS-CoV-2. Biochimie 2022;201:139–47. https://doi.org/10.1016/j.biochi.2022.06.005.

[29] Gobeil SM-C, Janowska K, McDowell S, Mansouri K, Parks R, Manne K, et al. D614G Mutation Alters SARS-CoV-2 Spike Conformation and Enhances Protease Cleavage at the S1/S2 Junction. Cell Rep 2021;34:108630. https://doi.org/10.1016/j.celrep.2020.108630.

[30] Pachetti M, Marini B, Benedetti F, Giudici F, Mauro E, Storici P, et al. Emerging SARS-CoV-2 mutation hot spots include a novel RNA-dependent-RNA polymerase variant. J Transl Med 2020;18:179. https://doi.org/10.1186/s12967-020-02344-6.

[31] Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 2009;25:1754–60. https://doi.org/10.1093/bioinformatics/btp324.

[32] Cingolani P, Platts A, Wang LL, Coon M, Nguyen T, Wang L, et al. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff. Fly (Austin) 2012;6:80–92. https://doi.org/10.4161/fly.19695.

[33] Rodrigues CHM, Pires DEV, Ascher DB. DynaMut2: Assessing changes in stability and flexibility upon single and multiple point missense mutations. Protein Sci 2021;30:60–9. https://doi.org/10.1002/pro.3942.

[34] Rosas-Lemus M, Minasov G, Shuvalova L, Inniss NL, Kiryukhina O, Wiersum G, et al. The crystal structure of nsp10-nsp16 heterodimer from SARS-CoV-2 in complex with S-adenosylmethionine. bioRxiv 2020:2020.04.17.047498. https://doi.org/10.1101/2020.04.17.047498.