RESEARCH ARTICLE

REVISED Importation, circulation, and emergence of variants of SARS-CoV-2 in the South Indian state of Karnataka [version 2; peer review: 2 approved]

Chitra Pattabiraman[1], Pramada Prasad[1], Anson K. George[1], Darshan Sreenivas[1], Risha Rasheed[1], Nakka Vijay Kiran Reddy[1], Anita Desai[1], Ravi Vasanthapuram[2]

[1]Neurovirology, National Institute of Mental Health and Neurosciences, India, Bangalore, Karnataka, 560029, India
[2]Nodal Officer Genetic Confirmation of SARS-CoV-2, Government of Karnataka, Bengaluru, India

v2 **First published:** 13 May 2021, **6**:110
https://doi.org/10.12688/wellcomeopenres.16768.1
Latest published: 07 Feb 2022, **6**:110
https://doi.org/10.12688/wellcomeopenres.16768.2

Abstract

Background: As the coronavirus disease 2019 (COVID-19) pandemic continues, the selection of genomic variants of severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) associated with higher transmission, more severe disease, re-infection, and immune escape are a cause for concern. Such variants have been reported from the UK (B.1.1.7), South Africa (B.1.351) and, Brazil (P.1/B.1.1.28). We performed this study to track the importation, spread, and emergence of variants locally.

Methods: We sequenced whole genomes of SARS-CoV-2 from international travellers (n=75) entering Karnataka, South India, between Dec 22, 2020 and Jan 31, 2021, and from positive cases in the city of Bengaluru (n=108), between Nov 22, 2020- Jan 22, 2021, as well as a local outbreak. We present the lineage distribution and analysis of these sequences.

Results: Genomes from the study group into 34 lineages. Variant B.1.1.7 was introduced by international travel (24/73, 32.9%). Lineage B.1.36 and B.1 formed a major fraction of both imported (B.1.36: 20/73, 27.4%; B.1: 14/73, 19.2%), and circulating viruses (B.1.36: 45/103; 43.7%,. B.1: 26/103; 25.2%). The lineage B.1.36 was also associated with a local outbreak. We detected nine amino acid changes, previously associated with immune escape, spread across multiple lineages. The N440K change was detected in 45/162 (27.7%) of the sequences, 37 of these were in the B.1.36 lineage (37/65, 56.92%)

Conclusions: Our data support the idea that variants of concern

Open Peer Review

Approval Status ✓ ✓

	1	2
version 2 (revision) 07 Feb 2022		✓ view
version 1 13 May 2021	✓ view	? view

1. **Nei-Yuan Hsiao**, University of Cape Town, Cape Town, South Africa
National Health Laboratory Service, Johannesburg, South Africa

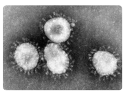
2. **Richard Orton**, MRC-University of Glasgow Centre for Virus Research, Glasgow, UK

Any reports and responses or comments on the article can be found at the end of the article.

spread by travel. Viruses with amino acid replacements associated with immune escape are already circulating. It is critical to check transmission and monitor changes in SARS-CoV-2 locally.

Keywords
SARS-CoV-2, variants, Variants of Concern, VOC, COVID-19, COVID-19 India, Karnataka, India, genomic epidemiology, outbreak investigation, SARS-CoV-2 spread, SARS-CoV-2 variants of concern

This article is included in the Wellcome Trust/DBT India Alliance gateway.



This article is included in the Coronavirus (COVID-19) collection.

Corresponding authors: Chitra Pattabiraman (chitra.nimhans@gmail.com), Anita Desai (anitasdesai@gmail.com)

Author roles: Pattabiraman C: Conceptualization, Data Curation, Formal Analysis, Funding Acquisition, Writing – Original Draft Preparation; **Prasad P**: Data Curation, Investigation, Methodology, Writing – Review & Editing; **George AK**: Formal Analysis, Investigation, Writing – Review & Editing; **Sreenivas D**: Data Curation, Formal Analysis, Writing – Review & Editing; **Rasheed R**: Investigation; **Reddy NVK**: Data Curation; **Desai A**: Conceptualization, Supervision, Writing – Original Draft Preparation, Writing – Review & Editing; **Vasanthapuram R**: Conceptualization, Funding Acquisition, Supervision, Writing – Original Draft Preparation, Writing – Review & Editing

Competing interests: No competing interests were disclosed.

Grant information: This work was supported by core funds of NIMHANS to the Department of Neurovirology, funds from the Government of Karnataka for genomic surveillance of SARS-CoV-2 and the DBT/Wellcome Trust India Alliance Fellowship IA/E/15/1/502336 awarded to Chitra P.
The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

How to cite this article: Pattabiraman C, Prasad P, George AK *et al.* **Importation, circulation, and emergence of variants of SARS-CoV-2 in the South Indian state of Karnataka [version 2; peer review: 2 approved]** Wellcome Open Research 2022, **6**:110 https://doi.org/10.12688/wellcomeopenres.16768.2

First published: 13 May 2021, **6**:110 https://doi.org/10.12688/wellcomeopenres.16768.1

REVISED **Amendments from Version 1**

Additional information has been added to the manuscript to provide the context for the sequences analyzed in the study. Location information and the epi curve for SARS-CoV-2 in the region during the study period has been added to Figure 1. Tables 2–4 have been moved to extended data. Modifications to the Abstract, Methods, Results and Discussion have been made in response to the comments from the reviewers. Extended data have been modified to include new tables as well as 2 additional figures.

Modified extended data is available at the following location: DOI 10.17605/OSF.IO/S56BR

Any further responses from the reviewers can be found at the end of the article

Introduction

The coronavirus disease 2019 (COVID-19) pandemic caused by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) has claimed millions of lives and has affected people living in all parts of the globe[1]. The evolution of the virus did not initially alarm public health specialists or those involved in vaccine development[2]. However, the emergence of variants with distinct biological properties which include one or more mutations that confer higher infectivity, increased transmission, severe disease, re-infection, and immune escape are a cause for concern[3–9]. Such variants may influence the trend of the pandemic and are therefore broadly known as Variants of Concern (VOCs)[3–8].

In India, the COVID-19 pandemic began with the importation of the virus in January 2020[10]. It is only after 11 million cases and over 150,000 deaths that the numbers declined, signalling the end of the first wave of SARS-CoV-2 in the country[1,10,11]. As with other countries in the world, India too started vaccination campaigns in January 2021, at about the same time that reports of VOCs were communicated from the United Kingdom (UK), Brazil, and South Africa[3,4,6,11]. The primary concern is that they may herald the second wave of SARS-CoV-2 in the county and/or undermine the vaccination drive.

Genomic studies in India have shown that several lineages of SARS-CoV-2 have been introduced, have spread, and fallen below the limit of detection since January 2020[12–22]. We have previously performed detailed genomic epidemiology of SARS-CoV-2 in the South Indian state of Karnataka, with a population of 64.1 million (Census 2014)[22]. We found multiple introductions of SARS-CoV-2 into the state and at least seven distinct lineages were already circulating in the state by May 2020. Detailed analysis of the contact network of COVID-19 cases to look at transmission within the state emphasized the role of symptomatic individuals in spreading the virus[23]. These data have contributed to our understanding of how the virus enters, spreads, and evolves in a population. In the genomic epidemiology study, no particular lineages were associated with disease severity[22]. Studies of sequences from India juxtaposed with sequences from all over the world, suggest that mutations associated with immune escape and re-infection are already circulating in the population[2,24–26].

Multiple lineages of SARS-CoV-2 have been reported from across the world and in India[12,13,15–17,19–22,27]. There are two ancestral lineages of SARS-CoV-2 in the PANGO classification system, A and B[28]. While viruses of both lineages are circulating across the world, viruses of lineage B are more widespread and prominent in number. The viruses responsible for the outbreak in Italy, in early 2020, with an amino acid change in the spike protein D614G were classified into lineage B.1[28]. This lineage is now the dominant lineage across the world. Several studies have now shown that viruses in this lineage transmit better, with increased infectivity in cell culture[29–32].

Viruses of the lineage B.1 have acquired several other amino acid replacements in the Receptor Binding Domain of the Spike protein – specifically in the lineages which have been designated as VOCs, namely -B.1.1.7 (N501Y), B.1.351 (N501Y, E484K, K417T) and P.1 from the lineage B.1.1.28 (N501Y, E484K, K417T). Some of these amino acid replacements either singly or in combination have been shown to influence transmission of the virus, interfere with neutralization of the virus, and are associated with an increase in the number of hospitalizations[2,5,7,8]. The spread of these lineages, therefore, has global implications[5,33]. Early data suggests that some variants may escape neutralization by both therapeutic antibodies and antibodies induced by previous infection and vaccination[8,9,34,35]. This has implications for the efficacy of Spike sequence-based vaccines and suggests that re-infection is possible[7,36].

Rapid sharing of genomic information enabled the global community to pick-up cases of VOCs and implement relevant public health measures[3,4,6]. A concentrated, ongoing, local approach to genomic surveillance is critical for the identification of variants and establishing epidemiological links with the trend of the outbreak[5,7,12,22]. This has also proved critical for local outbreak management and informed policy decisions across the world[5,7,37,38].

It is in this context that we conducted genomic surveillance of COVID-19 positive international travellers to the south Indian state of Karnataka between Dec 22, 2020- Jan 31, 2021 (n=75). We also performed sequencing of SARS-CoV-2 (n=108), in samples collected between Nov 22, 2020- Jan 22, 2021) in Bengaluru city (Bengaluru Urban District) to identify and track locally circulating variants and potential VOCs.

Methods

Study setting and ethical considerations

The Department of Neurovirology, at the National Institute of Mental Health and Neurosciences (NIMHANS), Bengaluru, is an ICMR (Indian Council of Medical Research) approved COVID-19 diagnostic centre. The Government of Karnataka and the Government of India designated our lab as a nodal centre for genomic sequencing. This study was granted a waiver

by the Institutional Ethics Committee of NIMHANS in light of the public health emergency. All samples were collected for routine diagnosis for COVID-19, as part of the State's requirement for epidemiological investigation of variants; and de-identified before analysis of data.

Samples for sequencing

On December 23, 2020, the Government of India established surveillance for detecting the importation of Variants of Concern. As part of this surveillance, nasopharyngeal and oro-pharyngeal swabs were collected from international travellers arriving at the international airport in Bengaluru between Dec 22, 2020- Jan 31, 2021. Samples testing positive by reverse transcription polymerase chain reaction (RT-PCR) and having Ct value < 30 (n=75) were included in the study. Further, the surveillance was also designed to include at least 5% of RT PCR positive samples received for routine diagnosis of COVID-19 from Nov 2020 -Jan 2021. To fulfil this samples from COVID-19 cases in Bengaluru city (n=108, 16.25% (108/664) collected between Nov 22, 2020- Jan 22, 2021) through routine surveillance and from a local outbreak in a nursing college in Bengaluru city in Feb 2021, n=14 were included in the study. Of the 42 samples collected from the local outbreak, 14 were suitable for sequencing (RT-PCR positive, Ct value < 30) and were analysed further. From previous experiments in our laboratory using a similar sequencing approach, we have ascertained that a Ct value of < 30 can inform on lineage of the virus and a Ct of < 25 was correlated with recovery of complete genomes. This was used to set the cut-offs for the two sample types.

Data for epidemiological curve

The data for confirmed new cases and numbers tested for SARS-CoV-2 in Bengaluru Urban district during the study period (Nov 22, 2020 – Jan 31, 2021) were obtained from covid19india.org, a dashboard that collated information released by the Government of Karnataka.

Nucleic Acid extraction and RT-PCR

Nucleic acid extraction was performed with automated magnetic bead-based extraction method, using the Chemagic Viral DNA/RNA special H96 kit (PerkinElmer, CMG-1033-S) following manufacturer's instruction. SARS-CoV-2 detection was done using ICMR approved diagnostic kits. A total of 197 RT-PCR positive samples fulfilling the following criteria – (i) Ct values less than 30 in the case of international travellers (n=75), and local outbreak (n=14) or (ii) Ct value less than 25 for local cases (n=108), were taken for whole genome sequencing. Samples and RNA were stored at 4C for <1 week and -80C for long term storage.

Whole genome sequencing of SARS-CoV-2

Whole genome sequencing was performed using the amplicon sequencing approach described in the ARTIC Network protocol using the V3 primer set[39]. The resulting amplicons from 12–24 samples were barcoded using the native barcoding kits (NBD104/114, Oxford Nanopore Technology (ONT)) and sequencing libraries were prepared using the ligation sequencing kit (SQK-LSK109, ONT). The barcoded library was loaded on to FLO-MIN-106 flow cells and sequenced on the MinION (ONT). An average of 0.12 million (median) sequencing reads were acquired per sample with a median coverage of 1737x (see extended data S1[40]). Raw sequencing reads have been deposited within BioProject ID: PRJNA670824.

Analysis of sequencing data and data sharing

Analysis of sequencing reads was performed as described previously[22]. Briefly, sequences were basecalled and demultiplexed using guppy (v3.6) from ONT, alternatively open source basecallers such as Bonito can be used Sequences of length 100–600bp were considered for further analysis. Primer sequences were removed from the sequencing reads by trimming 25bp at the ends and additional trimming based on alignment using BBDuk (v38.37). Resulting reads were mapped to SARS-CoV-2 reference genome (NC_045512) using Minimap2 (v2.17) within Geneious Prime (Geneious Prime 2020.0.3). A consensus was created by calling the majority base (most common base) at each position, resulting in a consensus with the lowest ambiguities. Regions with coverage lower than 10 reads were called Ns. The consensus was aligned to the reference to ensure the correct reading frame, edited by manual inspection of the contigs, followed by transfer of annotations from the reference sequence. Of the 183 samples from international travellers and local cases, 176 (73/75 imported, 103/108 circulating) genomes could be used for the determination of lineage using the PANGO web application (Pangolin v2.2.2 lineages version 2021-02-12)[28]. Of the 176 genomes,162 were complete (>92% at 1X and >85% at 10X) and were deposited into the GISAID Database[41], accession numbers for the sequences are provided as extended data (see extended data S2[40]). Sequences have also been deposited in GenBank with accession numbers OM073810 – OM073973. Complete sequences (162) were analysed for SNPs and amino acid replacements with reference MN908947.3 (Wuhan-Hu-1) using the CoV-Glue Web Application[42].

Phylogenetic analysis

A total of 168 genomes, including the 162 described above, and an additional 6 complete genomes from a local outbreak, were used for phylogenetic analysis with the reference NC_045512 as an outgroup. Multiple sequence alignment was performed using MUSCLE and a maximum likelihood tree was constructed using iqtree[43,44]. The GTR+F+I+G4 substitution model was found to be the best-fit model (of the 88 models tested) using the Bayesian Information Criterion. The consensus tree was constructed from 1000 bootstraps and bootstrap values over 70 were interpreted.

Results

In the study period (Nov 22, 2020 – Jan 21, 2021) an average of 510 SARS-CoV-2 cases were detected daily in the district of Bengaluru Urban in the South Indian state of Karnataka (Figure 1A, B). We sequenced SARS-CoV-2 genomes from 197 SARS-CoV-2 positive individuals, including international travellers (n=75), local cases (n=108), and a local outbreak (n=14).

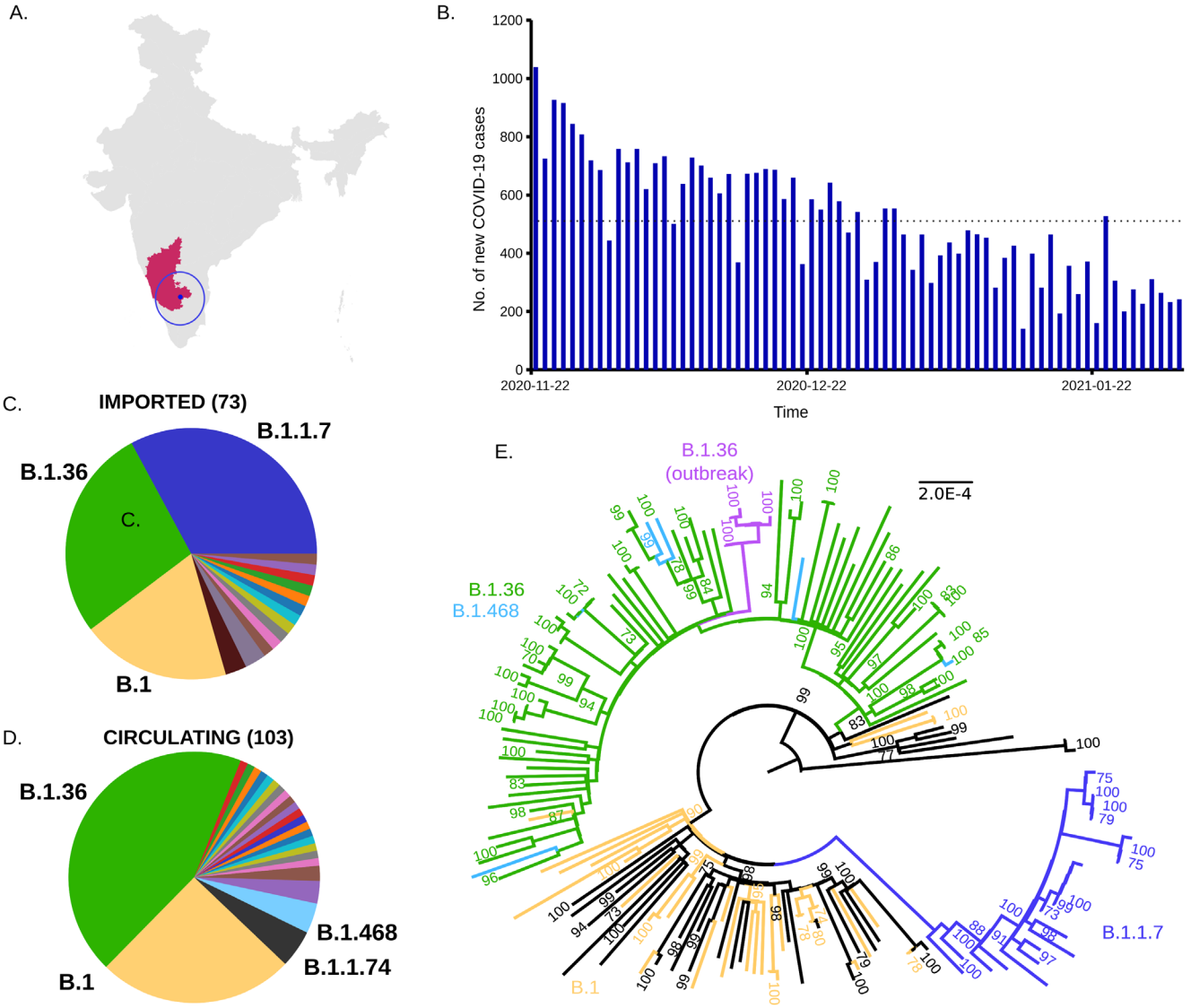


Figure 1. Distribution of SARS-CoV-2 Lineages in the state of Karnataka (Nov 22, 2020- Jan 31, 2021). A) Geographical location of Bengaluru city (blue dot, highlighted by blue circle) in the South Indian state of Karnataka (in pink) is shown on the map of India. (**B**) Epidemiological curve of COVID-19 cases in Bengaluru Urban district in the study period. Dotted horizontal line is the average daily cases for the time period. (**C**) Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) lineages imported by international travel into Karnataka (n= 73). (**D**) SARS-CoV-2 lineages circulating in Bengaluru city (n= 103). Colours represent different lineages. Lineages with greater > 4 sequences are labelled. (**E**) Maximum Likelihood Phylogenetic tree of 168 complete SARS-CoV-2-genomes from Karnataka, rooted by the refer-ence genome (NC_045512). The scale (length of branches) is in substitutions per nucleotide site. Predominant lineages are coloured. Sequences from a local outbreak of SARS-CoV-2 are coloured in magenta. Numbers on the nodes indicate bootstrap support values (in %), values above 70 are shown.

Lineage classification using the PANGO scheme was possible for 176 genomes which were either imported (73/75) or circulating (103/108), and for all 14 genomes from the local outbreak (extended data S3[40]). The genomic surveillance for the local outbreak was carried out to identify the lineage/lineages responsible for the outbreak (Figure 1CE).

A total of 34 lineages were detected from the 176 genomes in this study. A complete list of lineages and their frequencies are tabulated in Table 1. Briefly, genomes from imported and circulating viruses belong to both A (3/176) and B (173/176) lineages. Within A, two (2/103) circulating genomes were classified into A.23.1. Of the 173 genomes in lineage B, two genomes were classified into lineage B (2/173), the rest were derived from B.1 (130/173) or B.1.1 (41/173).

The genomes from imported cases grouped into 16 distinct lineages (Figure 1AC, Table 1) including B.1.1.7 (24/73, 32.9%), B.1.36 (20/73, 27.4%) and B.1 (14/73, 19.2%). The first introduction of B.1.1.7 was noted in the last week of December

Table 1. PANGO lineage assignments for SARS-CoV-2 genomes. Frequency of SARS-CoV-2 lineages (PANGO classification) and percentages of sequenced samples in the imported (by international travel, n =73) or already circulating (in Bengaluru city, n=103) are tabulated. Lineages in black are present in both categories, blue were found in the imported group and orange in circulating viruses. Sub-lineages of B.1.36 are highlighted.

	Lineage	IMPORTED		CIRCULATING	
	(PANGO)	Dec 22, 2020 – Jan 31, 2021		Nov 22, 2020 – Jan 22, 2021	
		Frequency	Percentage	Frequency	Percentage
1	B.1.1.7	24	32.9	1	1.0
2	B.1.36	20	27.4	45	43.7
3	B.1	14	19.2	26	25.2
4	B.1.177	2	2.7		
5	B.1.468	2	2.7	4	3.9
6	A	1	1.4		
7	B	1	1.4	1	1.0
8	B.1.1.194	1	1.4		
9	B.1.1.317	1	1.4		
10	B.1.1.74	1	1.4	5	4.9
11	B.1.177.19	1	1.4		
12	B.1.177.4	1	1.4		
13	B.1.2	1	1.4		
14	B.1.258	1	1.4		
15	B.1.308	1	1.4		
16	B.1.36.18	1	1.4		
17	A.23.1			2	1.9
18	B.1.1.106			1	1.0
19	B.1.1.130			1	1.0
20	B.1.1.184			1	1.0
21	B.1.1.197			1	1.0
22	B.1.1.216			3	2.9
23	B.1.1.306			1	1.0
24	B.1.197			1	1.0
25	B.1.216			1	1.0
26	B.1.221			1	1.0
27	B.1.256			1	1.0
28	B.1.36.10			1	1.0
29	B.1.36.13			1	1.0
30	B.1.36.17			1	1.0
31	B.1.36.23			1	1.0
32	B.1.456			1	1.0
33	B.1.509			1	1.0
34	B.1.188			1	1.0

2020, and by January 31, 2021, this lineage made up 32.9% (24/73) of all imported cases (Figure 1AC, Table 1). Circulating genomes grouped into 24 distinct lineages, dominated by the lineages B.1.36 (45/103; 43.7%), B.1 (26/103; 25.2%), B.1.1.74 (5/103; 4.9%) and B.1.468 (4/103; 3.9%) (Figure 1BD, Table 1). Only a single sequence of B.1.1.7 was detected during the study period as part of this surveillance effort in a non-traveller. Sequences from the lineage B.1.36 and derived lineages (70/176) grouped into a distinct phylogenetic clade together with sequences belonging to lineage B.1.468 (6/176) (Figure 1CE). Phylogenetic analysis of a sub-set of B.1.36 genomes from across India and across the world showed that most of the sequences both from people with a history of international travel as well as circulating viruses clustered together (extended data SF1[40]).

Genomic investigation of an outbreak of SARS-CoV-2 in the city of Bengaluru in early Feb 2021, revealed that 14/14 sequences (all 14 genomes had 1X coverage >74%.) from the outbreak could be classified into lineage B.1.36. Complete genome sequences (with >92% coverage of reference at 1X and >85% at 10X) could be recovered from 6/14 cases. All six viruses grouped into a clade within the largely B.1.36+ B.1.468 clade (Figure 1CE).

Of the 176 genomes from travellers and in circulation, for which lineage classification was possible,162 complete genomes (with coverage > 92% at 1X and > 85% at 10X) were used for the analysis of SNPs and amino acid replacements. A total of 968 SNPs (extended data, S4[40]) and 529 amino acid replacements (extended data S5[40]) were identified. Of these amino acid replacements 61 were in the Spike protein of circulating viruses, and 32 in Spike protein of imported viruses (extended data S6[40]). The B.1.36 lineage had 226 amino acid replacements, 31 of these were in the Spike protein. Although only the D614G and N440K were present in an appreciable number (>50%) of sequences (extended data S7[40]. The N440K was found in 37/65 (56.92%) of B.1.36 sequences (extended data S7[40]).

We carried out further analysis of the amino acid replacements in the receptor binding domain (RBD) of the spike protein (Figure 2A, extended data, S6[40]) and mapped them on the Maximum-Likelihood tree (Figure 2B). We identified mutations leading to nine amino acid replacements in the RBD (Figure 2A). Of these, five (S477N, E484K, E484Q, S494L, S494P) were found in viruses circulating in Bengaluru, and the amino acid replacement V483A was from an imported case. The N501Y change was confined to the B.1.1.7 lineage. The N440K mutation was present in 45/162 (27.7%) of the sequences. All 45 sequences with N440K are grouped into the B.1.36+B.1.468 clade (Figure 2B). Of the six sequences from a cluster of cases (Outbreak), only a single sequence carried the mutation resulting in the N440K change (Figure 2B). A single branch of the B.1.36+B.1.468 clade (n=4, 3 of which were imported) had an additional amino acid replacement F490S in the RBD (Figure 2B). The mutations in the RBD were seen across the phylogenetic tree and clades (Figure 2B).

Discussion

In this study, we found 34 lineages of SARS-CoV-2 circulating and imported into Bengaluru city in Karnataka, India, between Nov 22, 2020 – Jan 31, 2021. We aimed to detect the introduction of the global VOCs (lineages B.1.1.7, B.1.351, P.1/ B.1.1.28), as well as genotype the variants of SARS-CoV-2, circulating since our last study, which highlighted the introduction and spread of seven lineages of SARS-CoV-2 in Karnataka, between March-May 2020[22].

We found no evidence suggesting that the B.1.1.7 lineage was present in Karnataka before late-Dec 2020. We first detected the B.1.1.7 variant in Karnataka, in an international traveller from a sample collected on Dec 22, 2020 (extended data, S3[40]). The first and only case of non-travel related B.1.1.7, in our study, was detected in the middle of Jan 2021 in an individual who was in contact with an international traveller (extended data, S3[40]). These data together suggest that B.1.1.7 in Karnataka was limited to travel-associated cases and was not in the community during the study period. At the end of the study period, the B.1.1.7 lineage was detected in 32.9% of all imported cases (Table 1). We did not detect the variants P.1/B.1.1.28 or B.1.351 reported from Brazil and South Africa respectively in this study.

We found that B.1.36 and B.1 lineages dominated in both the imported (20/73; 27.4%,14/73, 19.2%) and circulating viruses (45/103; 43.7%, 26/103; 25.2%) in our study (Table 1). The clustering of B.1.36 lineage from travellers with locally circulating viruses suggests either a common source of infection or infection with circulating lineages post-arrival (extended data SF1). B.1.36 was first reported from Saudi Arabia in Feb 2020 (extended data S8[40] Table 1). The B.1.36 lineage was both imported by international travel (20/73) and circulating (45/103) in Bengaluru city (Table 1). The lineage is characterized by the following amino acid replacements- nsp12-P323L(95.38%), S-D614G (93.85%), S-N440K (56.92%), ORF 3a-Q57H (90.77%), ORF 3a-E261*(81.54%), nsp3-T183I (81.54%), nsp16-L126F(80%), N-S2P (72.31%), ORF 8-S97I (72.31%) (extended data, Supplementary Table 7[40]). The immune escape associated amino acid change, N440K has been reported from the states of Andhra Pradesh, Maharashtra, Telangana, and Karnataka, and is also associated with reinfection[24,36,45]. This change was found in 37/65 (56.92%) of the sequences clustering to B.1.36 (extended data S7[40]). Sequencing from Bengaluru Urban district, which encompasses Bengaluru city, during this time period is sparse. Analysis of a limited number of sequences (n= 649 between 23 Nov 2020 - May 2 2021) from our laboratory suggest that B.1.36 and its sub lineages dominated in late 2020, co-circulated with B.1.1.7/Alpha and B.1.617.1/Kappa between Jan-Mar 2021 and were displaced by B.1.617.2/Delta by April-May 2021 (extended data SF2).

An outbreak of SARS-CoV-2 occurred in Bengaluru in early Feb 2021, raising concerns about the spread of variants, the threat of a second wave, and reduction in the efficacy of vaccines. This outbreak in a college where students were returning from different states within India was driven by related viruses

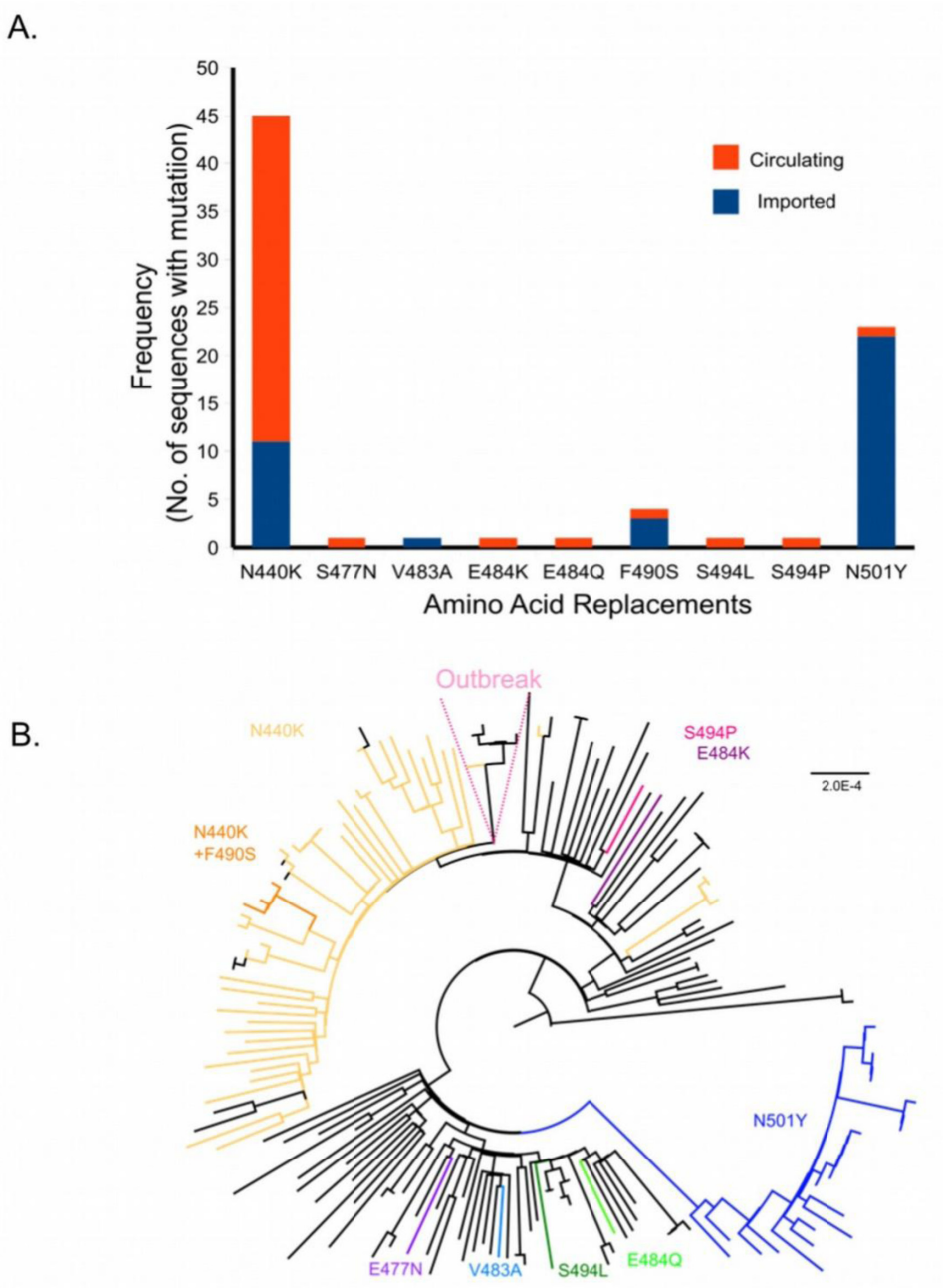


Figure 2. Amino acid replacements in the Receptor Binding Domain (RBD) of severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2). (**A**) Frequency of amino acid replacements in the RBD (amino acids- 387- 516 in the Spike protein) is shown as a bar graph. Frequencies are plotted against the amino acid replacement. Orange and blue represent the circulating and imported genomes respectively. (**B**) Maximum Likelihood phylogenetic tree highlighting branches with the indicated amino acid substitutions.

belonging to the B.1.36 lineage (Figure 1CE and extended data, S3[40]). Only one of the six sequences from the outbreak cluster had the mutation resulting the N440K replacement in the Spike protein (Figure 2C, extended data S6).

Apart from the introduction and spread of known VOCs, the emergence of variants locally is also a cause for concern. Early in the pandemic, a single mutation in the gene encoding the Spike protein of SARS-CoV-2 resulting in a D614G amino acid change was identified to increase infectivity and transmission[2,29,32]. Viruses with this amino acid replacement dominate across the globe[31,46]. Mutations in the gene encoding the Spike protein are of particular concern due to the role of this protein and its Receptor Binding Domain (RBD) in viral binding and entry[47].

Some of these mutations have been shown to increase infectivity, affinity to the angiotensin converting enzyme 2 (ACE-2) receptor or affect neutralization by antibodies i*n vitro*. Viral genomes with these mutations were already circulating viruses by mid-2020[2,25,26,45,48,49].

In the sequences from this study, nine amino acids replacements were noted in the RBD domain of the Spike protein (Figure 2B and extended data S6[40]). They occurred singly or in pairs (N440K+F490S) (Figure 2). All nine amino acid changes, namely N440K, S477N, V483A, E484K/Q, F490S, S494L/P, N501Y are associated with immune escape[24,25]. Viruses with some of these amino acid changes were already known to be circulating in other parts of India[16,17,24].

Mutations in the gene encoding Spike protein that do not map to the RBD have also been described; particularly near the polybasic cleavage site at the S1/S2 boundary of the Spike protein. Towards the end of the year 2020, multiple lineages with amino acid replacements at position 677 were noted[50]. Four viruses in our study have mutations resulting in amino acid changes at this position (Q677H (n=3), Q677P (n=1)) (extended data, S6[40]).

It is to be noted that in this study we have only included samples with Ct values less than 25 for surveillance of circulating SARS-CoV-2 genomes and Ct values less than 30 for sequencing of international travel-related cases. We have also sequenced only a fraction of cases in a limited geographical area. This may therefore present an incomplete view of circulating viruses and inflate the ones that are more readily sequenced. Also, as we have used the amplicon sequencing approach, not all regions of all lineages are well covered by sequencing reads. Others have also noted homoplasy in SARS-CoV-2, this highlights the need to be cautious while interpreting the phylogenetic relationships between SARS-CoV-2 sequences, especially in the context of outbreaks[51].

In summary, our data highlight an increase in the frequency of the lineage B.1.36 in Bengaluru Urban, in Karnataka, and indicate an underappreciated global presence of this lineage (Figure 1, Table 1, extended data S8[40]). Whether this increase is because of epidemiological linkages such as increased travel, continued local transmission chains or super-spreader events remains to be determined. It is beyond the scope of this work to examine whether the lineage, contributing mutations, and amino acid changes impact transmission/infectivity of the virus. Our data emphasize that a consolidated and local approach to genomic surveillance which includes sequencing of SARS-CoV-2 from travellers, circulating variants, and outbreaks, in a continuous manner is necessary to detect VOCs. We believe that in regions where sequencing capacity is limited and sustained and continuous sequencing of local cases is a challenge, rapid and focussed sequencing of travellers and local outbreaks may serve as early warning systems for novel variants. Even as this conjecture remains to be tested, it is clear that rapid identification of such variants can aid in preparing the healthcare system for a surge in cases, suggest revisions to vaccines and diagnostic tests, inform the international community, and guide public health measures.

Data availability

Underlying data

NCBI BioProject: SARS-CoV-2 Genome Sequencing. Accession number PRJNA670824; https://identifiers.org/NCBI/bioproject:PRJNA670824.

Extended data

Open Science Framework: SARS-CoV-2 Sequencing.

DOI 10.17605/OSF.IO/S56BR[40].

This project contains the following extended data:

S1: Summary of sequencing results

S2: GISAID Accession ID for sequences

S3: Lineage, source and collection date of sequenced samples

S4: Position and frequency of single nucleotide polymorphisms

S5: Position and frequency of amino acid replacements

S6: Amino acid replacement in Spike protein

S7: Frequency of amino acid replacements in lineage B.1.36

S8: Location and Sequence counts for B.1.36 between Nov 22, 2020 – Jan 31, 2021

S9: B.1.36 Sequence GISAID Submitters_acknowledgement

SF1: Phylogenetic tree of a subset of B.1.36 sequences (Nov 2020 - Jan 2021)

SF2: Distribution of SARS-CoV-2 lineages in Bengaluru Urban (Nov 23, 2020 – May 2, 2021)

Acknowledgements

This work would not have been possible without the support of the Government of Karnataka, State Surveillance team for COVID-19. We would like to thank all the labs and Primary Health Care centres that collected samples for testing and genomic surveillance. We would like to thank the COVID testing lab in NIMHANS. We would also like to acknowledge the National Centre for Biological Sciences (NCBS) for support with reagents (Prof. Sudhir Krishna's laboratory) and computational resources for carrying out our analysis, and Dr. Farhat Habib for custom scripts used in data analysis. We gratefully acknowledge the contributions of all the laboratories that have submitted their sequences to GISAID, in particular laboratories across India that have been involved in sequencing efforts and the INSACOG consortium of which NIMHANS is a member.

An earlier version of this article can be found on medRxiv (doi: https://doi.org/10.1101/2021.03.17.21253810).

References

1. Dong E, Du H, Gardner L: **An interactive web-based dashboard to track COVID-19 in real time.** *Lancet Infect Dis.* 2020; **20**(5): 533–534.
PubMed Abstract | Publisher Full Text | Free Full Text

2. Li Q, Wu J, Nie J, *et al.*: **The Impact of Mutations in SARS-CoV-2 Spike on Viral Infectivity and Antigenicity.** *Cell.* 2020; **182**(5): 1284–1294.e9.
PubMed Abstract | Publisher Full Text | Free Full Text

3. Rambaut A, Loman N, Pybus O, *et al.*: **Preliminary genomic characterisation of an emergent SARS-CoV-2 lineage in the UK defined by a novel set of spike mutations.** *VirologicalOrg.* 2020.
Reference Source

4. Faria NR, Mellan TA, Whittaker C, *et al.*: **Genomics and epidemiology of a novel SARS-CoV-2 lineage in Manaus, Brazil.** *medRxiv.* 2021; 2021.02.26.21252554.
PubMed Abstract | Publisher Full Text | Free Full Text

5. Volz E, Mishra S, Chand M, *et al.*: **Transmission of SARS-CoV-2 Lineage B.1.1.7 in England: Insights from linking epidemiological and genetic data.** *medRxiv.* 2021.
Publisher Full Text

6. Tegally H, Wilkinson E, Giovanetti M, *et al.*: **Emergence and rapid spread of a new severe acute respiratory syndrome-related coronavirus 2 (SARS-CoV-2) lineage with multiple spike mutations in South Africa.** *medRxiv.* 2020.
Publisher Full Text

7. Sabino EC, Buss LF, Carvalho MPS, *et al.*: **Resurgence of COVID-19 in Manaus, Brazil, despite high seroprevalence.** *Lancet.* 2021; **397**(10273): 452–455.
PubMed Abstract | Publisher Full Text | Free Full Text

8. Cele S, Gazy I, Jackson L, *et al.*: **Escape of SARS-CoV-2 501Y.V2 from neutralization by convalescent plasma.** *medRxiv.* 2021.
Publisher Full Text

9. Wang P, Nair MS, Liu L, *et al.*: **Antibody Resistance of SARS-CoV-2 Variants B.1.351 and B.1.1.7.** *Nature.* 2021 ; 593(**7857**):130–135.
PubMed Abstract | Publisher Full Text

10. ICMR, Team CE& DM, Team CL: **Laboratory surveillance for SARS-CoV-2 in India: Performance of testing & descriptive epidemiology of detected COVID-19, January 22 - April 30, 2020.** *Indian J Med Res.* 2020; **151**(5): 424–437.
PubMed Abstract | Publisher Full Text | Free Full Text

11. **Coronavirus outbreak in Karnataka.** (accessed March 11, 2021).
Reference Source

12. Kumar P, Pandey R, Sharma P, *et al.*: **Integrated genomic view of SARS-CoV-2 in India [version 1; peer review: 3 approved].** *Wellcome Open Res.* 2020; **5**: 184.
PubMed Abstract | Publisher Full Text | Free Full Text

13. Srivastava S, Banu S, Singh P, *et al.*: **SARS-CoV-2 genomics: An Indian perspective on sequencing viral variants.** *J Biosci.* 2021; **46**(1): 22.
PubMed Abstract | Publisher Full Text | Free Full Text

14. Banu S, Jolly B, Mukherjee P, *et al.*: **A Distinct Phylogenetic Cluster of Indian Severe Acute Respiratory Syndrome Coronavirus 2 Isolates.** *Open Forum Infect Dis.* 2020; **7**(11): ofaa434.
PubMed Abstract | Publisher Full Text | Free Full Text

15. Maitra A, Sarkar MC, Raheja H, *et al.*: **Mutations in SARS-CoV-2 viral RNA identified in Eastern India: Possible implications for the ongoing outbreak in India and impact on viral structure and host susceptibility.** *J Biosci.* 2020; **45**(1): 76.
PubMed Abstract | Publisher Full Text | Free Full Text

16. Rani PR, Imran M, Lakshmi JV, *et al.*: **Insights from Genomes and Genetic Epidemiology of SARS-CoV-2 isolates from the state of Andhra Pradesh.** *bioRxiv.* 2021; 2021.01.22.427775.
Publisher Full Text

17. Gupta A, Sabarinathan R, Bala P, *et al.*: **Mutational landscape and dominant lineages in the SARS-CoV-2 infections in the state of Telangana, India.** *medRxiv.* 2020.
Publisher Full Text

18. Yadav PD, Potdar VA, Choudhary ML, *et al.*: **Full-genome sequences of the first two SARS-CoV-2 viruses from India.** *Indian J Med Res.* 2020; **151**(2 & 3): 200–209.
PubMed Abstract | Publisher Full Text | Free Full Text

19. Radhakrishnan C, Divakar MK, Jain A, *et al.*: **Initial insights into the genetic epidemiology of SARS-CoV-2 isolates from Kerala suggest local spread from limited introductions.** *bioRxiv.* 2020.
Publisher Full Text

20. Jain A, Rophina M, Mahajan S, *et al.*: **Analysis of the potential impact of genomic variants in global SARS-CoV-2 genomes on molecular diagnostic assays.** *Int J Infect Dis.* 2021; **102**: 460–462.
PubMed Abstract | Publisher Full Text | Free Full Text

21. Joshi M, Puvar A, Kumar D, *et al.*: **Genomic variations in SARS-CoV-2 genomes from Gujarat: Underlying role of variants in disease epidemiology.** *bioRxiv.* 2020.
Publisher Full Text

22. Pattabiraman C, Habib F, Harsha PK, *et al.*: **Genomic epidemiology reveals multiple introductions and spread of SARS-CoV-2 in the Indian state of Karnataka.** *PLoS One.* 2020; **15**(12): e0243412.
PubMed Abstract | Publisher Full Text | Free Full Text

23. Kumar N, Hameed SKS, Babu GR, *et al.*: **Descriptive epidemiology of SARS-CoV-2 infection in Karnataka state, South India: Transmission dynamics of symptomatic vs. asymptomatic infections.** *EClinicalMedicine.* 2021; **32**: 100717.
PubMed Abstract | Publisher Full Text | Free Full Text

24. Jolly B, Rophina M, Shamnath A, *et al.*: **Genetic epidemiology of variants associated with immune escape from global SARS-CoV-2 genomes.** *bioRxiv.* 2020.
Publisher Full Text

25. Greaney AJ, Starr TN, Gilchuk P, *et al.*: **Complete Mapping of Mutations to the SARS-CoV-2 Spike Receptor-Binding Domain that Escape Antibody Recognition.** *Cell Host Microbe.* 2021; **29**(1): 44–57.e9.
PubMed Abstract | Publisher Full Text | Free Full Text

26. Starr TN, Greaney AJ, Hilton SK, *et al.*: **Deep Mutational Scanning of SARS-CoV-2 Receptor Binding Domain Reveals Constraints on Folding and ACE2 Binding.** *Cell.* 2020; **182**(5): 1295–1310.e20.
PubMed Abstract | Publisher Full Text | Free Full Text

27. Bhoyar RC, Jain A, Sehgal P, *et al.*: **High throughput detection and genetic epidemiology of SARS-CoV-2 using COVIDSeq next generation sequencing.** *bioRxiv.* 2020.
Publisher Full Text

28. Rambaut A, Holmes EC, O'Toole Á, *et al.*: **A dynamic nomenclature proposal for SARS-CoV-2 lineages to assist genomic epidemiology.** *Nat Microbiol.* 2020; **5**(11): 1403–1407.
PubMed Abstract | Publisher Full Text | Free Full Text

29. Plante JA, Liu Y, Liu J, *et al.*: **Spike mutation D614G alters SARS-CoV-2 fitness.** *Nature.* 2021; **592**(7852): 116–121.
PubMed Abstract | Publisher Full Text | Free Full Text

30. Jackson CB, Zhang L, Farzan M, *et al.*: **Functional importance of the D614G mutation in the SARS-CoV-2 spike protein.** *Biochem Biophys Res Commun.* 2021; **538**: 108–115.
PubMed Abstract | Publisher Full Text | Free Full Text

31. Korber B, Fischer WM, Gnanakaran S, *et al.*: **Tracking Changes in SARS-CoV-2 Spike: Evidence that D614G Increases Infectivity of the COVID-19 Virus.** *Cell.* 2020; **182**(4): 812–827.e19.
PubMed Abstract | Publisher Full Text | Free Full Text

32. Volz E, Hill V, McCrone JT, *et al.*: **Evaluating the Effects of SARS-CoV-2 Spike Mutation D614G on Transmissibility and Pathogenicity.** *Cell.* 2021; **184**(1): 64–75.e11.
PubMed Abstract | Publisher Full Text | Free Full Text

33. O'Toole Á, Hill V, Pybus OG, *et al.*: **Tracking the international spread of SARS-CoV-2 lineages B.1.1.7 and B.1.351/501Y-V2.** *VirologicalOrg.* 2021.
Reference Source

34. Hoffmann M, Arora P, Groß R, *et al.*: **SARS-CoV-2 variants B.1.351 and B.1.1.248: Escape from therapeutic antibodies and antibodies induced by infection and vaccination.** *bioRxiv.* 2021.
Publisher Full Text

35. Wibmer CK, Ayres F, Hermanus T, *et al.*: **SARS-CoV-2 501Y.V2 escapes neutralization by South African COVID-19 donor plasma.** *bioRxiv.* 2021; 2021.01.18.427166.
PubMed Abstract | Publisher Full Text | Free Full Text

36. Gupta V, Bhoyar RC, Jain A, *et al.*: **Asymptomatic reinfection in two healthcare workers from India with genetically distinct SARS-CoV-2.** *Clin Infect Dis.* 2020; ciaa1451.
PubMed Abstract | Publisher Full Text | Free Full Text

37. Page AJ, Mather AE, Le-Viet T, *et al.*: **Large scale sequencing of SARS-CoV-2 genomes from one region allows detailed epidemiology and enables local outbreak management.** *medRxiv.* 2020.
Publisher Full Text

38. Tegally H, Wilkinson E, Lessells RJ, *et al.*: **Sixteen novel lineages of SARS-CoV-2 in South Africa.** *Nat Med.* 2021; **27**(3): 440–446.
PubMed Abstract | Publisher Full Text

39. Quick J: **nCoV-2019 sequencing protocol v3 (LoCost).** Protocols.io. 2020.
Reference Source

40. Pattabiraman C: **SARS-CoV-2 Sequencing**. 2021.
http://www.doi.org/10.17605/OSF.IO/S56BR

41. **GISAID**.
Reference Source

42. Singer J, Gifford R, Cotten M, *et al.*: **CoV-GLUE: A Web Application for Tracking SARS-CoV-2 Genomic Variation**. *Preprints.* 2020; 2020060225.
Publisher Full Text

43. Nguyen LT, Schmidt HA, von Haeseler A, *et al.*: **IQ-TREE: A fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies.** *Mol Biol Evol.* 2015; **32**(1): 268–74.
PubMed Abstract | Publisher Full Text | Free Full Text

44. Edgar RC: **MUSCLE: multiple sequence alignment with high accuracy and high throughput.** *Nucleic Acids Res.* 2004; **32**(5): 1792–7.
PubMed Abstract | Publisher Full Text | Free Full Text

45. Weisblum Y, Schmidt F, Zhang F, *et al.*: **Escape from neutralizing antibodies by SARS-CoV-2 spike protein variants.** *eLife.* 2020; **9**: e61312.
PubMed Abstract | Publisher Full Text | Free Full Text

46. Biswas NK, Majumder PP: **Analysis of RNA sequences of 3636 SARS-CoV-2 collected from 55 countries reveals selective sweep of one virus type.** *Indian J Med Res.* 2020; **151**(5): 450–458.
PubMed Abstract | Publisher Full Text | Free Full Text

47. Shang J, Ye G, Shi K, *et al.*: **Structural basis of receptor recognition by SARS-CoV-2.** *Nature.* 2020; **581**(7807): 221–224.
PubMed Abstract | Publisher Full Text | Free Full Text

48. Baum A, Fulton BO, Wloga E, *et al.*: **Antibody cocktail to SARS-CoV-2 spike protein prevents rapid mutational escape seen with individual antibodies.** *Science.* 2020; **369**(6506): 1014–1018.
PubMed Abstract | Publisher Full Text | Free Full Text

49. Starr TN, Greaney AJ, Addetia A, *et al.*: **Prospective mapping of viral mutations that escape antibodies used to treat COVID-19.** *Science.* 2021; **371**(6531): 850–854.
PubMed Abstract | Publisher Full Text | Free Full Text

50. Hodcroft EB, Domman DB, Snyder DJ, *et al.*: **Emergence in late 2020 of multiple lineages of SARS-CoV-2 Spike protein variants affecting amino acid position 677.** *medRxiv.* 2021; 2021.02.12.21251658.
PubMed Abstract | Publisher Full Text | Free Full Text

51. Turakhia Y, de Maio N, Thornlow B, *et al.*: **Stability of SARS-CoV-2 phylogenies.** *PLoS Genet.* 2020; **16**(11): e1009175.
PubMed Abstract | Publisher Full Text | Free Full Text

Open Peer Review

Current Peer Review Status: ✔ ✔

Version 2

Reviewer Report 18 February 2022

https://doi.org/10.21956/wellcomeopenres.19557.r48535

✔ **Richard Orton**
MRC-University of Glasgow Centre for Virus Research, Glasgow, UK

The authors have addressed all my previous comments. Two minor points:

1. References should be used for the tools used such as minimap2, etc. - there are papers for these tools.

2. The bioinformatics section should have the "manual inspection/curation" line expanded a little bit further to state that you specifically checking indwells and manually inspecting for underlying read support.

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: viral bioinformatics, inter and intra-host viral evolution

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Version 1

Reviewer Report 09 November 2021

https://doi.org/10.21956/wellcomeopenres.18493.r46594

? **Richard Orton**

MRC-University of Glasgow Centre for Virus Research, Glasgow, UK

Overall, I found this to be a very well written and interesting paper. I do have a number of minor points to address:

- Abstract - (B.136: 20/73, 27.4%; B.1: 14/73, 19.2%) - B.1.36 not B.136?

- Introduction - "The viruses responsible for the catastrophic outbreak in Italy," - I have not seen this outbreak in Italy described as "catastrophic" before, and almost suggests this is the cause of the pandemic rather than the initial outbreak in China.

- Methods - Analysis of seq data and data sharing - this point is quite important:

- The ONT sequencing was done using the artic protocol and artic v3 primers. The bioinformatics, however, does not use the recommended artic protocol: https://artic.network/ncov-2019/ncov2019-bioinformatics-sop.html. There needs to be more information on how the bioinformatics was done, specifically this step "A consensus genome was generated with a coverage cut-off of 10x and the 0% majority rule."

- What tool was used to generate this? What does a 0% majority's rule mean?

- To put into context the artic pipeline using a 40x cutoff for consensus calling, and the variant calling goes through nanopolish to address the highly problematic indels from ONT data, there are other steps as well to remove reads of abnormal length, and remove reads aligning to unexpected regions - were any of these steps done? If not are you sure of the reliability of your consensus seqs? How many of your consensus sequences have indels which broke coding sequences?

- Results - "Genomic investigation of an outbreak of SARS-CoV-2 in the city of Bengaluru in early Feb 2021, revealed that 14/14 sequences from the outbreak could be classified into lineage B.1.36. Complete genome sequences could be recovered from 6/14 cases." How can all 14 be classified as B.1.36 if only 6/14 could have a complete genome recovered?

- Overall - Lineage B.1.36 - I found it quite surprising that B.1.36 formed a large proportion of importations. This seems a relatively small lineage (and just checking it out quickly seems only really observed at a decent freq in India, Canada, and Hong Kong - and to a lesser degree the UK) - so it seems quite surprising that this is the dominant lineage imported from international travellers? I think it would be very useful to have a breakdown of the country of origin the travellers have come from - is this possible? How many of the travellers are tourists and how many are returning travellers (would they have been tested before travelling out) And/Or - A table of B.1.36 counts from countries around the world for the weeks preceding and during the study.

- Context - some extra elements for general context might be useful - a map showing the geographical location of the region, an epidemic curve showing the number of COVID-19 cases (& genomes sequenced) in India and the region in question, and the dominant lineages over time - in particular, B.1.36 but also B.1.1.7 and later delta - did B.1.36 persist after the introduction of B.1.1.7 and then the subsequent emergence of delta, etc

- As the mutation N440K is a main point of the paper - how frequently is this mutation observed across lineages and within B.136 (all seqs)? The N440K mutation does not seem be a lineage defining mutation of B.1.36 - but the majority of your seqs seem to have it.

Is the work clearly and accurately presented and does it cite the current literature?
Yes

Is the study design appropriate and is the work technically sound?
Yes

Are sufficient details of methods and analysis provided to allow replication by others?
Yes

If applicable, is the statistical analysis and its interpretation appropriate?
Not applicable

Are all the source data underlying the results available to ensure full reproducibility?
Yes

Are the conclusions drawn adequately supported by the results?
Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: viral bioinformatics, inter and intra-host viral evolution

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Author Response 30 Jan 2022

Chitra Pattabiraman, National Institute of Mental Health and Neurosciences, India, Bangalore, India

We thank the reviewer for accepting and taking the time to review this manuscript during the pandemic and for their considered comments. We have made changes to the manuscript based on their recommendations. Please find a point-wise response to the comments below with changes to the manuscript indicated

1.Abstract - (B.136: 20/73, 27.4%; B.1: 14/73, 19.2%) - B.1.36 not B.136?

This has been corrected.

2. Introduction - "The viruses responsible for the catastrophic outbreak in Italy," - I have not seen this outbreak in Italy described as "catastrophic" before, and almost

suggests this is the cause of the pandemic rather than the initial outbreak in China.

The word catastrophic has been removed.

3. Methods - Analysis of seq data and data sharing - this point is quite important:

The ONT sequencing was done using the artic protocol and artic v3 primers. The bioinformatics, however, does not use the recommended artic protocol: https://artic.network/ncov-2019/ncov2019-bioinformatics-sop.html. There needs to be more information on how the bioinformatics was done, specifically this step "A consensus genome was generated with a coverage cut-off of 10x and the 0% majority rule."
The methods section has been changed to include a detailed bioinformatic workflow. “Briefly, sequences were basecalled and demultiplexed using guppy (v3.6) from ONT, alternatively open source basecallers such as Bonito can be used Sequences of length 100–600bp were considered for further analysis. Primer sequences were removed from the sequencing reads by trimming 25bp and additional trimming based on alignment using BBDuk (v38.37). Resulting reads were mapped to SARS-CoV-2 reference genome (NC_045512) using Minimap2 (v2.17) within Geneious Prime (Geneious Prime 2020.0.3). A consensus was created by calling the majority base (most common base) at each position, resulting in a consensus with the lowest ambiguities. Regions with coverage lower than 10 reads were called Ns. The consensus was aligned to the reference to ensure the correct reading frame, edited by manual inspection of the contigs, followed by transfer of annotations from the reference sequence.”

4. What tool was used to generate this? What does a 0% majority's rule mean?
A mapping based assembly was performed using minimap2 to the SARS-CoV-2 reference genome (NC_045512). The consensus generation was performed using Geneious Prime® 2020.0.3. The 0% majority rule within Geneious Prime is the following - the most common base at position is called, resulting in a consensus with the lowest ambiguities. A coverage cut-off of 10X was set in order to call the consensus. Regions with coverage lower than 10 were called Ns.

We have changed this sentence in the Methods (described above) for clarity.

5. To put into context the artic pipeline using a 40x cutoff for consensus calling, and the variant calling goes through nanopolish to address the highly problematic indels from ONT data, there are other steps as well to remove reads of abnormal length, and remove reads aligning to unexpected regions - were any of these steps done? If not are you sure of the reliability of your consensus seqs? How many of your consensus sequences have indels which broke coding sequences?
We agree with the reviewer that these are important considerations to generate a good quality consensus. As described above in the changes to the methods section - adapter and barcode removal, primer trimming and length filters were used in the analysis workflow.

Due to the computational resources needed for nanopolish, we chose to manually inspect the contigs for indels. The following rules were followed during manual inspection-

The indels that were due to sequencing errors tended to be poorly supported/regions of heterogeneity and caused frameshifts. We manually inspected the contig for frameshifts. If the frameshift was due to an indel, we inspected the supporting reads. If >70% of the reads had the indel, the consensus was left unchanged. If the indel was not well supported and there were at least 10 reads without the indel, then the majority base was called. In case of a tie, the consensus was resolved to the wild-type/reference base. If less than 10 reads were present, the consensus was corrected to an N. If the indel was present at the end of reads, only the bridging reads were considered for calling the consensus using the above rules.

We have previously validated this approach using a few samples with nanopolish and are therefore confident of the consensus calls. We also noted that the indels occurred at specific positions in the genome, depending on the genomic context - usually homopolymer repeats or end of reads. These positions that needed manual editing were conserved for a particular lineage and were roughly about 29-30 positions per assembly. Also, both the phylogenetics and Nextclade QC did not suggest major issues with our consensus sequences. Additionally we also provide the sequencing reads in the SRA database for others to rebuild the consensus.

6. Results - "Genomic investigation of an outbreak of SARS-CoV-2 in the city of Bengaluru in early Feb 2021, revealed that 14/14 sequences from the outbreak could be classified into lineage B.1.36. Complete genome sequences could be recovered from 6/14 cases." How can all 14 be classified as B.1.36 if only 6/14 could have a complete genome recovered?

We call genomes complete if consensus covers >85% of reference at 10X and >92% of reference at 1X. Some genomes were not complete by this criteria (extended data S1 and S3), however they had enough information to be assigned a lineage using the Pangolin tool. All 14 genomes had 1X coverage >74%. This now indicated in the text.

7. Overall - Lineage B.1.36 - I found it quite surprising that B.1.36 formed a large proportion of importations. This seems a relatively small lineage (and just checking it out quickly seems only really observed at a decent freq in India, Canada, and Hong Kong - and to a lesser degree the UK) - so it seems quite surprising that this is the dominant lineage imported from international travellers? I think it would be very useful to have a breakdown of the country of origin the travellers have come from - is this possible? How many of the travellers are tourists and how many are returning travellers (would they have been tested before travelling out) And/Or - A table of B.1.36 counts from countries around the world for the weeks preceding and during the study.

We thank the reviewer for the opportunity to discuss this further. We do not have a detailed travel history for the international travellers, however, people with a history of travel to the

UK were prioritized for sequencing at that time. The lineage B.1.36 has been detected in 67 countries (with >5% prevalence Saudi Arabia, Iran and Afghanistan) based on GISAID data collated by outbreak.info and has been circulating since February 2020. A cumulative table B.1.36 sequences based on location between Nov 22, 2020 - Jan 21, 2021 is provided in extended data S8. This point is addressed further in response to comment 2 from reviewer 1 and in extended data SF1.

8. As the mutation N440K is a main point of the paper - how frequently is this mutation observed across lineages and within B.136 (all seqs)? The N440K mutation does not seem be a lineage defining mutation of B.1.36 - but the majority of your seqs seem to have it.

The 440K is not a lineage defining mutation for B.1.36. In our study 45/176 genomes had this mutation. The 440K substitution was observed in 37/65 (56.92%) B.1.36 genomes. The following changes have been made to highlight this in the results section-
“The N440K change was detected in 45/162 (27.7%) of the sequences, 37 of these were in the B.1.36 lineage (37/65, 56.92%).”
“The N440K was found in 37/65 (56.92%) of B.1.36 sequences (extended data S7).”
“The N440K mutation was present in 45/162 (27.7%) of the sequences. All 45 sequences with N440K are grouped into the B.1.36+B.1.468 clade (Figure 2B).”

9. Context - some extra elements for general context might be useful - a map showing the geographical location of the region, an epidemic curve showing the number of COVID-19 cases (& genomes sequenced) in India and the region in question, and the dominant lineages over time - in particular, B.1.36 but also B.1.1.7 and later delta - did B.1.36 persist after the introduction of B.1.1.7 and then the subsequent emergence of delta, etc

Fig1 has been modified to include a map of India, showing Karnataka and Bengaluru city. An epi curve for Bengaluru Urban (No. of new cases detected daily during the study period) has also been included as Fig 1B. The sequencing in this region has not been uniform and continuous, however, the data indicate displacement of B.1.36 by Delta by April 2021 (extended data SF2).

The following line has been added to the text to provide the context
“Sequencing from Bengaluru Urban district, which encompasses Bengaluru city, during this time period is sparse. Analysis of a limited number of sequences (n= 649 between 23 Nov 2020 - May 2 2021) from our laboratory suggest that B.1.36 and its sub lineages dominated in late 2020, co-circulated with B.1.1.7/Alpha and B.1.617.1/Kappa between Jan-Mar 2021 and were displaced by B.1.617.2/Delta by 'April-May 2021 (extended data SF2).”

Competing Interests: No competing interests were disclosed.

Reviewer Report 26 July 2021

https://doi.org/10.21956/wellcomeopenres.18493.r44897

Nei-Yuan Hsiao
[1] Division of Medical Virology, Department of Pathology, University of Cape Town, Cape Town, South Africa
[2] National Health Laboratory Service, Johannesburg, South Africa

Thank you for the opportunity to review this manuscript.

In this study, Nimhans and colleagues performed a whole genome sequencing on virus identified among returning travellers, local surveillance effort and one outbreak in college students. The main purpose of the study is to describe the circulation of variants in Karnataka state and identify key mutations that may have clinical and public health importance.
I think the work is of high quality and the manuscript both clearly presented the research conducted and correctly cites the international literature around this issue. As the study is descriptive in nature, I found the study design to be appropriate and the methodology widely adopted in the field so readily reproducible.
The key interesting finding of the study is the identification of the B.1.36 lineage which I admit have not come across before. This lineage was both present in high enough proportion and contain significant mutations in spike to warrant attention and investigation. The authors did well to describe the extent of spread of this variant at the time, given the difficulty of the constantly changing nature SARS-CoV-2 lineage circulation.
There are only a few minor suggested revisions on my part:

1. The local epidemic situation at the time of study could provide important context for which the genomic data can be interpreted. A simple epidemic curve or a description of number of people tested and positive in the region would assist readers further gauge the significance of the findings around novel variants.

2. The trees and phylogenetic analyses were solely based on local sequences. Even though for description purposes this is sufficient, adding related national and international sequences, especially those from potential sources of introduction, could vastly improve the interpretation.

3. Table 3 and 4 are superfluous and can probably just be described in text.

4. I would like to see the author discuss how this work shape their future priorities around genomic surveillance. The discussion suggested that the returning travellers are more readily sequenced than routinely diagnosed local cases. Would sequencing returning travellers from an important part of genomic surveillance? I think it would be good to add more voices in this discussion.

Is the work clearly and accurately presented and does it cite the current literature?
Yes

Is the study design appropriate and is the work technically sound?
Yes

Are sufficient details of methods and analysis provided to allow replication by others?
Yes

If applicable, is the statistical analysis and its interpretation appropriate?
I cannot comment. A qualified statistician is required.

Are all the source data underlying the results available to ensure full reproducibility?
Yes

Are the conclusions drawn adequately supported by the results?
Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Clinical virology, Epidemiology, molecular diagnostics

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Author Response 30 Jan 2022

Chitra Pattabiraman, National Institute of Mental Health and Neurosciences, India, Bangalore, India

We thank the reviewer for accepting and taking the time to review this manuscript during the pandemic and for their considered comments. We have made changes to the manuscript based on their recommendations. Please find a point-wise response to the comments below with changes to the manuscript indicated

1. The local epidemic situation at the time of study could provide important context for which the genomic data can be interpreted. A simple epidemic curve or a description of the number of people tested and positive in the region would assist readers further gauge the significance of the findings around novel variants

We thank the reviewer for this suggestion. We have modified Figure1 to include an epi curve showing the daily confirmed cases in Bengaluru Urban district in the study period (Figure 1A). The average number of cases was 510, with a median test positivity of 1.2 %. The methods section now includes a data source for this graph. The following line has been added to the results section.
“In the study period an average of 510 SARS-CoV-2 cases were detected daily in the district of Bengaluru Urban in the South Indian state of Karnataka (Figure 1).”

2.The trees and phylogenetic analyses were solely based on local sequences. Even though for description purposes this is sufficient, adding related national and international sequences, especially those from potential sources of introduction, could vastly improve the interpretation.

We have performed a phylogenetic analysis using limited sequences of the B.1.36 lineage from all over India and across the world in this time period. This is included as part of extended data SF1. The following changes have been made-
"Phylogenetic analysis of a subset of B.1.36 genomes (n=183) from across India and across the world showed that sequences from the study (n=70) both from circulating viruses as well as travel associated cases largely clustered together (extended data SF1[40])."
"The clustering of B.1.36 lineage from travellers with locally circulating viruses suggests either a common source of infection or infection with circulating lineages post-arrival (extended data SF1).

3. **Table 3 and 4 are superfluous and can probably just be described in text.**

Table 2 and 3 have been described in the text and moved to extended data table S6. Table 4 has been removed, and a more detailed summary count of Lineage B.1.36 from across the world is provided in extended data S8.

4. I would like to see the author discuss how this work shape their future priorities around genomic surveillance. The discussion suggested that the returning travellers are more readily sequenced than routinely diagnosed local cases. Would sequencing returning travellers from an important part of genomic surveillance? I think it would be good to add more voices in this discussion.

The resources for sequencing are not available equally in all parts of the world and given the resource limitations amplified by the pandemic, it has been hard to argue for an investment in genomic surveillance when it is not always clear what to do with the data or who truly benefits from it. Nevertheless, we believe this is important and of local and global relevance.

We have modified the last paragraph of the discussion as follows-
"We believe that in regions where sequencing capacity is limited and sustained and continuous sequencing of local cases is a challenge, rapid and focussed sequencing of travellers and local outbreaks may serve as early warning systems for novel variants. Even as this conjecture remains to be tested, it is clear that rapid identification of such variants can aid in preparing the healthcare system for a surge in cases, suggest revisions to vaccines and diagnostic tests, inform the international community, and guide public health measures."

Competing Interests: No competing interests were disclosed.