Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID-19. The COVID-19 resource centre is hosted on Elsevier Connect, the company's public news and information website.

Virus Research 308 (2022) 198642

Contents lists available at ScienceDirect

Virus Research

journal homepage: www.elsevier.com/locate/virusres

Cross-sectional genomic perspective of epidemic waves of SARS-CoV-2: A pan India study

Sanjeet Kumar [a], Kanika Bansal [b,*]

[a] School of Biotechnology, Gangadhar Meher University, Sambalpur, India
[b] CSIR- Institute of Microbial Technology, Chandigarh 160036, India

ARTICLE INFO

Keywords:
SARS-CoV-2
COVID-19
Genome-wide
Evolution
Deadly variants
VOC
VOI
SNP
Mutation
Non-synonymous
Silent mutation
Spike
RNA dependent RNA polymerase
NSP
UTR

ABSTRACT

Background: COVID-19 has posed unforeseen circumstances and throttled major economies worldwide. India has witnessed two waves affecting around 31 million people representing 16% of the cases globally. To date, the epidemic waves have not been comprehensively investigated to understand pandemic progress in India.
Objective: Here, we aim for pan Indian cross-sectional evolutionary analysis since inception of SARS-CoV-2.
Methods: High quality genomes, along with their collection date till 26th July 2021, were downloaded. Whole genome-based phylogeny was obtained. Further, the mutational analysis was performed using SARS-CoV-2 first reported from Wuhan (NC_045512.2) as reference.
Results: Based on reported cases and mutation rates, we could divide the Indian epidemic into seven phases. The average mutation rate for the pre-first wave was <11, which elevated to 17 in the first wave and doubled in the second wave (~34). In accordance with mutation rate, VOCs and VOIs started appearing in the first wave (1.5%), which dominated the second (~96%) and post-second wave (100%). Nation-wide mutational analysis depicted >0.5 million mutation events with four major mutations in >19,300 genomes, including two mutations in coding (spike (D614G), and NSP 12b (P314L) of rdrp), one silent mutation (NSP3 F106F) and one extragenic mutation (5′ UTR 241).
Conclusion: Whole genome-based phylogeny could demarcate post-first wave isolates from previous ones by point of diversification leading to incidences of VOCs and VOIs in India. Such analysis is crucial in the timely management of pandemic.

1. Introduction

Coronavirus represents a large family of RNA viruses causing upper and lower respiratory tract infections to humans ranging from mild to lethal. Previously reported outbreaks of coronaviruses causing significant public health threats include Severe Acute Respiratory Syndrome (SARS) and Middle East Respiratory Syndrome (MERS) (Memish et al., 2014; Peiris et al., 2004). In late December 2019, the ongoing outbreak was caused by novel coronavirus epi-centered in Hubei province of People's Republic of China (Chen et al., 2020; Wu et al., 2020). Patients were epidemiologically linked to a wet animal and seafood wholesale market in Wuhan (Lu et al., 2020; Bogoch et al., 2020). Based on phylogeny and taxonomic analysis Coronavirus Study Group of International Committee on Taxonomy of Viruses recognized this as a sister to SARS-CoV (Peiris et al., 2004) and named it as SARS-CoV-2 (Gorbalenya et al., 2020). SARS-CoV-2 has the largest genome (26.4 to 31.7 kb) among all known RNA viruses with a variable GC content ranging from 32 to 43% (Woo et al., 2010). On 30th January 2020, a global health emergency was declared by the WHO Emergency Committee. Due to substantial human-to-human transmissions, SARS-CoV-2 has spread to many countries, and till now affected more than 195 million people with more than 4 million casualties worldwide (Lu et al., 2020; Worldometer, 2020).

The first case of SARS-CoV-2 from India was reported in Kerala between 27 and 31 January 2020 from individuals with a travel history of Wuhan, China (Andrews et al., 2020). In order to contain further spread of SARS-CoV-2 strict restrictions were imposed, like banning the flights to/from the affected countries, despite that, continuous local transmission of the virus resulted in a considerable surge in COVID-19 cases (https://www.worldometers.info/coronavirus/country/india/;

Abbreviations: VOC, variant of concern; VOI, variant of interest; RDRP, RNA dependent RNA polymerase; NSP, non-structural protein; UTR, untranslated region.

* Corresponding author.
E-mail address: kanikabansal@imtech.res.in (K. Bansal).

https://doi.org/10.1016/j.virusres.2021.198642
Received 3 September 2021; Received in revised form 20 October 2021; Accepted 18 November 2021
Available online 22 November 2021

https://www.covid19india.org/). Therefore, a nationwide lockdown from 26th March to 11th May 2020 was imposed to contain the spread. These were one of the most rigid lockdown restrictions in the world which helped in controlling infectivity rate in India (Mitra et al., 2020; Maitra et al., 2020). After unlocking, India again witnessed a surge in cases resulting in the first wave from July to December 2020. Nationwide first wave was at its peak with 93,732 cases on 17th September 2020. However, after six weeks, the toll had come to half. In the mid of March 2021, a second wave started in India which peaked on the 8th May 2021 with 391,236 cases per day. This wave witnessed a steep rise in COVID-19 cases, which over-burdened the healthcare system in the country. The strict lockdown restrictions, proper identification of containment zones played a crucial role in controlling the second wave. The second wave was effectively controled in record time of around three weeks, contrary to the first wave that lasted for several months. The second wave is reported ending by June 2021. During both the waves, Delhi and Maharashtra were the most badly affected states with several localized outbursts due to rampant community transmission. As of 26th July 2021, India reported 30,820 cases per day with total cases of 31,440,492 and the death toll reached to 421,414. Nevertheless, a nationwide third wave was also predicted, which was a great concern for the policymakers and public governance.

We have witnessed the generation of unprecedented genomic resources of SARS-CoV-2 worldwide such as by the UK Consortium (Gorbalenya et al., 2020), African union (Salyer et al., 2021), Indian SARS-CoV-2 Genomic Consortia (INSACOG) (Maitra et al., 2020; Alai et al., 2021), etc. GISAID (https://www.gisaid.org/), is a global initiative for a public repository for the genomic data of SARS-COV-2 storage and analysis. Such a vast genomic resource has been investigated in detail based on mutation in SARS-CoV-2 into various lineages (Rambaut et al., 2020). Based on these studies, World Health organization has announced variants of concern (VOC) (alpha, beta, gamma, and delta) and variants of interest (VOI) (eta, lota, kappa, lambda, and mu) (https://www.who.int/en/activities/tracking-SARS-CoV-2-variants/). These VOCs and VOIs are known to pose an increased risk to public health globally and will aid in monitoring the evolution of deadly variants worldwide.

In order to understand the genetic diversity, transmission, and cure in human, several large-scale genome-based studies have been conducted (Salyer et al., 2021; Bajaj and Purohit, 2020; Phan, 2020; Helmy et al., 2020; Yadav et al., 2021). Pan India study of 1000 sequences across ten states suggests the widespread presence of the several lineages of SARS-CoV-2 (Maitra et al., 2020). Pan India sero survey suggested average seropositivity to be 10.14% (among 10,427 subjects) (Naushin et al., 2021). Unfortunately, the spread of lineages across India and seropositivity rate is very complex due to the vast population and landmass. Genomic diversity of Indian isolates compared to the global lineages is supposed to evolve further, which needs to be closely monitored (Alai et al., 2021).

To date, pan India genome-based studies focus on the evolution of SARS-CoV-2 only upto first wave (Maitra et al., 2020; Alai et al., 2021; Yadav et al., 2021). However, since then, India has witnessed devastating second wave with more than 0.4 million cases per day which are four times the cases reported during the first wave (https://www.worldometers.info/coronavirus/country/india/; https://www.covid19india.org/). This created a lacuna in understanding the evolution of deadly variants of SARS-CoV-2. Since the first epidemic wave, there has been an upsurge in public genomic resources of SARS-CoV-2 from India, which is available from the global GISAID initiative. Current scenario provides the scope of cross-sectional genome-based monitoring of the deadly variants across India. In the present study, we have analyzed 20,086 high-quality genomes to understand the dominance of VOCs and VOIs in the second wave. We could identify 0.52 million mutational events, out of which 90% were intergenic. Single nucleotide polymorphism (SNP) ($n = 0.46$ million) was the major player in the evolution of SARS-CoV-2 in India. Overall, we could identify four major mutation events in more than 19,300 genomes in spike, RNA dependent RNA polymerase, and extragenic 5′UTR. Pan India study based on the mutation can open a gateway to understand the hotspots of mutations in SARS-CoV-2.

2. Results and discussion

2.1. Evolutionary timeline of epidemic waves of SARS-CoV-2 in India

In India, the first case of COVID-19 was reported at the end of January 2020. Till then, the number of cases has increased abruptly twice, which we call the first and second waves. In order to understand the peak and plateau of incidences, we have differentiated the period from the first incidence (January 2020) to 26th July 2021 in seven different phases (Fig. 1C). Here, phase I indicates the early days of infection, i.e., the introductory phase from January 2020 to 25th March 2020. Phase II refers to the nationwide lockdown from 26th March 2020 to 11th May 2020, which was imposed to contain the spread of COVID-19. Once the lockdown was relaxed, the incidences of cases rose gradually, termed as pre-first wave period or phase III from 12th May 2020 to 31st June 2020. Phase IV, or the first wave of COVID-19 was demarcated for quite a long duration from 1st July 2020 to 31st December 2020. India had witnessed a peak incidence of 93,732 cases of COVID-19 on 17th September 2020 (https://www.worldometers.info/coronavirus/country/india/; https://www.covid19india.org/) during phase IV (Fig. 1A). With the gradual decrease in cases across India, phase V was demarcated from 1st January 2021 to 15th March 2021 as post-first or pre-second wave. Once again major leap in incidences was observed, resulting in the second wave, which we refer to as phase VI, from 16th March 2021 to 30th June 2021. Strikingly, the rise and fall of incidences during the second wave (phase VI) was steep compared to the first wave (phase IV). Currently, India is undergoing phase VII with a drastic reduction in overall incidences from 1st July 2021 up to 26th July 2021.

2.2. Nation-wide phylogenetic network of SARS-CoV-2

India has witnessed two epidemic waves of SARS-CoV-2, the number of cases and genome resources have also increased accordingly. India reported 6525 high-quality genomes up to first wave and additionally 13,531 high-quality genomes by the end of the second wave and still continuing (Fig. 1B). Overall, up to 26th July 2021, India has reported 31,725,450 cases and 20,086 high-quality genomes (Supplementary Figure 1 and Supplementary Table 1) (https://www.worldometers.info/coronavirus/country/india/; https://www.covid19india.org/).

Pan India phylogeny based on whole-genome sequences ($n = 20,086$) has revealed major lineages of SARS-CoV-2 in India (Table 1). It demarcated post-first wave isolates (phase V, VI, and VII) from the earlier isolates (phase I, II, III, and IV) (Fig. 2). The post-first wave represents the recent introduction of deadly variants in the Indian population of SARS-CoV-2 (Figs. 1E and 2). We could identify the clade representing the point of diversification (marked as a red dot in Fig. 2) as the significant event in the evolution of SARS-CoV-2.

In India outbreak of the second wave witnessed a sudden rise in infection from 0.7 to 1.06% of the total population within two months. During this time maximum per day cases reported were around 0.4 million, which were more than recorded worldwide. According to the pangolin lineages, out of 20,086 strains used in the present study, 7421 were delta variants (B.1.617.2, AY.1, AY.2, and AY.3), 95 were beta (B.1.351) and 1 gamma (P.1) constituting VOCs and 42 eta (B.1.525), 2300 kappa (B.1.617.1) constituting VOIs (Table 2). None of the VOCs or VOIs were reported in India up to phase III, i.e., before the first wave. Deadly variants were first reported during the first wave with VOCs (alpha=55 and delta=1)) and VOIs (kappa=7) constituting 1.5% of the phase IV genomes analysed (Fig. 1E, Table 2 and Supplementary Table 2). After the first wave 41% of the genomes in phase V could be linked to the deadly variants with VOCs (alpha=509, beta=26, and delta=97) and VOIs (eta=6 and, kappa=703). While, second wave or

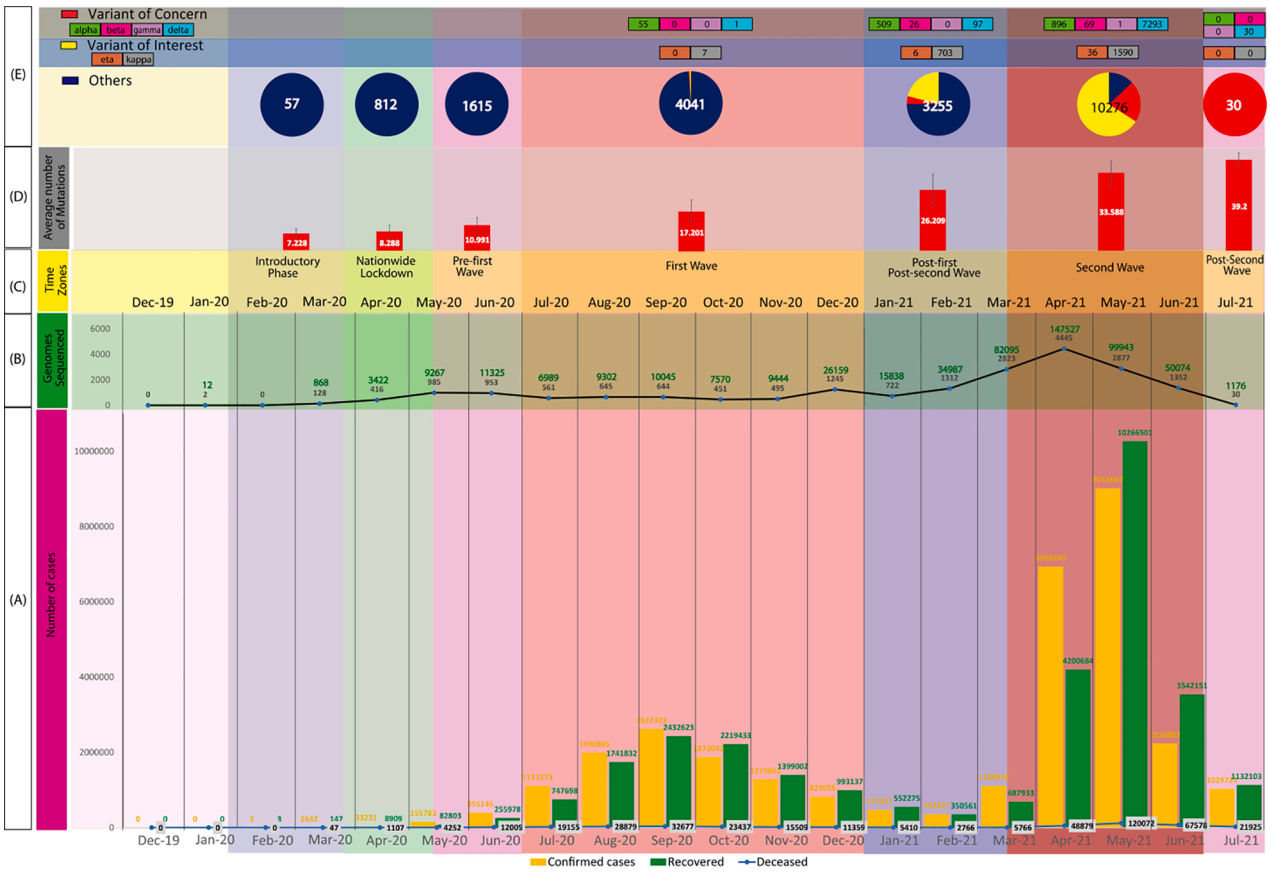

Fig. 1. Pan Indian overview of SARS-CoV-2 across seven phases. (A). Bar graph plot of number of confirmed cases (yellow), recovered (green) and deceased (black) during each month since their first incidence in India. (B). Number of genome sequences submitted in the public repository of GISAID and mutations detected in the present study are labeled in black and green color respectively. (C). Seven phases of SARS-CoV-2 pandemic and their time zones are represented as phase I: introductory phase, phase II: nationwide lockdown, phase III: pre-first wave, phase IV: first wave, phase V: post-first wave/ pre-second wave, phase VI: second wave and phase VII: post-second wave. (D). Average number of mutations during each phase. Standard deviation in the mutation rate is indicated by a vertical line. (E). Distribution of variants of concern (VOC) and variants of interest (VOI) and other lineages defined in accordance with pangolin lineage. The total number of genomes included in each phase is marked in the center of the pie chart. Number of VOCs (alpha, beta, gamma, and delta) and VOIs (eta and kappa) are designated according to the color codes indicated in left side of panel E.(For interpretation of the references to color in this figure, the reader is referred to the web version of this article.)

phase VI was dominated (96%) with VOCs (alpha=896, beta=69, gamma=1 and, delta=7293) and VOIs (eta=36 and, kappa=1590). In comparison, the post-second wave is dominated by delta variants only (based on only 30 genomes sequenced till 26th July 2021) (Table 2). Lota, lambda and mu variants were not detected in India based on the genome sequencing data available to date.

2.3. Mutations driving emergence of VOCs and VOIs

To understand the evolution of SARS-CoV-2, we looked at the mutational events occurring since COVID-19 pandemic inception in India. Nationwide mutational analysis depicted more than 0.52 million mutations upto 26th July 2021 (Supplementary Table 3, Supplementary Fig. 2). Overall, 0.47 million mutations were intergenic and remaining (0.054 million) were in the extergenic region. Mutational count for all the SARS-CoV-2 genes is provided in (Table 3). Among the intergenic mutations, majority contributed towards SNPs of synonymous (0.11 million) and non-synonymous (0.34 million) nature. However, remaining 10,976 mutations in the intergenic region were due to insertion/ deletion events out of which, 1725 indels resulted in frameshift. We have looked at the top twenty prevalent mutations (Table 4). Here, the top twenty mutations were basically found in spike, RNA dependent RNA polymerase, nucleocapsid, ORF3, ORF7, and extragenic regions (5′ UTR and 3′ UTR). Interestingly, there were four widespread mutations in ~97% of the genomes analyzed, essentially representing all over India since its emergence. Two mutations in the protein coding region i.e., D614G in spike and P314L in NSP 12b; one extragenic mutation at 241 position of 5′UTR and one silent mutation at F106 NSP3 (Fig. 3, Table 4). Such prevalent non-synonymous or silent mutations in spike protein and rdrp and 5′ UTR is likely to improve pathogenicity of the virus, in evasion from host immune system and risk of human-to-human transmissions, etc. (Shishir et al., 2021; Plante et al., 2021).

During the initial months of COVID-19 pandemic, number of reported cases and mutational events were not widespread. While, with the increase in number of cases, mutational events also started accumulating before the first wave (Fig. 1B). During the initial months of the first wave (June-September 2020), number of cases and mutational events were in the rising trend. While, during later months of the first wave (October-December 2020) number of cases were at decline, yet, mutational events were rising and reached to a maximum of 26,159 events in December 2020. This also coincided with the emergence of deadly variants during first wave and their increasing dominance after the first wave of pandemic. Interestingly, at the later months of pandemic (January-June 2021), higher mutations were accumulating in the viral genome irrespective of the exponential increase or decrease in the number of cases. Cascade of all these events led to the evolution of deadly variants, outcompeting others. Large-scale genomic analysis of SARS-CoV-2 also concluded that accumulation of mutations over time affects the severity and spread of SARS-CoV-2 globally (Laamarti et al., 2020).

Table 1
Lineage distribution among the genomes of different time zones. Here, pangolin and GISAID lineages are indicated in the first column and number of strains in a lineage in a seven phases are indicated.

	Phase I (Introductory Phase)	Phase II (Nationwide lockdown)	Phase III (Pre-first wave)	Phase IV (First wave)	Phase V (Post-first wave/ Pre-second wave)	Phase VI (Second wave)	Phase VII (Post-second wave)
PANGOLIN LINEAGE							
A	4	17	15	6	–	–	–
A.1	2	–	–	–	–	–	–
A.2	–	1	–	1	4	3	–
A.7	–	12	–	–	–	–	–
A.9	–	16	6	1	–	–	–
AE.1	–	–	–	–	1	1	–
AE.2	–	–	–	1	–	–	–
AM.3	–	–	–	2	1	–	–
AY.1	–	–	–	–	–	6	–
AY.3	–	–	–	–	–	2	1
B	1	5	7	12	1	–	–
B.1	31	476	1520	4002	3239	10,258	29
B.4	13	3	4	5	–	–	–
B.53	1	–	–	–	–	–	–
B.6	5	282	63	9	3	1	–
C.36	–	–	–	1	3	3	–
L.3	–	–	–	–	1	–	–
P.1	–	–	–	–	–	1	–
P.2	–	–	–	1	–	–	–
R.1	–	–	–	–	2	–	–
GISAID LINEAGE							
G	14	214	372	522	1316	8382	13
GH	8	103	439	1576	849	178	–
GK	–	–	–	–	–	777	17
GR	9	154	675	1847	739	426	–
GRY	–	–	–	23	331	460	–
GV	–	–	–	5	3	6	–
L	3	10	3	6	1	–	–
O	15	288	103	54	12	44	–
S	6	42	23	8	4	3	–
V	2	1	–	–	–	–	–

In addition to the confirmed cases, seven phases of the pandemic in India can also be distinguished based on of mutations detected in the rapidly evolving virus. For instance, the average mutation detected before the first wave was less than 11, which elevated to 17.2 and 33.6 in the first and second waves, respectively (Fig. 1D). Hence, average mutations were doubled in the second wave compared to the first wave and still have the rising trend.

Strikingly, in accordance with the phylogeny, the rise in mutations is directly correlated with the emergence of deadly variants (VOCs and VOIs) in India. For instance, genomic analysis reported these deadly variants during the first wave, yet cases due to them started alarmingly incriminating only post-first wave (Fig. 1E). However, the second wave was dominated by these deadly viruses. The post-second wave is hauntingly related to these deadly variants only (this is based on just 30 genomes available till 26th July for phase VII).

The outbreak the origin of SARS-CoV-2 is a heated topic among the scientific community. Lack of direct evidence of zoonotic transfer had shifted the lab leak conspiracy theory to the mainstream. Further, in hunt of its origin, SARS-CoV-2 various aspects of genomics are investigated (Casadevall et al., 2021; Sallard, 2021; Thacker, 2021; Bansal and Patil, 2020). Our comprehensive genome-based study will allow to track and understand the highly evolving SARS-CoV-2.

3. Methods

3.1. Procurement of SARS-CoV-2 genome from public repository

We have considered 20,086 high quality genomes from India encompassing the country's length and breadth. The source of the genomes we considered in this study is the EpiCoV database maintained under GISAID initiative. Here, we have included high quality genomes according to standards of GISAID having information of collection date to aid in time zone depiction. A detailed state-wise information is provided in supplementary information (Supplementary Table 1) indicating patients details such as geographical location, age, sex, pangolin lineage, GISAID lineage etc. First strain reported from Wuhan (China) was taken as a reference strain for all the analysis in the study. We have considered all high-quality complete genomes submitted until 26th of July 2021.

3.2. Phylogenetic analysis

In order to obtain a pan India phylogeny of SARS-CoV-2, we have used all high-quality genomes ($n = 20,086$) available by 26th July 2021 in the public repository of GISAID. Multiple sequencing alignment (MSA) was performed using (MAFFT v7.467) (Nakamura et al., 2018) which is based on fast fourier transform keeping NC 045,512.2 (Wuhan-Hu-1) strain as reference. MSA obtained was used for the phylogenomic tree generation using fasttree v2.1.8 with double precision (Price et al., 2010) with gamma time reversal method (gtr). Visualization of the phylogenomic tree was performed using a web server of iTOL v6 (Letunic and Bork, 2019). The isolates were marked in accordance with their phase of isolation and pangolin lineage (Rambaut et al., 2020).

3.3. Pan India mutation analysis across different phases

In order to understand the rate of mutation across the genome procured for seven different phases, we have first aligned all the sequences ($n = 20,086$) against NC 045,512.2 (Wuhan-Hu-1) strain (reference strain) using nucmer v3.1 (Delcher et al., 2002). Also, we have separately aligned all the strains from several phases against the reference sequence. In order to translate all the alignments scores into mutational

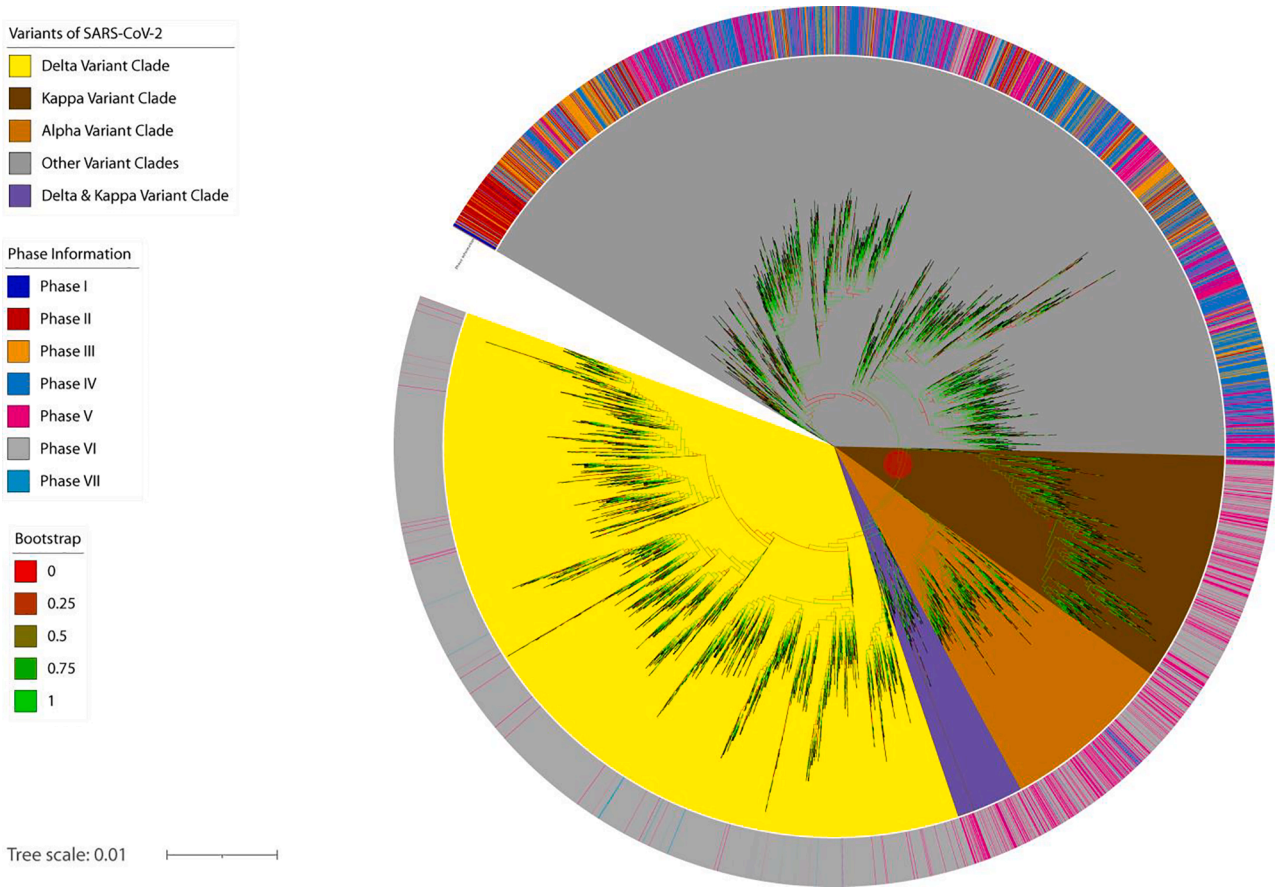

Fig. 2. Pan India whole genome-based phylogeny of SARS-CoV-2. Here, bootstrap values are represented in the color range from 0 (minimum value clade marked with red) to 1 (maximum value clade marked with green). Clades representing variants of SARS-CoV-2 (delta, kappa, alpha etc.) are marked with respective colors as indicated. Isolates reported in different phases are marked with the color strip against their respective leaf in the phylogenetic tree. The point of diversification is indicated by a red dot in the phylogenetic tree.(For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 2
Distribution of variant of concern (VOC) and variant of interest (VOI) down the timeline in India.

	High quality genomes	VOC (alpha + beta + gamma + delta)	VOI (eta + lota + kappa + lambda)
Phase I (Introductory phase)	57	0	0
Phase II (Nation-wide lockdown phase)	812	0	0
Phase III (Pre-first wave)	1615	0	0
Phase IV (First wave)	4041	56 (55 + 0 + 0 + 1)	7 (0 + 0 + 7 + 0)
Phase V (Pre-second or post-first wave)	3255	632 (509 + 26 + 0 + 97)	709 (6 + 0 + 703 + 0)
Phase VI (Second wave)	10,276	8259(896 + 69 + 1 + 7293)	1626 (36 + 0 + 1590 + 0)
Phase VII (Post-second wave)	30	30 (0 + 0 + 0 + 30)	0 (0 + 0 + 0 + 0)

events, we have implemented a well-documented method earlier described by Mercatelli and Giorgi (2020). This approach uses a gff3 annotation file and reference sequence of NC_045512.2 to extract the genomic coordinates of SARS-CoV-2 proteins. R library package seqinr (https://cran.r-project.org/web/packages/seqinr/index.html) and biostring package of bioconductor (https://bioconductor.org/packages/release/bioc/html/Biostrings.html) was used to get the list of mutational

Table 3
Gene-wise mutational count among the pan Indian SARS-CoV-2 isolates.

Gene annotation	Gene	Count of mutational events
RNA dependent RNA polymerase	NSP1	3618
	NSP2	18,508
	NSP3	69,623
	NSP4	19,232
	NSP5	2845
	NSP6	19,438
	NSP7	923
	NSP8	1568
	NSP9	4765
	NSP10	804
	NSP12a	18
	NSP12b	39,778
	NSP13	16,487
	NSP14	12,014
	NSP15	11,825
	NSP16	3966
Spike protein	S	113,339
ORF3a protein	ORF3a	22,608
Envelope	E	876
Membrane	M	16,119
ORF6 protein	ORF6	2104
ORF7a protein	ORF7a	22,018
ORF7b protein	ORF7b	3144
ORF8 protein	ORF8	14,679
Nucleocapsid protein	N	50,947
ORF10 protein	ORF10	745
	Total	471,991

Table 4
Top twenty mutations among the pan Indian SARS-CoV-2 isolates.

Nucleotide mutation	Protein mutation	Number of strains	Gene	Mutation type	Annotation
A23403G	D614G	19,571	Spike	Non-synonymous SNP	Spike
C3037T	F106F	19,527	NSP3	Synonymous mutation	Predicted phosphoesterase, papain-like proteinase
C241T	–	19,583	5′UTR	Extragenic	NA
C14408T	P314L	19,353	NSP 12b	Non-synonymous SNP	RNA-dependent RNA polymerase, post-ribosomal frameshift
G210T	–	9985	5′UTR	Extragenic	NA
C23604G	P681R	9887	Spike	Non-synonymous SNP	Spike
G29402T	D377Y	9878	N	Non-synonymous SNP	Nucleocapsid protein
T22917G	L452R	9875	Spike	Non-synonymous SNP	Spike
G29742T	–	9836	3′UTR	Extragenic	NA
G28881T	R203M	9835	N	Non-synonymous SNP	Nucleocapsid protein
T27638C	V82A	9813	ORF7a	Non-synonymous SNP	ORF7a protein
C25469T	S26L	9713	ORF3a	Non-synonymous SNP	ORF3a protein
C21618G	T19R	7473	Spike	Non-synonymous SNP	Spike
T26767C	I82T	7429	M	Non-synonymous SNP	Membrane
C27752T	T120I	7397	ORF7a	Non-synonymous SNP	ORF7a protein
C22995A	T478K	7376	Spike	Non-synonymous SNP	Spike
A28461G	D63G	6999	N	Non-synonymous SNP	Nucleocapsid protein
G15451A	G662S	6785	NSP12b	Non-synonymous SNP	RNA-dependent RNA polymerase, post-ribosomal frameshift
C16466T	P77L	6762	NSP13	Non-synonymous SNP	Helicase
G24410A	D950N	5470	Spike	Non-synonymous SNP	Spike

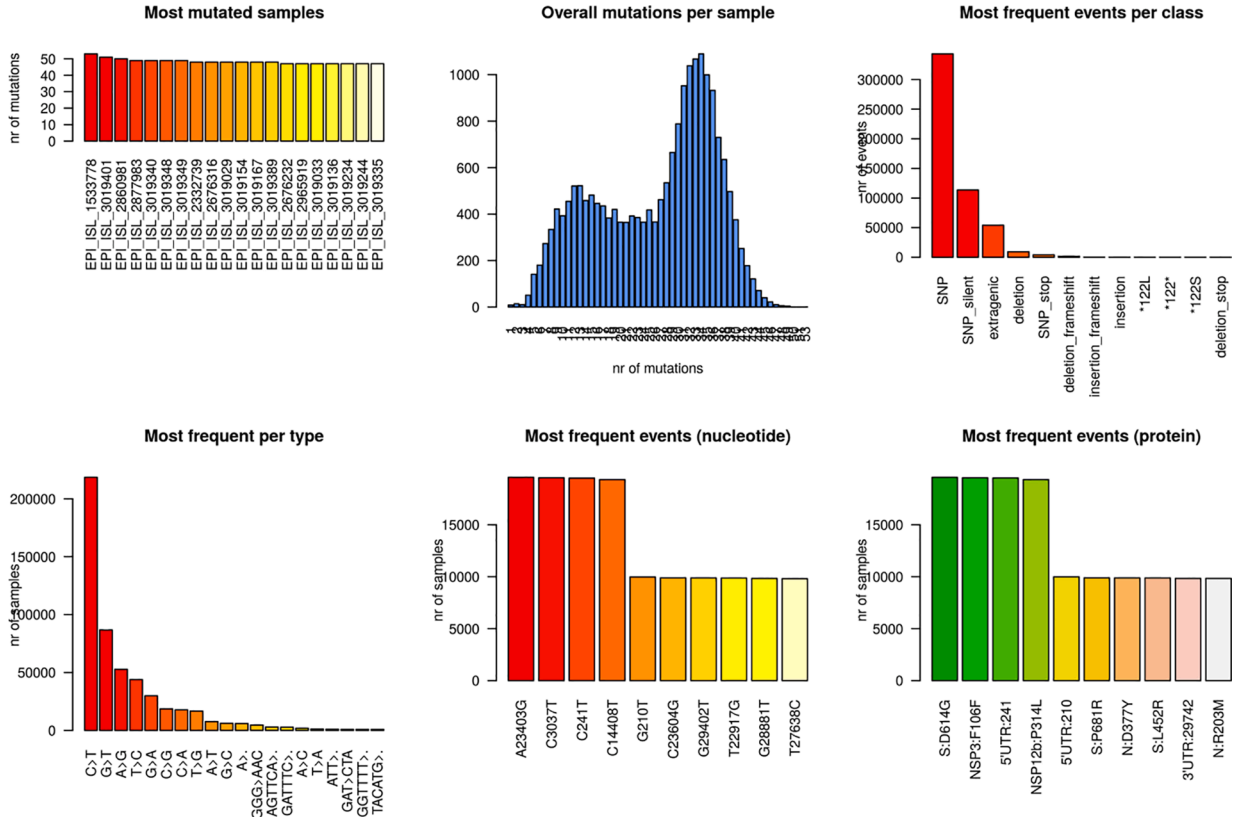

Fig. 3. Pan India mutation analysis. Six panel image displays the most mutated samples, overall mutations per samples, most frequent events per class of mutation category, changes of nucleotide per type, nucleotide wise most frequent events and protein level most frequent events for 20,086 genomes used in the study.

events in terms of nucleotide and protein. Frequency and rate of mutation per sample was also obtained for all the samples and across each phase. Overall number of mutations, coordinates of mutations with respect to the reference strain were also calculated using same R script.

Supplementary Fig. 1: (A). Total number of cases reported state wise in India until 26th July 2021. B). Genome sequence data available from each state from India until 26th July 2021.

Financial support and sponsorship

Nil.

Data availability

All the metadata files generated in this study can be accessed through https://figshare.com/s/0a81433867e6e6df2cec.

CRediT authorship contribution statement

Sanjeet Kumar: Data curation, Formal analysis, Conceptualization, Writing – original draft. **Kanika Bansal:** Data curation, Formal analysis, Conceptualization, Writing – original draft.

CRediT authorship contribution statement

Sanjeet Kumar: Data curation, Formal analysis, Conceptualization, Writing – original draft. **Kanika Bansal:** Data curation, Formal analysis, Conceptualization, Writing – original draft.

Declaration of Competing Interest

The authors declare no competing interests.

Acknowledgment

Authors whole heartedly acknowledge motivation and advice from Dr. Prabhu B. Patil – CSIR- Institute of Microbial Technology. We do gratefully acknowledge GISAID for sharing the genomic sequences in public domain and several contributors of SARS-CoV-2 genomic data. We would also like to convey our acknowledgement to Government of India for enriching the genomic resources through initiatives like INSACOG. Further, authors would like to acknowledge Mr. J. B. Bansal for extensively proofreading the manuscript and Ms. Anu Singh for tea time chit-chat over SARS-CoV-2.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.virusres.2021.198642.

References

Memish, Z.A., et al., 2014. Human infection with MERS coronavirus after exposure to infected camels, Saudi Arabia, 2013. Emerg. Infect. Dis. 20 (6), 1012.
Peiris, J.S., Guan, Y., Yuen, K.Y., 2004. Severe acute respiratory syndrome. Nat. Med. 10 (12), S88–S97.
Chen, L., et al., 2020. RNA based mNGS approach identifies a novel human coronavirus from two individual pneumonia cases in 2019 Wuhan outbreak. Emerg. Microbes Infect. 9 (1), 313–319.
Wu, F., et al., 2020. A new coronavirus associated with human respiratory disease in China. Nature 579 (7798), 265–269.
Lu, H., Stratton, C.W., Tang, Y.W., 2020a. Outbreak of pneumonia of unknown etiology in Wuhan, China: the mystery and the miracle. J. Med. Virol. 92 (4), 401.
Bogoch, I.I., et al., 2020. Pneumonia of unknown aetiology in Wuhan, China: potential for international spread via commercial air travel. J. Travel Med. 27 (2), taaa008.
Gorbalenya, A.E., et al., 2020a. Coronaviridae Study Group of the International Committee on Taxonomy of Viruses. The species Severe acute respiratory syndrome-related coronavirus: classifying 2019-nCoV and naming it SARS-CoV-2. Nat Microbiol 5, 536–544. https://doi.org/10.1038/s41564-020-0695-z.
Woo, P.C., et al., 2010. Coronavirus genomics and bioinformatics analysis. Viruses 2 (8), 1804–1820.
Lu, R., et al., 2020b. Genomic characterisation and epidemiology of 2019 novel coronavirus: implications for virus origins and receptor binding. Lancet 395 (10224), 565–574.
Worldometer, D., 2020. COVID-19 Coronavirus Pandemic. World Health Organization. www.worldometers. info.
Andrews, M., et al., 2020. First confirmed case of COVID-19 infection in India: a case report. Indian J. Med. Res. 151 (5), 490.
Mitra, P., Misra, S., Sharma, P., 2020. COVID-19 Pandemic in India: What Lies Ahead. Indian J. Clin. Biochem. 35, 257–259. https://doi.org/10.1007/s12291-020-00886-63.
Maitra, A., et al., 2020. PAN-INDIA 1000 SARS-CoV-2 RNA genome sequencing reveals important insights into the outbreak. BioRxiv. 2020.08.03.233718.
Gorbalenya, A.E.B., Susan, C., Baric, R., Groot, R.J.D., C, D., Gulyaeva, A.A., Haagmans, B.L., Lauber, C., Leontovich, A.M., Neuman, B.W., Dmitry, P., Stanley, P., Leo, P., Dmitry, S., Igor, S., Solá, A., Isabel, G., John, Z., 2020b. An integrated national scale SARS-CoV-2 genomic surveillance network. Lancet Microbe 1 (3), e99–e100. https://doi.org/10.1016/S2666-5247(20)30054-9.
Sallard, E., Halloy, J., Casane, D., Decroly, E., van Helden, J., 2021. Tracing the origins of SARS-COV-2 in coronavirus phylogenies: a review. Environ. Chem. 1–17.
Salyer, S.J., et al., 2021. The first and second waves of the COVID-19 pandemic in Africa: a cross-sectional study. Lancet 397 (10281), 1265–1275.
Alai, S., et al., 2021. Pan-India novel coronavirus SARS-CoV-2 genomics and global diversity analysis in spike protein. Heliyon 7 (3), e06564.
Rambaut, A., et al., 2020. A dynamic nomenclature proposal for SARS-CoV-2 lineages to assist genomic epidemiology. Nat. Microbiol. 5 (11), 1403–1407.
Bajaj, A., Purohit, H.J., 2020. Understanding SARS-CoV-2: genetic diversity, transmission and cure in human. Indian J. Microbiol. 60 (3), 398–401.
Phan, T., 2020. Genetic diversity and evolution of SARS-CoV-2. Infect. Genet. Evol. 81, 104260.
Helmy, Y.A., et al., 2020. The COVID-19 pandemic: a comprehensive review of taxonomy, genetics, epidemiology, diagnosis, treatment, and control. J. Clin. Med. 9 (4), 1225.
Yadav, P.D., et al., 2021. An epidemiological analysis of SARS-CoV-2 genomic sequences from different regions of India. Viruses 13 (5), 925.
Naushin, S., et al., 2021. Insights from a Pan India sero-epidemiological survey (phenome-India cohort) for SARS-CoV2. Elife 10, e66537.
Shishir, T.A., Naser, I.B., Faruque, S.M., 2021. In silico comparative genomics of SARS-CoV-2 to determine the source and diversity of the pathogen in Bangladesh. PLoS ONE 16 (1), e0245584.
Plante, J.A., et al., 2021. Spike mutation D614G alters SARS-CoV-2 fitness. Nature 592 (7852), 116–121.
Laamarti, M., et al., 2020. Large scale genomic analysis of 3067 SARS-CoV-2 genomes reveals a clonal geo-distribution and a rich genetic variations of hotspots mutations. PLoS ONE 15 (11), e0240345.
Casadevall, A., Weiss, S.R., Imperiale, M.J., 2021. Can science help resolve the controversy on the origins of the SARS-CoV-2 pandemic? Am. Soc. Microbiol., e01948-21
P.D. Thacker J.b., The COVID-19 lab leak hypothesis: did the media fall victim to a misinformation campaign? 2021. 374.
K. Bansal and P.B. Patil, J.b., Codon pattern reveals SARS-CoV-2 to be a monomorphic strain that emerged through recombination of replicase and envelope alleles of bat and pangolin origin. 2020.
Nakamura, T., et al., 2018. Parallelization of MAFFT for large-scale multiple sequence alignments. Bioinformatics 34 (14), 2490–2492.
Price, M.N., Dehal, P.S., Arkin, A.P., 2010. FastTree 2–approximately maximum-likelihood trees for large alignments. PLoS ONE 5 (3), e9490.
Letunic, I., Bork, P., 2019. Interactive tree of life (iTOL) v4: recent updates and new developments. Nucleic Acids Res. 47 (W1), W256–W259.
Delcher, A.L., et al., 2002. Fast algorithms for large-scale genome alignment and comparison. Nucleic Acids Res. 30 (11), 2478–2483.
Mercatelli, D., Giorgi, F.M., 2020. Geographic and genomic distribution of SARS-CoV-2 mutations. Front. Microbiol. 11, 1800.