bioRxiv preprint doi: https://doi.org/10.1101/2020.08.24.265827; this version posted August 25, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Unique mutational changes in SARS-CoV2 genome of different state of India

Priti Prasad[1,2*], Shantanu Prakash[3*], Kishan Sahu[1,2], Babita Singh[1], Suruchi Shukla[3], Hricha Mishra[3], Danish Nasar Khan[3] , Om Prakash[3] ,MLB Bhatt[3], SK Barik[1], Mehar H.Asif[1,2] ,Samir V. Sawant[1,2],Amita Jain[3#], Sumit Kr. Bag[1,2#]

***Equal Contribution**

[1]CSIR-National Botanical Research Institute, 436, Rana Pratap Marg, Lucknow - 226001, India

[2]Acadaemy of Scientific and Innovative Industrial Research (AcSIR), Ghaziabad -201002, India

[3]King Georges Medical University, Lucknow-226001, India

#Corresponding author

Amita Jain

Professor and Head, Department of Microbiology

King Georges Medical University

Lucknow-226001, India

Email: amita602002@yahoo.com

Sumit Kr. Bag

Principal Scientist, Computational Biology

CSIR-National Botanical Research Institute

Lucknow-226001, India

Email: sumit.bag@nbri.res.in

bioRxiv preprint doi: https://doi.org/10.1101/2020.08.24.265827; this version posted August 25, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Abstract

COVID-19 leads to a global emergency that causes more than 7 million causalities until mid-August throughout the world. In India alone, 2 million confirmed cases were reported that increased abruptly day by day with the lowest fatality rate. The availability of a large number of Severe Acute Respiratory Syndrome Coronavirus-2 (SARS-CoV2) genome in the public domain provides a great opportunity to study mutational changes in COVID-19 genomes in Indian populations. In this study, we sequenced the genomes of SARS-CoV2 viruses isolated from 47 individuals from 13 districts of Uttar Pradesh (UP), the largest state of India using Third Generation Sequencing Technology. We further did the phylogenetic clustering of UP state of Indian SARS-CoV2 genomes revealed a perceivable point that no UP samples were aligned on the USA defined clade where the fatality rate is high. We also identified 56 distinctive Single Nucleotide Polymorphism variations in UP state that majorly clustered into two groups which shows the deleterious effects on the genome. Additionally, we conducted the mutation analysis of the 2323 SARS-CoV2 genome of different states of India from the Global Initiative on Sharing All Influenza Data (GISAID) where we find ~80% unique mutations rate in each sample of the Indian population. Thus, this is the first extensive mutational study of the largest state of Indian populations in which we report the novel deleterious SNPs in virus genome along with the other states which access the less infectious form of SARS-CoV2 genome through synonymous to nonsynonymous mutation variation.

Keywords

Third Generation Sequencing, COVID-19, India, Uttar Pradesh, SNP, Unique

bioRxiv preprint doi: https://doi.org/10.1101/2020.08.24.265827; this version posted August 25, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Introduction

The Coronavirus disease (COVID-19) emerged out as a global pandemic[1] and has become a center point for many researchers to identify the potential region/s of its pathogenicity. The causative agent, SARS-CoV2 belongs to the Coronaviridae family comprises of single-stranded, positive-sense RNA with a genome size of approximately 30 kb [2]. SARS-CoV2 genome consists of 29 open reading frames (ORFs), among which four are structural proteins (envelope protein, membrane protein, nucleocapsid protein, and Spike protein), some accessory protein translated by ORF3a, ORF6, ORF7a, ORF7b, ORF8, and ORF10 and 16 nonstructural proteins (nsps) enclosed the complete genome [3].

At the end of December 2019, China reported its first case in Wuhan city, after then it grew exponentially through human to human transmission. During its spreading, it evolved continuously which causes changes in fatality rates in different countries [4][5]. As of 20th August, approximately 23 million confirmed cases were reported worldwide with ~800 thousand registered deaths. At present India is ranked the 3rd position in the total number of cases, however, the fatality rate is declining to 1.9 %, which is considered as the one of the lowest fatality rates [6] [7] [8]. Many reports correlate the fatality rate to the mutational changes in the virus genome in different geographical regions [9][10], [11][12] in which D614G substitution at the spike protein cause a higher mortality rate in human [4]. A study on western Indian populations linked the mutational changes to the specific age group and their mode of infection. To delineate the different mutations rate and understanding the evolution in the SARS-CoV2 genome of different continents, many genomes were sequenced by employing different techniques and are submitted at GISAID [13]. More than 84 thousands genome sequences of the full-length SARS-CoV2 were available on the GISAID platform where India has sequenced >2400 viral genomes and from where Indian specific clade was retrieved [14].

bioRxiv preprint doi: https://doi.org/10.1101/2020.08.24.265827; this version posted August 25, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

India comprises of 29 states with the 7 Union Territories in which Uttar Pradesh is ranked 5th in confirmed cases and surprisingly, only 11 sequenced samples were available on the database. Therefore in this study, we first report the detailed genome sequences of SARS-CoV2 in different districts of Uttar Pradesh (UP) through the third generation sequencing technology (Oxford Nanopore PromethION) and generate the phylogenetic clustering . We have also investigated the unique mutational features in a different state of India in comparison to other continents of the world to find out the mutation rate during the transmission. We also correlated the fatality rate of different state with their synonymous to nonsynonymous mutation ratio.

Methods

RNA Preparation and Sequencing of the viral genome

Total viral nucleic acid was extracted from the clinical sample of SARS-CoV-2 using PureLink DNA/RNA mini kit (Invitrogen, USA) as per manufacturer's instructions and Indian Council of Medical Research (ICMR) guidelines [15] . The Human RnaseP gene was tested in all samples to check the quality of samples and to validate the process of nucleic acid extraction. Real time RT-PCR assays were performed using Superscript III One-Step Real-Time PCR kit (ThermoFisher Scientific, USA) as per the protocol[16]. All samples were tested by 2 stage protocol consisting of screening and confirmation. E-gene and RnaseP were used for screening while RdRp and ORF1b were used as a confirmatory assay for SARS-CoV-2. All the samples positive for E-gene, RdRp, and ORF1ab with Ct $\leq$ 25 were further taken for whole-genome sequencing.

Virus genomes were generated by using ARTIC COVID-19 multiplex PCR primers followed by nanopore sequencing on two ONT PromethION flow cells[17]. For PromethION sequencing, libraries were prepared using the ligation sequencing kit (SQK-LSK109) and

bioRxiv preprint doi: https://doi.org/10.1101/2020.08.24.265827; this version posted August 25, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Native barcoding (EXP-NBD 104 and EXP-NBD 114) of 24 samples in a single flow cell. The selected samples were converted to first-strand cDNA using reverse transcriptase as suggested in Nanopore protocol "PCR tiling of COVID-19 virus" (Version: PTC_9096_v109_revD_06Feb2020). The multiplex PCR was performed with two pooled primer mixture as recommended and the cDNA was amplified. After 35 rounds of amplification, the PCR products were collected, purified with AMPure XP beads (Beckman coulter), and quantified using Qubit4 Fluorometer (Invitrogen). The double-stranded cDNA was further subjected to end-repair, A-tailing, and Native barcode ligation individually. The barcoded samples were pooled together and ligated with sequencing adapters followed by loading of 50ng of pooled barcoded material and sequencing on 2 PromethION flow cells having 24 samples each.

Genome Assembly -

Quality passed nanopore reads were subjected to remove the chimeric reads between the 300 to 500 base pair length according to the amplicon size. With the help of the Artic network pipeline [18], we individually assembled the high coverage (approx. 9000X) filtered long reads of each sample. During the assembly --normalize options were used to normalizes the data up to 200 bp coverage. This normalization is useful for the minimization of the computational requirements of our high coverage data. The SARS-CoV2 reference genome (MN908947.3) [19] was used for generating the consensus sequence from the assembled reads. Low complexity or primer binding regions were masked during the consensus sequence generation and variants calling to avoid the ambiguity.

Data Preparation-

For comparative and explanatory analyses of our assembled genome form Uttar Pradesh (UP) district of India, we downloaded the whole genome data of all Indian samples from the

bioRxiv preprint doi: https://doi.org/10.1101/2020.08.24.265827; this version posted August 25, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

GISAID database that was available until the 12th of August. As an outlier, and take the representation of all the continents of the world, we randomly selected 100 samples from Asia, Europe, South America, North America, Central America, and Eurasia. A total of 2923 processed samples (2323 from India and 600 from different continents) were further screened based on unrecognized nucleotides (Ns) with the full length of the genome. We kept only those samples whose unrecognized base pair number (number of Ns) is less than 1 % of the total sequence followed by the masking of the low complexity region.

Phylodynamics clustering of Indian sample –

We sequenced and assembled a total of 47 genome sequences from the Uttar Pradesh state of India followed by the phylodynamic clustering according to the nextstrain clade [20]. Whole-genome alignment has been carried out by the mafft aligner [21] by taking the Wuhan nCov2019 Hu-1(MN908947.3) sample as a reference genome. IQTREE software [22] was used for the phylogenetic tree with the GTR substitution model. The refinement of the tree was done by the --timetree parameters to adjust the branch lengths and filter out the interquartile ranges from regression. Thus, the refined time-resolved tree was further screened to derive the ancestral traits, infer nucleotide mutation as well as an amino acid variation from the root. The resulting phylodynamics tree was then used to defining the clade based on a new classification that signifies the different geographic spreading of nCoV-2019 in India. The final tree was further visualized through the nextstrain shell by using the auspice interface [23].

Divergence estimates of the newly assembled genome-

Divergence estimation of the newly assembled genome and the most recent common ancestor (tMRCA) was examined using the BEAST v1.10.4 program [24]. Through the MRCA analysis, the origin of the ancestral virus has been identified that was present in our samples.

bioRxiv preprint doi: https://doi.org/10.1101/2020.08.24.265827; this version posted August 25, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

With the strict molecular clock and the exponential growth model, we processed the Markov chain Monte Carlo algorithm on the masked alignment file for 100 million steps, where the first 30 million steps were used for burn-in to get the effective sample size. Bayesian coalescent analysis of the UP samples was conducted using the GTR substitution model with the kappa scale of 1 to 1.25. Molecular clock and root divergence time were reconfirmed with the treetime [25] using the same substitution model.

Variant identification and functional evaluation

We scrutinized the unique mutations in our assembled genome of UP district that was called during the assembly. The SNP variants between all processed samples (downloaded and assembled) and to ensure the uniqueness of variants in our assembled data, we aligned each downloaded samples separately to the reference genome (MN908947.3) by use of mafft aligner [21] and find the variation with the snp-sites program and [26] nextstrain. The resulting Variant Call Format (VCF) file was annotated further to find out the position of the mutation that affects the nCoV-2019 genome though the snpEff program [27]. Finally, manual screening of mutation has been carried out to define the unique mutation in our processed samples (2370) along with the other continents of the world. Functional validation of the emerging new mutations in UP samples was also assessed through the Sorting Intolerant from Tolerant (SIFT) [28]. SIFT evaluated the functional consequences of the variants where the SIFT score of 0.0 to 0.5 is used to explain the deleterious effect while >0.05 was interpreted as a toleratable mutation. Visualization of identified variants of Uttar Pradesh samples was processed through the R package [29].

Results and Discussion:-

Demographic Representation

bioRxiv preprint doi: https://doi.org/10.1101/2020.08.24.265827; this version posted August 25, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

In this study, a total of 47 genomes of SARS-CoV2 have been sequenced using the long reads PromethION sequencer. All the samples are from the Uttar Pradesh (UP) state, India encompassing over 13 different districts (**Fig 1a**). The samples were collected from symptomatic and asymptomatic patients in which the male-female ratio of patients was 3:1. The age of the patients varied from 2 to 86 years with an average age of 35years. The maximum number of samples was collected in May month while 2 and 13 samples were collected in March and April respectively (**Supplementary File 1**). The average coverage of all the samples was 9000X which is correlated with the ct value of the samples (**Fig. 1b**).

Assembly of the whole genome

The resulting fastq and fast5 files of the nanopore PromethION sequencer were quality checked by the artic network pipeline [18]. The amplicon reads of 300 to 700 base pair length was removed from the raw fastq file to avoid the generation of chimeric reads. Filtered fastq file and raw fast5 format files of each sample were taken as input for whole-genome assembly. The MN908947.3 genome was used as a reference to generated the consensus sequence for each sample individually. Additionally, medaka parameters were also used to identify the variants or SNP based on assembled reads during the generation of the consensus sequence. To avoid the ambiguity at the time of SNP calling, the primer binding region in the genome was masked. The resulting Single-nucleotide Polymorphism (SNPs) were considered as true SNPs in our sequenced samples data because they have high-quality read depth (~300) at each base pair. These High-quality SNP reads were used to define the unique mutation features in the SARS-CoV2 genome samples of UP state in comparison to the samples from other states of India as well as from different continents. The assembled whole-genome sequence was submitted to the Global Initiative on Sharing All Influenza Data with accession ID (EPI_ISL_516940-EPI_ISL_516986) and their metadata information is available in **Supplementary File 1**.

bioRxiv preprint doi: https://doi.org/10.1101/2020.08.24.265827; this version posted August 25, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Phylodynamic analyses of the assembled genome

Phylodynamic clustering is used for clade assignment in our assembled genome based on nextstrain where we can trace the origin of the clade concerning the specific mutation on the genomic region. A total of 47 assembled genome of SARS CoV2 from the 13 different districts of UP in India and 856 genomes from 21 different states of India (available on 10th June) was used to determine the phylodynamics clustering according to the new clade classification by taking the Wuhan SARS-CoV2 genome reference genome (EPI_ISL_406798) as an outlier.All the information regarding the samples including accession number, submitting institution name, collection date is mentioned in **Supplementary File 2**.

The augur pipeline [30] was used for the alignment of all the samples, their phylogenetic clustering, and deciphering the time resolving tree, using SARS-CoV2 (MN908947.3) genome as the root. The resulting tree was subjected to defining the potential emerging clades based on nucleotide mutation(s). Indian samples were clustered into all 5 clusters namely 19A, 19B, 20A, 20 B, 20C (**Fig. 2a**) as it is represented in the nextstrain nCoV global page and described in [31]. The two clades 19A and 19B are primitive ones that are dominated by the Asia continents. These two clades are delineated by C to T mutation at 8782 positions at ORF1ab region and T to C mutation at 28144 positions. A total of 43% of our processed sample aligned on these two clades of which samples from the UP states aligned only on clade 19A.

The clade 20A, 20B, and 20C are well known for the European and North American clades respectively, which deviated from the root (19A) in early 2020. Most of the Indian samples (around 41%) were established under 20A clade with C14408T and A23403G mutations in the orf1ab region followed by the 19A clade were 332 samples were superimposed (**Fig. 2b**).

bioRxiv preprint doi: https://doi.org/10.1101/2020.08.24.265827; this version posted August 25, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

The smallest cluster of India samples has been seen in the case of a 20C clade. The USA is the primary source for 20C clade and only 4 Indian samples (2 from Delhi and Gujarat each) were aligned. This 20C clade formed by Delhi samples (EPI_ISL_435063, EPI_ISL_435064) doesn't show any nucleotide or amino-acid mutation whereas Gujarat samples (EPI_ISL_435051, EPI_ISL_435052) recorded ORF1ab variations (mention the variations). The SARS-CoV2 genome samples of UP state doesn't align on the 19B and 20C clade which is a matter of investigation that can be solved with the increasing number of COVID genome sequences of Uttar Pradesh state with the probable mutation of clade defining mutation like 1059T for 19B and 25563T for the 20C clade.

Estimation of the date of origin

The mutation rate of 47 sequenced genome in UP state was estimated using BEAST software and by using the coalescent model as the prior of the trees [24]. The estimated rate of mutation is predicted as 3.279e-04 substitutions per site per year for the given SARS-CoV2 genomes samples. The 95% of Highest Posterior Density intervals (HPD) are in the range of 2.24e-04 to 4.459e-04, which was calculated for each sample (**Supplementary File 3**). The substitution rate is further confirmed through the treetime software [25] where root to tip regression rate is calculated as 3.078e-04 with the 0.05 correlation value (**Fig. 3**). This correlation value indicates the informative behavior of inputs for temporal information to rationalize the molecular clock approach. The root date of the SARS-CoV2 genome from UP states is evaluated as 29 December 2019 which substantiated the reported timing of the SARS-CoV2 genome in Wuhan city of China [32].

Assessment of unique Mutational features in different states of India

We identified the mutational features in the SARS-CoV2 genome of other states of India also wherein the unique mutation rate is around 80% of the total sequenced genome. The unique

bioRxiv preprint doi: https://doi.org/10.1101/2020.08.24.265827; this version posted August 25, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

mutation features hypothesize that at each transmission, the virus mutated at the higher speed with insynonymous to nonsynonymous mutation rate is around 0.5. For the isolation of unique SNPs features, we selected only those states whose SARS-CoV2 genome samples were available in more than 10 samples. Thus, total 12 states; Gujarat, Maharashtra, Telangana, Delhi, Odisha, Karnataka, West Bengal, Uttarakhand, Madhya Pradesh, Tamil Nadu, Haryana, and Uttar Pradesh were selected **(Table 1)** whose average active case is around 2 lacs according to the WHO India tally [7]. We strike off the common mutation position from different continents (Asia, Africa, Eurasia, North America, Oceania, and South America) to check the mutated version of the SARS-CoV2 genome in the Indian strain. A higher number of unique mutations were seen in Maharashtra followed by the Gujarat, Odisha, Telangana, West Bengal, and Karnataka where an abundant amount of sequenced data were also available **(Table 1)**. These unique SNPs information were annotated on the SARS-CoV2 genome and classified it into synonymous, missense, stop gain, stop loss, upstream gene variant and downstream gene variant **(Fig. 4)**. Surprisingly, Delhi has very low number of unique mutation rate (23%) and its mortality rate is comparably high (2.66) which suggest that Delhi was infected with a fatal version of the virus genome. The phylogenetic clustering by nextstrain also supplemented the information where genome from Delhi was clustered on 20C clade [20] which is specifically defined as the North American clade. On the other hand, Maharashtra and Gujarat were top listed according to the fatality rate where the unique mutation rate is ~111% and ~89% respectively**.** These two states have the equal number of missense variant which affect almost all the genomic region with the stop codon gain at the envelope protein in Maharashtra state and ORF7a in Gujarat state **(Supplementary File 4) (Fig. 4)**. Odisha state has also higher number of mutation rate with the synonymous to nonsynonymous mutation rate is below (0.39) than average (0.5) and its fatality rate is very low. Similar pattern is observed in the Telangana state as well as in

bioRxiv preprint doi: https://doi.org/10.1101/2020.08.24.265827; this version posted August 25, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Harayana state so through this we can postulated that fatality rate could be estimatd through the synonymous to the nonsynonymous ratio of unique mutational features.

Dynamic characterization of Unique SNP variation in UP state

The identification of clade is defined by at least two mutations changes from its parent clade which is ubiquitous for all SARS-CoV2 genome. Besides the mutations that are shaping the clade, we identified 56 distinctive SNPs in our UP district sequence samples, referred to as unique SNPs. Among these uniquely identified SNPs, 36 SNPs are annotated as a missense variant which causes to change in the amino acids while 2 mutations (T18126A, C18431T) show the premature stop gain in the ORF1ab regions (**Table 2, Fig 5**). The functional evaluation of the nonsynonymous mutations was investigated through Sorting Intolerant From Tolerant SIFT) prediction server which revealed the deleterious effects of 37% of total mutations. A Maximum number of deleterious mutations have been seen in the non-structural protein of ORF1ab regions while 2 stop gain mutations have also been seen in the NSP11 margin [2]. The Threonine to Isoleucine intolerable mutations in ORF1ab regions, at 6056th position, affects the NSP11 protein in the NBRI-N21 genome which might change the polymerase activity of the genomes. The different type of identified mutation in non-structural protein 11(NSP11) of orf1ab in UP states gives a clue for conducting the detailed study on it, as it is associated with the programmed cell death evasion, cell cycle and DNA replication [35][36]. Only 1 each deleterious mutation (**Table 3**) was found in ORF3a, ORF7a, and ORF8 regions respectively which functions as an accessory factor in the viral genome [37] and also assisted to virus-host interactions [38]. Accessory factor, ORF3a regulates the interferon signaling pathway and cytokines production [39] and thus is involved in the virus pathogenesis. Such kind of deleterious mutation in the accessory protein has been found in the latest evolving SARS-CoV2 genome from Indian samples [40]. A structural protein, N encoded the nucleocapsid protein shows the deleterious effects by the mutations at

bioRxiv preprint doi: https://doi.org/10.1101/2020.08.24.265827; this version posted August 25, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

D371V, K370N, D371Y, P279Q, and T362I positions in 4 UP state genomes. SARS CoV2 N proteins have a highly conserved domain [41] that plays a crucial function to complete the viral life cycle by regulating the RNA transcription, replication, and modulating the infected cell biological processes [42]. Another nonsynonymous, deleterious mutation is noticeable in NSP3 protein in the NBRI-N42 sample. This protein has papain-like viral protease activity to generate the other replicase subunits from nsp1 to nsp16 [43]. This intolerable mutation from Valine to Leucine at 1570 position might affect the functionality of the papain activity which leads to the inactivation of replicase subunits from nsp1 to nsp16.

Conclusion:

The present study is the first one to report the SARS-CoV2 genome sequence of the Uttar Pradesh state with a large scale of high coverage third-generation sequencing data. We identified 56 peculiar SNPs in our sequenced genome in which more than 40% of SNPs in the ORF1ab region showed the deleterious effects. These SNP variations in the ORF1ab region might affect the replication of the virus genome during the infection which ultimately would be the reason for the less fatality rate. The identified mutations in SARS-CoV2 genomes of 47 individuals could be used as the potential target for personalized medications or effective vaccine doses to combat the effects of the COVID-19. Additionally, the relation with the synonymous to nonsynonymous unique mutation ratio with the fatality rate could be studied further to understand the putative region/SNPs that cause fatal to the human being.

Author contributions:

Priti Prasad: Conceptualization; Data curation; Supervision; Formal analysis; Investigation; Methodology; Resources; Software; Visualization; Writing - original draft; Writing - review & editing, Shantanu Prakash: Conceptualization; Methodology; Writing - review & editing, Mehar H. Asif: Conceptualization; Supervision; Writing - original draft; Writing - review &

bioRxiv preprint doi: https://doi.org/10.1101/2020.08.24.265827; this version posted August 25, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

editing, Kishan Sahu: Resources; Writing - review & editing, Suruchi Shukla, Babita Singh, Hricha Mishra, Danish Nasar Khan, Om Prakash, MLB Bhatt: Resources, SK Barik: Conceptualization; Resources; Writing - review & editing, Samir V. Sawant: Conceptualization; Resources; Supervision; Formal analysis; Writing - review & editing, Amita Jain: Conceptualization; Resources; Writing - review & editing, Sumit Kr. Bag: Conceptualization; Resources; Supervision; Writing - review & editing

Conflicts of Interest

Authors declare no conflicts of interest

Acknowledgments

Authors acknowledge the GISAID team and all those who submitted the genome to the GISAID database without which it's impossible to conduct the research. PP and KS acknowledged the University of Grant Commission for providing the Senior Research fellowship. Institute manuscript number is CSIR-NBRI_MS/2020/08/04.

Figure legends:

Fig. 1

a) The detailed map of the SARS-CoV2 genome sequence of different districts of the Uttar-Pradesh state of India. The scale denoted the number of genome sequencing of the virus genome. Grey color means no genome was sequenced while the darker blue color represents the highest number of genome sequencing. b) The Co-linear plot of the Ct value of 47 positive COVID-19 patients to the coverage of the whole genome sequence long-read data. Each sample was represented through the blue dots where higher Ct value shows high coverage data.

bioRxiv preprint doi: https://doi.org/10.1101/2020.08.24.265827; this version posted August 25, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Fig. 2

a) The phylogenetic clustering of 889 samples of India into 5 different clades according to the nextstrain based mutation. The 19A and 19B clades are the primitive clades that were sequenced in early 2020; whereas 20A, 20B and 20C clades are emerging clades. The reference genome of the Wuhan sample was implemented in the clustering trees which aligned in 19A clade before the starting of the Jan-2020. b) The distribution of genome sequenced in different clades was visualized through the piechart. In the 20A clade, 366 genomes were clustered followed by the 19A clade. The lowest number of clustering has been seen in the 20C clade which is well known for the USA defining clade.

Fig. 3

The root-to-tip regression rate of sequenced SARS-CoV2 genome of Uttar Pradesh state of India to investigate the origin of infection. X-axis describes the timing of the sample collection of COVID-19 infected persons. Y-axis describes the regression rate in comparison to the sample collection date. The origin of infection in Uttar Pradesh samples is estimated by 2019.5.

Fig. 4

Unique mutation features in 12 different states of India whose SARS-CoV2 genome sequence is present in large amounts in GISAID. The pie chart shows the different types of variants that were annotated on the SARS-CoV2 genome with their unique mutation variation per high-quality genome sequenced were available on the public domain. In the pie chart of Uttar Pradesh, the publiclly available genome and the assembled genome in this study were mentioned separately.

bioRxiv preprint doi: https://doi.org/10.1101/2020.08.24.265827; this version posted August 25, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Fig. 5

The nonsynonymous mutation of the SARS-CoV2 genome of the Uttar Pradesh 47 samples. The scale bar shows the length of the genome on which the dashed line with a brown color number represents the mutational points that putatively altered the functionality of the genome. The circled nucleotides are the mutated nucleotide bases, in which black color bases are the missense variation while red color shows the stop gain at the genomic regions.

Table 1

The statistics of the different selected states of India along with their unique mutation rate and the fatality rate for the respective states according to the mid-August, 2020 COVID cases. The sequences data is downloaded from GISAID, available till 12th of August.

Table 2

Detail of Nonsynonymous SNP mutation of all 47 sequenced UP samples with their genomic and amino acid position on the genome. The different colored Gene color shows different genomic regions while stop codon affecting mutation wee highlighted in green in color.

Table 3

The Putative nonsynonymous mutation with their SIFT score. The SIFT score value less than 0.05 is depicted as a deleterious mutation that is highlighted in yellow whereas unhighlighted SIFT score value for the particular SNPs is predicted as a tolerable mutation.

References -

[1] P. Yang and X. Wang, "COVID-19: a new challenge for human beings," *Cell. Mol. Immunol.*, vol. 17, no. 5, pp. 555–557, 2020, doi: 10.1038/s41423-020-0407-x.

[2] J. F. W. Chan *et al.*, "Genomic characterization of the 2019 novel human-pathogenic coronavirus isolated from a patient with atypical pneumonia after visiting Wuhan," *Emerg. Microbes Infect.*, vol. 9, no. 1, pp. 221–236, 2020, doi: 10.1080/22221751.2020.1719902.

bioRxiv preprint doi: https://doi.org/10.1101/2020.08.24.265827; this version posted August 25, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

[3] R. A. Khailany, M. Safdar, and M. Ozaslan, "Genomic characterization of a novel SARS-CoV-2," *Gene Reports*, vol. 19, no. April, p. 100682, 2020, doi: 10.1016/j.genrep.2020.100682.

[4] M. Becerra-Flores and T. Cardozo, "SARS-CoV-2 viral spike G614 mutation exhibits higher case fatality rate," *Int. J. Clin. Pract.*, no. May, pp. 4–7, 2020, doi: 10.1111/ijcp.13525.

[5] B. Korber *et al.*, "Spike mutation pipeline reveals the emergence of a more transmissible form of SARS-CoV-2," *bioRxiv*, p. 2020.04.29.069054, 2020, doi: 10.1101/2020.04.29.069054.

[6] H. Ritchie *et al.*, "India : Coronavirus Pandemic Country Pro le," pp. 1–24, 2020.

[7] World Health Organization, "Coronavirus disease," *Coronavirus Dis. Situat. Rep. – 119*, vol. 2019, no. May, p. 2633, 2020, doi: 10.1001/jama.2020.2633.

[8] F. Welfare, "Latest Updates 27.04.2020," pp. 3–5, 2020.

[9] D. Mercatelli and F. M. Giorgi, "Geographic and Genomic Distribution of SARS-CoV-2 Mutations," *Front. Microbiol.*, vol. 11, no. July, pp. 1–13, 2020, doi: 10.3389/fmicb.2020.01800.

[10] P. Kumar *et al.*, "Integrated genomic view of SARS-CoV-2 in India," *Wellcome Open Res.*, vol. 5, p. 184, 2020, doi: 10.12688/wellcomeopenres.16119.1.

[11] A. Maitra *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.*, vol. 45, no. 1, pp. 1–18, 2020, doi: 10.1007/s12038-020-00046-1.

[12] R. C. Bhoyar *et al.*, "High throughput detection and genetic epidemiology of SARS-CoV-2 using COVIDSeq next generation sequencing," *bioRxiv*, p. 2020.08.10.242677, 2020, doi: 10.1101/2020.08.10.242677.

[13] "Pandemic coronavirus causing COVID-19," p. 2020, 2020.

[14] S. Banu *et al.*, "A distinct phylogenetic cluster of Indian SARS-CoV-2 isolates," *bioRxiv*, p. 2020.05.31.126136, 2020, doi: 10.1101/2020.05.31.126136.

[15] "ICMR_Guidelines.pdf." .

[16] V. M. Corman *et al.*, "Detection of 2019 novel coronavirus (2019-nCoV) by real-time RT-PCR," *Eurosurveillance*, vol. 25, no. 3, pp. 1–8, 2020, doi: 10.2807/1560-7917.ES.2020.25.3.2000045.

[17] K. Itokawa, T. Sekizuka, M. Hashino, R. Tanaka, and M. Kuroda, "A proposal of alternative primers for the ARTIC Network's multiplex PCR to improve coverage of SARS-CoV-2 genome sequencing," *bioRxiv*, no. December 2019, p. 2020.03.10.985150, 2020, doi: 10.1101/2020.03.10.985150.

[18] N. Loman, W. Rowe, and A. Rambaut, "nCoV-2019 novel coronavirus bioinformatics protocol," pp. 1–5, 2020, [Online]. Available: https://artic.network/ncov-2019/ncov2019-bioinformatics-sop.html.

[19] C. Wang *et al.*, "The establishment of reference sequence for SARS-CoV-2 and variation analysis," *J. Med. Virol.*, vol. 92, no. 6, pp. 667–674, 2020, doi: 10.1002/jmv.25762.

[20] Nextstrain, "Genomic epidemiology of novel coronavirus - Global subsampling," *Nextstrain.Org*, p. 1, 2020.

[21] K. Katoh, K. Misawa, K. I. Kuma, and T. Miyata, "MAFFT: A novel method for rapid multiple sequence alignment based on fast Fourier transform," *Nucleic Acids Res.*, vol. 30, no. 14, pp. 3059–3066, 2002, doi: 10.1093/nar/gkf436.

[22] L. T. Nguyen, H. A. Schmidt, A. Von Haeseler, and B. Q. Minh, "IQ-TREE: A fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies," *Mol. Biol. Evol.*, vol. 32, no. 1, pp. 268–274, 2015, doi: 10.1093/molbev/msu300.

[23] "Auspice : An Open-source Interactive Tool for Visualising Phylogenomic Data," pp. 7–9, 2020.

bioRxiv preprint doi: https://doi.org/10.1101/2020.08.24.265827; this version posted August 25, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

[24] A. J. Drummond, A. Rambaut, and R. R. Bouckaert, "Divergence Dating Tutorial with BEAST 2.0," *Beast 2.0*, p. 19, 2013.

[25] P. Sagulenko and R. Neher, "TreeTime Documentation," 2020.

[26] A. J. Page *et al.*, "SNP-sites: rapid efficient extraction of SNPs from multi-FASTA alignments," *Microb. genomics*, vol. 2, no. 4, p. e000056, 2016, doi: 10.1099/mgen.0.000056.

[27] P. Cingolani *et al.*, "A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3," *Fly (Austin).*, vol. 6, no. 2, pp. 80–92, 2012, doi: 10.4161/fly.19695.

[28] N. L. Sim, P. Kumar, J. Hu, S. Henikoff, G. Schneider, and P. C. Ng, "SIFT web server: Predicting effects of amino acid substitutions on proteins," *Nucleic Acids Res.*, vol. 40, no. W1, pp. 452–457, 2012, doi: 10.1093/nar/gks539.

[29] R. C. Team, "The R Project for Statistical Computing," *Http://Www.R-Project.Org/*, vol. 3, pp. 1–12, 2013.

[30] C. Issues, A. Projects, and W. Secu, "About Nextstrain About Augur Quickstart," pp. 1–4, 2020.

[31] E. B. Hodcroft, J. Hadfield, R. A. Neher, and T. Bedford, "Year-letter genetic clade naming for SARS-CoV-2 on nextstrain.org," *Nextstrain*, p. 2020, 2020, [Online]. Available: https://virological.org/t/year-letter-genetic-clade-naming-for-sars-cov-2-on-nextstain-org/498.

[32] K. G. Andersen, A. Rambaut, W. I. Lipkin, E. C. Holmes, and R. F. Garry, "The proximal origin of SARS-CoV-2," *Nat. Med.*, vol. 26, no. 4, pp. 450–452, 2020, doi: 10.1038/s41591-020-0820-9.

[33] D. E. Gordon *et al.*, "A SARS-CoV-2 protein interaction map reveals targets for drug repurposing," *Nature*, vol. 583, no. July, 2020, doi: 10.1038/s41586-020-2286-9.

[34] A. Pekosz, S. R. Schaecher, M. S. Diamond, D. H. Fremont, A. C. Sims, and R. S. Baric, "Structure, expression, and intracellular localization of the SARS-CoV accessory proteins 7a and 7b," *Adv. Exp. Med. Biol.*, vol. 581, pp. 115–120, 2006, doi: 10.1007/978-0-387-33012-9_20.

[35] Q. He, Y. Li, L. Zhou, X. Ge, X. Guo, and H. Yang, "Both Nsp1β and Nsp11 are responsible for differential TNF-α production induced by porcine reproductive and respiratory syndrome virus strains with different pathogenicity in vitro," *Virus Res.*, vol. 201, pp. 32–40, 2015, doi: 10.1016/j.virusres.2015.02.014.

[36] Y. Sun, D. Li, S. Giri, S. G. Prasanth, and D. Yoo, "Differential host cell gene expression and regulation of cell cycle progression by nonstructural protein 11 of porcine reproductive and respiratory syndrome virus," *Biomed Res. Int.*, vol. 2014, 2014, doi: 10.1155/2014/430508.

[37] L. Mousavizadeh and S. Ghasemi, "Genotype and phenotype of COVID-19: Their roles in pathogenesis," *J. Microbiol. Immunol. Infect.*, no. xxxx, pp. 0–4, 2020, doi: 10.1016/j.jmii.2020.03.022.

[38] D. X. Liu, T. S. Fung, K. K. L. Chong, A. Shukla, and R. Hilgenfeld, "Accessory proteins of SARS-CoV and other coronaviruses," *Antiviral Res.*, vol. 109, no. 1, pp. 97–109, 2014, doi: 10.1016/j.antiviral.2014.06.013.

[39] M. Frieman and R. Baric, "Mechanisms of Severe Acute Respiratory Syndrome Pathogenesis and Innate Immunomodulation," *Microbiol. Mol. Biol. Rev.*, vol. 72, no. 4, pp. 672–685, 2008, doi: 10.1128/mmbr.00015-08.

[40] D. S. S. Hassan, P. P. Choudhury, B. Roy, and S. S. Jana, "Missense Mutations in SARS-CoV2 Genomes from Indian Patients," *SSRN Electron. J.*, 2020, doi: 10.2139/ssrn.3609566.

[41] R. McBride, M. van Zyl, and B. C. Fielding, "The coronavirus nucleocapsid is a multifunctional protein," *Viruses*, vol. 6, no. 8, pp. 2991–3018, 2014, doi: 10.3390/v6082991.

[42] Y. Cong *et al.*, "Nucleocapsid Protein Recruitment to Replication-Transcription Complexes Plays a Crucial Role in Coronaviral Life Cycle," *J. Virol.*, vol. 94, no. 4, pp. 1–21, 2019, doi:

bioRxiv preprint doi: https://doi.org/10.1101/2020.08.24.265827; this version posted August 25, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

10.1128/jvi.01925-19.

[43] I. Imbert, E. J. Snijder, M. Dimitrova, J. C. Guillemot, P. Lécine, and B. Canard, "The SARS-Coronavirus PLnc domain of nsp3 as a replication/transcription scaffolding protein," *Virus Res.*, vol. 133, no. 2, pp. 136–148, 2008, doi: 10.1016/j.virusres.2007.11.017.

Fig. 1a

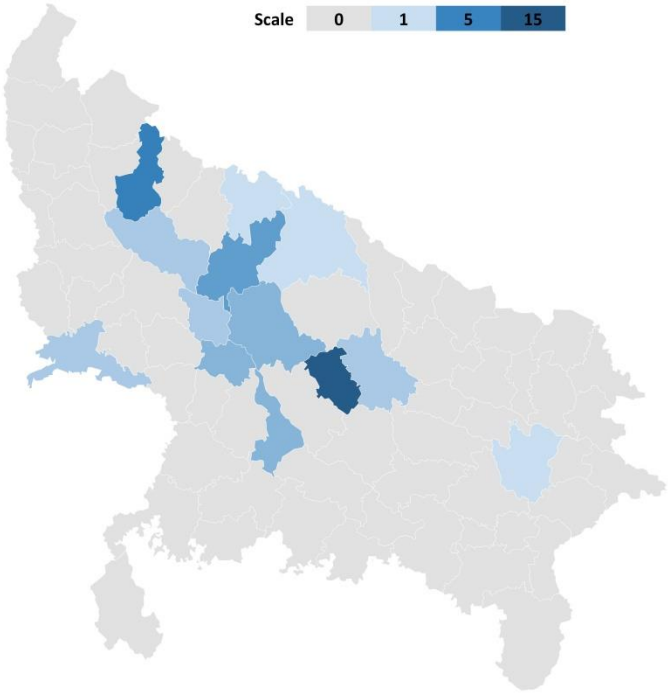
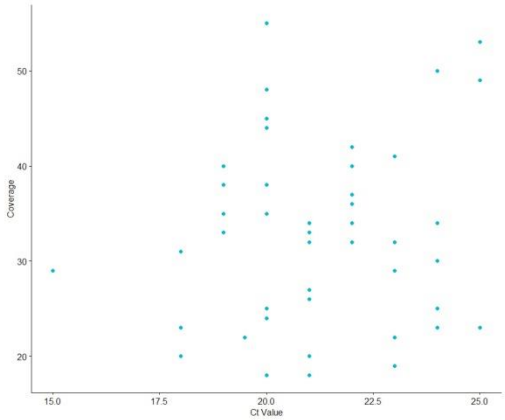


Fig. 1b



bioRxiv preprint doi: https://doi.org/10.1101/2020.08.24.265827; this version posted August 25, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Fig. 2a

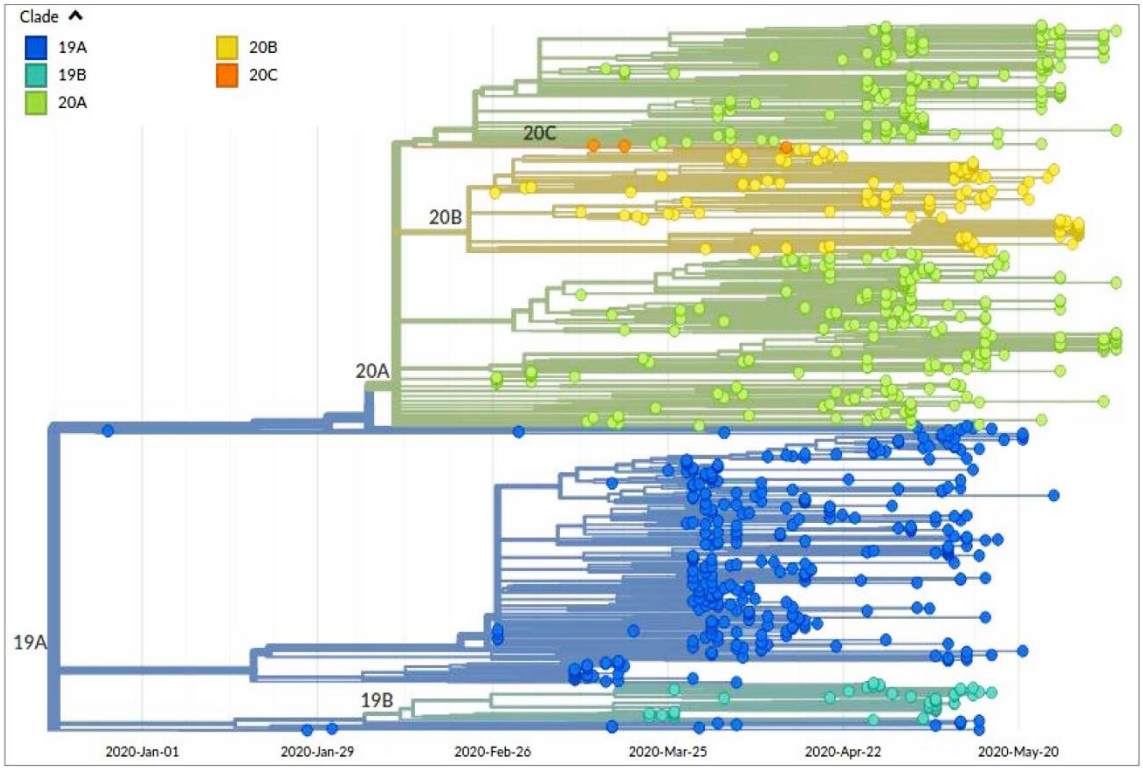
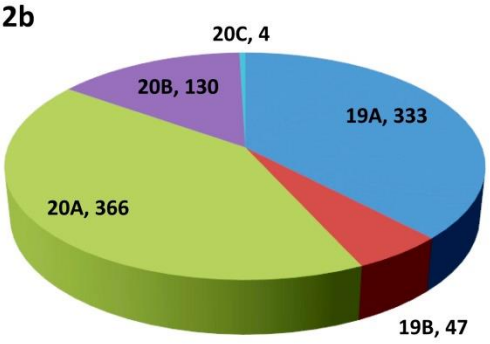
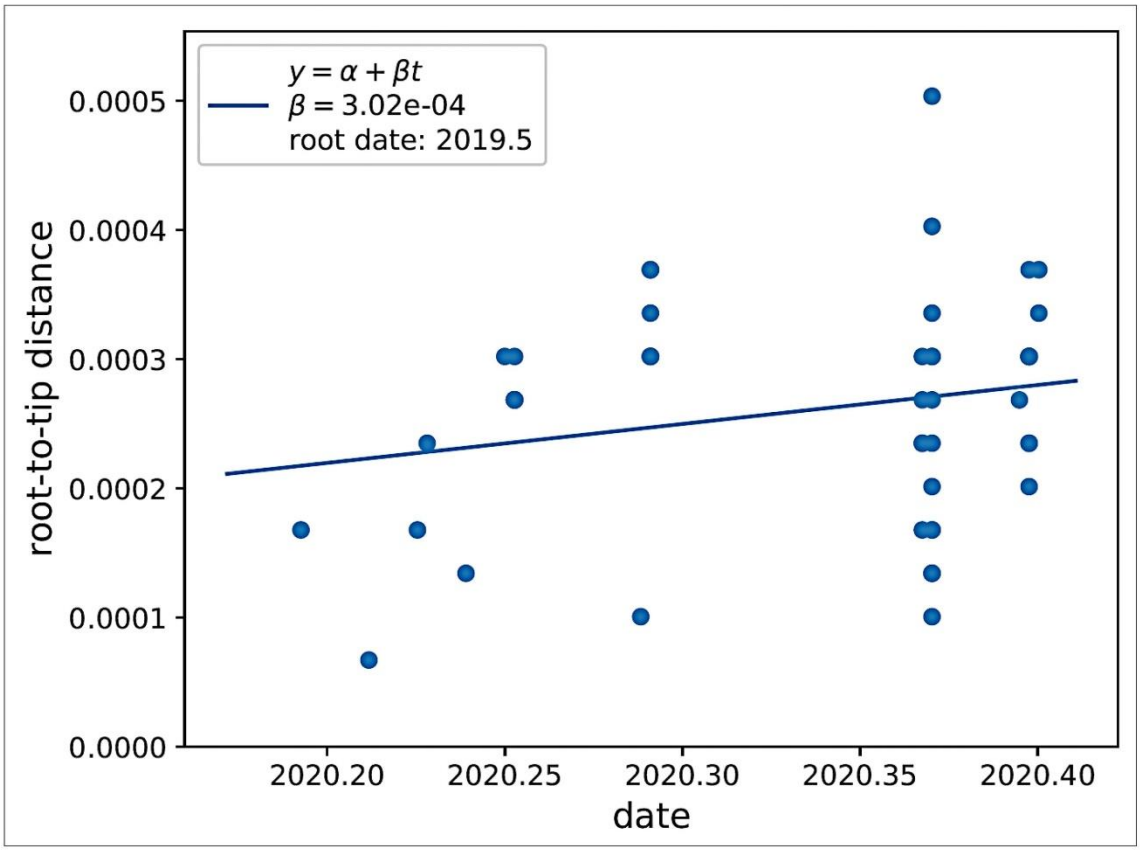


Fig. 2b



bioRxiv preprint doi: https://doi.org/10.1101/2020.08.24.265827; this version posted August 25, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Fig. 3



491

492

bioRxiv preprint doi: https://doi.org/10.1101/2020.08.24.265827; this version posted August 25, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Fig. 4

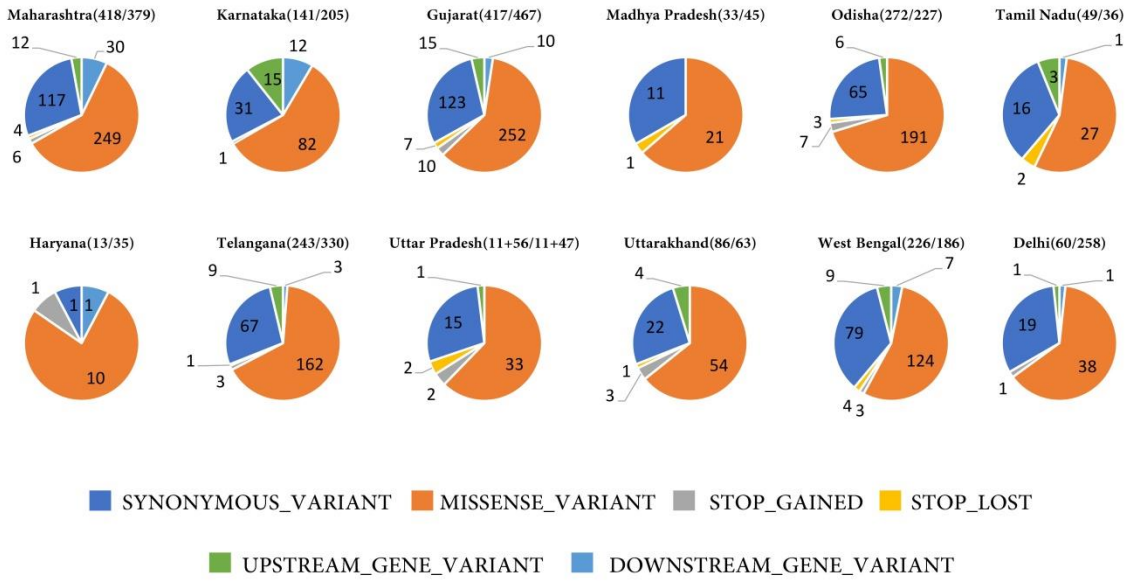
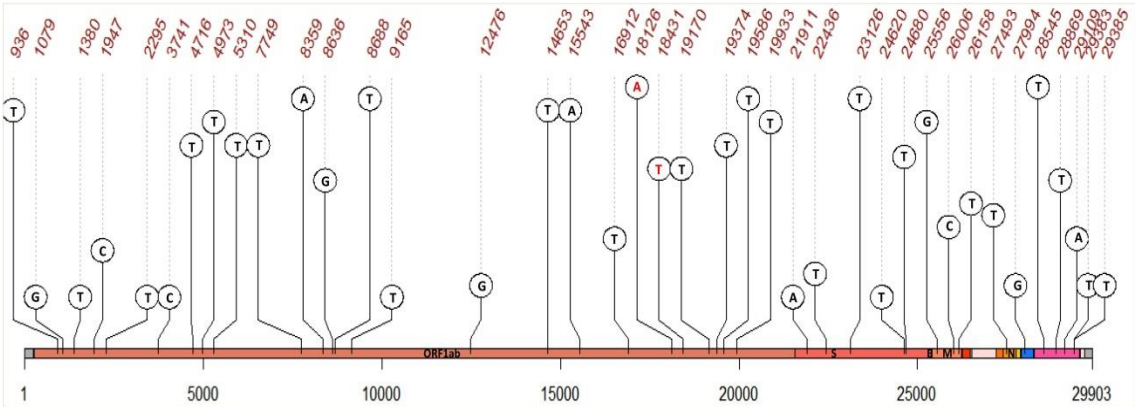


Fig. 5



bioRxiv preprint doi: https://doi.org/10.1101/2020.08.24.265827; this version posted August 25, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Table -1

State	Genome Sequenced	Unique mutation	Mutation rate	Fatality rate
Maharashtra	379	422	111.3456464	3.261071861
Gujarat	467	417	89.29336188	3.336060568
Odisha	227	273	120.2643172	0.5883102
Telangana	330	243	73.63636364	0.717308726
West Bengal	186	227	122.0430108	2.011953626
Karnataka	205	141	68.7804878	1.68566019
Uttarakhand	63	87	138.0952381	1.322401481
Delhi	258	60	23.25581395	2.663099352
Tamil Nadu	36	49	136.1111111	1.717780091
Madhya Pradesh	45	33	73.33333333	2.431820632
Haryana	35	13	37.14285714	1.10874122
Uttar Pradesh	11 + 47	11 + 56	96.55172414	1.558198114

Table – 2

Sample_ID	POS	REF	ALT	Gene	Mutation	Type of mutation
NBRI-N24	936	C	T	ORF1ab	Thr224Ile	missense_variant
NBRI-N25	1079	A	G	ORF1ab	Asn272Asp	missense_variant
NBRI-N38	1380	C	T	ORF1ab	Ala372Val	missense_variant
NBRI-N48	1947	T	C	ORF1ab	Val561Ala	missense_variant
NBRI-N51	2295	C	T	ORF1ab	Thr677Ile	missense_variant
NBRI-N22	3741	T	C	ORF1ab	Ile1159Thr	missense_variant
NBRI-N55	4716	C	T	ORF1ab	Thr1484Ile	missense_variant
NBRI-N42	4973	G	T	ORF1ab	Val1570Leu	missense_variant
NBRI-N47	5310	C	T	ORF1ab	Thr1682Ile	missense_variant
NBRI-15	7749	C	T	ORF1ab	Thr2495Ile	missense_variant
NBRI-N39	8359	T	A	ORF1ab	His2698Gln	missense_variant
NBRI-N24	8636	A	G	ORF1ab	Thr2791Ala	missense_variant
NBRI-N26	8688	G	T	ORF1ab	Gly2808Val	missense_variant
NBRI-N42	9165	C	T	ORF1ab	Thr2967Ile	missense_variant
NBRI-N44	12476	A	G	ORF1ab	Met4071Val	missense_variant
NBRI-N30	14653	G	T	ORF1ab	Met4796Ile	missense_variant
NBRI-N48	15543	G	A	ORF1ab	Arg5093Gln	missense_variant
NBRI-N21	16912	G	T	ORF1ab	Leu5549Phe	missense_variant

bioRxiv preprint doi: https://doi.org/10.1101/2020.08.24.265827; this version posted August 25, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

NBRI-N49	19170	C	T	ORF1ab	Ser6302Leu	missense_variant
NBRI-N50	19374	C	T	ORF1ab	Ser6370Phe	missense_variant
NBRI-N21	19586	C	T	ORF1ab	Leu6441Phe	missense_variant
NBRI-N58	21911	C	A	S	Leu117Ile	missense_variant
NBRI-N32	22436	G	T	S	Ala292Ser	missense_variant
NBRI-N49	23126	G	T	S	Ala522Ser	missense_variant
NBRI-N48	24620	G	T	S	Ala1020Ser	missense_variant
NBRI-N35	24680	G	T	S	Val1040Phe	missense_variant
NBRI-N48	25556	T	G	ORF3a	Val55Gly	missense_variant
NBRI-N39	26006	G	C	ORF3a	Ser205Thr	missense_variant
NBRI-N44	26158	G	T	ORF3a	Val256Phe	missense_variant
NBRI-N24	27493	C	T	ORF7a	Pro34Ser	missense_variant
NBRI-N34	27994	A	G	ORF8	Asp34Gly	missense_variant
NBRI-N20	28545	C	T	N	Thr91Ile	missense_variant
NBRI-N50	28869	C	T	N	Pro199Leu	missense_variant
NBRI-N57	29109	C	A	N	Pro279Gln	missense_variant
NBRI-N29	29383	G	T	N	Lys370Asn	missense_variant
NBRI-N34	29385	A	T	N	Asp371Val	missense_variant
NBRI-N20	18126	T	A	ORF1ab	Leu5954*	stop_gained
NBRI-N50	18431	C	T	ORF1ab	Gln6056*	stop_gained
NBRI-N28	19933	A	T	ORF1ab	Ter6556Cysext*?	stop_lost

Table -3

Sample_ID	Gene	Mutation	SIFT Score
NBRI-N34	ORF1ab	P5496S	0.81
NBRI-N57		T6056I	0.03
NBRI-N15		T2495I	0.55
NBRI-N16		T2016K	0.59
NBRI-N21		V4691F	0.01
		V5550L	0.02
		V5894I	0.59
		T6441I	0.85
NBRI-N22		S5483F	0.02
NBRI-N23		T5300I	0.13
NBRI-N25		N272D	0.06
NBRI-N27		P3395S	1
NBRI-N28		T6557S	0
NBRI-N30		V4797F	0
NBRI-N36		R24S	0.47
		G7063S	0.7
NBRI-N38		A372V	0.05
NBRI-N40		S558F	0.46
NBRI-N42		V1570L	0.02
		T2967I	0.25
NBRI-N47		T1682I	0.2

bioRxiv preprint doi: https://doi.org/10.1101/2020.08.24.265827; this version posted August 25, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

NBRI-N51		T677I	0.36
NBRI-N55		T1484I	0.05
		V4980L	0
NBRI-N39		H2698Q	0.18
NBRI-N24		T224I	0.03
NBRI-N26		G2808V	0.14
NBRI-N17	N	D343H	0.19
NBRI-N34		D371V	0.01
NBRI-N29		K370N	0.02
		D371Y	0.01
NBRI-N50		P199L	0.11
NBRI-N57		P279Q	0
NBRI-N41		T362I	0.03
NBRI-N48	S	A1020S	0.39
NBRI-N32		A292S	0.64
NBRI-N35		V1040F	0.12
NBRI-N49		L5F, A522S	0.10,0.73
NBRI-N58		L117I	0.76
NBRI-N39	ORF3a	S205T	0.06
NBRI-N48		V55G	0
NBRI-N24	ORF7a	P34S	0
NBRI-N29	ORF8	D34G	0
NBRI-N26		V62L	0.54

509