Supplementary Materials for

Genomic characterization and Epidemiology of an emerging SARS-CoV-2 variant in Delhi, India

Mahesh S Dhar[#], Robin Marwal[#], Radhakrishnan VS[#], Kalaiarasan Ponnusamy[#], Bani Jolly[#], Rahul C. Bhoyar[#], Viren Sardana, Salwa Naushin, Mercy Rophina, Thomas A Mellan, Swapnil Mishra, Charlie Whittaker, Saman Fatihi, Meena Datta, Priyanka Singh, Uma Sharma, Rajat Ujjainiya, Nitin Bhateja, Mohit Kumar Divakar, Manoj K Singh, Mohamed Imran, Vigneshwar Senthivel, Ranjeet Maurya, Neha Jha, Priyanka Mehta, Vivekanand A, Pooja Sharma, Arvinden VR, Urmila Chaudhary, Lipi Thukral, Seth Flaxman, Samir Bhatt, Rajesh Pandey, Debasis Dash, Mohammed Faruq, Hemlata Lall, Hema Gogia, Preeti Madan, Sanket Kulkarni, Himanshu Chauhan, Shantanu Sengupta, Sandhya Kabra, The Indian SARS-CoV-2 Genomics Consortium (INSACOG), Ravindra K. Gupta, Sujeet K Singh, Anurag Agrawal*, Partha Rakshit*

[#]Equal contribution

*Correspondence to: partho_rakshit@yahoo.com (Partha Rakshit), a.agrawal@igib.in (Anurag Agrawal)

This PDF file includes:

Materials and Methods

Genome Sequencing and Analysis

Nasopharyngeal and throat swab samples from COVID-19 confirmed cases with Ct value < 25 was collected and transported to Biotechnology Division, NCDC from the various testing sites across the States as per the sampling strategy of Central Surveillance Unit (CSU) of Integrated Disease Surveillance Programme (IDSP), NCDC. Twenty-seven post-vaccinated COVID-19 positive samples were also included in this study. All patient details were filled on the patient identification form and were accompanied with the samples. A total of 11,335 samples were received for genomic sequencing at NCDC from Nov 2020 till May 2021. The RNA was isolated using MagNA Pure RNA extraction system (Roche). The samples were processed as per the CovidSeq (Illumina) protocol for sequencing on the NextSeq 550 (Illumina) instrument according to the manufacturer's instructions. A total of 376 samples per lot were processed in batches of 94 with indexes A-D and 1.4pmol of the library was loaded on the 75 cycle High Output flow cell. Approximately 20Gb of data was generated and was processed on the Dragen COVID Lineage Application v3.5.1 (Illumina). The raw data generated in binary base call (BCL) format from NextSeq 550 was demultiplexed to FASTQ files using bcl2fastq v2.20. The raw reads were aligned to the SARS-CoV-2 reference genome (NC_045512) (*1*). The minimum acceptable alignment score was set to 12, so alignment results scoring lower were discarded. The coverage threshold and virus detection threshold were set to 20 and 5 respectively. The variant calling target coverage was set to 50 which specifies the maximum number of reads with a start position overlapping any given position. Out of the 9,557 genome sequences generated, 7,858 sequences with complete metadata were used for further analysis.

Genome Datasets and Lineage Analysis

The lineage/clade analysis was performed using pangolin (v3.0.5, pangoLEARN 2021-06-05) (*2*) and Nextclade (https://clades.nextstrain.org/). The lineage data was segregated according to date of collection (DOC), State, Age, Sex. State wise distribution of B.1.617.2 compared with B.1.617.1, B.1.1.7 and 'other variants' over a period of five months (January-May, 2021) was performed to ascertain the prevalence of each variant within these geographical areas. The number of tests, confirmed cases and positivity rate were accessed from the national databases.

Phylogenetic analysis was performed following the Nextstrain protocol (https://nextstrain.org/) using a subset of genomes sequenced from the states of Delhi, Punjab and Maharashtra. Genome sequences having more than 5% ambiguous bases were excluded from the phylogenetic analysis. A total of 10,189 genome sequences including additional publicly available sequences from GISAID (up to 2021-06-09) were used to calculate lineage proportions for the different states.

Serosurvey

The serosurvey was conducted through a voluntary participation wherein personnel working at CSIR labs/centers and their family members gave their blood samples in June-Sep 2020 (Phase 1) and Feb-March 2021 (Phase 2). The study was approved by the Institutional Human Ethics Committee of CSIR-IGIB vide approval CSIR-IGIB/IHEC/2019–20 and carried out in over 40 CSIR laboratories and centers spread across the country. In this study, samples from 10,427 adult subjects were obtained of which; 1,399 samples were from Delhi based laboratories

and offices in Phase 1 and 9,918 samples were obtained of which 1,115 samples were from Delhi in Phase 2. Blood samples (6 ml) were collected in EDTA vials from each participant and analyzed on site or transported to CSIR-IGIB, New Delhi for Analysis. Elecsys Anti-SARS-CoV-2 kit from Roche Diagnostics was used to detect antibodies to SARS-CoV-2 NucleoCapsid antigen. It is a qualitative kit which was used for screening and a Cut-off index COI >1 was considered seropositive. Positive samples were further tested for quantitative antibody titers using the same manufacturer's kit directed against the spike protein (S-antigen). An antibody levels >0.8 U/ml was considered sero-positive as per manufacturer's protocol. The detection range of this kit is from 0.4 U/ml to 250 U/ml. For samples, where values of >250 U/ml were obtained; appropriate dilutions were made. Neutralizing antibody (NAB) response directed against the spike protein (RBD site) was assessed using GENScript cPass kit which is a surrogate virus neutralization test (sVNT). A value of 30% or above was considered to have neutralizing ability.

Epidemiological Model

The model described here builds on a previously published model of SARS-CoV2 transmission introduced in Flaxman et al, 2020 (1), subsequently extended into a two-category framework in Faria et al, 2020 (2).

The model describes two categories, denoted $s \in \{1,2\}$. The population unadjusted reproduction number for the first category is defined as

$$R_{s=1,t} = \mu_0 2\,\sigma(X_t), \qquad (1)$$

where μ_0 is a scale parameter (3.3), σ is a logistic function, and X_t is a second order

autoregressive process with weekly time innovations, as specified in earlier work (3). The population-unadjusted reproduction number of the second category is modelled as

$$R_{s=2,t} = \rho \mathbf{1}_{[t_2,\infty)} R_{1,t}, \qquad (2)$$

with

$$\rho \sim \text{Gamma}(5,5) \in [0,\infty), \qquad (3)$$

where ρ is a parameter defining the relative transmissibility of category 2 compared to category 1 and $\mathbf{1}_{[t_2,\infty)}$ is an indicator function taking the value of 0 prior to t_2, and 1 thereafter, highlighting that category 2 does not contribute to the observed epidemic evolution before its emergence.

Infections arise for each category according to a discrete renewal process (4, 5)

$$i_{s,t} = \left(1 - \frac{n_{s,t}}{N}\right) R_{s,t} \sum_{\tau<t} i_{s,\tau} g_{t-\tau}, \qquad (4)$$

where N is the total population size, $n_{s,t}$ is the total extent of population immunity to category s present at time t, and g is the generation interval distribution.

The susceptible depletion term for category s is modelled as

$$n_{s,t} = \sum_{\tau<t} i_{s,\tau} W_{t-\tau} + \beta_s \left(1 - \alpha_{s,t}\right) \sum_{\tau<t} i_{\setminus s,\tau} W_{t-\tau}\,. \qquad (5)$$

under assumptions of symmetric cross-immunity with prior

$$\beta \sim \text{Beta}(2,1). \qquad (6)$$

$W_{t-\tau}$ is the time-dependent waning of immunity elicited by previous infection, which is modelled as a Rayleigh survival-type function with Rayleigh parameter of sigma = 310, which produces 50% of individuals still immune after 1 year. The cross-immunity susceptible term $\alpha_{s,t}$ is modelled as

$$\alpha_{s,t} = \frac{(1-\beta_s)\sum_{\tau<t} i_{s,\tau} W_{t-\tau}}{N - \beta_s \sum_{\tau<t} i_{s,\tau} W_{t-\tau}}. \quad (7)$$

Infections in Delhi are seeded for six days at the start of the epidemic as

$$i_{1,t_{1..6}} \sim \text{Exponential}(1/\tau), \quad (8)$$

with

$$\tau \sim \text{Exponential}(0.03), \quad (9)$$

and the second category for one day (t_2 = 14-02-2021) as

$$i_{2,t_2} \sim \text{Normal}(1,20) \in [1,\infty). \quad (10)$$

Non-unit seeding of the B.1.617.2 variant and the diffuse prior represent our uncertainty in the precise date and magnitude of B.1.617.2's introduction/importation into Delhi.

The model generates deaths via the following mechanistic relationship:

$$d_t = \sum_s \text{ifr}_s \sum_{\tau<t} i_{s,\tau} \pi_{t-\tau}. \quad (11)$$

The infection fatality ratios (ifr_s) of each of the categories are given moderately informative priors:

$$\text{ifr}_s \sim \text{Normal}(0.25, 0.02^2) \in [0, 100] \qquad (12)$$

allowing some variation in the IFR while suggesting it is unlikely to be less than 0.15 or greater than 0.3.

The observation model uses three types of data from four sources. In the first, the likelihood for the expected deaths D_t, is modelled as negative-binomially distributed,

$$D_t \sim \text{Negative Binomial} \quad (d_t, d_t + \frac{d_t^2}{\phi}), \qquad (13)$$

with mortality data d_t and dispersion prior

$$\varphi \sim \text{Normal}(0, 5) \in [0, \infty). \qquad (14)$$

The second likelihood is based on genomic data from individuals where infections were sequenced and where the sequence was uploaded to GISAID. Specifically, the proportion of sequenced genomes identified as B.1.617.2 at time t are modelled with a binomial likelihood

$$G^+_t \sim \text{Binomial}(G^+_t + G^-_t, \theta_t), \qquad (15)$$

with positive counts for B.1.617.2 denoted G^+_t and counts for lineages not belonging to B.1.617.2 recorded as G^-_t. The success probability for B.1.617.2 positivity is modelled as the infection ratio

$$\theta_t = \frac{\tilde{i}_{2,t}}{\tilde{i}_{1,t} + \tilde{i}_{2,t}} \quad , \qquad (16)$$

where $\tilde{i}_{s,t}$ is given by

$$\tilde{i}_{s,t} = \sum_{\tau \leq t} i_{s,\tau}\, \kappa_{t-\tau}\,, \qquad (17)$$

to account for the time varying PCR positivity displayed over the natural course of a COVID-19 infection. The distribution κ describes the probability of being PCR positive over time following infection and is based on Hellewell J et al. (6).

Thirdly, serological data from two sources are incorporated in our modelling framework. The observed seropositivity (S_t) on a given day, t, is modelled as follows

$$S_t \sim \text{Normal}\left(\sum_{\tau \leq t} i_{s,\tau} C_{t-\tau}, \sigma_s\right), \quad (18)$$

where $C_{t-\tau}$ is the cumulative probability of an individual infected on day τ having seroconverted by time t. This distribution is empirical and based on (7). The σ_s is given a Gamma$(5,1)$ prior.

Eq (4) can be modified to account for population effects (decreasing susceptible population over time) such that no over-shooting happens due to discretization as follows (8, 9):

$$i_{s,t} = (N - n_{s,t})\left(1 - \exp\left(-\frac{i_{s,t}}{N}\right)\right), \quad (19)$$

The formula for $i_{s,t}$ is derived from a continuous time model on $[t-1,t]$. This is to avoid discrete time effects such as infections going above the total population

N. Specifically, we assume that the infections $i(\Delta t)$ in $[t-1, t-1+\Delta t]$ are given by the differential equation $\partial i(\Delta t)/\partial \Delta t = i_t(1-(n_{s,t}+i(\Delta t))/N)$, which has the solution $i(1) = i_t$ as above.

Data Availability

The consensus fasta sequences generated as a part of this study have been submitted in GISAID.

Protein annotation and modelling

In total, 8,477 SARS-CoV-2 genomes were annotated for amino-acid substitutions by SnpEff version 4.5 (10). The annotation was done according to the SARS-CoV-2 reference genome (NC_045512) (11). The structural model of the spike in 1 RBD-up state was generated using cryoEM structure of the spike 1 RBD-up state (PDB ID: 6VSB) as a template (12). To generate ACE2 bound structure, we took the X-ray structure of human ACE2 bound to the RBD domain with PDB ID: 6M0J (13). Detailed modelling methodology is mentioned in our previous work (14). The structural mutant model of B.1.617.2 variant was generated using the structural model of ACE2-bound 1 RBD-up spike conformation as a reference. Each chain was mutated for missense mutations using ChimeraX (15) whereas deletions in each chain were introduced by employing Coot (16).

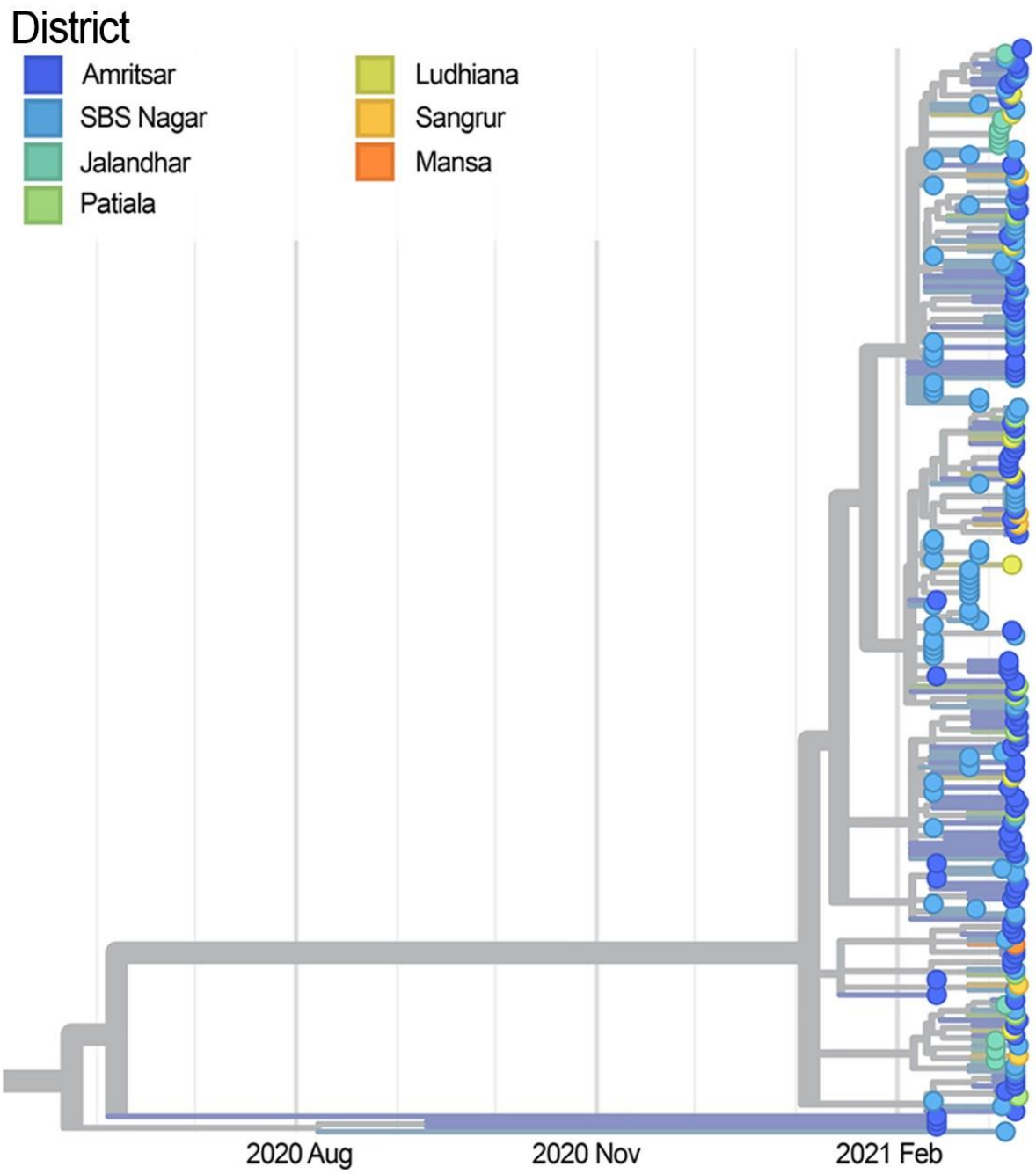

fig. S1. Molecular signature of super-spreader event in Punjab. Time-resolved phylogenetic tree for genome sequences from Punjab using samples collected during February-March 2021. Strong identity can be seen between sequences from various districts, corresponding to known social events during this period

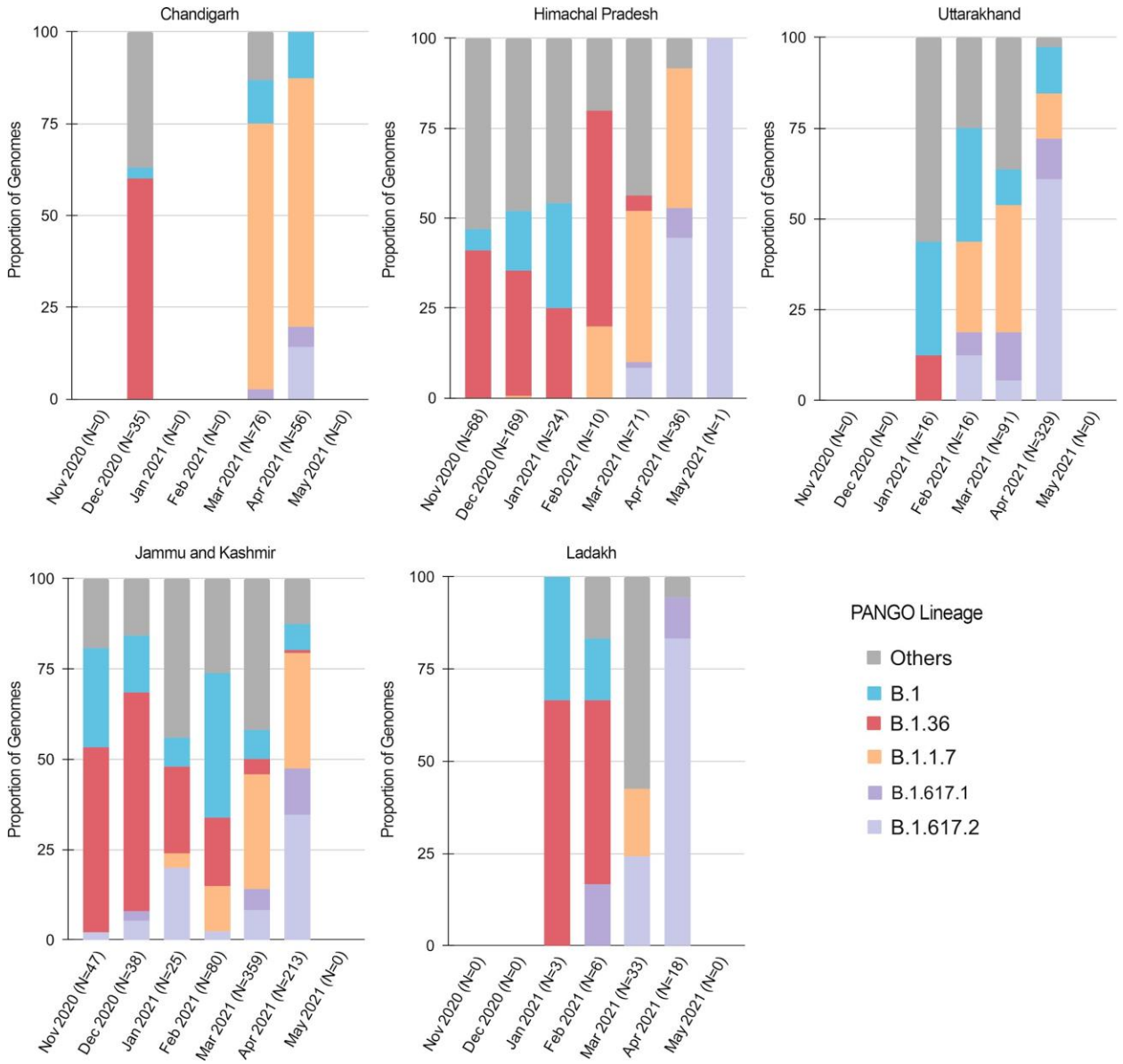

fig. S2. Displacement of Alpha by Delta strain all over North India. Normalized stacked bar graphs of main lineages for the states of Chandigarh, Himachal Pradesh, Uttarakhand, Jammu and Kashmir, and Ladakh. Outbreaks were seen in these states during April and May 2021, coincident with the rise of Delta.

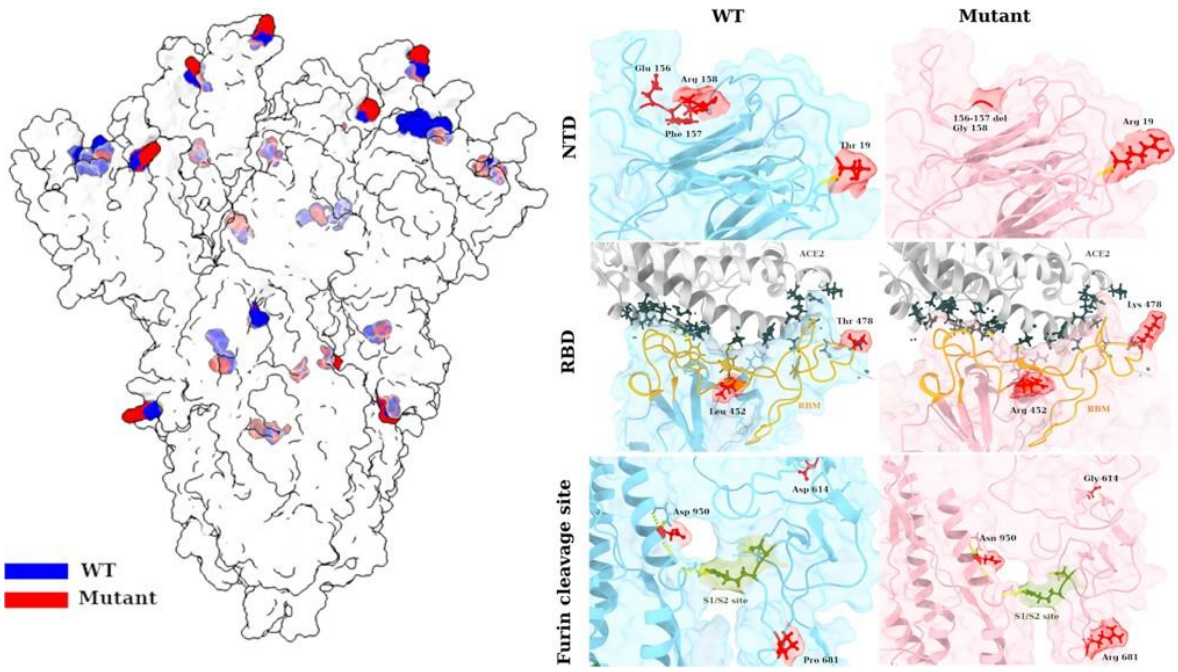

fig. S3. Mutant spike protein of B.1.617.2 lineage has critical mutations at furin cleavage and RBD sites that may enhance binding and cleavage. The structural model of spike protein with seven mutations was generated and side-chains are highlighted in blue and red color to illustrate amino-acid substitutions. On the right panel, we show a zoomed snapshot of three critical regions namely, NTD, RBD, and furin cleavage sites. The location of mutated residues is marked in red and, and critical regions such as RBM and S1/S2 site are highlighted for clarity.

Table S1.

Inferred changes in epidemiological characteristics of B.1.617.2, depending on the timing of introduction assumed in the model, and level of under-ascertainment present in Delhi mortality data. Results presented are the median, with the 50% Bayesian Credible Interval, bCI, in brackets. Note that "Immune escape" refers specifically to immunity conferred by prior infection with other variants, rather than escape from immunity acquired through vaccination. It is further important to note that immune escape and transmissibility increase inferred for B.1.617.2 are values given with reference to the composition of earlier and co-circulating variants in Delhi, from the start of the epidemic to 25 May 2021.

		Inferred epidemiological characteristic	
Timing of B.1.617.2 introduction	**Mortality under-ascertainment**	**Immune escape**	**Transmissibility increase**
14 Jan 2021	**10%**	0.36 (0.16-0.59)	1.47 (1.36-1.57)
14 Feb 2021	**10%**	0.40 (0.19-0.61)	1.49 (1.40-1.59)
28 Feb 2021	**10%**	0.43 (0.21-0.65)	1.58 (1.50-1.66)
14 Jan 2021	**33%**	0.19 (0.06-0.46)	1.53 (1.35-1.62)
14 Feb 2021	**33%**	0.25 (0.09-0.51)	1.51 (1.36-1.63)
28 Feb 2021	**33%**	0.37 (0.17-0.64)	1.55 (1.41-1.68)
14 Jan 2021	**50%**	0.12 (0.06-0.25)	1.56 (1.45-1.62)
14 Feb 2021	**50%**	0.13 (0.06-0.33)	1.59 (1.43-1.66)
28 Feb 2021	**50%**	0.25 (0.09-0.52)	1.60 (1.42-1.74)
14 Jan 2021	**66%**	0.41 (0.33-0.52)	1.24 (1.15-1.31)
14 Feb 2021	**66%**	0.42 (0.34-0.57)	1.28 (1.16-1.35)
28 Feb 2021	**66%**	0.56 (0.40-0.75)	1.31 (1.19-1.43)

References

1. Flaxman S, Mishra S, Gandy A, Unwin HJT, Mellan TA, Coupland H, et al. Estimating the effects of non-pharmaceutical interventions on COVID-19 in Europe. Nature. 2020;584(7820):257-61.

2. Faria NR, Mellan TA, Whittaker C, Claro IM, Candido DdS, Mishra S, et al. Genomics and epidemiology of the P.1 SARS-CoV-2 lineage in Manaus, Brazil. Science. 2021;372(6544):815-21.

3. Unwin HJT, Mishra S, Bradley VC, Gandy A, Mellan TA, Coupland H, et al. State-level tracking of COVID-19 in the United States. Nature Communications. 2020;11(1):6189.

4. Feller W. On the Integral Equation of Renewal Theory. The Annals of Mathematical Statistics. 1941;12(3):243-67, 25.

5. Bellman R, Harris T. On Age-Dependent Binary Branching Processes. Annals of Mathematics. 1952;55(2):280-95.

6. Hellewell J, Russell TW, Matthews R, Severn A, Adam S, Enfield L, et al. Estimating the effectiveness of routine asymptomatic PCR testing at different frequencies for the detection of SARS-CoV-2 infections. BMC Medicine. 2021;19(1):106.

7. Borremans B, Gamble A, Prager KC, Helman SK, McClain AM, Cox C, et al. Quantifying antibody kinetics and RNA detection during early-phase SARS-CoV-2 infection by time since symptom onset. eLife. 2020;9:e60122.

8. James A. Scott AG, Swapnil Mishra, Juliette Unwin, Seth Flaxman, and Samir Bhatt. epidemia: Modeling of Epidemics using Hierarchical Bayesian Models. R package version 0.5.3. 2020. url: https:// imperialcollegelondon.github.io/epidemia/. epidemia: Modeling of Epidemics using Hierarchical Bayesian Models. R package version 0.5.3. 2020 [Available from: https://

imperialcollegelondon.github.io/epidemia/.

9. Samir Bhatt NF, Seth Flaxman, Axel Gandy, Swapnil Mishra, James A. Scott. Semi-Mechanistic Bayesian Modeling of COVID-19 with Renewal Processes. arXiv. 2020.

10. Cingolani P, Platts A, Wang le L, Coon M, Nguyen T, Wang L, 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. 2012;6(2):80-92.

11. Wang C, Liu Z, Chen Z, Huang X, Xu M, He T, et al. The establishment of reference sequence for SARS-CoV-2 and variation analysis. Journal of medical virology. 2020;92(6):667-74.

12. Wrapp D, Wang N, Corbett KS, Goldsmith JA, Hsieh C-L, Abiona O, et al. Cryo-EM structure of the 2019-nCoV spike in the prefusion conformation. Science. 2020;367(6483):1260-3.

13. Lan J, Ge J, Yu J, Shan S, Zhou H, Fan S, et al. Structure of the SARS-CoV-2 spike receptor-binding domain bound to the ACE2 receptor. Nature. 2020;581(7807):215-20.

14. Fatihi S, Rathore S, Pathak A, Gahlot D, Mukerji M, Jatana N, et al. A rigorous framework for detecting SARS-CoV-2 spike protein mutational ensemble from genomic and structural features. bioRxiv. 2021:2021.02.17.431625.

15. Goddard TD, Huang CC, Meng EC, Pettersen EF, Couch GS, Morris JH, et al. UCSF ChimeraX: Meeting modern challenges in visualization and analysis. Protein science : a publication of the Protein Society. 2018;27(1):14-25.

16. Emsley P, Lohkamp B, Scott WG, Cowtan K. Features and development of Coot. Acta crystallographica Section D, Biological crystallography. 2010;66(Pt 4):486-501.

INSACOG

INSACOG CONSORTIUM MEMBERS

NIBMG: Saumitra Das, Arindam Maitra, Sreedhar Chinnaswamy, Nidhan Kumar Biswas;

ILS: Ajay Parida, Sunil K Raghav, Punit Prasad;

InSTEM/ NCBS: Apurva Sarin, Satyajit Mayor, Uma Ramakrishnan, Dasaradhi Palakodeti, Aswin Sai Narain Seshasayee;

CDFD: K Thangaraj, Murali Dharan Bashyam, Ashwin Dalal;

NCCS: Manoj Bhat, Yogesh Shouche, Ajay Pillai;

IGIB: Sridhar Sivasubbu, Vinod Scaria;

NIV: Priya Abraham, Potdar Varsha Atul, Sarah S Cherian;

NIMHANS: Anita Sudhir Desai, Chitra Pattabiraman, M. V. Manjunatha, Reeta S Mani, Gautam Arunachal Udupi;

NCDC: Tanzin Dikid

CCMB: Vinay Nandicoori, Karthik Bharadwaj Tallapaka, Divya Tej Sowpati