Spike protein mutational landscape in India: Could Muller's ratchet be a future game-changer for COVID-19?

Rachana Banerjee[a], Kausik Basak[a], Anamika Ghosh[b], Vyshakh Rajachandran[c], Kamakshi Sureka[a], Debabani Ganguly[a,1], Sujay Chattopadhyay[a,1]

[a]Centre for Health Science and Technology, JIS Institute of Advanced Studies and Research Kolkata, JIS University, 700091, West Bengal, India. [b]Department of Chemistry, Indian Institute of Engineering Science and Technology, Shibpur, 711103, West Bengal, India. [c]School of Biotechnology, Amrita Vishwa Vidyapeetham, Kollam, 690525, Kerala, India.

[1]To whom correspondence may be addressed. Email: sujayc@jisiasr.org or debabani@jisiasr.org

Classification

Major: Biological Sciences,

Minor: Evolution

Keywords:

SARS-CoV-2, Muller's ratchet, mutational meltdown, molecular docking, viral evolutionary dynamics

Abstract

The dire need of effective preventive measures and treatment approaches against SARS-CoV-2 virus, causing COVID-19 pandemic, calls for an in-depth understanding of its evolutionary dynamics with attention to specific geographic locations, since lockdown and social distancing to prevent the virus spread could lead to distinct localized dynamics of virus evolution within and between countries owing to different environmental and host-specific selection pressures. To decipher any correlation between SARS-CoV-2 evolution and its epidemiology in India, we studied the mutational diversity of spike glycoprotein, the key player for the attachment, fusion and entry of virus to the host cell. For this, we analyzed the sequences of 630 Indian isolates as available in GISAID database till June 07, 2020, and detected the spike protein variants to emerge from two major ancestors – Wuhan-Hu-1/2019 and its D614G variant. Average stability of the docked spike protein – host receptor (S-R) complexes for these variants correlated strongly (R^2=0.96) with the fatality rates across Indian states. However, while more than half of the variants were found unique to India, 67% of all variants showed lower stability of S-R complex than the respective ancestral variants, indicating a possible fitness loss in recently emerged variants, despite a continuous increase in mutation rate. These results conform to the sharply declining fatality rate countrywide (>7-fold during April 11 – June 28, 2020). Altogether, while we propose the potential of S-R complex stability to track disease severity, we urge an immediate need to explore if SARS-CoV-2 is approaching mutational meltdown in India.

Significance

Epidemiological features are intricately linked to evolutionary diversity of rapidly evolving pathogens, and SARS-CoV-2 is no exception. Our work suggests the potential of average stability of complexes formed by the circulating spike mutational variants and the human host receptor to track the severity of SARS-CoV-2 infection in a given region. In India, the stability of these complexes for recent variants tend to decrease relative to their ancestral ones, following countrywide declining fatality rate, in contrast to an increasing mutation rate. We hypothesize such a scenario as nascent footprints of Muller's ratchet, proposing large-scale population genomics study for its validation, since this understanding could lead to therapeutic approaches for facilitating mutational meltdown of SARS-CoV-2, as experienced earlier for influenza A virus.

Introduction

The emergence and rapid global spread of novel severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), the causative agent of the coronavirus disease 2019 (COVID-19), has led to an unprecedented worldwide public health crisis. SARS-CoV-2 is the third virus of the Betacoronavirus genus, after the SARS-CoV and the MERS-CoV, to cross the species barrier and to disseminate rapidly through the global population (1). Like any other rapidly evolving emerging pathogens, effective preventive measures and treatment approaches call for an in-depth understanding of the evolutionary dynamics of SARS-CoV-2 and its association with the epidemiological data (2). The assessment of interactions between evolutionary and ecological processes should not be restricted to just at the global level, but also needs to be localized with distinct attention to specific geographic regions, as the nature and strength of selection pressures acting on the pathogen can vary to a great extent across different ethnicities, countries, or even among specific administrative regions (such as states or provinces) within a country like India, China, USA. The need of localized assessment is specifically important for the COVID-19 scenario which has instigated us to opt for social distancing via lockdown as a preventive measure against the geographic spread of the pathogen between and within countries.

The transmission and epidemiological features of any newly emerging or re-emerging pathogens are driven by their genetic diversity (3). The success of SARS-CoV-2 infection lies in the accumulation of specific mutations enabling it to make a leap from some reservoir species to humans. Furthermore, being an RNA virus, SARS-CoV-2 possesses a high mutation rate which contributes to viral adaptation, evolvability and virulence modulation (4). Therefore, an in-depth investigation of molecular divergence and evolutionary patterns of SARS-CoV-2 would help to understand its role in pathogenicity, leading to development of novel interventions.

The two most notable genetic features of SARS-CoV-2 lies within the transmembrane spike glycoprotein protruding from the viral surface and is responsible for viral attachment, fusion and entrance into the host cells and thereby establish its infection (5, 6). The surface unit (S1) of spike protein binds to the human cellular receptor angiotensin converting enzyme 2 (hACE2) and the transmembrane unit (S2) then fuses the viral membrane with a host cell membrane. The S1 subunit contains the receptor-binding domains (RBD) supporting the stabilization of the membrane-anchored state of the S2 subunit along with its fusion machinery (7-12). Some previous works have acknowledged that extensive and irreversible conformational changes stimulate the cleavage of spike protein, thereby activating it for membrane fusion (12-18). Though it is not yet confirmed if the differences owing to conformational changes are aiding to the expansion of SAR-CoV2 varieties improving its infectivity or transmissibility, but the researchers confirmed the spike proteins to be the key pathogenic determinant that differentiates SAR-CoV2 from other SARS-related coronaviruses. Consequently, by examining the architecture of the spike glycoprotein and its mechanics, one can reveal the susceptibility of the virus and can foresee the facts helping in the discovery of the SARS-CoV2 antidote (18). Although the RBD in the spike protein is known as the most important player to recognize the host receptor, it is highly likely that the region outside of this C-terminal domain in S1 subunit as well as the domains of S2 subunit also influence the host-receptor binding allosterically.

The present study aims to understand the evolutionary patterns and driving forces behind the emerging SARS-CoV-2 infection and its potential epidemiological footprints, based on a comprehensive analysis of the spike protein mutational landscape of SARS-CoV-2 strains circulating in India. Genome sequences of 630 Indian isolates (collected from GISAID database till June 07, 2020) have been analyzed to identify the countrywide as well as region-specific

distribution of the mutational variants. To study possible functional implications of these variations, we next measured the stability of complex formed by the circulating spike protein variants and the host receptor (S-R complex). Interestingly, we detected a strong direct correlation between average stability contributed by the circulating variants and the disease severity of a given location, suggesting the S-R complex stability as a potential marker to assess the severity of the disease. On the other hand, majority of the variants emerging from the ancestral ones showed decreased stability, which conforms to the declining fatality rates of the disease countrywide. However, the rates of mutations, especially the nonsynonymous (amino acid replacement) ones, were found to be only increasing over the period of three months' data collection. Together, our findings indicate an accumulation of mutations in the emerging spike protein variants offering reduced fitness to the organism. Could the fixation of these deleterious mutations in the population lead to mutational meltdown, following Muller's ratchet dynamics of evolution?

Results

Two major ancestors circulating in India lead to a burst of spike protein variants.

We identified a total of 630 isolates with complete gene sequences encoding the spike protein based on the submissions of Indian SARS CoV-2 genome sequences to GISAID till June 7, 2020 (*SI appendix*, Table S1). The samples analyzed were isolated from 17 states and 2 union territories of India, collectively called as 'states' hereafter. We found a countrywide average pairwise nucleotide diversity (π) of 0.048±0.02%. The rates of synonymous (or silent) and nonsynonymous (or amino acid replacement) mutations were found to be 0.097±0.05% and 0.033±0.02% respectively. Phylogenetic analysis showed the Wuhan-Hu-1/2019 variant of the spike protein as

the most ancestral one, as expected, while the D614G variant emerged from the Wuhan-Hu-1/2019 variant appeared to be another stable variant circulating in the Indian population. Since both these variants have established themselves in the worldwide population as two major ancestors of SARS CoV-2 spike protein variants, we here onwards will refer the Wuhan-Hu-1/2019 and D614G variants as ancestor 1 and ancestor 2 respectively.

Apart from giving rise to ancestor 2, the ancestor 1 led to a total of 20 variants (Fig. 1 and *SI appendix*, Table S2). Of these, the variant K77M evolved further to yield three more variants isolated from three different states (Bihar, Tamil Nadu and Telengana), suggesting the emergence of K77M as another stable variant. On the other hand, the ancestor 2 showed about twice more diversity by giving rise to 47 variants (Fig. 1 and *SI appendix*, Table S2). In this ancestor 2 clade, several variants (L5F, T22I, L54F, G261S, T572I, E583D, Q677H, A706S, H1083Q) indicated their stability in the population via mutating further, giving rise to additional variants. Besides, although the ancestral variants were predominant in the population circulating in India, we detected a total of 16 variants of spike proteins that were represented by multiple isolates (Fig. 1), from 2 to as many as 9 isolates, indicating the possible fixation of some of these variants in the population irrespective of the stability of S-R complex. Of these, an array of mutations in 8 variants (at positions 5, 54, 78, 558, 574, 583, 677, 1243) showed their convergent nature, where those mutations at the same positions were phylogenetically unlinked, i.e. repeated independently (*SI appendix*, Table S2). Interestingly, 53% of the total set of variants detected in Indian population was found unique, i.e. not found in 17529 worldwide genomic isolates analyzed from the GISAID database till May 09, 2020 (*SI appendix*, Table S2). We believe that this considerable level of uniqueness could be an expected scenario in almost all geographical regions where a newly emerging, rapidly evolving viral pathogen tries to adapt to a new host, and many of these variants

might be detrimental to the fitness of the organism. However, specific positive selection pressures could also play a role in this mutational pattern which needs to be studied separately in some greater depth.

The average stability index of S-R complex correlates strongly with the fatality rates in a given location.

We next looked into the distribution of these spike protein variants across Indian states (*SI appendix*, Table S3). As we computed the diversity based on both richness and evenness of spike variants, some of the states like Maharashtra, Odisha, West Bengal and Gujarat demonstrated significantly higher ($P<0.05$) diversity of circulating variants than most of the remaining states. However, while we had 196, 97, 75 and 73 sequenced isolates from Gujarat, Telengana, Delhi and Maharashtra respectively, the remaining states were represented by even lower sample size (ranging from 1 to 48 isolates. Of these, Delhi variants showed lowest diversity, significantly different from both Maharashtra ($P=0.0002$) and Gujarat ($P=0.0006$), though not from Telengana ($P=0.16$). On the other hand, Maharashtra variants presented much higher diversity than Telengana ($P=0.017$) or Gujarat ($P=0.21$).

It is expected that these variations might be in response to strong selection pressures acting on the spike protein, especially its S1 subunit which, being a major immunogenic target for the host, plays the pivotal role to evade the host immune response and to offer a successful viral entry. Therefore, the mutational variations in spike proteins can essentially affect the stability of the S-R complex. To assess this, we modeled each of the spike protein variants, and then docked to the binding site of host receptor, hACE2, using HADDOCK webserver (19, 20) of data-driven docking algorithm by providing binding site information as the same was already established from the

crystal structure (21, 22). In the circulating variants, we detected a significant excess (P=0.035) of mutations in the S1 subunit (with 41 mutations) compared to the S2 subunit (with 22 mutations). The mutation positions in the secondary structures of the analyzed variants are detailed in the *SI appendix*, Dataset S1.

The docking score (HADDOCK score) of each variant on hACE2 is hereafter designated as the stability index of S-R complex. More negative is the docking (HADDOCK) score, higher is the stability of the S-R complex (23). Under the assumption that better stability would lead to better invasion of the virus into the host, we hypothesize that such a stability index of a given spike protein variant could be linked to the severity of viral pathogenicity. To test this hypothesis, we measured the severity as the 'fatality rate' calculated simply as the ratio of the number of deceased people to the number of recovered in a given state (available from the Government of India website: www.mygov.in/corona-data/covid19-statewise-status/).

While we aimed to estimate an average stability index of a given state based on the stability indices of circulating variants in that location, we were handicapped with the available sample size and the information of collection diversity. Considering this issue, we restricted our study to the states having 50 or more sequenced samples for analysis. Therefore, we could assess the association of average stability index with fatality rates for three states and one union territory (Maharashtra, Gujarat, Telengana and Delhi) which qualified our sample size threshold. Importantly, these four regions harbored 70% of all samples analyzed across 19 states, while their variant diversity ranged from the highest to one of the lowest (*SI appendix*, Table S3).

We detected a strong exponential correlation (R^2=0.96) between the average stability index of circulating spike variants of the region with the fatality rate in that region (Fig. 2). While Telengana and Delhi showed comparable average stability index values with ~7% fatality rates,

Maharashtra and Gujarat had exponentially higher stability index values (i.e., more negative docking or HADDOCK scores) with 8% and 9% fatality rates respectively. It is highly plausible that the spike protein, as the primary controller of both the attachment to the host cell surface and the initiation of infection by fusing the viral and the host cell membranes, would be represented by variants with varying efficiency of the virus to enter human cells, and to get transmitted among people (18). However, our conclusions based on the available initial data are premature due to low sample size per location and lack of direct evidence for the correlation between spike protein variant's docking score and the pathogen's contribution to host fatality, thereby warranting population-level robust association analysis and experimental validations.

The emerging spike protein variants showing reduced stability of S-R complex are significantly abundant.

Of 630 isolates analyzed, Ancestors 1 and 2 were represented by 253 and 248 isolates, suggesting their steady circulation across India. However, the remaining ones, i.e. more than 20% isolates represented relatively recently emerged variants out of two ancestral variants. Interestingly, a quick look at Fig. 1 showed that, for majority of the variants derived from the Ancestor 1 and Ancestor 2, the stability was reduced (having less negative stability index values) relative to their respective ancestors. We therefore plotted the trend of those emerging variants with reference to their ancestral backgrounds (Fig. 3). Significant majority (χ^2 P=0.03) of the variants that emerged from the two ancestral variants showed reduced stability (having less negative docking scores) from their respective ancestors. This picture got even more prominent (χ^2 P=0.009) when we looked into exclusively the variants which were detected multiple times in the dataset, i.e. represented by more than one isolate (sometimes collected from different states), thereby

suggesting possible fixation of those variants in the population (*SI appendix*, Table S2). Interestingly, 13 of these variants with multiple occurrences accumulated mutations exclusively in S1 subunit, while only 3 variants showed all mutations in S2 subunit, which might be suggesting an increased selection pressure in the S1 subunit because of its key role in the viral entry and the presence of the receptor binding domain.

At this point, we can propose that the relatively recent variants emerging from the ancestors in India are losing their ability on an average to form a stable complex with the human receptor as compared with their ancestors, which could possibly result in lower countrywide fatality rate, if we combine our earlier observation of the direct correlation between average stability index and fatality rate. Conforming to this proposal, we found that, after an initial steady increase of fatality rate for the first four weeks, the fatality rate reached a peak at 38.2% on April 11, 2020, followed by a continuous sharp decline at 5.3% on June 28, 2020 until when the data were available (Fig. 4), fitting a power law function (R^2=0.94) (data source: http://covidindiaupdates.in/). This was despite the fact that the rate of mutations was found to be increasing during the three-month period from March till May 2020 during which our analyzed samples were collected. Interestingly, the increase in nonsynonymous (i.e. amino acid replacement) mutation rate (from 0.027% in March to 0.033% in May 2020) was detected to be about 14% higher than the increase in synonymous mutation rate (from 0.091% in March to 0.097% in May 2020), suggesting a stronger selection pressure for amino acid changes in the spike protein.

Could Muller's ratchet be a player in shaping SARS-CoV-2 evolutionary dynamics in India?

In evolutionary genetics, Muller's ratchet signifies accumulation of deleterious mutation in a population leading to 'mutational meltdown', leading to a gradual extinction of that population

(24). In general, genetic mutations that provide adaptive advantages are fixed in the population by natural selection whereas deleterious ones are wiped off from the population. However, an accelerated mutation rate puts huge mutational pressure, and natural selection is unable to wash out these deleterious mutations, retaining the newly formed variants within the population and thereby leading to their fixation. This irretrievable evolutionary mechanism is coined as Muller's ratchet by evolutionary biologists (25). When more and more deleterious mutations are accumulated and become permanent in the population, this results in 'mutational meltdown' or ultimate loss of the population (26).

It is known that the apparent tendency to directly correlate the high mutation rate of virus with its infectivity and transmissibility is without merit (27). On the other side, Jensen and Lynch (26) suggested that Muller's ratchet, via mutational meltdown, could be a key player in leading the SARS-CoV-2 population to gradual extinction due to accumulation and fixation of deleterious mutations in future. On one hand, we observed that the stability of S-R complex is directly linked to the fatality rates, while the continuous emergence of variants from the ancestral ones were found to be less stable compared to their ancestors. As we combine these findings with the countrywide data of a sharp decline in the fatality rate over time, we propose the possibility of mutational meltdown in action for SARS-CoV-2 in India, indicating Muller's ratchet as a plausible game-changer for COVID-19 scenario here in near future.

We understand that, only in three months, we cannot expect a radical increase in the mutation rate to give rise to any significant accumulation of deleterious mutations that could have offered a prominent picture of mutational meltdown. However, our results altogether point toward the trend, thereby suggesting the potential of future studies in this otherwise overlooked domain

of microbial dynamics, which could in turn lead to a possibility of a successful therapeutic approach, as suggested by Jensen & Lynch (26).

Discussion

Our work has combined genetic and epidemiological data of SARS-CoV2 in India to decipher a direct correlation between the average stability of S-R complexes for the circulating spike protein variants and the fatality rate of a geographic region. The docking score of S-R complex, designated here as the stability index, is estimated by protein-protein docking based on intermolecular interactions, such as electrostatic and van der Waals interactions, desolvation and restraint violation energies along with buried surface area upon binding via detection of the correct binding pose. This score, in essence, quantifies the stability of the docked complex by optimizing global minimum energy conformation of the complex (28, 29). On the other side, while it is probable that a better stability of S-R complex could lead to an increased viral load and maybe an increased infectivity, previous works showed that the spike protein – hACE2 complex is crucial for viral pathogenesis by causing acute lung damage (30, 31), which suggests a direct link between S-R complex stability and fatality rate. However, the robustness of the potential of S-R complex stability index for the spike protein variants as a tracker of fatality rate or disease severity needs to be studied in greater depths with more structured region-specific patient data (that are primarily available to selected government/non-government agencies) in connection with larger population-level sequence datasets for the given locations.

As expected for any fast-moving pathogen outbreaks (27), we find an increasing rate of mutations, while the extent of increase is much higher for the nonsynonymous (amino acid

replacement) changes. Interestingly, the S-R complex of a significant majority of variants tend to lose stability relative to their two most stable ancestral variants. Alongside, the countrywide data show a continuous sharp decline in the fatality rate after an initial surge. Therefore, keeping in mind the observed correlation between the S-R complex stability and fatality rate across Indian states, here we pose an open question for future research: Does the phylodynamics of SARS-CoV-2 in India indicate any nascent action of Muller's ratchet where the otherwise deleterious mutations tend to get fixed in the population as natural selection remains unable to purge them due to excessive mutational pressures? If so, there lies an immense potential of using therapeutics that could facilitate such a process of mutational meltdown, as was demonstrated earlier for influenza A virus (32, 33). Therefore, future large-scale population genomics analysis supported by epidemiological information is of high importance to explore this question for India as well as for other countries across the globe to develop efficient analytical methods, thereby guiding better surveillance programs, prevention and treatment management of COVID-19.

Materials and Methods

Analysis of sequence diversity and reconstruction of phylogeny

The average pairwise nucleotide diversity (π) and the rates of synonymous (dS) and nonsynonymous (dN) mutations for the spike protein encoding gene sequences were calculated using MEGA version X (34). TimeZone software (35) was used to reconstruct the maximum-likelihood based phylogeny of spike protein encoding gene sequences to map the corresponding protein variants and identify the convergent amino acid changes (i.e. repeated independent or phylogenetically unlinked mutations at the same amino acid positions) in the spike protein variants

circulating in India. The spike gene sequence from Wuhan-Hu-1/2019 genome was used as reference to detect the orthologs in the sequenced Indian genomes based on a threshold value of 95% for both nucleotide sequence diversity and gene length coverage.

Identification of spike protein variants unique to India

All 17529 spike protein sequences from worldwide isolates available till May 9, 2020 in the GISAID database (https://www.gisaid.org/) were downloaded. We implemented CD-HIT Suite (36, 37) to cluster all the spike protein sequences considering 100% amino acid sequence identity as ortholog clustering criteria, and detected a total of 3706 clusters. Of these clusters, we considered only 3577 clusters which matched the complete length of the spike protein (1273 amino acids), using the spike protein sequence from Wuhan-Hu-1/2019 (GenBank accession number MN908947) as reference. These 3577 spike protein sequences were aligned using ClustalW program (38, 39). The resulting alignment was compared with Indian mutational variants mapped in the previous step to distinguish the variants unique to Indian isolates.

Analysis of state-wise diversity of spike protein variants

For each Indian state, we computed the number of spike protein variants and the frequency of each of those variants. The state-wise calculation of variant diversity was performed using Simpson's index (40) considering both richness and evenness in the distribution of spike protein variants in the given state.

Modeling of spike protein – hACE2 complex variants

Mutant variants of the ancestral spike protein (Wuhan-Hu-1/2019) were built by homology modeling using Swiss modeler (41) with the aid of available templates (residue range: 27-1147) based on cryoelectron microscopy structures (18, 42) for spike protein using Wuhan-Hu-1/2019 isolate as reference. X-ray crystal structure of the human ACE2 was used from the complex of receptor binding domain (RBD) of spike protein with hACE2 (PDB code: 6lzg) (22). We have docked the RBD (residue range: 331-524) of spike protein (residue range: 27-1147) mutants to the binding site of hACE2 using HADDOCK webserver by providing binding site information (21, 42). HADDOCK score (often mentioned as docking score in the text), with some arbitrary unit, signifies a measure determined by weighted sum of intermolecular interactions, such as electrostatic and van der Waals interactions between protein and ligand, desolvation energy, restraint violation energy and the buried surface area upon binding. For each docked complex HADDOCK score was estimated and VMD (43) was used to visualize the structures.

Acknowledgements

DG acknowledges Department of Biotechnology, Govt. of India for Ramalingaswami Fellowship (BT/RLF/Re-entry/52/2012) and research grant BT/PR25670/BRB/10/1657/2018). SC acknowledges Department of Biotechnology, Govt. of India for research grant BT/PR26891/BID/7/816/2017.

Author Contributions

D.G., and S.C. designed research; R.B., K.B., A.G., V.R., D.G., and S.C. performed research; R.B., K.S., D.G., and S.C. analyzed data; and R.B., K.S., D.G., and S.C. wrote the paper.

Competing Interest

The authors declare no competing interest.

References

1. V. Coronaviridae Study Group of the International Committee on Taxonomy of, The species Severe acute respiratory syndrome-related coronavirus: classifying 2019-nCoV and naming it SARS-CoV-2. *Nat Microbiol* **5**, 536-544 (2020).
2. O. G. Pybus, A. Rambaut, Evolutionary analysis of the dynamics of viral infectious disease. *Nat Rev Genet* **10**, 540-550 (2009).
3. E. C. Holmes, S. Nee, A. Rambaut, G. P. Garnett, P. H. Harvey, Revealing the history of infectious disease epidemics through phylogenetic trees. *Philos Trans R Soc Lond B Biol Sci* **349**, 33-40 (1995).
4. S. Duffy, L. A. Shackelton, E. C. Holmes, Rates of evolutionary change in viruses: patterns and determinants. *Nat Rev Genet* **9**, 267-276 (2008).
5. K. G. Andersen, A. Rambaut, W. I. Lipkin, E. C. Holmes, R. F. Garry, The proximal origin of SARS-CoV-2. *Nat Med* **26**, 450-452 (2020).
6. M. A. Tortorici, D. Veesler, Structural insights into coronavirus entry. *Adv Virus Res* **105**, 93-116 (2019).
7. M. Gui *et al.*, Cryo-electron microscopy structures of the SARS-CoV spike glycoprotein reveal a prerequisite conformational state for receptor binding. *Cell Res* **27**, 119-129 (2017).
8. R. N. Kirchdoerfer *et al.*, Pre-fusion structure of a human coronavirus spike protein. *Nature* **531**, 118-121 (2016).
9. J. Pallesen *et al.*, Immunogenicity and structures of a rationally designed prefusion MERS-CoV spike antigen. *Proc Natl Acad Sci U S A* **114**, E7348-E7357 (2017).
10. W. Song, M. Gui, X. Wang, Y. Xiang, Cryo-EM structure of the SARS coronavirus spike glycoprotein in complex with its host cell receptor ACE2. *PLoS Pathog* **14**, e1007236 (2018).
11. A. C. Walls *et al.*, Cryo-electron microscopy structure of a coronavirus spike glycoprotein trimer. *Nature* **531**, 114-117 (2016).
12. A. C. Walls *et al.*, Tectonic conformational changes of a coronavirus spike glycoprotein promote membrane fusion. *Proc Natl Acad Sci U S A* **114**, 11157-11162 (2017).
13. S. Belouzard, V. C. Chu, G. R. Whittaker, Activation of the SARS coronavirus spike protein via sequential proteolytic cleavage at two distinct sites. *Proc Natl Acad Sci U S A* **106**, 5871-5876 (2009).
14. J. K. Millet, G. R. Whittaker, Host cell entry of Middle East respiratory syndrome coronavirus after two-step, furin-mediated activation of the spike protein. *Proc Natl Acad Sci U S A* **111**, 15214-15219 (2014).
15. J. K. Millet, G. R. Whittaker, Host cell proteases: Critical determinants of coronavirus tropism and pathogenesis. *Virus Res* **202**, 120-134 (2015).
16. J. E. Park *et al.*, Proteolytic processing of Middle East respiratory syndrome coronavirus spikes expands virus tropism. *Proc Natl Acad Sci U S A* **113**, 12262-12267 (2016).
17. T. Heald-Sargent, T. Gallagher, Ready, set, fuse! The coronavirus spike protein and acquisition of fusion competence. *Viruses* **4**, 557-580 (2012).
18. A. C. Walls *et al.*, Structure, Function, and Antigenicity of the SARS-CoV-2 Spike Glycoprotein. *Cell* **181**, 281-292 e286 (2020).

19. G. C. P. van Zundert *et al.*, The HADDOCK2.2 Web Server: User-Friendly Integrative Modeling of Biomolecular Complexes. *J Mol Biol* **428**, 720-725 (2016).
20. M. van Dijk, T. A. Wassenaar, A. M. Bonvin, A Flexible, Grid-Enabled Web Portal for GROMACS Molecular Dynamics Simulations. *J Chem Theory Comput* **8**, 3463-3472 (2012).
21. J. Lan *et al.*, Structure of the SARS-CoV-2 spike receptor-binding domain bound to the ACE2 receptor. *Nature* **581**, 215-220 (2020).
22. Q. Wang *et al.*, Structural and Functional Basis of SARS-CoV-2 Entry by Using Human ACE2. *Cell* **181**, 894-904 e899 (2020).
23. T. Pantsar, A. Poso, Binding Affinity via Docking: Fact and Fiction. *Molecules* **23** (2018).
24. H. J. Muller, The Relation of Recombination to Mutational Advance. *Mutat Res* **106**, 2-9 (1964).
25. J. Felsenstein, The evolutionary advantage of recombination. *Genetics* **78**, 737-756 (1974).
26. J. D. Jensen, M. Lynch, Considering mutational meltdown as a potential SARS-CoV-2 treatment strategy. *Heredity (Edinb)* **124**, 619-620 (2020).
27. N. D. Grubaugh, M. E. Petrone, E. C. Holmes, We shouldn't worry when a virus mutates during disease outbreaks. *Nat Microbiol* **5**, 529-530 (2020).
28. I. Halperin, B. Ma, H. Wolfson, R. Nussinov, Principles of docking: An overview of search algorithms and a guide to scoring functions. *Proteins* **47**, 409-443 (2002).
29. I. S. Moreira, P. A. Fernandes, M. J. Ramos, Protein-protein docking dealing with the unknown. *J Comput Chem* **31**, 317-342 (2010).
30. K. Kuba *et al.*, A crucial role of angiotensin converting enzyme 2 (ACE2) in SARS coronavirus-induced lung injury. *Nat Med* **11**, 875-879 (2005).
31. E. R. Lumbers, S. J. Delforce, K. G. Pringle, G. R. Smith, The Lung, the Heart, the Novel Coronavirus, and the Renin-Angiotensin System; The Need for Clinical Trials. *Front Med (Lausanne)* **7**, 248 (2020).
32. L. Ormond *et al.*, The Combined Effect of Oseltamivir and Favipiravir on Influenza A Virus Evolution. *Genome Biol Evol* **9**, 1913-1924 (2017).
33. S. Penisson, T. Singh, P. Sniegowski, P. Gerrish, Dynamics and Fate of Beneficial Mutations Under Lineage Contamination by Linked Deleterious Mutations. *Genetics* **205**, 1305-1318 (2017).
34. S. Kumar, G. Stecher, M. Li, C. Knyaz, K. Tamura, MEGA X: Molecular Evolutionary Genetics Analysis across Computing Platforms. *Mol Biol Evol* **35**, 1547-1549 (2018).
35. S. Chattopadhyay, S. Paul, D. E. Dykhuizen, E. V. Sokurenko, Tracking recent adaptive evolution in microbial species using TimeZone. *Nat Protoc* **8**, 652-665 (2013).
36. W. Li, A. Godzik, Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* **22**, 1658-1659 (2006).
37. Y. Huang, B. Niu, Y. Gao, L. Fu, W. Li, CD-HIT Suite: a web server for clustering and comparing biological sequences. *Bioinformatics* **26**, 680-682 (2010).
38. P. Rice, I. Longden, A. Bleasby, EMBOSS: the European Molecular Biology Open Software Suite. *Trends Genet* **16**, 276-277 (2000).
39. D. G. Higgins, P. M. Sharp, CLUSTAL: a package for performing multiple sequence alignment on a microcomputer. *Gene* **73**, 237-244 (1988).
40. E. H. Simpson, Measurement in Diversity. *Nature* **163**, 688 (1949).
41. A. Waterhouse *et al.*, SWISS-MODEL: homology modelling of protein structures and complexes. *Nucleic Acids Res* **46**, W296-W303 (2018).
42. D. Wrapp *et al.*, Cryo-EM structure of the 2019-nCoV spike in the prefusion conformation. *Science* **367**, 1260-1263 (2020).
43. W. Humphrey, A. Dalke, K. Schulten, VMD: visual molecular dynamics. *J Mol Graph* **14**, 33-38, 27-38 (1996).

Figure Legends

Fig. 1. Schematic representation of the diversity of spike protein variants circulating in India, using maximum likelihood-based phylogeny reconstruction. Each node represents a specific spike protein variant, while the node-size and the number inside depict the frequency of that variant. The red or green color of each arrow indicates the higher or lower stability index respectively of the S-R complex for each variant than the major ancestral variant it emerged from (either Ancestor 1 or Ancestor 2). The black arrows lead to the variants for which the docking scores could not be determined either because of the presence of at least one variation outside the available template region for docking (18, 42), or due to non-existing isolate in the lone hypothetical node with H1083Q mutation denoted by black color. This black node signifies a variant with no available isolate in our dataset, while it gives rise to two derived variants, H1083Q:R78M and H1083Q:E583D, for which representative isolates were available.

Fig. 2. Heat map distribution across four Indian states with >50 sequenced isolates based on average stability index. The average stability index for a particular state denotes the averaged value of docking scores / HADDOCK scores of S-R complexes for all circulating variants. The values of average stability index and fatality rate in Indian states are plotted to fit an exponential function (R^2=0.96).

Fig. 3. Stability index (i.e., docking score or HADDOCK score) plot of S-R complexes for spike protein variants emerging from (a) Ancestor 1 (Wuhan-Hu-1/2019 variant) and (b) Ancestor 2 (D614G variant). The blue dotted line is used as a reference to denote the stability index value for respective ancestral variants. More negative is the value, higher is the stability level. The red dotted rectangular block includes the variants that are represented by multiple isolates in our dataset (*SI appendix*, Table S2).

Fig. 4. Countrywide fatality rates (the number of deaths / the number of recovered cases) over time in India. The date of first available data (March 13, 2020), the date after which the decline of fatality rate started (April 11, 2020), and the date until which the analyzed samples were collected (May 27, 2020) are denoted by blue dotted lines. The declining fatality rate curve is best fitted by a power law function (R^2=0.95).

Fig. 1. Schematic representation of the diversity of spike protein variants circulating in India, using maximum likelihood-based phylogeny reconstruction. Each node represents a specific spike protein variant, while the node-size and the number inside depict the frequency of that variant. The red or green color of each arrow indicates the higher or lower stability index respectively of the S-R complex for each variant than the major ancestral variant it emerged from (either Ancestor 1 or Ancestor 2). The black arrows lead to the variants for which the docking scores could not be determined either because of the presence of at least one variation outside the available template region for docking(18, 42), or due to non-existing isolate in the lone hypothetical node with H1083Q mutation denoted by black color. This black node signifies a variant with no available isolate in our dataset, while it gives rise to two derived variants, H1083Q:R78M and H1083Q:E583D, for which representative isolates were available.

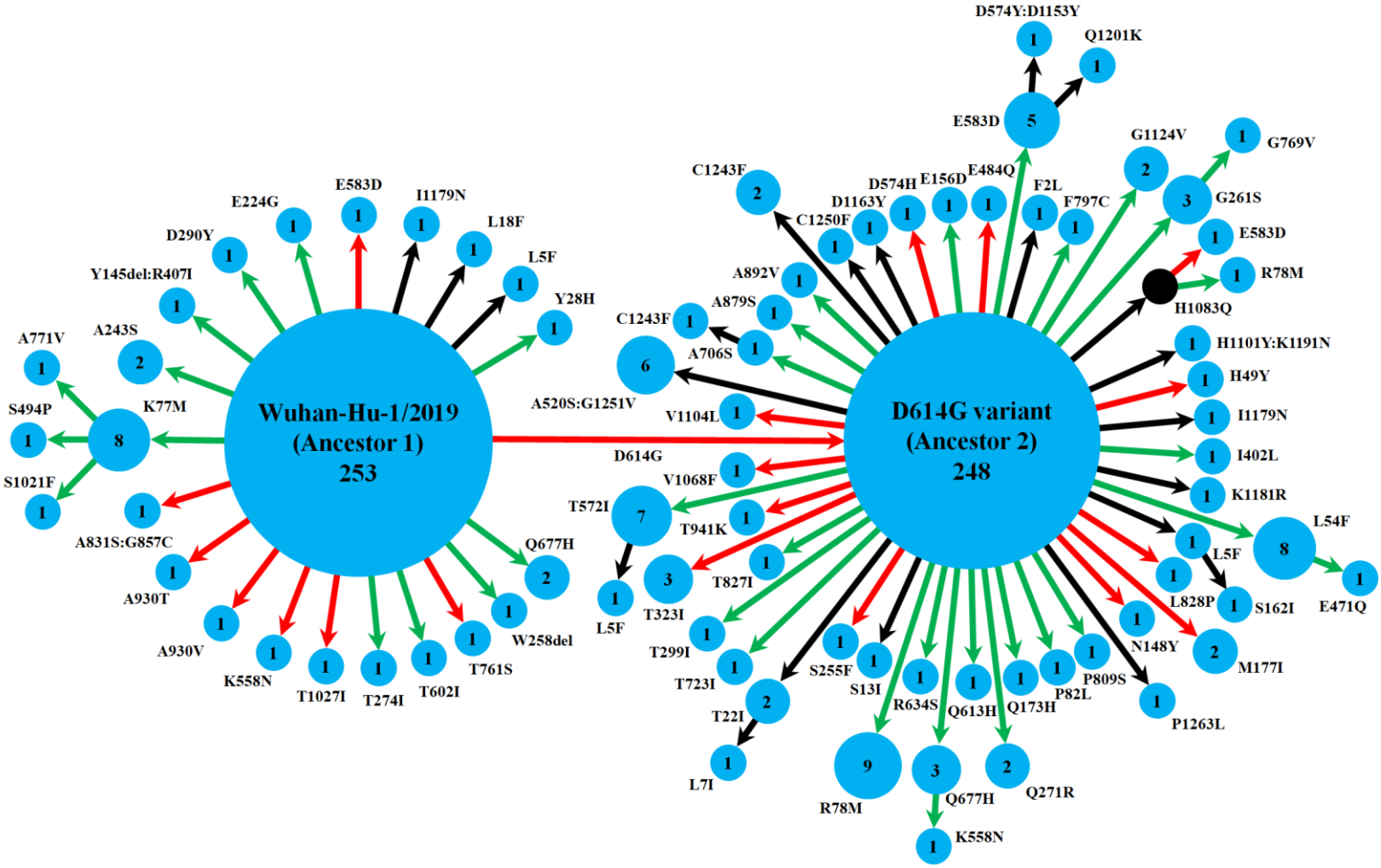

Fig. 2. Heat map distribution across four Indian states with >50 sequenced isolates based on average stability index. The average stability index for a particular state denotes the averaged value of docking scores / HADDOCK scores of S-R complexes for all circulating variants. The values of average stability index and fatality rate in Indian states are plotted to fit an exponential function (R^2=0.96).

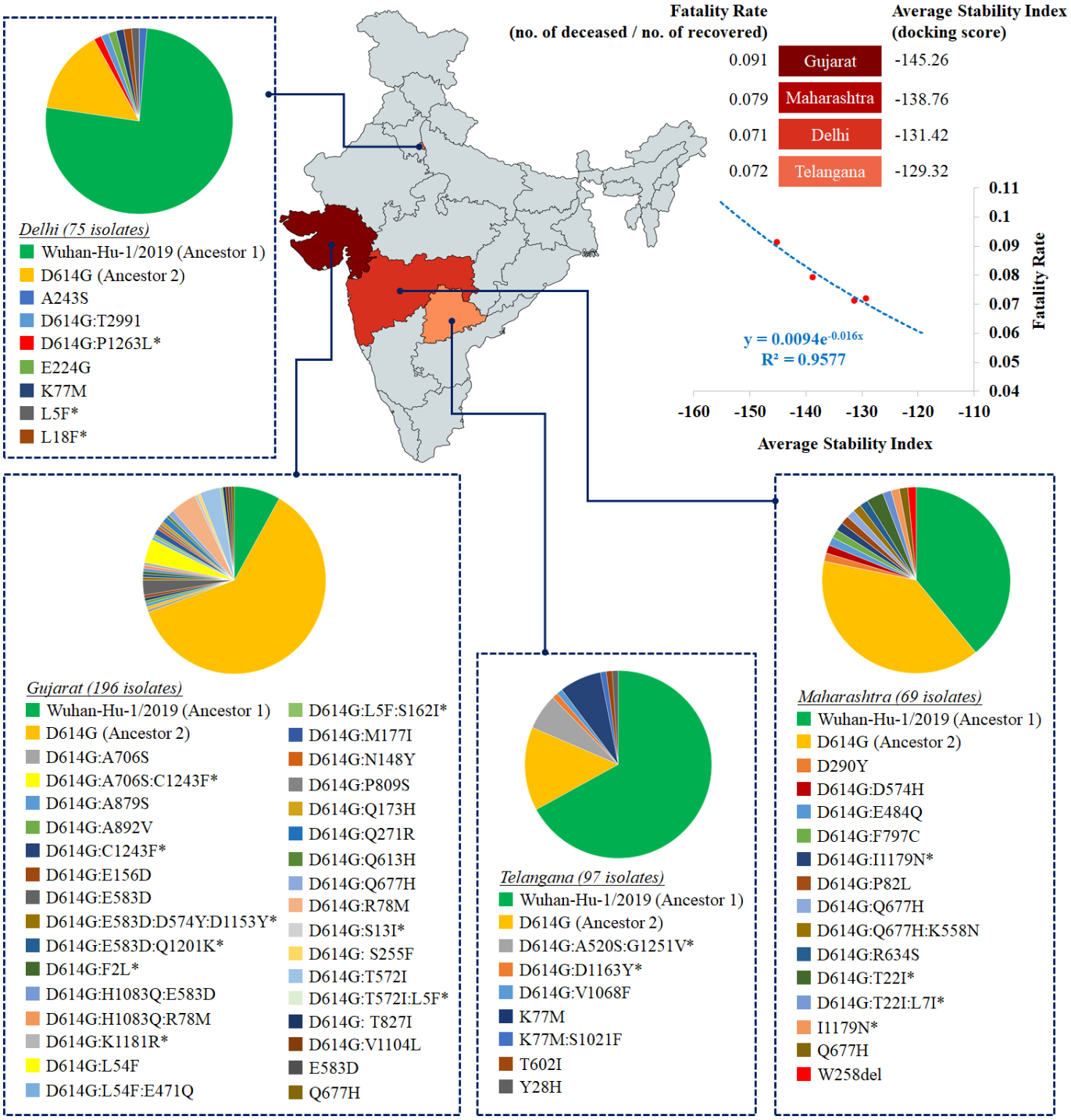

Fig. 3. Stability index (i.e., docking score or HADDOCK score) plot of S-R complexes for spike protein variants emerging from (a) Ancestor 1 (Wuhan-Hu-1/2019 variant) and (b) Ancestor 2 (D614G variant). The blue dotted line is used as a reference to denote the stability index value for respective ancestral variants. More negative is the value, higher is the stability level. The red dotted rectangular block includes the variants that are represented by multiple isolates in our dataset (*SI appendix*, Table S2).

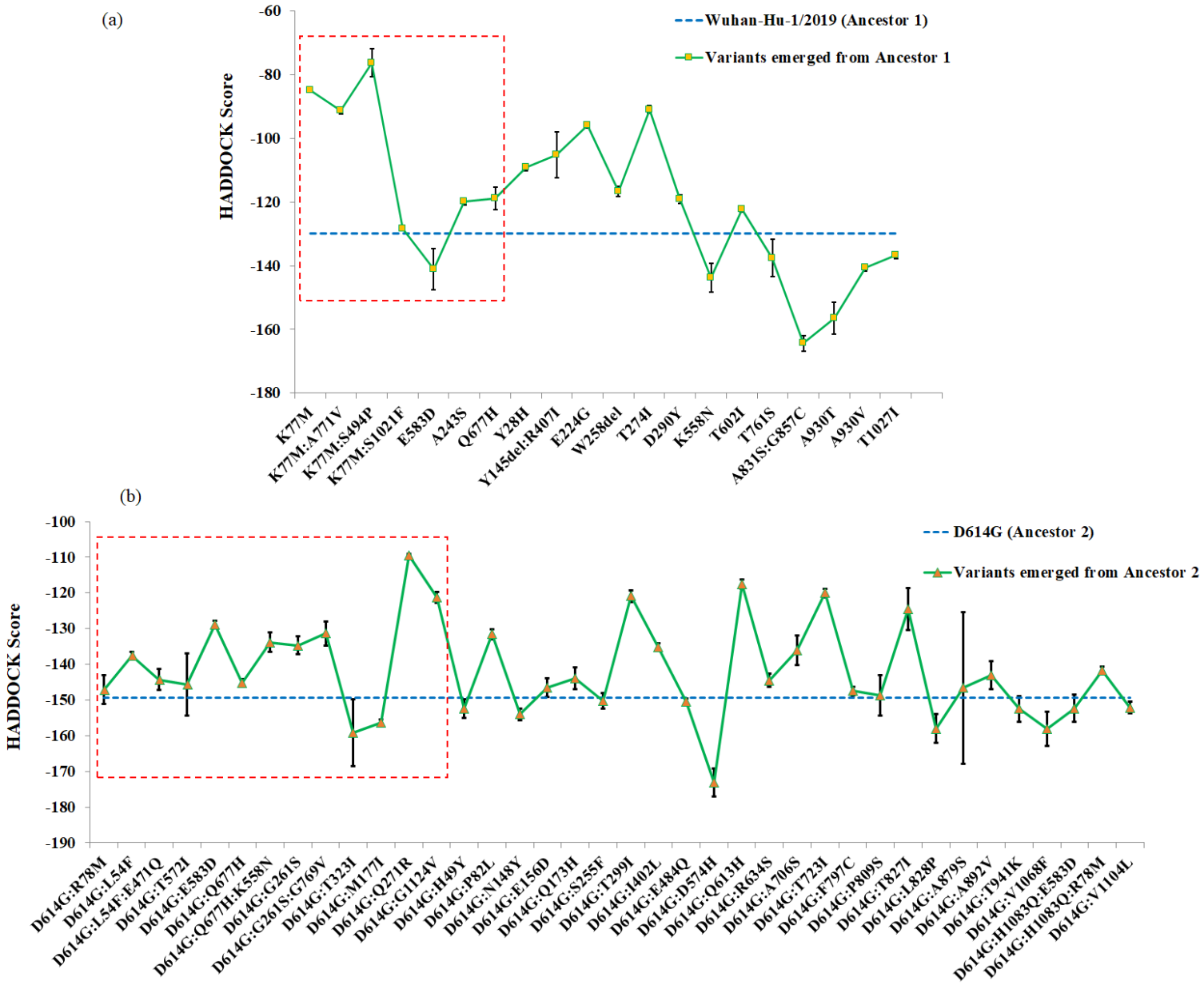

Fig. 4. Countrywide fatality rates (the number of deaths / the number of recovered cases) over time in India. The date of first available data (March 13, 2020), the date after which the decline of fatality rate started (April 11, 2020), and the date until which the analyzed samples were collected (May 27, 2020) are denoted by blue dotted lines. The declining fatality rate curve is best fitted by a power law function (R^2=0.95).