The Frequency of G614 SARS-CoV-2 Variant in India

Babu V. Bassa* and Rao M. Uppu

Author Affiliations

Department of Environmental Toxicology, College of Sciences and Engineering, 108 Fisher Hall, James L. Hunt Street, Southern University and A&M College, Baton Rouge, LA 70813

* Corresponding author, Email address: bbassa9824@gmail.com

Abstract

As reported by us and others previously (1, 2), the D614G mutation appeared in the spike glycoprotein (SPG) of the SARS-CoV-2 (the pathogen behind COVID-19) at the early stages of the pandemic and then G614 containing variant of SARS-CoV-2 became the predominant strain in most human populations across the world. However, one of the most recent reports from India (3) stated the incidence of G614 to be only 26% in the Indian population. This report is contradictory to the information available through the GenBank (4) SARS-CoV-2 sequence deposits made by various laboratories from India. The above stated report currently circulating in the Indian media is likely to create a public perception that the Indian strain is less contagious and such a notion could be harmful to people's welfare. In view of this concern we have re-evaluated, updated and recalculated the incidence of the G614 variant in the Indian population by analyzing 395 Indian SARS-CoV-2 genomic sequences available in the GenBank as of June 26, 2020. In our analysis we have categorized the samples by the month in which the samples were collected. We have used an alignment-free software tool named *Compare* (5, 6), and the Basic Local Alignment Search Tool (BLAST) (7) in the present analysis. We finally inspected each of the 395 sequences physically for the presence of aspartic acid (D) or glycine (G) at the 614th position of the spike glycoprotein. We analyzed an Australian cohort in parallel for comparison. We have found that the prevalence of G614 variant in the Indian samples for the month of June 2020 is 90.6%. The trends are similar with the Australian samples.

Keywords

Alignment-free software tool, Coronavirus, COVID-19, D614G mutation, Sarbecovirus, SARS-CoV, SARS-CoV-2, Spike glycoprotein

Introduction

The first cases of the present Severe Acute Respiratory Syndrome (SARS) outbreak was reported in December 2019 from the Wuhan City, China. The pathogen behind the disease, a coronavirus strain, was quickly identified and designated as SARS-CoV-2 to distinguish it from the SARS-CoV strain of the 2002 outbreak (8). The World Health Organization (WHO) later declared the outbreak as a pandemic and gave it the name Coronavirus Disease 2019 or COVID-19. The virus is so far responsible for about 17 million infections and 675k fatalities across the world (9). The virus is a positive strand RNA virus, and thus has exhibited relatively higher rates of mutation. One such mutation leading to the substitution of aspartic acid (D) by glycine (G) at the 614th position of the spike glycoprotein (SPG) of SARS-CoV-2 is of special interest because the variant containing this mutation (D614G) spread quickly and became predominant in most geographical COVID–19 hotspots across the world. As reported by us recently (2), the mutation was completely absent in the GenBank annotated sequences, from China with the sample collection dates of December 2019 and January and February 2020. First samples containing D614G appeared in late March in China and many other countries simultaneously as per the GenBank records. We have previously demonstrated that the 614th position is located in an extraordinarily conserved and highly hydrophobic 11-amino acid motif (11-aa) in the SGP of SARS-CoV-2 virus (2). By April 2020 greater than 50% of sequenced samples from various geographical hotspots of SARS-CoV-2 contained the G614 variant. In our previous analysis (2) we found that over 92% of the SARS-CoV-2 sequences annotated in the GenBank by the Indian laboratories contained the G614 variant. However, a recent report from a research Institution in India has claimed that the frequency of G614 in India was only 26% (3). In view of this contradiction we re-analyzed a much bigger cohort of 395 sequences of Indian origin, annotated in the GenBank. All these virus samples had been collected, sequenced and deposited in the GenBank by various laboratories in India. After the initial analysis with certain software tools we physically verified each of the 395 sequences for the presence of either D or G at the 614th position. We analyzed over 400 Australian sequences in parallel for comparison. We still find that like the Australian samples, over 90% of the Indian samples with the sample collection dates falling in the month of June 2020, contain G614 variant. Our findings are consistent with our previous report on the general trends in India and across the world (2).

Materials and Methods

All the SARS-CoV-2 spike glycoprotein sequences used in this study were obtained from the GenBank. NCBI’s SARS-CoV-2 Resources Database (10) was used to sort the sequences based on sample collection dates and the geographical locations.

Many of the sequences were analyzed with the software tools *Compare* (5), and the Basic Alignment Search Tool (BLAST) (7). *Compare* compares the query sequences by identifying common permutations larger than two amino acids between them. As reported previously (2), our analysis using compare identified an eleven amino acid highly conserved, hydrophobic motif into which D614G is embedded. The motif is ***vavlyq<u>d</u>vnct*** which is present practically in all coronavirus strains belonging to the Sarbecovirus subgenus (SARS-related strains) as an identical permutation. This finding gave us a handle to use BLAST in our analysis, because queries with ***vavlyqdvnct*** or ***vavlyqgvnct*** produced distinct scores in the BLAST analysis. We analyzed SGP sequences for the presence of D614 or G614 with respect to time and geographic locations. We also used BLAST to determine the total incidence of the two variants till July 26th, 2020, in the pandemic. Finally we inspected all the relevant GenBank annotations for the presence of D614G mutation and categorically recorded the accession numbers in Excel files, which will be made available to the parties interested, upon request.

Results

There are only seven SARS-CoV-2 sequences available for the month of March from India in the GenBank and four of them are of D614 type. Only the data points represented by sufficiently big sample size are included in Figure 1. As shown in the cases of both India and Australia, there are significant proportions of D614 variants in the samples with April 2020 collection dates. The proportion of D614 variant has fallen dramatically in the May 2020, samples and it is nearly absent in the samples from June 2020. As reported by us previously this was the general trend at all geographical locations (2). Our BLAST search with ***vavlyqgvnct*** retrieved 1282 SGP sequences containing ***vavlyqgvnct*** in its full form. Similarly, the BLAST search with ***vavlyqdvnct*** retrieved 615 SGP sequences containing ***vavlyqdvnct*** as an identical permutation. This translates to an all-time G614 variant proportion of 67.6 %. A global search of GenBank SGP sequences for the month of July 2020, using the NCBI SARS-CoV-2 resources database found only one SARS-CoV-2 isolate containing D614 and this sample originated from Egypt. Our investigations also identified globally 19 isolates containing a substitution of serine (S) by glutamic acid (N) at the 477th position (S477N) of SGP. Interestingly 16 of the 19 samples originated from Australia and the rest three are from the United States. There were two samples from the Kerala state of India with the sample collection month of January, 2020 and both of them of D614 type.

Discussion

The DNA and RNA replication errors are the main sources of genomic mutations in nature. These errors are more likely to occur frequently in RNA viruses like the coronaviruses, where the proof-reading mechanisms are either absent or less robust. When an error in the RNA replication results in the replacement of an amino acid, the variant containing the substitution gets amplified only if the substitution has positive effect on the overall reproductive (in this case replicative) fitness of the virus. Otherwise, many of those sequences are like needles in a haystack when they first appear, and will be lost quickly. The influences on the fitness include effects on the replication efficiency, invasiveness, outside survival, and spread of the virus. The general principles governing the spread of a mutation with a positive influence on the reproductive fitness was well developed in the work of Sewall Wright an American mathematician (11). These principles apply to all situations where a mutation gets fixed in populations including the virus populations.

As reported by us previously, the G614 variant of the SARS-CoV-2 was completely absent in the samples collected in December 2019 and in January and February 2020 in China. The first G614 variants appeared in the Chinese samples in late March as per the GenBank records. They were only three at that time. Interestingly around the same time G614 variant surged at several geographic locations across the world. As shown in Figure 1 the G614 variant became the predominant variant through the months of March, April, May, and June 2020 at almost all global locations. A global search through NCBI SARS-CoV-2 Resources Database has revealed only one isolate of the D614 kind annotated for the month of July in the GenBank and the isolate is from Egypt. Otherwise, D614 variant of the SARS-CoV-2 is absent at all other hotspots. The rate at which the G614 variant has surged globally suggests that the D614G mutation conferred selective advantage on the resulting G614 variant. As shown previously (2) the 11-aa motif in which the mutation is located is present in all *Sarbecovirus* strains in an identical form except in the new D614G SARS-CoV-2 variants where D614 is substituted by G. These *Sarbecovirus* strains include all human SARS related strains and the civet, and bat-SARS-like strains. The degree of conservation again underscores the importance of the 11-aa motif to the reproductive fitness of the virus. As also shown previously, replacement of D by G only increased the hydrophobicity of already densely hydrophobic motif. Furthermore, the motif is in close proximity to both the receptor binding domain and the proteolytic cleavage site in the S1 subunit of the SGP. This "lipid raft" like 11-aa region in the SGP molecule is likely to offer many advantages to the virus. Firstly, they would enhance the attachment of the virus particles to the plasma membranes of the host cells, in the immediate vicinity of the virus particles. Secondly, purely hypothetically, we suggest that the hydrophobic motifs help the virus to get attached to new hydrophobic surfaces, like the plastic surfaces, outside the host's body before the virus travels to a new host, a kind of hitch-hiking.

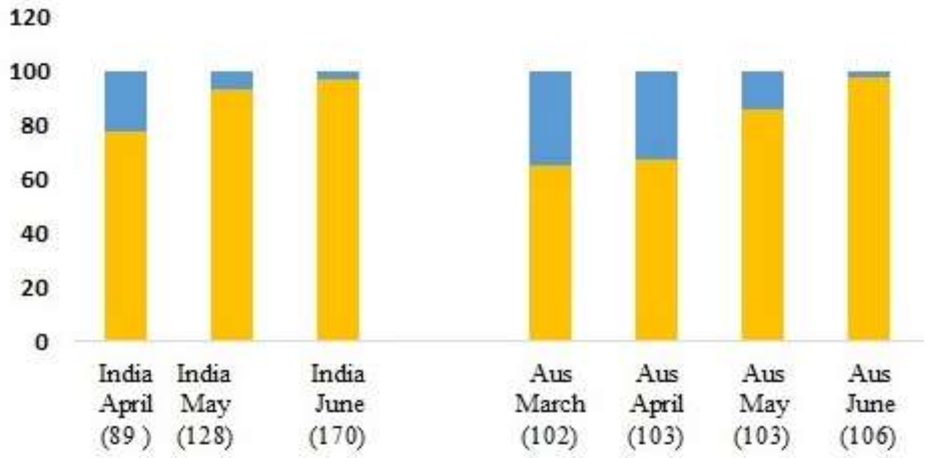
In the current investigation employing bioinformatics approach we analyzed 395 genomic sequences from India as they are available in the GenBank. As of July 26, 2020 there were 409 SARS-CoV-2 genomic annotations from India available in the GenBank. We analyzed 395 of them after eliminating the incomplete ones. Initially a smaller number of samples were analyzed with *Compare*, but finally, for confirmation all the 395 SGP sequences were physically inspected for the presence or absence of the mutation. Information was also collected in parallel on 344 Australian isolates of SARS-CoV-2. In both the cases the information was collected month-wise with sample collection date as the basis, from the NCBI SARS-CoV-2 Resources database. The number of samples in each category is indicated in the parentheses in Figure 1. The G614 variant accounted to about 50% of the samples for March in the Australian samples and it increased gradually to 97.2% by 30th June, 2020. There are only seven isolates available for the month of March, 2020 from India and three of them are of the G614 variants. Samples from India with the collection dates falling in April, May, and June 2020 contained 77.5%, 93%, and 96.5% of the G614 variant respectively (Figure 1). Additionally, the two samples from Kerala, India with the collection dates of January, 2020 are of D614 type. The first cases of COVID-19 in India appeared in the state of Kerala and the Kerala state government earned worldwide reputation for containing the outbreak effectively. However, it is now clear that at the beginning of the pandemic both China and the Kerala state in India dealt with a less contagious variant of SARS-CoV-2. As indicated under the results section our BLAST analysis shows that the proportion of

G614 variant in the total samples (covering the entire pandemic duration till July 30, 2020) recorded in the GenBank is 67.6%. This observation prompted us to evaluate the current overall prevalence of the D614G mutant SARS-CoV-2. For this purpose, we searched the NCBI database and inspected all globally available sequences for the month of July 2020 for the presence of the G614 variant. We found that 146 of the 147 samples so retrieved carried the G614 variant. It is however, necessary to analyze samples for the month of August 2020 as well to firmly conclude the disappearance of the original D614 SARS-CoV-2 variant.

Based on our analysis we conclude that, the most recent report from India (3) that has claimed only 26% prevalence of G614 in the Indian population is completely at odds with the genomic data presented to the GenBank by various Indian laboratories. The report has been circulated in the Indian media creating an inaccurate sense of security among the Indian masses, that the Indian variant is less infectious. We suggest that the authors re-analyze their data and make all their sequences available to the rest of the world through a publicly accessible database like the GenBank.

Competing interests: None.

Figure 1. Frequency of G614 (Yellow) vs D614 (blue) in India and Australia



Proportions of G614 and D614 variants of SARS-CoV-2 in the COVID-19 patients of India and Australia. All the sequences were derived from the Gen Bank and categorized based on the month the samples were collected. The year is 2020. Number of samples in each case is given in the parentheses. All the sequences were physically verified for the presence of glycine (G) or aspartic acid (D) at the 614 position of the spike glycoprotein. Australian data was chosen for comparison bcause comparable number of genomic sequences are available in the Gene Bank from Australia

References

1. Korber, B., Fischer, W.M., Gnanakaran, S., *et al*. (2020). Tracking changes in SARS-CoV-2 Spike: Evidence that D614G increases infectivity of the COVID-19 virus. *Cell* **182**, 1-16. Published online 2 July 2020; doi: https://doi.org/10.1016/j.cell.2020.06.043

2. Bassa, B. and Brown, O. (2020). D614 Residue belongs to a highly conserved peptide motif in *Sarbecovirus* group and the D614G mutation of SARS-CoV-2 spike protein

appeared once in SARS-CoV. *Preprints* 2020070488; Published online: 21 July 2020; doi: 10.20944/preprints202007.0488.v1.

3. Kumar, P., Pandey, R., Sharma, P., et al. (2020). Integrated genomic view of SARS-CoV-2 in India. *bioRxiv* preprint doi: https://doi.org/10.1101/2020.06.04.128751. Published online 4 June 2020.

4. National Center for Biotechnology Information (NCBI) [Internet]. Bethesda (MD): National Library of Medicine (US), National Center for Biotechnology Information; [1988] – [cited 2020 July 2020]. Available from: https://www.ncbi.nlm.nih.gov/

5. Babu, B. V. and Brown, O.R. (2020). Comparative analysis of coronaviridae nucleocapsid and surface glycoprotein sequences. *Front. Biosci*. (Landmark Edn.) **25**, 1894-1900. Published online 31 May 2020; doi: 10.2741/4883

6. B.V. Bassa, B.V. and R.M. Uppu, R.M. (2020). SARS and HIV inhibitory peptides with therapeutic potential against Covid-19 [eLetter response to "Rapid repurposing of drugs for Covid-19", R.K. Guy *et al*., *Science* 368, 829-830, 2020] Published online June, 6 2020.

7. https://blast.ncbi.nlm.nih.gov/Blast.cgi?PAGE=Proteins, Accessed July 25, 2020

8. Anderson, R.M., Fraser, C., Ghani, A.C. *et al*. (2004). Epidemiology, transmission dynamics and control of SARS: The 2002-2003 epidemic. *Phil. Trans. R. Soc. Lond. B* **359**, 1091-1105. Published online 15 June 2004; doi: 10.1098/rstb.2004.1490.

9. https://covid19.who.int/?gclid=Cj0KCQjwyJn5BRDrARIsADZ9ykE1w2D4mJ9woB3SzK1ay_tu26qsIrC96L8TJQMREX1VBrbvV79wy1YaAltTEALw_wcB. Accessed August 1st 2010.

10. https://www.ncbi.nlm.nih.gov/labs/virus/vssi/#/virus?SeqType_s=Nucleotide&VirusLineage_ss=SARS-CoV-2,%20taxid:2697049. Accessed July 28th 2020.

11. Wright, S. (1942). Statistical genetics and evolution. *Bull. Amer. Math. Soc*. **48**, 223–246. doi:10.1090/S0002-9904-1942-07641-5. MR 0006700.