 doi:10.20944/preprints202007.0488.v1

Brief Report: 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

[1*]Babu V. Bassa, Ph.D.
[2]Olen R. Brown, Ph.D. (Emeritus)

Author Affiliations:

1. Department of Environmental Toxicology, 108 Fisher Hall, P.O. Box 9264, Southern University and A&M College, Baton Rouge, LA 70813.

2. Dalton Cardiovascular Research Center, University of Missouri, Columbia, MO 65211

*Correspondence to: bbassa9824@gmail.com

Keywords

SARS-CoV, SARS-CoV, COVID-19, Sarbecovirus, D614G, Spike glycoprotein, Coronavirus, Alignment-free.

Summary

Conservation history of D614 residue is valuable in predicting the consequences of D614G mutation in the SARS-CoV-2 spike glycoprotein (SGP). We report here that the D614 belonged to an extraordinarily conserved, densely hydrophobic eleven amino acid peptide-motif ***vavlyqdvnct*** (11-aa), in the *Sarbecovirus* group and the variant carrying ***vavlyqdvnct*** had appeared in Chinese samples and became predominant in several geographical hotspots in late March, 2020. Interestingly a 2009 annotation of SARS-CoV contained the same mutation.

Background

A genomic mutation resulting in the substitution of aspartic acid by glycine at the 614th position, of SARS-CoV-2 spike glycoprotein was identified by scientists in the early stages of the COVID-19 pandemic. An abrupt increase in the frequency of SARS-CoV-2 variant carrying D614G, has sounded alert recently with the speculation that the substitution confers selective advantage on the replication and spread of the virus[1]. In this regard knowledge of the conservation of aspartic acid at the position 614 (D614) is valuable in evaluating the impact of the mutation on the transmission and virulence of the virus.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 21 July 2020 doi:10.20944/preprints202007.0488.v1

Several strains of SARS-CoV (coronavirus from the 2002 outbreak), SARS-CoV-2 (the pathogen behind COVID-19), human ExoN 1, and witc - MB, civet, pangolin, and bat SARS-Like coronaviruses are all grouped under *Sarbecovirus* in the gene bank (NCBI)[2]. In order to sketch the conservational map of D614, we scanned the SGP sequences of the *Sarbecovirus* strains with an alignment-free software tool (*Compare*) described previously[3, 4]. We have found that D614 is located in an 11 amino acid motif of the SGP, and the motif is present in practically 100% of the *Sarbecovirus* strains as an identical amino acid permutation except in the currently predominant SARS-CoV-2 variants and one 2009 isolate of SARS-CoV.

Methods

About 1000 spike glycoprotein sequences belonging to *Sarbecovirus* group were used in the analysis. They were obtained from the NCBI database. Many of the sequences were analyzed by *Compare* as referenced earlier. Many other sequence annotations were physically inspected for the presence of the mutation as well as the sample collection dates.

The hydropathicity indices were calculated using our custom made software program into which the Kyte and Doolittle[5] amino acid side chain hydrophobicity values were incorporated.

Results and Discussion

Compare reports common amino acid permutations greater than two amino acids long in any given pair of queried protein sequences. As a non-alignment method, it is non-subjective, and common motifs can be used as markers for tracing mutations. The alignment-free comparison of SARS-CoV and SARS-CoV-2 is presented in Figure 1. As shown, D614 is part of ***vavlyqdvnct,*** an 11 amino acid peptide motif present at 608 location in the SGP of SARS-CoV-2. The same motif is present at 594 in SARS-CoV with “D” being at 600. Our identification of the ***vavlyqdvnct*** in SARS-CoV is based on the assumption that the probability of the 11 amino acid permutation to have been formed independently in these two sequences is infinitely small and that they both originated from a common ancestor. Tracing the mutation using a molecular signature is important in view of the differences in the sizes of SARS-CoV and SARS-CoV-2 spike glycoproteins. We next surveyed over 1000 *Sarbecovirus* SGP sequences, including SARS-CoV-2, for the presence of ***vavlyqdvnct*** using *Compare*. Compare returns whole or parts of the motif if present in the larger sequence against which it is queried. Thus as shown in Figure 2, the ***vavlyqdvnct*** motif is present as an identical permutation in all of the *Sarbecovirus* strains except the D614G variants of SARS-CoV-2 and one variant of SARS-CoV (Accession: FJ882963), where the aspartic acid is substituted by glycine in the ***vavlyqdvnct*** motif resulting in ***vavlyqgvnct.*** The frequencies of ***vavlyqdvnct*** and ***vavlyqgvnct*** in the *Sarbecovirus* group including SARS-CoV-2 is shown in Figure 2.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 21 July 2020 doi:10.20944/preprints202007.0488.v1

The implication of the unprecedented conservation of ***vavlyqdvnct*** is that the D614G mutation would not have achieved fixation unless the mutation has conferred a superior degree of additional fitness (in terms of the copy numbers) on the virus. The exact role played by ***vavlyqdvnct*** is not clear but ***vavlyqdvnct*** is highly hydrophobic and the mutation increases the hydrophobicity of this block even further by about 25%. Hydropathicity indices around the ***vavlyqdvnct*** region determined by the Kyte and Doolittle method[5] are depicted in Fig 1b. Hydropathicity index is proportional to the total amino acid side chain hydrophobicity of proteins and peptides. It is notable that the 11-aa is located in close proximity to both the receptor binding domain and the protease cleavage sites of SGP in both SARS-CoV and SARS-CoV-2. The densely hydrophobic regions are likely to increase the binding of the virus to the host cell plasma membrane. The same can also improve the adherence of the virus to plastic surfaces outside the host's body. Our survey also revealed that the SARS-CoV genomic sequence annotated in 2009 in the GeneBank with the accession number, FJ882963, carried D/G mutation in the ***vavlyqdvnct*** motif with the resultant ***vavlyqgvnct.*** This strain is likely to be useful in the *in vitro* determination of the role played by the D614G switch.

We have scanned several SGP sequences available in the gene bank to determine the frequency of ***vavlyqgvnct*** in SARS-CoV-2 strains from different geographical locations. D614G is completely absent in the Chinese SARS-CoV-2 samples of December 2019, and January 2020, and ***vavlyqgvnct*** appeared in three Chinese samples with the collection date of 3-24-2020 (Accession numbers: MT407656, MT 407657, MT407659). As shown in Figure 2, 86% of the New York samples carried "D614G' whereas only 43.7% of California samples carried it. The trend is similar with European and Australian genomes (Figure 2). Most of these samples were originally collected in late March and April, 2020 as per the GeneBank annotations. In the case of Louisiana (USA), the sample collection dates skewed more towards April, 2020. In the case of India where the infection rates seem to be peaking currently the sample collection dates spanned from early April to mid-June, 2020.

Conclusions

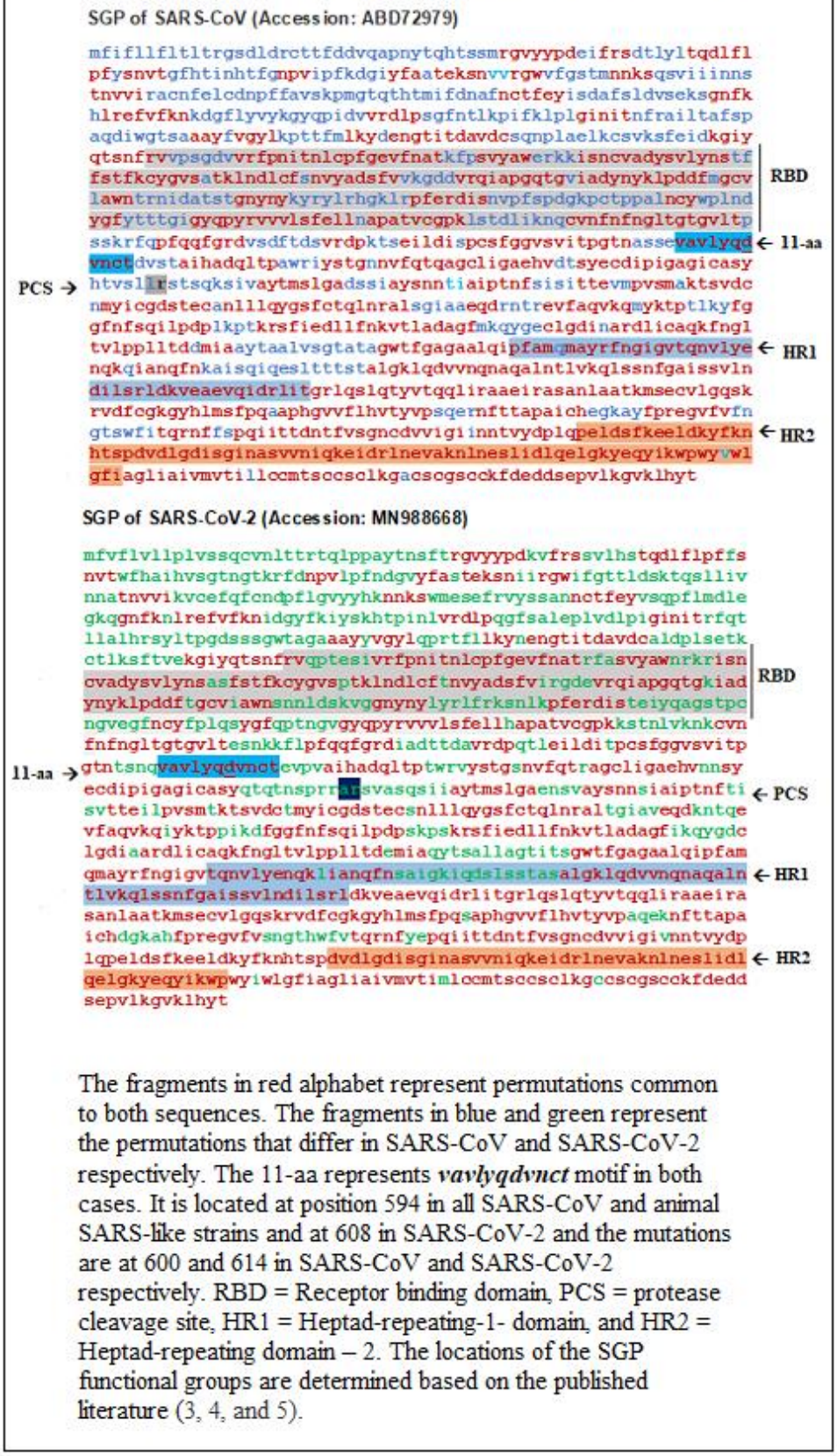
In conclusion the mode of action of the D614G is likely through its association with the highly conserved and highly hydrophobic 11- amino acid motif described in the present report. The more plausible explanation is that the mutation originated in China in early 2020 (Jan/Feb) and it is mechanistically responsible for the rapid spread of the virus. Evolutionary principles are incompatible with the idea that the same mutation redundantly occurred at multiple geographical locations in such a short time span. The mutation must have originated much early in China in order to appear in the sequenced samples in late March. The frequencies of the mutation do not seem to correlate with the mortality rates across the world. For example, Japan and India have the highest frequencies of D614G as per the GeneBank annotations, however their mortality rates are some of the lowest as per the World Health Organization records[6]. The relationship of the

 doi:10.20944/preprints202007.0488.v1

mutation to the spreading rates of the virus probably could never be solved because of widely varied rates of testing and the fact that the original D614 SARS-CoV-2 lasted only for a short time in the course of the pandemic. We have not attempted to statistically correlate the fatality rates in view of non-availability of the necessary data collected under uniform conditions.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 21 July 2020 doi:10.20944/preprints202007.0488.v1

Fig 1. An alignment-free comparison of SARS-CoV and SARS-CoV-2



The fragments in red alphabet represent permutations common to both sequences. The fragments in blue and green represent the permutations that differ in SARS-CoV and SARS-CoV-2 respectively. The 11-aa represents ***vavlyqdvnct*** motif in both cases. It is located at position 594 in all SARS-CoV and animal SARS-like strains and at 608 in SARS-CoV-2 and the mutations are at 600 and 614 in SARS-CoV and SARS-CoV-2 respectively. RBD = Receptor binding domain, PCS = protease cleavage site, HR1 = Heptad-repeating-1- domain, and HR2 = Heptad-repeating domain – 2. The locations of the SGP functional groups are determined based on the published literature (3, 4, and 5).

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 21 July 2020 doi:10.20944/preprints202007.0488.v1

Figure 2. A. Frequencies of *vavlyqdvnc* (blue) vs *vavlyqgvnct* (yellow) across the sarbecovirus group

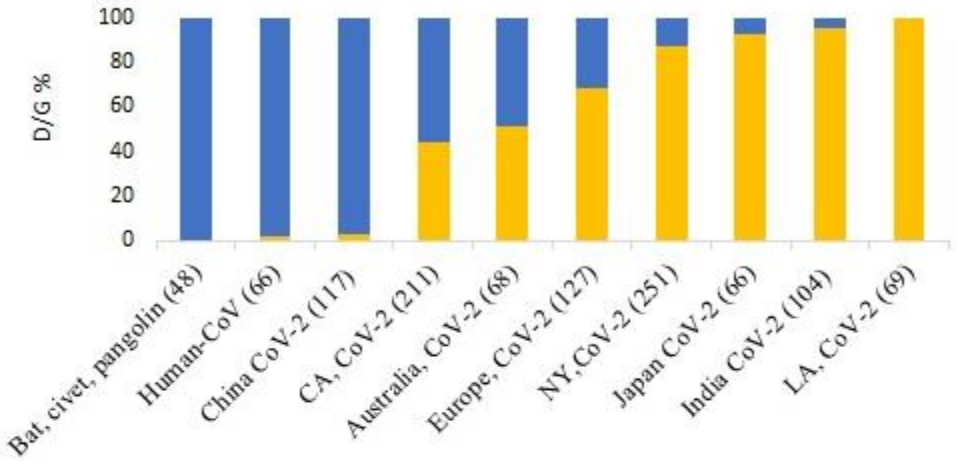


Fig 2. B. Hydropathic index of the conserved 11-aa domain and its vicinity

-vitpgtnassevavlyqdvnctdvstaihadql- CoV
(-1.5) (8.2) (0.9)

-vitpgtntsnqvavlyqdvnctevpvaihadql- CoV2
(-6.7) (8.2) (5)

-vitpgtntsnqvavlyqgvnctevpvaihadql- CoV2 D614G
(-6.7) (11.3) (5)

A. The *vavlyqdvnct* (11-aa) motif is highly conserved in the *Sarbecovirus* group. A variant of this motif *vavlyqgvnct* carrying D614G mutation appeared only in SARS-CoV-2 (CoV-2) strains. One isolate of SARS-CoV (CoV) (Accession: FJ882963) annotated in 2009 has been identified by us to carry the same mutation. The number of samples analyzed in each case is shown in parentheses. Most of the SARS-CoV-2 sample collections were carried out in March/April 2020 as per the genebank records and coincided with the peak mortality periods in each geographical location. For Europe data are combined from France, Italy, and Spain.

B. The hydropathicity indexes of 11-aa, and 11 amino acids preceding and 11 amino acids following 11-aa are calculated by the Kyte and Doolittle method. The D614G mutation likely has relevance mainly as a component of a conserved densely hydrophobic peptide domain.

References

1. Korber, B. et al. on behalf of Sheffield COVID-19 Genomics Group, Tracking changes in SARS-CoV-2 Spike: evidence that D614G increases infectivity of the COVID-19 virus, *Cell* (2020), doi: https://doi.org/10.1016/j.cell.2020.06.043

2. NCBI, 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/

3. Babu, B. V., Brown, O.R. Comparative analysis of coronaviridae nucleocapsid and surface glycoprotein sequences. Front. Biosci. 25, 1894-1900 (2020).

4. B.V. Bassa and R.M. Uppu (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 (6493): 829-830, 2020] Published online June 6, 2020.

5. Kyte, J. And Doolittle, R.F. (1982). A simple method for displaying the hydrophobic character of a protein. J. Mol. Biol. 157: 105-132.

6. https://covid19.who.int/?gclid=Cj0KCQjw3s_4BRDPARIsAJsyoLO90ri_1rDZT5Ffxl8oY0LkT4N9wXAgZatG3EN50X1ce2UvBbn4uCcaAivWEALw_wcB

7. Berend, J.B. et al. Severe acute respiratory syndrome coronavirus (SARS-CoV) infection inhibition using spike protein heptad repeat-derived peptides. PNAS. 101: 8455-8460 (2004).

8. Stephanie, B. et al. Cleavage and Activation of the Severe Acute Respiratory Syndrome Coronavirus Spike Protein by Human Airway Trypsin-Like Protease. J Virol. 85: 13363–13372 (2011).

9. Xia, S. et al. Inhibition of SARS-CoV2 (previously 2019-nCoV) infection by a highly potent pan-coronavirus fusion inhibitor targeting its spike protein that harbors a high capacity to mediate membrane fusion. Cell Research 30: 343 -355 (2020)