Medical Research Archives

ESMED
European Society of Medicine

OPEN ACCESS

Published: August 31, 2022

Citation: Huang J, Mi Y, et al., 2022. SARS-Cov-2 Variants: Biological and Mathematical Considerations for Nomenclature, Medical Research Archives, [online] 10(8). https://doi.org/10.18103/mra.v10i8.2939

DOI
https://doi.org/10.18103/mra.v10i8.2939

ISSN: 2375-1924

RESEARCH ARTICLE

SARS-Cov-2 Variants: Biological and Mathematical Considerations for Nomenclature

Jie Huang[1], **Yi Mi**[2], **Junxi Li**[3], **Gary R McLean***[4,5]

[1] School of Public Health and Emergency Management, Southern University of Science and Technology, Shenzhen, Guangdong, China
[2] Beijing No.4 High School International Campus
[3] Shenzhen College of International Education, Shenzhen, Guangdong, China
[4] School of Human Sciences, Cellular Molecular and Immunology Research Centre, London Metropolitan University
[5] National Heart and Lung Institute, Imperial College London, London, UK

* g.mclean@londonmet.ac.uk

ABSTRACT

Coronavirus (CoV) is one of the most widely used words during the past two years. If it were announced that Delta CoV only affects animals such as pigs and wigeons while Omicron CoV does not even exist, surely people would be offended and question the credibility of whoever stated this. But both statements are true, scientifically. Of note, it was stated Delta CoV and Omicron CoV, not Delta variant or Omicron variant of SARS-CoV-2. Such potentially confusing naming of a globally important virus therefore warrants further analyses. At the subfamily level, CoVs are divided into four genera (Alpha, Beta, Gamma, Delta) and only viruses of the Alpha and Beta branch infect humans. Now that the Omicron variant of SARS-CoV-2 have taken over from the Delta variant globally, the issue of the double use of genus labels (Alpha, Beta, Gamma, Delta) for variant naming is mitigated. However, we can still pause and ponder whether the Greek symbols alone are indeed ideal for labeling waves of SARS-CoV-2 variants. Here we propose additional criteria for naming of variants that considers specific biological and molecular characteristics of the virus-cell interaction. Our aim is to define a biological and structurally defined metric that can be used to distinguish SARS-CoV-2 variants interactions with host cells. This metric could find utility with numerous human viruses and provide an additional parameter for improved naming of viruses.

Keywords: SARS-CoV-2, variants, structural biology

A brief history of SARS-CoV-2 naming

The new respiratory virus first reported in Wuhan, China during late 2019 and known to cause an atypical pneumonia was originally named 2019-nCoV. It was then renamed SARS-CoV-2 based on the genomic similarity to the severe acute respiratory syndrome (SARS) virus that first appeared in 2003. SARS-CoV-2 has efficiently spread worldwide and caused the Covid-19 pandemic at an unprecedented scale. As the pandemic progressed, the open sharing of genomic data triggered a plethora of bioinformatics tools for standardization of lineage nomenclature to characterize the growing number of strains and variants of this new RNA virus. The most frequently used lineage assignment and data visualization tools such as Global Initiative on Sharing All Influenza Data (GISAID)[1], Nextstrain[2] and Pango[3] have greatly aided this process. However, it has been challenging as development was required during an evolving pandemic with sometimes limited data.

The existing virus labels from these tools, remaining in force for the foreseeable future, are confusing to scientists and public health professionals, as well as to the public and politicians. Considering Pango, early genome sequences were designated using English letters. Lineage A, the ancestral type, being represented by Wuhan/WH04/2020 (first sampled 5/01/2020) is the original lineage. This was quickly replaced by lineage B represented by Wuhan-Hu-1 (first sampled 26/12/2019) that forms the basis of all current variants including lineages C and D. The non-chronological lineage identification dates represents the dynamic situation early in the Covid-19 pandemic. Subsequent lineages derived from these reference sequences were assigned an Arabic number. Thus, B.1, B.2 etc. appeared. Subsequently, more numbers and letters have been added to represent further lineage and clade development (e.g. B.1.1.7, P.1, C.37, AY.4.2). Whilst labels like this are relatively simple and are easy to follow, unfortunately they have limited biological meaning despite the relationships to viral genome sequence. For example, Pango labels for Delta and Omicron variants are B.1.617.2 and B.1.1.529 respectively. One could easily make a simple mistake and assume that B.1.617.2 came after B.1.1.529, despite clear knowledge currently that Omicron is more recent. Other labels such as GH/501Y.V2 do carry a biological meaning (i.e. GH clade amino acid 501 of spike mutated to Y), albeit rather limited, especially when many variants, and principally the Omicron variant, carry many more spike mutations[4].

To ease the public's learning curve and to scientifically categorize the appearance of new SARS-CoV-2 genomic sequences, the World Health Organization (WHO) introduced the simple terminologies of variants of interest (VOI) and variants of concern (VOC) in late 2020[5]. On 31st May 2021, it announced a variant naming scheme based on the Greek alphabet, with the intention to simplify, quell confusion and avoid geographical stigmas associated with the location of new variant first identification. As shown in **Table 1**, these overlapping naming systems, although providing genomic relationships, have created a complex network of interchangeable names. Hence the WHO simplification to Greek letters has proved of vital importance (**Figure 1 left panel**). Some complexities with the naming systems in use can be summarized as follows:

1. Labels cannot be generated with a single SARS-CoV-2 genome sequence. A phylogeny analysis comparing numerous reference genomes to determine the phylogenetic position of the target genome in relation to reference genomes is required to come up with a label.

2. Labels become complex in a dynamic pandemic situation that requires frequent updating. Based on **Table 1** and the WHO website Tracking SARS-CoV-2 variants (https://www.who.int/en/activities/tracking-SARS-CoV-2-variants/), Delta variants may be labeled either as B.1.617.2 (Pango lineage), G/478K.V1 (GISAID clade) and 21A, 21I, 21J (Nextstrain clade). Thus, several interchangeable names have been used for the previously dominant variant circulating during 2021.

3. Labels such as B.1.1.7 or Delta do not readily supply enough biological meaning. Although very few virus names do provide this kind of information, a more complex system is in place for influenza viruses that contains additional information (https://www.cdc.gov/flu/about/viruses/types.htm). An ideal name identifies the virus, provides a relationship to those that are closely related and could quantify transmissibility and/or severity of the virus.

Potential biological metrics for SARS-CoV-2 naming

There are two widely used and important numeric metrics to measure new virus outbreak epidemiology – the basic reproductive rate (R0) and the case fatality rate (CFR). These are not ideal quantification metrics on their own since they are based on consequence instead of cause. R0 estimates how many people are infected on average by one person. It assumes an entirely susceptible population (not immune) and equal opportunities for transmission. We know this is not possible, populations may be restricted by non-pharmaceutical interventions (NPI) such as lockdowns, quarantines, and border closures. NPI will differ substantially by region and by human behavior which will affect R0 considerably. Importantly, R0 can determine if the outbreak is growing (R0 > 1) or receding (R0 < 1), but it fluctuates over time, depending on NPI measures in place and levels of protective immunity in the population. CFR, an important measure of disease outcome according to case numbers, has similar drawbacks, depends on the availability of medical resources, suitable diagnostic tests, and the period of the pandemic. CFR is influenced by the proportion of asymptomatic cases in an outbreak – a large difficulty with Covid-19 where a proportion of untested cases with no obvious symptoms results in underestimates of positive caseloads [6]. R0 and CFR cannot simply be compared across different infectious diseases. HIV has a R0 like SARS-CoV-2 but a much higher CFR [7]. Indeed, for HIV the transmission route is now completely foreseeable and preventable, highlighting the importance of a parameter such as infectiousness when characterizing virus epidemics and pandemics. Next, can we apply metrics other than R0 to determine how infectious a virus is?

Ideally, we require additional biologically sound metrics that can assess the ability of a virus to infect a human cell in addition to rates of transmission between hosts. An analogy could be a metric that functions like the category 1 to 5 system for hurricanes. By using such a simple digit for a virus (linear scale with a maximum value), we would then not panic about huge numbers and statements such as: "*Delta viral loads up to 1,260 times higher than those in people infected with the original strain*"[8]. Hurricanes are seasonal natural crises with the disastrous hurricane Katrina in late August 2005 being a large category 5 Atlantic hurricane that primarily affected the city of New Orleans USA. Besides labels of Katrina, New Orleans, and 2005, there is the key word of category 5. If using the analogy of a hurricane, what would be the Saffir-Simpson Hurricane Wind Scale for SARS-CoV-2? Is the Delta variant a category 5 with wind more than 157 mph, or is the Category 5 virus yet to come? Is it Omicron?

Thus, which quantification metrics for SARS-CoV-2 variants would really reflect biology, whilst simultaneously being simple, and could this be incorporated into naming of new variants? We have relied on 2-dimensional plotting tools to derive phylogeny trees relating viral genomes, including Covid-19 Genotyping Tool (CGT)[9], and covid-miner. Since SARS-CoV-2 is thought to have emerged from nature[10], a recent study proposed a new framework based on natural vector convex hull method conducting alignment-free sequence analysis[11]. However, this approach still relies on comparing the 1-dimensional genetic sequencing data. To determine biological consequences of viruses, we could instead inspect the virus and entry receptor binding site structurally for an answer, at least for a quantification metric that can determine strength and specificity of interaction. Such a metric could explain advantages gained by new variants such as Omicron, where numerous mutations in the spike gene improve its interaction with entry receptors and enhance transmission ability over existing variants [12]. A quantification metric like this should allow more simple direct comparisons between variants and offer an early warning system of infection waves.

The rationale for adopting 3-dimensional molecular structure metrics

Both SARS-CoV and SARS-CoV-2 enter human cells by interacting with angiotensin-converting enzyme 2 (ACE-2)[13], however SARS-CoV-2 binds ACE-2 with higher affinity than SARS-CoV[14,15]. This level of fitness underlies the relative ability of the virus to enter cells, replicate and ultimately transmit between hosts. Most virus classification approaches do not provide information of the final folded protein structure or binding interactions with entry receptors but instead focus on the 1-dimensional nucleic acid sequence similarity and divergence. Despite tertiary protein structure being determined by the primary sequence, a high similarity at the primary level does not necessarily link to a high match at the tertiary and 3-dimensional level. Both SARS-CoV-2 and SARS-CoV spike proteins bind to human ACE-2 as the entry receptor, even though their nucleotide sequence homology is approximately 85%. However, the closest related CoV to SARS-CoV-2 by sequence analysis is the bat CoV RaTG13, which cannot infect human cells via ACE2, despite its spike sequence being 96% identical[16]. The match between the SARS-CoV-2 spike protein and the human ACE-2 receptor might be considered analogous to a claw toy grabber machine (**Figure 1 right panel**). The click and the claw (spike protein) and the toy (human cell ACE-2 receptor) thus determines the speed of grabbing the prize (virus transmission rate) from the machine. Therefore, we need to look beyond the 1-dimensional information of sequence and investigate the 3-dimensional structure of the virus spike and the human ACE-2. After all, that is "ground zero" where the virus catches and enters human cells to begin the replication process. One of the biggest scientific advances over the past two years is the emergence of artificial intelligence (AI) powered protein structure prediction, namely Alpha Fold[17-19]. Such structural models could then define quantification metrics based on the complementarity of the virus and the host cell receptor. The resulting quantification metric for SARS-CoV-2 would then add biological meaning to the simple label of variants with English or Greek letters and Arabic numbers (**Table 1 and Figure 1 left panel**).

Table 1: summary of naming systems for identified SARS-CoV-2 variants listed by WHO label. VOC = variant of concern; VOI = variant of interest.

WHO Label (Greek alphabet)	Pango lineage	GISAID clade	Nextstrain clade	Country & Date detected	Description
Alpha	B.1.1.7	GRY	20I (V1)	UK 09-2020	De-escalated variant Former VOC
Beta	B.1.351	GH/501Y.V2	20H (V2)	South Africa 09-2020	Remains a VOC
Gamma	P.1	GR/501Y.V3	20J (V3)	Brazil 12-2020	Remains a VOC
Delta	B.1.617.2	G/478K.V1	21A, 21I, 21J	India 12-2020	Dominant until emergence of Omicron, remains a VOC
Epsilon	B.1.427 B.1.429	GH/452R.V1	21C	USA 09-2020	De-escalated variant Former VOI
Zeta	P.2	GR/484K.V2	20B/S.484K	Brazil 01-2021	De-escalated variant Former VOI
Eta	B.1.525	G/484K.V3	21D	Nigeria 12-2020	De-escalated variant Former VOI
Theta	P.3	GR/1092K.V1	21E	Philippines 01-2021	De-escalated variant Former VOI
Iota	B.1.526	GH/253G.V1	21F	USA 12-2020	De-escalated variant Former VOI
Kappa	B.1.617.1	G/452R.V3	21B	India 12-2020	De-escalated variant Former VOI
Lambda	C.37	GR/452Q.V1	21G	Peru 12-2020	VOI with sporadic transmission
Mu	B.1.621	GH	21H	Colombia 01-2021	VOI with sporadic transmission
Nu	Not assigned				
Xi	Not assigned				
Omicron	B.1.1.529	GRA	21K, 21L 21M	South Africa and Botswana 11-2021	Current dominant VOC
Pi	Reserved for potential new variants				
Rho					
Sigma					
Tau					
Upsilon					
Phi					
Chi					
Psi					

From a virus evolutionary and survival perspective, the maximum match is represented by the virus' end goal (i.e., perfect match with entry receptor - arbitrarily assigned as ten). If this is the maximum value that a virus can reach, the minimum value of zero could be assigned to the 3-dimensional structure of a virus such as RaTG13 spike interaction with ACE-2, that is very close to SARS-CoV-2 at the 1-dimensional level but does not bind human ACE-2. Subsequently, by using the 3-dimensional structure references for minimum and maximum distance, biologically meaningful metrics can be derived to quantify a variant. Such a metric would allow reporting a variant with a simple value (e.g., eight = highly infectious and transmissible; two = weakly infective and unlikely to transmit). Utility of this type of structural approach has recently been shown for SARS-CoV-1 and SARS-CoV-2, potentially explaining interactions and mutations important for viral infection, pathogenesis, and transmission through differences in interaction with ACE-2[20]. In their published work, this group of researchers only mentioned Alpha-Fold as "*recent advances in protein-folding predictions*" that might ameliorate certain technological restrictions in the future. Nevertheless, they did present an illuminating visualization of the SARS-CoV spike protein and ACE-2 interface and furthermore a binding affinity (ΔΔG) quantification metric that could potentially lead to biologically meaningful nomenclature. Through this type of approach, extrapolation to any virus is theoretically possible, provided the entry receptor is defined and viral structural or sequence information exists. This simple number or interaction index, unpolitical and purely scientific, could assist in explaining viral molecular and biological fitness, when used in combination with existing labels already in use for the virus.

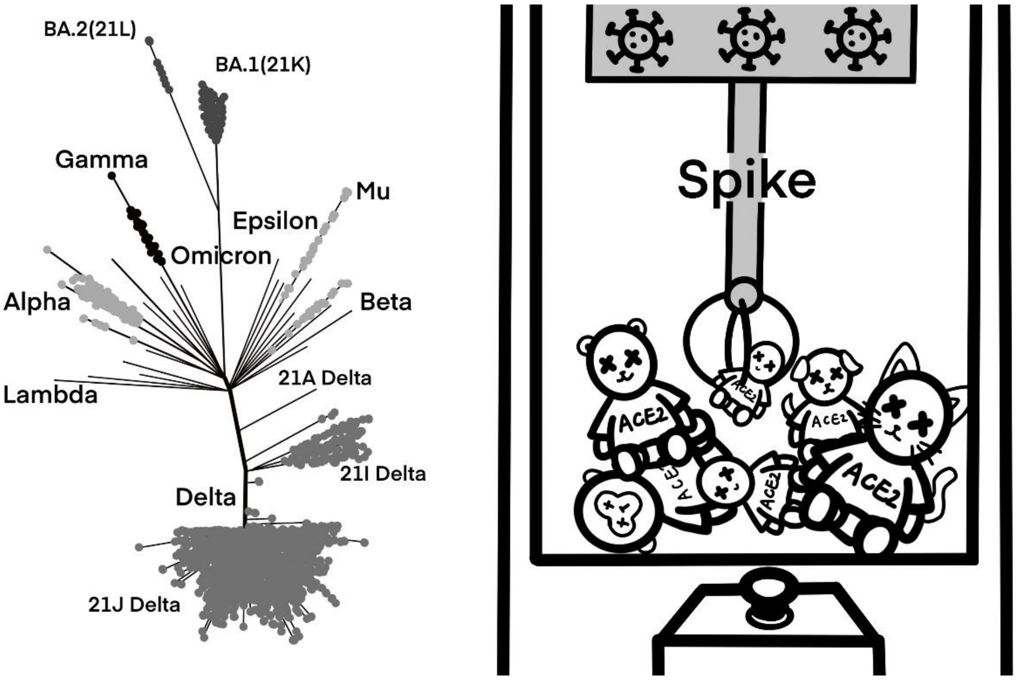

Figure 1. Phylogeny-based labels for SARS-CoV-2 (left panel; obtained from Nextstrain.org) and virus-host 3-dimensional structure quantification analogy (right panel).

Concluding remarks and perspectives

Continued reassessment is needed as new challenges arise with SARS-CoV-2 variants[21]. Labels based on the geographic location where new variants first appeared, such as Kent, South African or Indian variant, are inappropriate, create a country-associated stigma and have been replaced by the Greek letters. The WHO, instrumental in this renaming, has been very careful not to be viewed as political, for many reasons, and potentially due to state political pressure. The existing Pango, GISAID, and Nextstrain complex naming systems do place lineages and clades of SARS-CoV-2 into perspective but offer little functional information. In contrast, the WHO naming scheme for VOC and VOI that is based on the Greek alphabet, albeit simple, provides neither scientific relationships nor a biological context to these variants.

Because politicians are grappling with the pandemic effects and insist that decisions should be based on scientific evidence, the scientific community may wish to contemplate a scientifically sound naming scheme for SARS-CoV-2 variants that appear. Virus-specific information can move beyond the 1-dimensional level of genome or protein sequences and move into the 3-dimensional structures at the location where the virus attaches and attacks human cells. Developing a quantification metric based on the 3-dimensional structure between SARS-CoV-2 spike receptor binding domain and human ACE-2 is not as straightforward as the commonly used phylogenetic analysis based on 1-dimensional sequence. However, it is possible as the use of root-mean-square deviation in comparing 3-dimensional structures of other proteins was proposed decades ago[22] and would provide another level of biological context to the currently used virus labels.

Covid-19 has provided an unprecedented opportunity to investigate the interaction between viruses (SARS-CoV-2 variants) and entry receptors (human ACE-2) in a fast-changing dynamic situation. The additional use of the power of protein 3-dimensional structure prediction, by methods such as AlphaFold[23], is critical to quantify SARS-CoV-2 variants infectiousness with structural biology, simplifying the comparisons of variants alongside other metrics/labels already in use. World leaders and global citizens all point out the importance of science in studying the Covid-19 pandemic, for virus origin tracing, epidemiology of emergence and spread, and for vaccine or antiviral formulation updating. Here we identify the importance of creating a simple number that can consistently and accurately quantify the biological distance between SARS-CoV-2 variants and our human cells the virus needs to replicate and spread. Application of this number to existing SARS-CoV-2 variant labels/names will go some way towards quickly improving understanding of the relative importance of variants as they appear.

Conflict of interest: none

Funding statement: GRM received rescaling funds from the London Metropolitan University research and postgraduate office to further develop the structural bioinformatics aspects of this study

References

1. Shu Y, McCauley J. GISAID: Global initiative on sharing all influenza data - from vision to reality. *Euro Surveill.* Mar 30 2017;22(13)doi:10.2807/1560-7917.ES.2017.22.13.30494

2. Hadfield J, Megill C, Bell SM, et al. Nextstrain: real-time tracking of pathogen evolution. *Bioinformatics.* Dec 1 2018;34(23):4121-4123. doi:10.1093/bioinformatics/bty407

3. Rambaut A. *Nat Microbiol.* 2020// 2020;5doi:10.1038/s41564-020-0770-5

4. Callaway E, Ledford H. How bad is Omicron? What scientists know so far. *Nature.* Dec 2021;600(7888):197-199. doi:10.1038/d41586-021-03614-z

5. Konings F, Perkins MD, Kuhn JH, et al. SARS-CoV-2 Variants of Interest and Concern naming scheme conducive for global discourse. *Nat Microbiol.* Jul 2021;6(7):821-823. doi:10.1038/s41564-021-00932-w

6. Jefferson T, Spencer EA, Brassey J, et al. Transmission of Severe Acute Respiratory Syndrome Coronavirus-2 (SARS-CoV-2) from pre and asymptomatic infected individuals: a systematic review. *Clinical microbiology and infection : the official publication of the European Society of Clinical Microbiology and Infectious Diseases.* Oct 29 2021;doi:10.1016/j.cmi.2021.10.015

7. Powers KA, Kretzschmar ME, Miller WC, Cohen MS. Impact of early-stage HIV transmission on treatment as prevention. *Proc Natl Acad Sci U S A.* Nov 11 2014;111(45):15867-8. doi:10.1073/pnas.1418496111

8. Li Bea. Viral infection and transmission in a large, well-traced outbreak caused by the SARS-CoV-2 Delta variant. *medRxiv.* 2021; https://doi.org/10.1101/2021.07.07.21260122

9. Maan H, Mbareche H, Raphenya AR, et al. Genotyping SARS-CoV-2 through an interactive web application. *Lancet Digit Health.* Jul 2020;2(7):e340-e341. doi:10.1016/S2589-7500(20)30140-0

10. Andersen KG, Rambaut A, Lipkin WI, Holmes EC, Garry RF. The proximal origin of SARS-CoV-2. *Nat Med.* Apr 2020;26(4):450-452. doi:10.1038/s41591-020-0820-9

11. Zhao R, Pei S, Yau SST. New Genome Sequence Detection via Natural Vector Convex Hull Method. *IEEE/ACM Trans Comput Biol Bioinform.* Nov 25 2020;PPdoi:10.1109/TCBB.2020.3040706

12. Fantini J, Yahi N, Colson P, Chahinian H, La Scola B, Raoult D. The puzzling mutational landscape of the SARS-2-variant Omicron. *Journal of medical virology.* Jan 8 2022;doi:10.1002/jmv.27577

13. Lan J, Ge J, Yu J, et al. Structure of the SARS-CoV-2 spike receptor-binding domain bound to the ACE2 receptor. *Nature.* May 2020;581(7807):215-220. doi:10.1038/s41586-020-2180-5

14. Zhou P, Yang XL, Wang XG, et al. A pneumonia outbreak associated with a new coronavirus of probable bat origin. *Nature.* Mar 2020;579(7798):270-273. doi:10.1038/s41586-020-2012-7

15. Wang Y, Liu M, Gao J. Enhanced receptor binding of SARS-CoV-2 through networks of hydrogen-bonding and hydrophobic interactions. *Proc Natl Acad Sci U S A.* Jun 23 2020;117(25):13967-13974. doi:10.1073/pnas.2008209117

16. Shang J, Ye G, Shi K, et al. Structural basis of receptor recognition by SARS-CoV-2. *Nature.* May 2020;581(7807):221-224. doi:10.1038/s41586-020-2179-y

17. Jumper J, Evans R, Pritzel A, et al. Highly accurate protein structure prediction with AlphaFold. *Nature.* Jul 15 2021;doi:10.1038/s41586-021-03819-2

18. Tunyasuvunakool K, Adler J, Wu Z, et al. Highly accurate protein structure prediction for the human proteome. *Nature.* Jul 22 2021;doi:10.1038/s41586-021-03828-1

19. Baek M, DiMaio F, Anishchenko I, et al. Accurate prediction of protein structures and interactions using a three-track neural network. *Science.* Jul 15 2021;doi:10.1126/science.abj8754

20. Wierbowski SD, Liang S, Liu Y, et al. A 3D structural SARS-CoV-2-human interactome to explore genetic and drug perturbations. *Nature methods.* Dec 2021;18(12):1477-1488. doi:10.1038/s41592-021-01318-w

21. Callaway E. 'A bloody mess': Confusion reigns over naming of new COVID variants. *Nature.* Jan 2021;589(7842):339. doi:10.1038/d41586-021-00097-w

22. Maiorov VN, Crippen GM. Significance of root-mean-square deviation in comparing three-dimensional structures of globular proteins. *J Mol Biol.* Jan 14 1994;235(2):625-34. doi:10.1006/jmbi.1994.1017

23. Senior AW, Evans R, Jumper J, et al. Improved protein structure prediction using potentials from deep learning. *Nature.* Jan 2020;577(7792):706-710. doi:10.1038/s41586-019-1923-7