





## Emerging SARS-CoV-2 variants follow a historical pattern recorded in outgroups infecting non-human hosts

Kazutaka Katoh<sup>1</sup>   & Daron M. Standley<sup>1</sup> 

The ability to predict emerging variants of SARS-CoV-2 would be of enormous value, as it would enable proactive design of vaccines in advance of such emergence. We estimated diversity of each site on a multiple sequence alignment (MSA) of the Spike (S) proteins from close relatives of SARS-CoV-2 that infected bat and pangolin before the pandemic. Then we compared the locations of high diversity sites in this MSA and those of mutations found in multiple emerging lineages of human-infecting SARS-CoV-2. This comparison revealed a significant correspondence, which suggests that a limited number of sites in this protein are repeatedly substituted in different lineages of this group of viruses. It follows, therefore, that the sites of future emerging mutations in SARS-CoV-2 can be predicted by analyzing their relatives (outgroups) that have infected non-human hosts. We discuss a possible evolutionary basis for these substitutions and provide a list of frequently substituted sites that potentially include future emerging variants in SARS-CoV-2.

<sup>1</sup>Research Institute for Microbial Diseases, Osaka University, 3-1 Yamadaoka, Suita 565-0871, Japan. email: [katoh@ifrec.osaka-u.ac.jp](mailto:katoh@ifrec.osaka-u.ac.jp); [standley@biken.osaka-u.ac.jp](mailto:standley@biken.osaka-u.ac.jp)

In December 2020, three SARS-CoV-2 variants emerged with increased infectivity from England, South Africa, and Brazil. The fact that certain mutations in the Spike (S) protein had occurred independently prompted us to reexamine our September 2020 study of the evolution of this protein<sup>1</sup>. In our original study, we characterized the *importance* of each residue position in the S protein by comparing its diversity in SARS-CoV-2 with that in relatives (outgroups) that infected bats or pangolins by using a simple equation:

$$\text{Importance} = \text{diversity}(\text{SARS-CoV-2} + \text{outgroup}) - \text{diversity}(\text{SARS-CoV-2}), \quad (1)$$

where *diversity*(*x*) is defined as the number of different amino acids observed at the site in question in virus group *x*. This equation, which was meant to be descriptive rather than predictive, identified twenty positions of high *importance*. We were thus surprised to find that, of these 20 positions, four were characteristic of the above emerging variants: Histidine 69, Valine 70, Glutamine 484, and Asparagine 501. These sites coincide with four out of the five residues (69, 70, 417, 484, 501) that have mutated independently in two or more of the three emerging lineages or a lineage transmitted between human and mink<sup>2</sup>. We reanalyzed the underlying sequence data and found that the *importance* values of these sites were determined primarily by *diversity*(outgroup), rather than *diversity*(SARS-CoV-2). In hindsight, this is somewhat expected, as the latter term was close to unity at the time when we performed the analysis (i.e., before the emergence of new variants).

A natural question, then, is why a limited set of sites with high diversity in outgroups have also recently been substituted in SARS-CoV-2. As an evolutionary mechanism behind such frequent substitutions, two extreme scenarios, (i) neutral evolution and (ii) positive selection, are possible. These two scenarios give opposite predictions as to functionality of frequently substituted residues: scenario (i) predicts that the frequently substituted sites are not functionally important because they are under low functional constraints, while scenario (ii) predicts that functionally important sites have changed by being positively selected. Although the truth may lie in between these two extremes, we tested which scenario is more likely using the distribution of residues that are known to be important for infection to host cells.

## Results and discussion

**Known functionally important sites.** Currently available functional information supports scenario (ii). When viewing the distribution of residues with high *diversity*(outgroup) as a heatmap on the spike molecular surface (Fig. 1a, b), it is apparent that these residues are not evenly distributed, but form clusters in the N terminal domain (NTD), receptor binding domain (RBD) and S1/S2 cleavage site, which are thought to be important for interaction to human cells. More specifically, Glutamine 484 and Asparagine 501 are structurally close to the interface with the host cell receptor ACE2, which, in turn, is targeted by neutralizing antibodies. Histidine 69 and Valine 70, on the other hand, are far from the ACE2 binding site but proximal to a recently-reported epitope for infection-enhancing antibodies<sup>3,4</sup>. The 69/70 deletion mutant also occurred in an immunosuppressed individual who underwent convalescent plasma therapy<sup>5</sup>, suggesting that the mutation is a direct response to host antibodies. These two residues have also been reported to bind sialic acids<sup>6</sup>. There are also high diversity sites (around Alanine 684) adjacent to the S1/S2 cleavage site of SARS-CoV-2, as indicated in Table 1. The changes in this region seem to be host specific: HSMSS[LF]R in pangolin; QTQTNSR in two lineages of bat; QTQTNSPRRAR (which includes a polybasic insertion recognized by host's

protease<sup>7,8</sup>) in human. These changes might reflect adaptation to new hosts in the past. Further substitutions, such as Proline 681, could change infectivity in human<sup>9</sup> around this region.

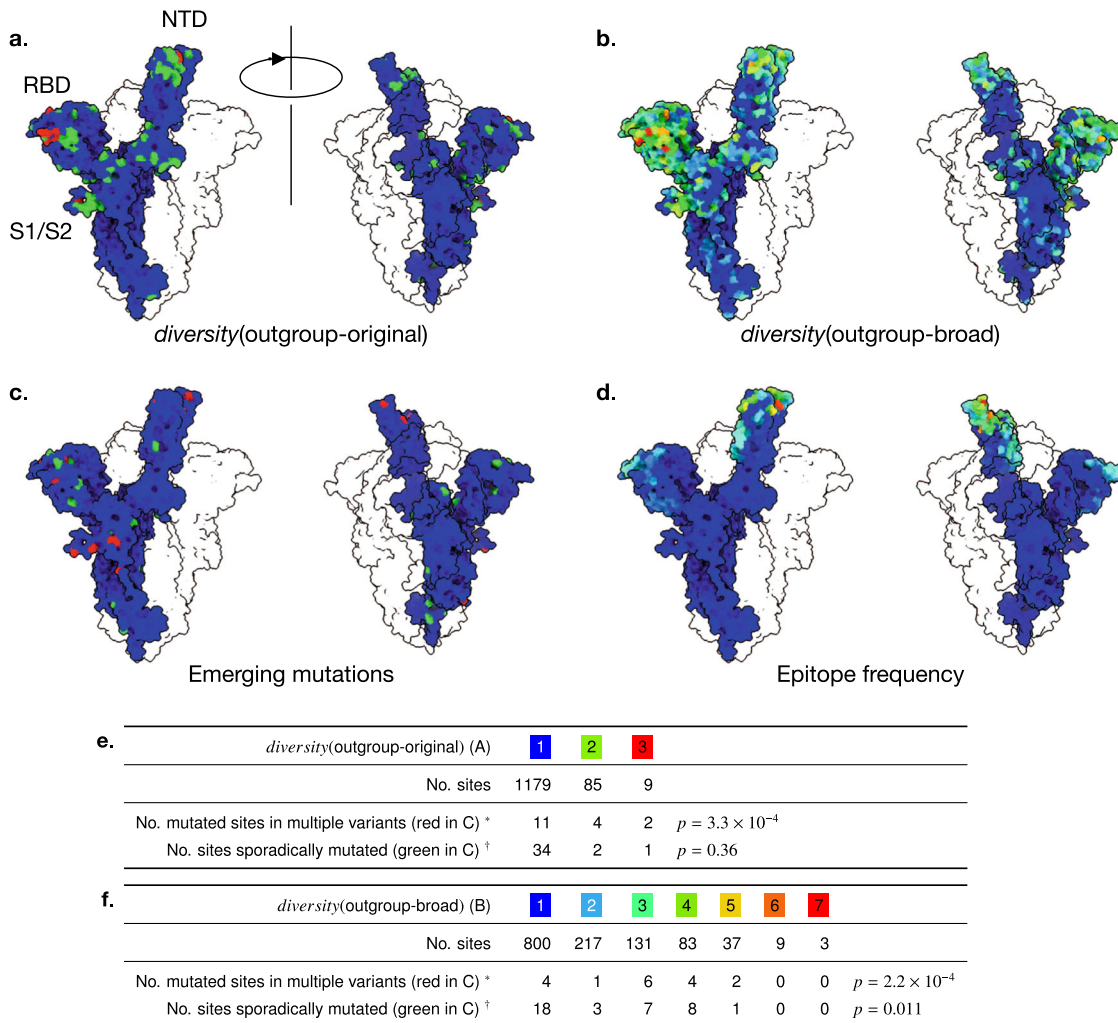
## Possible positive selection in SARS-CoV-2 and outgroups.

Modification of the regions discussed above could thus affect the infectivity or enable the virus to escape from the host's immune system, albeit temporarily, as the change will inevitably be counteracted by a shift in the antibody repertoire of the host, resulting in an effective "arms race", as reviewed in references<sup>10,11</sup>. In this scenario, the sites with higher diversity imply direct or indirect host-pathogen interactions and are thus in a constant state of flux. This interpretation is consistent with previous studies that reported the possibility of adaptive evolution to infect human in SARS-CoV-2<sup>12,13</sup> and in other coronaviruses<sup>14</sup>. More recently, reports suggest that mutations in the B.1.1.7/B.1.351/P.1 lineages result in reduced binding affinity to some but not all neutralizing antibodies<sup>15,16</sup>. For close outgroups of SARS-CoV-2, where functional information in non-human hosts is not available, we explored the possibility of positive selection using Bayes Empirical Bayes analysis<sup>17</sup> implemented in the PAML program<sup>18</sup>, excluding human-infecting lineages. The fourth column in <https://mafft.cbrc.jp/alignment/pub/sarscov2/fulllist.tsv> shows the sites estimated to have had more nonsynonymous substitutions than synonymous substitutions in the outgroup sequences, although this estimation is sensitive to sequence selection and alignment ambiguity.

## Correspondence of high diversity sites between SARS-CoV-2 and outgroups.

Assuming adaptive evolution, it is conceivable that different sites are positively selected in different hosts to "optimize" infectivity; however, the analysis of outgroups revealed that such sites can overlap, presumably being involved in a common mechanism of host-pathogen interaction in different lineages, and that frequent changes in these sites already occurred in outgroups before the pandemic. We note that the correspondence between the positions of emerging mutations found in multiple human-infecting variants and those with high *diversity*(outgroup) is significant by Fisher's exact test, regardless whether the original outgroup (Fig. 1a) or a broad outgroup (Fig. 1b) is used (see the lines marked with asterisk, \*, in Fig. 1e, f). By contrast, the positions of sporadic mutations that are found just in a single variant in human show less clear or no correspondence with *diversity*(outgroup) (see the lines marked with dagger, †, in Fig. 1e, f). The former type of mutations (found in multiple variants) are likely to affect interactions with host factors and to spread in humans, although it's difficult to differentiate between parallel evolution and recombination between lineages. The proposed simple method is suitable to predict such sites because they appear to be under positive selection in independent lineages including outgroups. There are some sites that have high diversity in outgroup but are not (yet) mutated in the current population of SARS-CoV-2. Such sites are regarded to be mis-predicted in this statistical test, but may mutate in the future. Indeed, residues close to the S1/S2 cleavage site were found to have high *diversity*(outgroup) in our initial analysis before emergence of several variants of concern (red sites in Fig. 1a) in 2020. Subsequently, substitutions in this region were indeed found in multiple variants infecting humans (red sites in Fig. 1c).

**Prediction of position of emerging mutations.** To anticipate new variants of SARS-CoV-2 as early as possible, a straightforward strategy would be to intensively collect a large amount of sequence data from human-infecting lineages<sup>19</sup>. Our observation above leads to a complementary strategy: prepare against new variants in advance by decoding the long history of host-pathogen interactions recorded in the outgroup sequences



**Fig. 1 Diversity and other indices mapped on structure of the S protein, visualized by ChimeraX<sup>28</sup>.** A movie of the structure rotating is available at <https://mafft.cbrc.jp/alignment/pub/sarscov2/structure.mp4>. For clarity, a single S protein is shown in the context of a spike trimer. **a, b** *diversity(outgroup)*, the number of different amino acids observed at each site in outgroup, was computed using the original and broad definitions of outgroup: Low diversity (blue); High diversity (red). **c** Emerging mutations found in single variant (green); in multiple variants (red) infecting humans. **d** Epitope frequency, the number of antibodies that contact each residue (<6 Å), was counted based on currently available Protein DataBank (PDB) entries of S protein-antibody complexes listed in <https://mafft.cbrc.jp/alignment/pub/sarscov2/epitopefrequency.txt>. 0 (blue); 15 (red). This value is not expected to represent all spike-targeting antibodies. **e** Correspondence between *diversity(outgroup-original)* (**a**) and emerging mutations (**c**). **f** Correspondence between *diversity(outgroup-broad)* (**b**) and emerging mutations (**c**). Positions of emerging mutations in **c, e**, and **f** were taken from reference<sup>27</sup>. Asterisk represents mutated sites in multiple variants = residues in bold in Table 2 in Peacock et al.<sup>27</sup>; dagger represents sporadic mutations = the other residues listed in the same table.

infecting non-human hosts. Unfortunately, currently efforts have focused almost exclusively on the former strategy and available outgroup sequences are limited. If richer sequence data of outgroups infecting bat, pangolin and other possible hosts becomes available, it would not only shed light on the origin of SARS-CoV-2<sup>20</sup>, but also give us an advantage in the arms race with this virus.

**Limitations.** This analysis has several limitations. First, genetic changes can be caused by recombinations, not only point mutations and insertions/deletions. Indeed the receptor binding motif of SARS-CoV-2 was reported to be acquired from a lineage infecting pangolin<sup>21</sup>. It is possible that some changes in outgroups are also caused by recombinations. Our analysis regards a recombination simply as simultaneous changes in successive sites in the recipient genome. As a result, in comparison with the number of evolutionary events, diversity is overestimated if recombination between lineages occurred and the donor is not

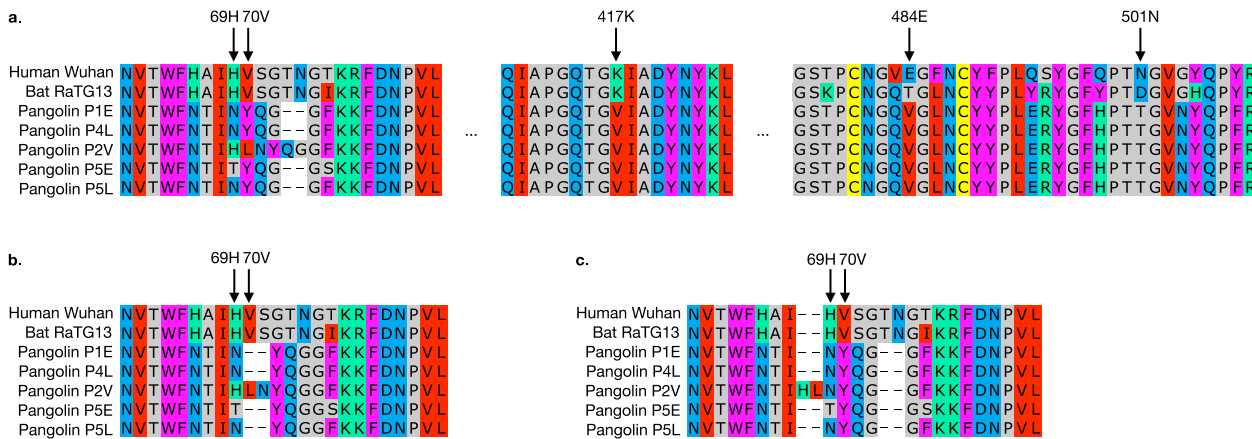
included in the alignment, while diversity is underestimated when the donor is included in the alignment. Second, the diversity in outgroup can be calculated only when the corresponding part exists in outgroups. This method cannot be applied to human-specific insertions. Third, the prediction of the position of a mutation can be ambiguous because of alignment ambiguity. Figure 2a shows the multiple sequence alignment (MSA) used in Saputri et al.<sup>1</sup>, and Fig. 2b shows an alternative MSA. Since the insertion in the P2V strain in pangolin should be independent from the insertion at human and RaTG13, gaps can be inserted in different positions between these two groups. Thus some other MSAs (eg, Fig. 2c) are also possible. By using different MSAs, the position of high diversity sites can shift. Even in this case, a high diversity region should exist nearby, because the alignment ambiguity itself is due to high diversity. This problem more frequently occurs when including a wider range of outgroups. Thus, to obtain a prediction at the residue-level resolution, a large amount of data from close outgroups is necessary.

**Table 1 High diversity residues.**

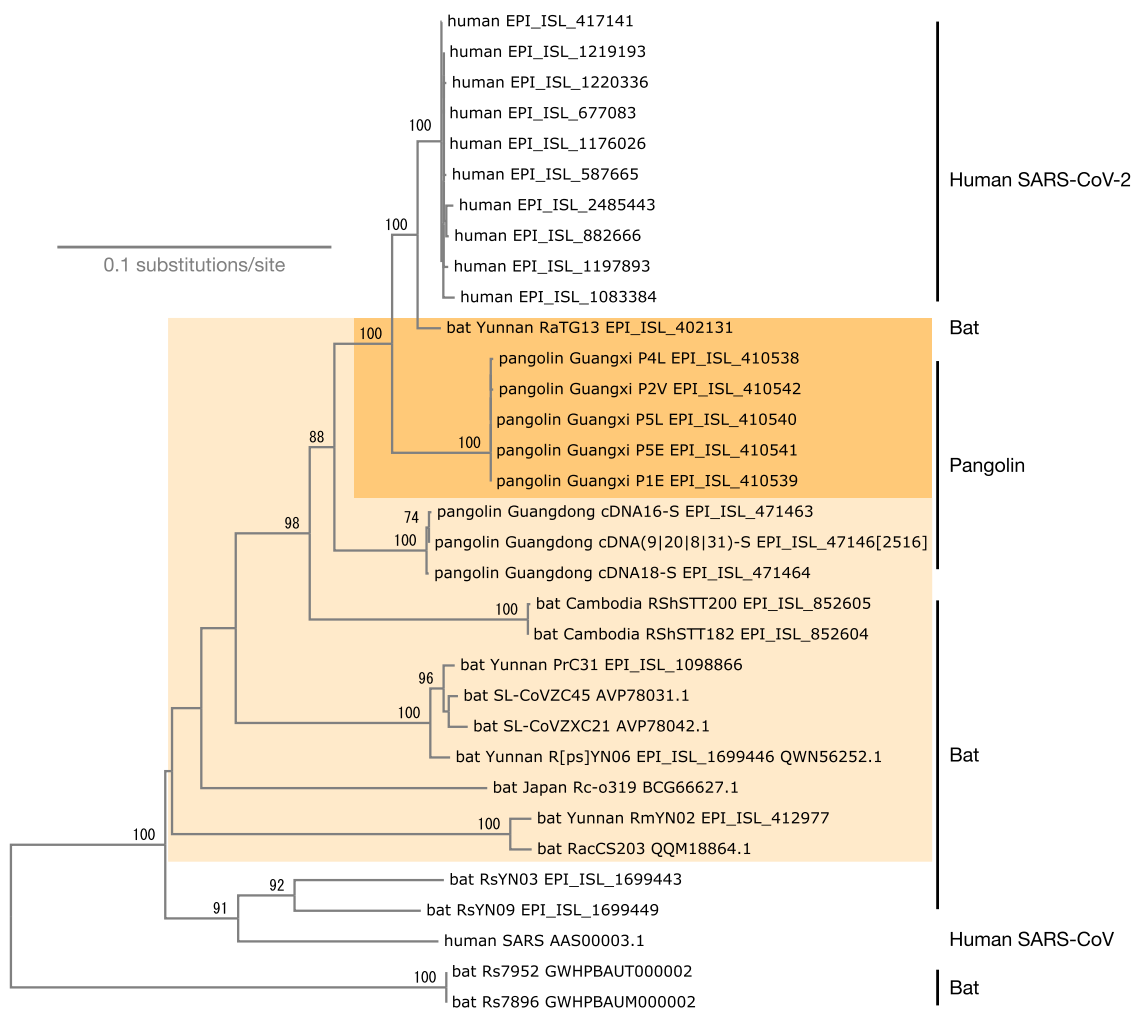
Res	AA	Orig	Broad	Epitope	ACE2	SialicAcid	Cleavage	Emerge
8	L	1	5	0	-	-	-	-
9	P	1	5	0	-	-	-	-
12	S	1	5	0	-	-	-	-
22	T	1	5	0	-	-	-	-
23	Q	2*	6	0	-	o	-	-
24	L	2	5	0	-	o	-	-
25	P	2	5	0	-	-	-	-
27	A	2	7	0	-	-	-	-
66	H	2	5	0	-	-	-	-
69	H	3*	2	2	-	o	-	m
70	V	3*	4	2	-	o	-	m
71	S	3*	7	1	-	-	-	-
72	G	2*	6	1	-	-	-	-
73	T	3*	4	1	-	-	-	-
74	N	3*	5	1	-	-	-	-
76	T	3*	7	2	-	-	-	-
85	P	1	5	0	-	-	-	-
137	N	2	6	0	-	-	-	-
140	F	1	5	0	-	-	-	-
147	K	2	5	4	-	-	-	-
164	N	1	5	0	-	-	-	-
169	E	1	5	0	-	-	-	-
176	L	1	6	0	-	-	-	-
183	Q	1	6	1	-	-	-	-
197	I	2	6	0	-	-	-	-
215	D	1	5	2	-	-	-	s
218	Q	2*	5	2	-	-	-	-
224	E	1	5	0	-	-	-	-
249	L	1	5	3	-	-	-	-
253	D	3*	4	2	-	-	-	s
255	S	2*	4	1	-	-	-	-
256	S	1*	3	0	-	-	-	-
260	A	2	5	1	-	-	-	-
272	P	2	5	0	-	-	-	-
324	E	2	5	0	-	-	-	-
417	K	2*	3	8	o	-	-	m
439	N	2*	6	0	-	-	-	-
440	N	2	5	0	-	-	-	-
441	L	2*	3	3	-	-	-	-
443	S	1	5	1	-	-	-	-
444	K	2*	5	0	-	-	-	-
445	V	2*	4	3	-	-	-	-
449	Y	3	3	2	o	-	-	-
450	N	2*	5	11	-	-	-	-
459	S	2	5	6	-	-	-	-
484	E	2	5	7	-	-	-	m
493	Q	2	5	10	o	-	-	-
501	N	2*	5	8	o	-	-	m
504	G	2	5	8	-	-	-	-
529	K	2*	3	0	-	-	-	-
532	N	2	5	0	-	-	-	-
554	E	2*	4	0	-	-	-	-
556	N	2	5	0	-	-	-	-
640	S	2	5	0	-	-	-	-
677	Q	2*	4	0	-	-	-	m
678	T	2	5	0	-	-	-	-
679	N	2*	5	0	-	-	-	-
680	S	2*	4	0	-	-	-	-
684	A	3	1	0	-	-	o	-
688	A	2	5	0	-	-	-	-
689	S	2*	4	0	-	-	-	-

The most diverse residue positions are listed along with several annotations.

Orig, *diversity*(outgroup-original); Asterisk, (nonsynonymous substitutions)/(synonymous substitutions) > 1 in outgroup-original; Broad, *diversity*(outgroup-broad); Epitope, epitope frequency. (See the caption of Fig. 1); ACE2, residue is within 6Å of ACE2 in PDB entry 7DF4; SialicAcid, reported sialic acid binding residue<sup>6</sup>; Cleavage, known protease cleavage site; Emerge, emerging variants in humans listed in Table 2 in Peacock et al. (2021)<sup>27</sup>. m, found in multiple variants; s, found in a single variant. See <https://mafft.cbrc.jp/alignment/pub/sarscov2/fullist.tsv> for a full list.



**Fig. 2** MSA around residues 69, 70, 417, 484, and 501, visualized by Jalview<sup>29</sup>. **a** MSA used in reference<sup>1</sup>. **b, c** Alternative MSAs around residues 69 and 70.



**Fig. 3** Outgroup sequences used as “original” (dark orange) and “broad” (light orange), displayed on a simple tree. Bootstrap values larger than 70%. Accession number in GISAID, genbank or National Genomics Data Center is given for each sequence.

**Methods**

**Sequence data to calculate diversity.** According to the interpretation of positive selection resulting in an “arms race”, it is possible that positions of mutations in future emerging variants can be predicted simply by identifying sites with high diversity in outgroups, where adversarial host-pathogen interactions have been occurring longer than for SARS-CoV-2 and humans. Because of their potential

importance in the design of vaccines against future emerging variants, we calculated *diversity*(outgroup) for each residue position considering two definitions of outgroups: one that is identical to that used in our original analysis in which 6 sequences were used and a broader definition (18 sequences) to increase the amount of data used in the calculation. Both datasets are available at <https://mafft.cbrc.jp/alignment/pub/sarscov2/>. The sequence data was taken from

GISAID<sup>22</sup> and genbank, to cover major lineages of outgroups appearing in recent reports<sup>23,24</sup>. As noted in Introduction,  $diversity(x)$  is defined as the number of different amino acids observed at the site in question, where  $x$  is either of the two outgroups. Amino acid residues with high  $diversity$ (outgroup) are listed in Table 1.

**Tree.** Figure 3 shows a phylogenetic tree of the S protein from SARS-CoV-2 and relatives including remote ones by the neighbor-joining method<sup>25</sup> applied to a distance matrix estimated with the Poisson correction based on an amino acid sequence alignment by MAFFT<sup>26</sup>. The outgroup sequences used here are highlighted in this figure. The accession numbers of the sequences are given in this tree.

**Statistics and reproducibility.** In Fig. 1e, f,  $p$  values were calculated by Fisher's exact test under the null hypothesis that the diversity of each site in outgroups (Fig. 1a or b) and the distribution of emerging mutations (Fig. 1c) are independent of each other. Positions of the mutations found in multiple variants and the sporadic mutations in Fig. 1c were taken from Peacock et al.<sup>27</sup>.

**Reporting summary.** Further information on research design is available in the Nature Research Reporting Summary linked to this article.

### Data availability

Sequence data used here are available at <https://mafft.cbrc.jp/alignment/pub/sarscov2/>.

### Code availability

A script to count the number of amino acids in each site is available at <https://mafft.cbrc.jp/alignment/pub/sarscov2/>.

Received: 22 April 2021; Accepted: 8 September 2021;

Published online: 22 September 2021

## References

- Saputri, D. S. et al. Flexible, functional, and familiar: characteristics of SARS-CoV-2 spike protein evolution. *Front. Microbiol.* **11**, 2112 (2020).
- Lassaunière, R. et al. Working paper on SARS-CoV-2 spike mutations arising in Danish mink, their 2 spread to humans and neutralization data. [https://files.ssi.dk/Mink-cluster-5-short-report\\_AFO2](https://files.ssi.dk/Mink-cluster-5-short-report_AFO2) (2021).
- Li, D. et al. The functions of SARS-CoV-2 neutralizing and infection-enhancing antibodies in vitro and in mice and nonhuman primates. *bioRxiv* <https://doi.org/10.1101/2020.12.31.424729> (2021).
- Liu, Y. et al. An infectivity-enhancing site on the SARS-CoV-2 spike protein is targeted by COVID-19 patient antibodies. *Cell* **184**, 3452–3466.e18 (2021).
- Kemp, S. A. et al. SARS-CoV-2 evolution during treatment of chronic infection. *Nature* **592**, 277–282 (2021).
- Baker, A. N. et al. The SARS-COV-2 spike protein binds sialic acids and enables rapid detection in a lateral flow point of care diagnostic device. *ACS Cent. Sci.* **6**, 2046–2052 (2020).
- Hoffmann, M. et al. A multibasic cleavage site in the spike protein of SARS-CoV-2 is essential for infection of human lung cells. *Mol. Cell* **78**, 779–784 (2020).
- Peacock, T. P. et al. The furin cleavage site in the SARS-CoV-2 spike protein is required for transmission in ferrets. *Nat. Microbiol.* **6**, 899–909 (2021).
- Lubinski, B. et al. Functional evaluation of proteolytic activation for the SARS-CoV-2 variant B.1.1.7: role of the P681H mutation. *bioRxiv* <https://doi.org/10.1101/2021.04.06.438731> (2021).
- Meyerson, N. R. & Sawyer, S. L. Two-stepping through time:mammals and viruses. *Trends Microbiol.* **19**, 286–294 (2011).
- Bonsignori, M. et al. Antibody-virus co-evolution in HIV infection: paths for HIV vaccine development. *Immunol. Rev.* **275**, 145–160 (2017).
- Tegally, H. et al. Detection of a SARS-CoV-2 variant of concern in South Africa. *Nature* **592**, 438–443 (2021).
- Kang, L. et al. A selective sweep in the Spike gene has driven SARS-CoV-2 human adaptation. *Cell* **184**, 4392–4400.e4 (2021).
- Kistler, K. E. & Bedford, T. Evidence for adaptive evolution in the receptor-binding domain of seasonal coronaviruses OC43 and 229e. *eLife* **10**, e64509 (2021).
- Zhou, D. et al. Evidence of escape of SARS-CoV-2 variant B.1.351 from natural and vaccine-induced sera. *Cell* **184**, 2348–2361 (2021).
- Hoffmann, M. et al. SARS-CoV-2 variants B.1.351 and P.1 escape from neutralizing antibodies. *Cell* **184**, 2384–2393 (2021).
- Yang, Z. et al. Bayes empirical Bayes inference of amino acid sites under positive selection. *Mol. Biol. Evol.* **22**, 1107–1118 (2005).
- Yang, Z. PAML 4: phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol.* **24**, 1586–1591 (2007).
- Pater, A. A. et al. Emergence and evolution of a prevalent new SARS-CoV-2 variant in the United States. *bioRxiv* <https://doi.org/10.1101/2021.01.11.426287> (2021).
- Andersen, K. G. et al. The proximal origin of SARS-CoV-2. *Nat. Med.* **26**, 450–452 (2020).
- Li, X. et al. Emergence of SARS-CoV-2 through recombination and strong purifying selection. *Sci. Adv.* **6**, eabb9153 (2020).
- Elbe, S. & Buckland-Merrett, G. Data, disease and diplomacy: GISAID's innovative contribution to global health. *Global Challenges* **1**, 33–46 (2017).
- Guo, H. et al. Identification of a novel lineage bat SARS-related coronaviruses that use bat ACE2 receptor. *bioRxiv* <https://doi.org/10.1101/2021.05.21.445091> (2021).
- Zhou, H. et al. Identification of novel bat coronaviruses sheds light on the evolutionary origins of SARS-CoV-2 and related viruses. *Cell* **184**, 1–12 (2021).
- Saitou, N. et al. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.* **4**, 406–425 (1987).
- Katoh, K. & Standley, D. M. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.* **30**, 772–380 (2013).
- Peacock, T. P. et al. SARS-CoV-2 one year on: evidence for ongoing viral adaptation. *J. Gen. Virol.* **201**, 001584 (2021).
- Pettersen, E. F. et al. UCSF ChimeraX: Structure visualization for researchers, educators, and developers. *Protein Sci.* **30**, 70–82 (2021).
- Waterhouse, A. M. et al. Jalview Version 2—a multiple sequence alignment editor and analysis workbench. *Bioinformatics* **25**, 1189–1191 (2009).

## Acknowledgements

This work was supported by JSPS KAKENHI Grant Number T20K067670 and by the Platform Project for Supporting Drug Discovery and Life Science Research [Basis for Supporting Innovative Drug Discovery and Life Science Research (BINDS)] from AMED under Grant Number 21am0101108j0005. We thank Keisuke Takahashi, Shunsuke Teraguchi, Tokiko Watanabe, and Songling Li for helpful discussions regarding the preparation of the manuscript.

## Author contributions

D.M.S. conceived the study. K.K. designed and performed the analysis. K.K. and D.M.S. wrote the manuscript. Both authors approved the manuscript.

## Competing interests

The authors declare no competing interests.

## Additional information

**Supplementary information** The online version contains supplementary material available at <https://doi.org/10.1038/s42003-021-02663-4>.

**Correspondence** and requests for materials should be addressed to Kazutaka Katoh or Daron M. Standley.

**Peer review information** *Communications Biology* thanks Mushtaq Hussain and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Primary Handling Editor: Karli Montague-Cardoso.

**Reprints and permission information** is available at <http://www.nature.com/reprints>

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2021