Frequent homologous recombination limited to the Omicron BA.1 lineage among SARS-CoV-2 variants

Hideki Kakeya

Institute of Systems and Information Engineering, University of Tsukuba,

1-1-1 Tennodai, Tsukuba, Ibaraki 305-8573, Japan

ORCID: 0000-0003-3788-9133

Corresponding author

Hideki Kakeya

1-1-1 Tennodai, Tsukuba, Ibaraki 305-8573, Japan

+81-29-853-5255

kake@iit.tsukuba.ac.jp

Keywords

COVID-19, Omicron variant, reverse mutation, homologous recombination, sequence exchange

Abstract

Since the emergence of SARS-CoV-2, variants carrying new mutations have played significant roles in global transmission. Among them, the Omicron variant, which surfaced in November 2021, stands out due to its unprecedented number of mutations, particularly in the spike protein. This paper investigates the anomaly of the BA.1 lineage of the Omicron variant, focusing on the presence of reverse mutations and potential homologous recombination. Using sequence data from GenBank, the study analyzes the mutation patterns of various Omicron subvariants and compares them with each other and with other major variants such as Alpha and Delta. The results reveal a notable peak of triple reverse mutations and relatively high frequencies in the emergence of quadruple or more reverse mutations in BA.1 lineage as traces of homologous recombination. The study also addresses the possibility of contamination by Delta variant sequences, ruling out this explanation based on the accompanying mutations. Given the unique recombination patterns observed only in the BA.1 lineage and the absence of such events in other variants, the study raises concerns about the potential lab origin of the Omicron BA.1 lineage, underscoring the need for rigorous oversight in facilities handling SARS-CoV-2.

Introduction

Since the emergence of SARS-CoV-2, a wide variety of mutations have been observed throughout the COVID-19 pandemic. D614G was the first major mutation observed in early 2020 [1,2]. It is known that D614 is unstable in vivo, not only in humans but also in hamsters [3]. Therefore, the early emergence of the D614G mutation through human-to-human transmission aligns with the expectation of natural evolution.

Following the D614G mutation, a series of additional mutations led to the emergence of key SARS-CoV-2 variants. Among these, B.1.1.7 (Alpha variant), which is believed to have originated in the United Kingdom, and B.1.617.2 (Delta variant), which is believed to have emerged in India, prevailed globally from late 2020 to mid-2021.

Most mutations among the major variants of concern (VOCs) are independent, with high dN/dS ratios [4,5] in the spike protein that remain consistent across all VOCs [6]. Hassan et al. highlighted that many mutations were location-specific [7], which could have contributed to the emergence of distinct variants across different regions.

Though SARS-CoV-2 is an RNA virus, which generally mutates quite often, the speed of mutation is relatively slow compared with other RNA viruses [8]. Indeed, the mutation speed of SARS-CoV-2 is 23.9 times slower than that of the Influenza A virus [9].

From this point of view, the Omicron variant, which emerged in November, 2021, is exceptional in its large amount of mutations. The Omicron variant includes 30 or more non-synonymous (N) mutations and only one

synonymous (S) mutation in the spike protein [10], whereas previous VOCs had only around 10 spike mutations. Phylogenetic analysis suggests that Omicron did not evolve directly from the preceding VOCs [11].

Initially, a few theories were proposed to explain the emergence of Omicron, including strong selective pressure to escape vaccine-induced immunity, prolonged incubation in an immunocompromised human host, or evolution in a non-human species before spilling over into humans [12]. However, the widespread vaccination in developed countries and close monitoring of populations make unnoticed community spread highly unlikely. Furthermore, reported mutation counts in immunocompromised individuals are typically around 10 or fewer [13-15], far less than the number observed in Omicron.

Regarding the theory of evolution in a non-human host, some researchers, including Wei et al., proposed that Omicron may have evolved in mice [16], a theory later supported by Zhang et al. [18]. However, the original strain of SARS-CoV-2 does not infect mice, casting doubt on this idea. Kakeya et al. suggested a lab origin for Omicron, potentially linked to a spillover from transgenic mice [19].

It is also noteworthy that various anomalous reversion mutants have been observed throughout the evolution of SARS-CoV-2. For example, reversions in D614G of the spike proteins were observed in the Delta variant and the Omicron variant BA.2 in the late stages of their community spread [21]. Additionally, sequence data from Omicron variants BA.1, BA.1.1, and BA.2 in the NCBI GenBank revealed a pattern of single reverse mutations without other changes in the spike protein [22]. Comparative analysis of spike protein mutations in major variants showed that BA.1 stood out for its high rate of reverse mutations and lower mutation

diversity [23].

In the previous studies, the author investigated the Omicron BA.1 lineage and found that pure reversion mutants, which contain only reverse mutations and no other mutations in the spike protein, were widespread from the early days of their emergence, showing statistically significant differences compared to the control group [24,25]. The regional and temporal anomalies in the Omicron BA.1 lineage are virtually impossible to explain by current theories of natural mutation and spread through human-to-human infection.

In this paper, the anomaly of the BA.1 lineage and its reverse mutations is analyzed in detail from a genetic perspective, building upon the previous study from an epidemiological viewpoint.

Methods

The sequence data of surface glycoproteins (spike proteins) of SARS-CoV-2 registered in GenBank were accessed in August, 2024. The analyses focused on predominant Omicron variants, as well as the Alpha variant (B.1.1.7) and the Delta variant (B.1.617.2), both of which were the major variants before the emergence of Omicron. For the Omicron variants, six major variants from the BA.1 lineage, each with more than 20,000 registered surface glycoprotein sequences (BA.1, BA.1.1, BA.1.15, BA.1.18, BA.1.20, and BA.1.1.18), were included. Additionally, six major variants that emerged after the surge of the Omicron BA.1 lineage were investigated (BA.2, BA.2.12.1, BA.5.2, BQ.1, XBB.1.5, and JN.1).

For each variant, the consensus sequence of the surface glycoprotein was calculated by first identifying the most frequent sequence length in the first 5,000 samples, then determining the most frequent amino acid at each position. To reduce computational costs, protein sequences with deletions and insertions relative to the consensus sequence were excluded from the analyses. Surface glycoprotein sequences, including only reverse mutations (referred to as pure reversion mutants), were retrieved from the dataset. The reverse mutations were counted for each mutation point, with sequences that had missing reads at any of the mutation points from the original Wuhan strain excluded from the analysis. The number of reverse mutations was tallied for each sequence, and the distribution of reverse mutations for each variant was calculated.

Regarding the Omicron JN.1 variant, which has approximately 60 mutations relative to the original Wuhan strain and an additional 30 mutations compared to Omicron BA.2, reverse mutation patterns of the pure reversion mutants from both BA.2 and the Wuhan strain were analyzed. For Omicron BA.1.1, which has the largest number of registrations, the mutation patterns of pure reversion mutants were analyzed in detail to identify any significant features.

Results

The distribution of reverse mutation counts in the pure reversion mutants of six variants in the BA.1 lineage (BA.1, BA.1.1, BA.1.15, BA.1.18, BA.1.20, BA.1.1.18) and eight other major variants (B.1.1.7, B.1.617.2, BA.2, BA.2.12.1, BA.5.2, BQ.1, XBB.1.5, JN.1) are shown in Figures 1 and 2, respectively. For JN.1, the distribution of reverse mutations with respect to Omicron BA.2 is also shown. As these figures illustrate, only

the variants in the BA.1 lineage exhibit a peak at the triple reversion mutant that exceeds or is equivalent to that at the single reversion mutant.

The mutation patterns of pure reversion mutants in BA.1.1, which has the largest number of entries in GenBank, are shown in Figure 3, with only those mutation patterns having more than 30 entries included. As shown in this figure, the pure reversion mutants of BA.1.1 reflect the trace of homologous recombination, with multiple reverse mutations observed in succession. These successive reverse mutations are responsible for the triple, quadruple, or higher numbers of reverse mutations significant for the BA.1 lineage in Figure 1. The other specific characteristics of reverse mutations found in the SARS-CoV-2 variants are listed in Table 1.

First, reversion mutants carrying the L452R mutation, which is found in the Delta variant, were surveyed. As shown in the table, reverse mutants carrying L452R always exhibit reversions at N440K and G446S, and almost always at K417N (except for two cases). This combination of mutations accounts for less than 4% of mutants exhibiting reversions at K417N, N440K, and G446S. Additionally, only 15% of L452R mutants are registered by the CDC (Centers for Disease Control and Prevention), a more reliable source, while over 60% of the BA.1.1 sequences in GenBank are registered by the CDC. It is also noteworthy that none of the reversion mutants with reverse mutations at spike amino acids 417, 440, 446, 477, and 478 include L452R mutations in between.

Second, the longest successive reversions were searched in the pure reversion mutant sequences of BA.1.1. While the longest successive reversions are observed from amino acids 417 to 505 in Figure 2 (which includes sequences with 30 or more entries), when sequences with fewer entries were examined, 20 entries comprising successive reversions from amino acids 371 to 505 were found.

Third, the peak of triple reversions in JN.1 from BA.2, shown in Table 1, was investigated. These triple reversions all exhibit mutation points at separate positions, indicating that they are not the result of homologous recombination. No successive reversions were found in BA.2.86.1, the ancestor of JN.1, which lacks the additional L455S mutation found in JN.1 as the only difference in the spike protein.

Discussion

It is quite anomalous that traces of homologous recombination are observed only in the variants in the BA.1 lineage and not in other variants. It is known that homologous recombination does not occur frequently in SARS-CoV-2 [26]. From that perspective, the successive reverse mutations observed in the BA.1 lineage are also conspicuous. Akashi et al. reports that the longest homologous recombination is 330 bases long [27]. As Table 1 shows, the longest homologous recombination in BA.1.1 is 135 amino acids long (405 bases long), which is also quite notable.

It is known that the sequences of Omicron BA.1 include many read errors due to contamination from the Delta variant, which is attributed to primer issues and appears as reverse mutations [28]. Since the Delta variant contains the L452R mutation, sequences contaminated by the Delta variant should carry the L452R mutation, along with reverse mutations relative to the Wuhan strain or the Delta variant itself. While the L452R mutation is prevalent in GISAID data, it is found less frequently in GenBank data, which this paper analyzes.

In GenBank data, BA.1.1 carrying L452R makes up less than 4% of the mutants exhibiting reversions of K417N, N440K, and G446S. Additionally, only 15% of L452R mutants are registered by the CDC, which is considered a more reliable source, while over 60% of the BA.1 lineage in GenBank is registered by the CDC. This suggests that GenBank data are less contaminated by the Delta variant.

If the reversions were mainly caused by read errors due to contamination from the Delta variant, the peak of reversions should have occurred at the outset. Indeed, Kakeya reports that worldwide reads with full spike mutations in BA.1, BA.1.1, and BA.1.1.18 registered in GISAID were rare in the early days and gradually increased, indicating that read errors were most frequent at the onset of their emergence. However, Kakeya also reports that the ratios and counts of reversions in BA.1, BA.1.1, and BA.1.1.18 in GenBank (most of which are from the United States) reached their peaks later, not in the early days when the Delta variant was

more prevalent than Omicron [25]. It is also noteworthy that this paper focuses only on data where all Omicron mutation sites are read. As Martin et al. [28] show, reversions increase as the amount of missing data at mutation sites increases.

From the large amount of circumstantial evidence presented above, the recombination in the BA.1 lineage is not considered to be the result of read errors. Therefore, the frequent recombination observed only in BA.1 should be regarded as a real phenomenon, and its mechanism of emergence should be investigated.

It is true that some variants have emerged through the recombination of prior variants. However, this does not explain the massive emergence of recombination within a single lineage. One possible explanation is the leakage of artificially created recombinants from a laboratory. As for the original Wuhan strain, it is reported that many scientists in governmental agencies, such as Jason Bannan of the Federal Bureau of Investigation, and Hardham, Robert Cutlip, and Jean-Paul Chretien of the National Center for Medical Intelligence of the Defense Intelligence Agency, all found evidence of human manipulation of the virus in SARS-CoV-2, which was silenced by the Director of National Intelligence, Avril Haines [29].

Many lab-leak accidents have occurred historically, and the number of such incidents has been increasing due to the spread of genetic engineering in recent years [30,31]. At the end of 2021, a researcher in Taiwan was bitten by a mouse in a biosafety level 3 laboratory and was infected with the Delta variant of SARS-CoV-2, unknowingly spreading the disease [32]. It is also true that the lab origin of the Omicron variants has been suggested by many studies in the past [19-25,33]. A thorough investigation, including oversight of laboratories dealing with SARS-CoV-2, is needed to prevent the emergence of highly infectious variants in the future.

Data Availability

The source code used for this study is available at

https://visual-media-lab.github.io/data/PRM_source_code/index.html

Conflicts of interest

The author declares no conflict of interest exists.

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(4), 812–827. https://doi.org/10.1016/j.cell.2020.06.043

[2] Volz, E., Hill, V., McCrone, J. T., et al. (2021). Evaluating the effects of SARS-CoV-2 spike mutation D614G on transmissibility and pathogenicity. Cell, 184(1), 64–75. https://doi.org/10.1016/j.cell.2020.11.020

[3] Hou, Y. J., Chiba, S., Halfmann, P., et al. (2020). SARS-CoV-2 D614G variant exhibits efficient replication ex vivo and transmission in vivo. Science, 370(6523), 1464–1468. https://doi.org/10.1126/science.abe8499

[4] Miyata, T., & Yasunaga, T. (1986). Molecular evolution of mRNA: A method for estimating evolutionary rates of synonymous and amino acid substitutions from homologous nucleotide sequences and its application. Journal of Molecular Evolution, 16(1), 23–36. https://doi.org/10.1007/BF01732067

[5] Li, W. H., Wu, C. I., & Luo, C. C. (1985). A new method for estimating synonymous and nonsynonymous rates of nucleotide substitution considering the relative likelihood of nucleotide and codon changes. Molecular Biology and Evolution, 2(2), 150–174. https://doi.org/10.1093/oxfordjournals.molbev.a040343

[6] Nikolaidis, M., Papakyriakou, A., Chlichlia, K., et al. (2022). Comparative analysis of SARS-CoV-2 variants of concern, including Omicron, highlights their common and distinctive amino acid substitution patterns, especially at the spike ORF. Viruses, 14(4), 707. https://doi.org/10.3390/v14040707

[7] Hassan, S. S., Kodakandla, V., Redwan, E. M., et al. (2022). Non-uniform aspects of the SARS-CoV-2 intraspecies evolution reopen question of its origin. International Journal of Biological Macromolecules, 222, 972–993. https://doi.org/10.1016/j.ijbiomac.2022.09.184

[8] Amicone, M., Borges, V., Alves, M. J., et al. (2022). Mutation rate of SARS-CoV-2 and emergence of mutators during experimental evolution. Evolution, Medicine, and Public Health, 2022(1), 142–155. https://doi.org/10.1093/emph/eoac010

[9] Kawasaki, Y., Abe, H., & Yasuda, J. (2023). Comparison of genome replication fidelity between SARS-CoV-2 and influenza A virus in cell culture. Scientific Reports, 13, 13105. https://doi.org/10.1038/s41598-023-40463-4

[10] Callaway, E. (2021). Heavily mutated Omicron variant puts scientists on alert. Nature, 600(7888), 21. https://doi.org/10.1038/d41586-021-03552-w

[11] Jung, C., Kmiec, D., Koepke, L., et al. (2022). Omicron: What makes the latest SARS-CoV-2 variant of concern so concerning? Journal of Virology, 96(3), 02077–21. https://doi.org/10.1128/jvi.02077-21

[12] Mallapaty, C. (2022). The hunt for the origin of Omicron. Nature, 602(7897), 26–28. https://doi.org/10.1038/d41586-022-00215-2

[13] Choi, B., Choudhary, M. C., Regan, J., et al. (2020). Persistence and evolution of SARS-CoV-2 in an immunocompromised host. The New England Journal of Medicine, 383(23), 2291–2293. https://doi.org/10.1056/NEJMc2031364

[14] Kemp, S. A., Collier, D. A., Datier, R. P., et al. (2021). SARS-CoV-2 evolution during treatment of chronic infection. Nature, 592(7852), 277–282. https://doi.org/10.1038/s41586-021-03291-y

[15] Truong, T. T., Ryutov, A., Pandey, U., et al. (2021). Increased viral variants in children and young adults with impaired humoral immunity and persistent SARS-CoV-2 infection: A consecutive case

series. EBioMedicine, 67, 103355.
https://doi.org/10.1016/j.ebiom.2021.103355

[16] Wei, C., Shan, K. J., Wang, W., et al. (2021). Evidence for a mouse origin of the SARS-CoV-2 Omicron variant. Journal of Genetics and Genomics, 48(12), 1111–1121.
https://doi.org/10.1016/j.jgg.2021.12.003

[17] Zhang, W., Shi, K., Geng, Q., et al. (2022). Structural basis for mouse receptor recognition by SARS-CoV-2 omicron variant. Proceedings of the National Academy of Sciences, 119(5), e2206509119.
https://doi.org/10.1073/pnas.2206509119

[18] Piplani, S., Singh, P. K., Winkler, D. A., et al. (2021). In silico comparison of SARS-CoV-2 spike protein-ACE2 binding affinities across species and implications for virus origin. Scientific Reports, 11, 13063. https://doi.org/10.1038/s41598-021-92388-5

[19] Kakeya, H., & Matsumoto, Y. (2022). A probabilistic approach to evaluate the likelihood of artificial genetic modification and its application to SARS-CoV-2 Omicron variant. ISPJ Transactions on Bioinformatics, 15, 22–29.
https://doi.org/10.2197/ipsjtbio.15.22

[20] Kakeya, H., Arakawa, H., & Matsumoto, Y. (2023). Multiple probabilistic analyses suggest non-natural origin of SARS-CoV-2 Omicron variant. Zenodo.
https://doi.org/10.5281/zenodo.7470652

[21] Kakeya, H., & Matsumoto, Y. (2023). Repeated emergence of probabilistically and chronologically anomalous mutations in SARS-CoV-2 during the COVID-19 pandemic. Zenodo.
https://doi.org/10.5281/zenodo.8216232

[22] Tanaka, A., & Miyazawa, T. (2023). Unnaturalness in the evolution process of the SARS-CoV-2 variants and the possibility of deliberate natural selection. Zenodo. https://doi.org/10.5281/zenodo.8361577

[23] Kakeya, H., & Kanazaki, T. (2023). Anomalous biases of reverse mutations in SARS-CoV-2 variants. Jxiv. https://doi.org/10.51094/jxiv.545

[24] Kakeya, H. (2024). Anomalous US-wide prevalence of reversion mutants in the emergence of Omicron BA.1. Research Square. https://doi.org/10.21203/rs.3.rs-4919461/v1

[25] Kakeya, H. (2024). Anomalies in regional and chronological distributions of SARS-CoV-2 Omicron BA.1.1 lineage in the United States. medRxiv. https://doi.org/10.1101/2024.08.14.24311991

[26] Akaishi, T., Fujiwara, K., & Ishii, T. (2023). Genetic recombination sites away from the insertion/deletion hotspots in SARS-related coronaviruses. Tohoku Journal of Experimental Medicine, 259(1), 17–26.

[27] Akaishi, T., Horii, A., & Ishii, T. (2022). Sequence exchange involving dozens of consecutive bases with external origin in SARS-related coronaviruses. Journal of Virology, 96(15), 1–4. https://doi.org/10.1128/jvi.01002-22

[28] Martin, D. P., Lytras, S., Lucaci, A. G., et al. (2022). Selection analysis identifies clusters of unusual mutational changes in Omicron lineage BA.1 that likely impact spike function. Molecular Biology and Evolution, 39(6), msac061. https://doi.org/10.1093/molbev/msac061

[29] Gordon, M. R. and Strobel, W. P. (2024) Behind Closed Doors: The Spy-World Scientists Who Argued Covid Was a Lab Leak. Wall Street Journal.
https://www.wsj.com/politics/national-security/fbi-covid-19-pandemic-lab-leak-theory-dfbd8a51

[30] Butler, D. (2011). Fears grow over lab-bred flu. Nature, 480(7378), 421–422.
https://doi.org/10.1038/480421a

[31] - (2014). Biosafety in the balance. Nature, 510(7506), 443.
https://doi.org/10.1038/510443a

[32] Silver, A. (2022). Taiwan's science academy fined for biosafety lapses after lab worker contracts COVID-19. Science.
https://doi.org/10.1126/science.ada0525

[33] Arakawa, H. (2024). The natural evolution of RNA viruses provides important clues about the origin of SARS-CoV-2 variants. SynBio, 2(3), 285–297.
https://doi.org/10.3390/synbio2030017

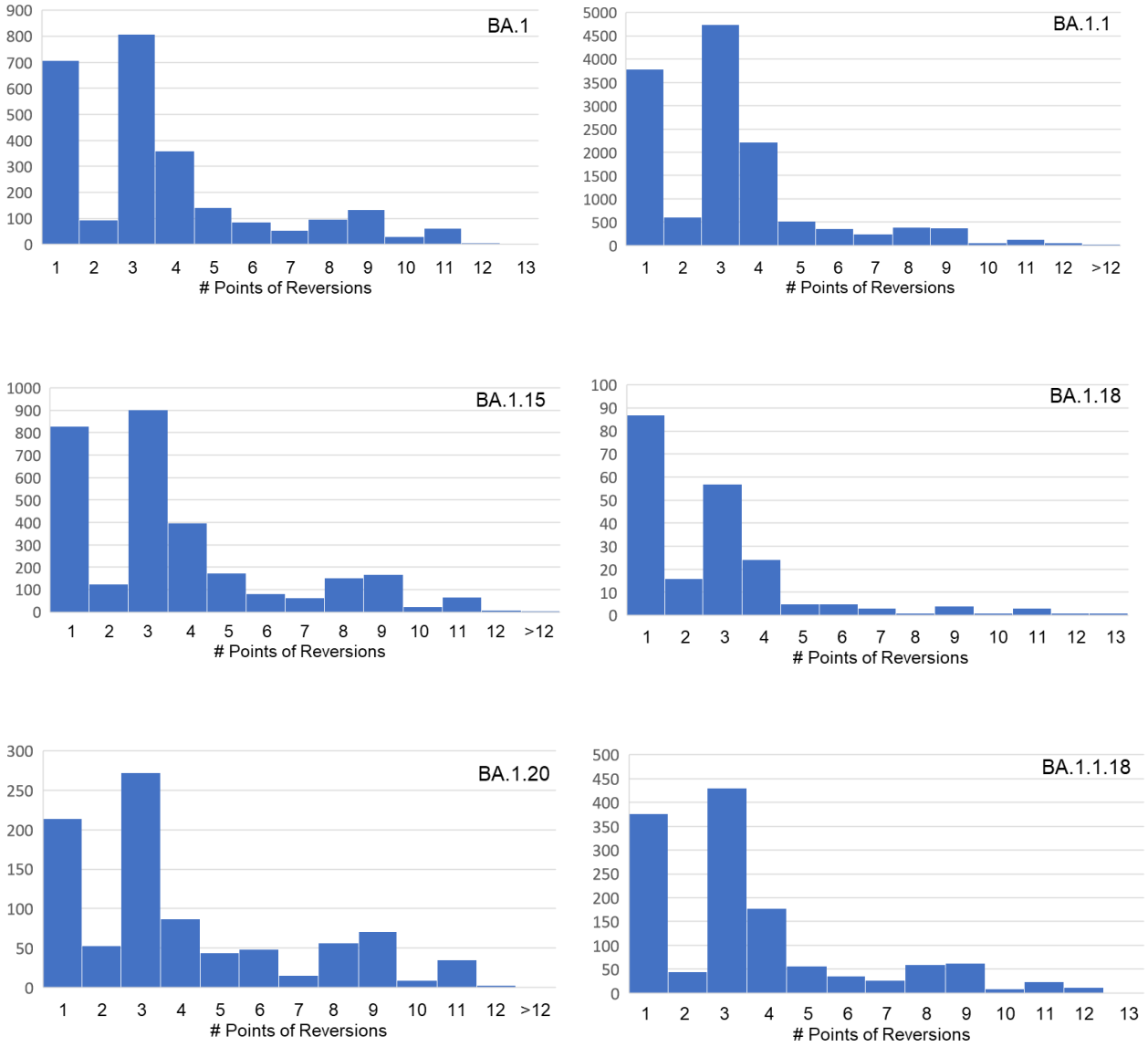

Figure 1. Distribution of reverse mutation counts in the pure reversion mutants of six variants in the BA.1 lineage (BA.1, BA.1.1, BA.1.15, BA.1.18, BA.1.20, BA.1.1.18).

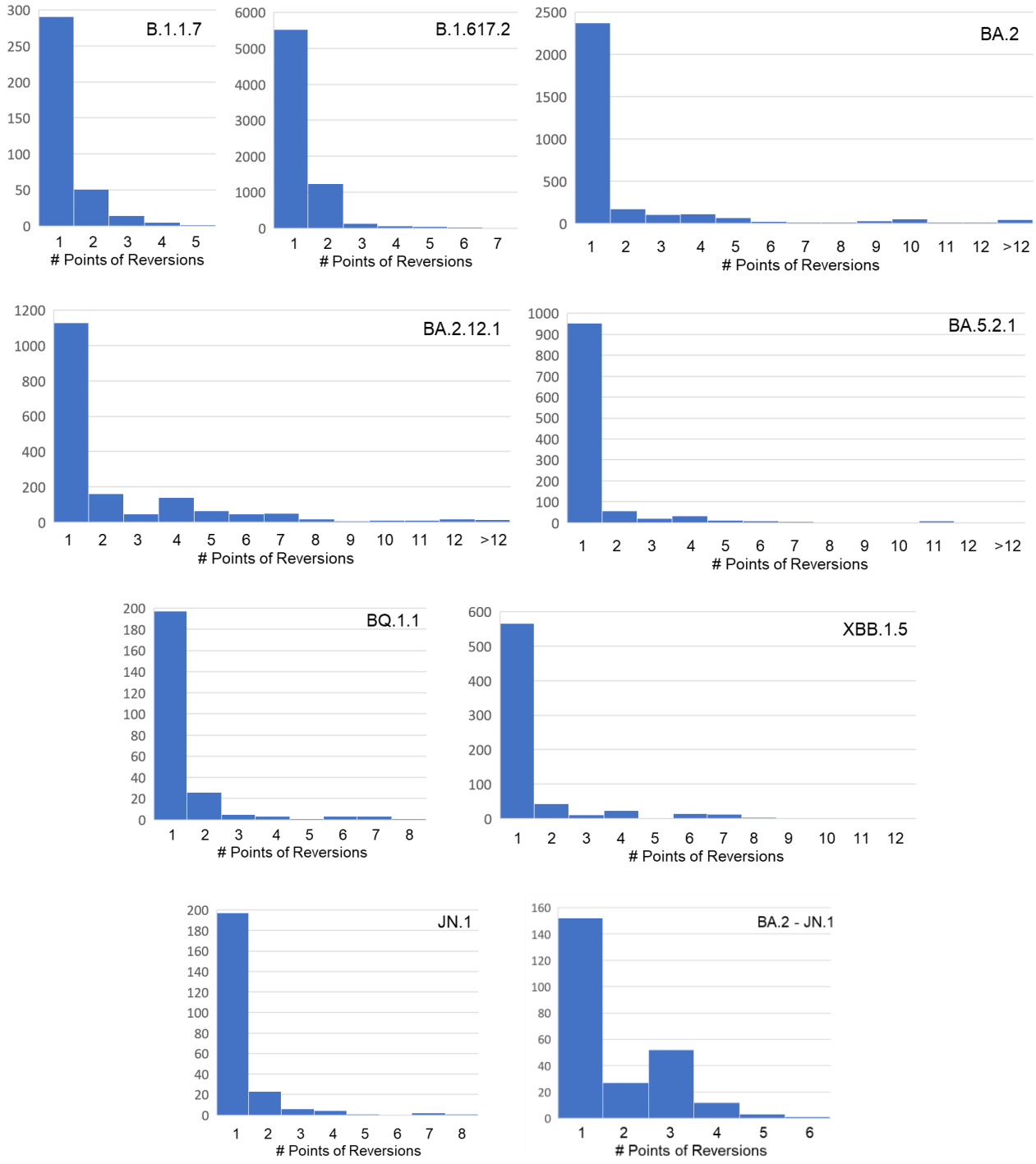

Figure 2. Distribution of reverse mutation counts in the pure reversion mutants of eight major variants (B.1.1.7, B.1.617.2, BA.2, BA.2.12.1, BA.5.2, BQ.1, XBB.1.5, JN.1) .

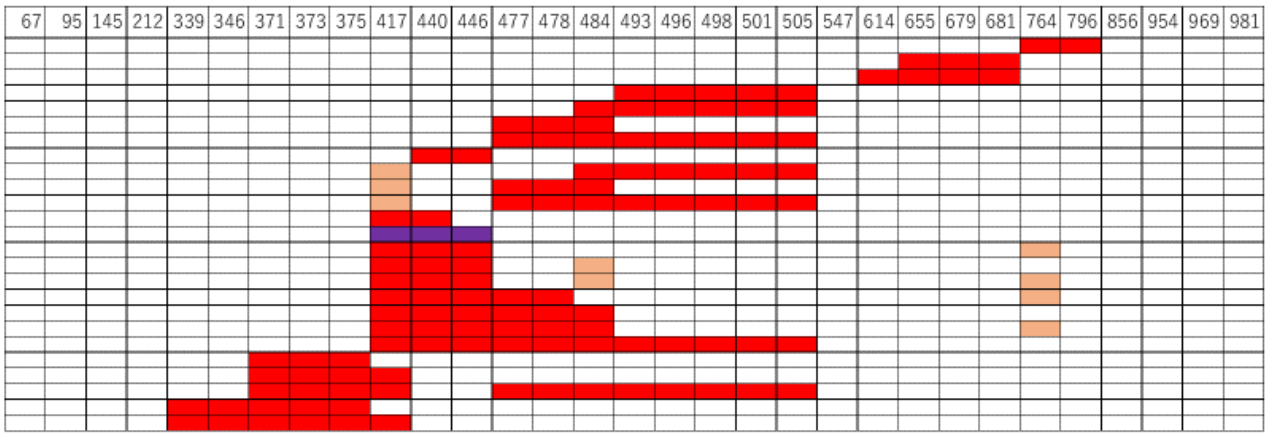

Figure 3. Mutation patterns of pure reversion mutants in BA.1.1 (more than 30 entries). Colored cells represent the reversions of amino acids to the original Wuhan strain. The purple color represents the most frequent pattern, while the light orange color represents a stand-alone mutation.

Table 1. Characteristics of reverse mutations found in the SARS-CoV-2 variants.

1	Reverse mutants carrying L452R are reversions at 417-446 (171 entries) or 440-446 (two entries). Among them, 15% of L452R mutants are registered by the CDC, while over 60% of the BA.1.1 sequences in GenBank are registered by the CDC. None of the reversion mutants that have reverse mutations at spike amino acids 417, 440, 446, 477, and 478, do not include L452R mutations in between.
2	Successive reversions from amino acids 371 to 505 were found (20 entries).
3	Triple reversion of JN.1 from BA.2 in Figure 2 all exhibit mutations points at separate positions. No successive reversions were found either in BA.2.86.1 (pure reversions from BA.2).