Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID-19. The COVID-19 resource centre is hosted on Elsevier Connect, the company's public news and information website.

Elsevier hereby grants permission to make all its COVID-19-related research that is available on the COVID-19 resource centre - including this research content - immediately available in PubMed Central and other publicly funded repositories, such as the WHO COVID database with rights for unrestricted research re-use and analyses in any form or by any means with acknowledgement of the original source. These permissions are granted for free by Elsevier for as long as the COVID-19 resource centre remains active.

International Journal of Biological Macromolecules 204 (2022) 356–363

Contents lists available at ScienceDirect

International Journal of Biological Macromolecules

journal homepage: www.elsevier.com/locate/ijbiomac

Human SARS-CoV-2 has evolved to increase U content and reduce genome size

Yong Wang [a,*,1], Xin-Yu Chen [a,1], Liu Yang [a], Qin Yao [b], K.P. Chen [b]

[a] School of Food and Biological Engineering, Jiangsu University, 301 Xuefu Road, Zhenjiang 212013, China
[b] School of Life Sciences, Jiangsu University, 301 Xuefu Road, Zhenjiang 212013, China

ARTICLE INFO

Keywords:
Coronavirus
Transmission
Mutation
Evolution
Variant
Nucleotide

ABSTRACT

Infections caused by SARS-CoV-2 have brought great harm to human health. After transmission for over two years, SARS-CoV-2 has diverged greatly and formed dozens of different lineages. Understanding the trend of its genome evolution could help foresee difficulties in controlling transmission of the virus. In this study, we conducted an extensive monthly survey and in-depth analysis on variations of nucleotide, amino acid and codon numbers in 311,260 virus samples collected till January 2022. The results demonstrate that the evolution of SARS-CoV-2 is toward increasing U-content and reducing genome-size. C, G and A to U mutations have all contributed to this U-content increase. Mutations of C, G and A at codon position 1, 2 or 3 have no significant difference in most SARS-CoV-2 lineages. Current viruses are more cryptic and more efficient in replication, and are thus less virulent yet more infectious. Delta and Omicron variants have high mutability over other lineages, bringing new threat to human health. This trend of genome evolution may provide a clue for tracing the origin of SARS-CoV-2, because ancestral viruses should have lower U-content and probably bigger genome-size.

1. Introduction

The pandemic of COVID-19 (coronavirus disease 2019) has brought over 370 million infection cases and over 5.6 million deaths worldwide by 30 January 2022 [1]. Its causative virus, SARS-CoV-2, has a single-strand genomic RNA of approximately 30,000 nucleotides [2]. Since outbreak of this disease, great efforts have been made to establish fast diagnostic methods [3,4], to develop effective therapeutic drugs and vaccines [5–8], to implement strict inspections on transportations of goods, and to enforce certain restrictions on activities of people [9,10]. All these efforts have helped reduce virus transmission effectively and treat infected people properly [11,12]. However, with fast emergence of new variants [13,14], the world is still facing big challenges in controlling this pandemic.

Fast emergence of new variants is largely due to high mutability of SARS-CoV-2 genome, which is prone to mutation though it encodes a proofreading enzyme to prevent replication error [15,16]. A great number of mutations have been recorded in SARS-CoV-2 genome [17–19], some of which affect structure and function of viral proteins [20–22]. Rapid and extensive transmission worldwide has provided more chances for SARS-CoV-2 genome to accumulate mutations. After frequent mutations for around two years, SARS-CoV-2 has diverged greatly and formed many different lineages [23,24]. While monitoring mutations in specific viral proteins facilitates the development of effective vaccines and antiviral drugs [25,26], understanding the evolution trend of SARS-CoV-2 genome could help foresee difficulties in controlling spread of the virus [27,28].

In GISAID (global initiative on sharing all influenza data) database (www.gisaid.gov), SARS-CoV-2 viruses are currently classified into nine clades and five variants of concern (VOCs). The nine clades are named based on presence of specific amino acid at particular site. For examples, clades S and L are named because amino acid 84 of their NSP8 (non-structural protein 8) is serine (S) and leucine (L), respectively. Clades V and G are derived from clade L, in which amino acid 251 of NSP3 protein is V (valine) and amino acid 614 of S (spike) protein is G (glycine), respectively. Clades GH, GK, GR and GV are all derived from clade G. Their names are based on presence of histidine (H), lysine (K), arginine (R) and valine (V) at specific sites, respectively. Clade GRY is derived from clade GR, because tyrosine (Y) is at amino acid position 501 in its S protein. The five VOCs are all derived from G-series clades due to

* Corresponding author.
E-mail addresses: ywang@ujs.edu.cn (Y. Wang), eric_chen_xy@sina.com (X.-Y. Chen), yangliu2129@163.com (L. Yang), yaoqin@ujs.edu.cn (Q. Yao), kpchen@ujs.edu.cn (K.P. Chen).
[1] Wang Y and Chen XY are jointly first authors.

https://doi.org/10.1016/j.ijbiomac.2022.02.034
Received 8 November 2021; Received in revised form 6 February 2022; Accepted 7 February 2022
Available online 8 February 2022
0141-8130/© 2022 The Authors. Published by Elsevier B.V. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

additional nucleotide mutations and deletions/insertions. They are named Alpha, Beta, Gamma, Delta, and Omicron, respectively [29].

The GISAID database has collected over 4.8 million full-genome sequences of SARS-CoV-2 by 31 January 2022. With these sequence resources, tracing mutations in SARS-CoV-2 can be conducted separately for different lineages along certain timelines. Therefore, in this study, we conducted an extensive monthly survey on variations of nucleotide, amino acid and codon numbers in twelve SARS-CoV-2 lineages. Further in-depth analyses reveal that, in all surveyed SARS-CoV-2 lineages, U content has been steadily increased, and genome stability has been slightly reduced. These two changes are considered to make the virus less virulent yet more infectious [30,31].

2. Materials and methods

2.1. Sampling and processing of genome sequences

Genome sequences were downloaded from GISAID (www.epicov.org) database for twelve SARS-CoV-2 lineages including seven clades (S, L, V, G, GH, GR and GV) and five variants (Alpha, Beta, Gamma, Delta and Omicron). Filters were set to retrieve sequences with human as host, and having high coverage and complete collection date. Sequences were downloaded separately for different lineages and different months. The downloaded sequences (in Fasta format) were analysed using Alignment Explorer of MEGA X [32] to exclude those having any ambiguous base (e.g. N, R and Y for any base, purine and pyrimidine, respectively). Then, each of them was trimmed to retain 29,769 nucleotides for the seven clades. For Alpha, Beta, Gamma, Delta and Omicron variants, 29,750, 29,751, 29,764, 29,756 and 29,742 nucleotides were retained respectively, because they have various deletions/insertions in the region for survey (Fig. 1).

2.2. Counting of nucleotide, amino acid and codon numbers

The trimmed sequences were loaded into a computer program to count the numbers of nucleotides in whole survey region and the numbers of amino acids and codons in ORFs (open reading frames). C++ scripts of computer programs are available upon request.

2.3. Measurement of free energy

Free energy of viral genomic sequence was measured using RNAstructure 5.7, which uses a dynamic programming algorithm to predict RNA secondary structures based on the principle of minimizing free energy [33]. The minimum free energy of an analysed viral sequence was used to estimate genome stability of viruses. For each viral sequence, the first 200 nucleotides (i.e. 5′-untranslated region) were loaded into the program directly. The rest of the nucleotides were segmented into 29 pieces of 1000 nucleotides plus the last one of 542–569 nucleotides, and then loaded into the program in order.

2.4. Statistical analysis

SPSS software (version 17.0) was used to conduct independent-sample *t*-test for comparing the overall variation in nucleotide, amino acid and codon numbers, and for comparing variation of a specific nucleotide at different codon positions. The variation is considered significant when $p < 0.05$.

3. Results

3.1. Variations in nucleotide numbers

Based on the survey of all 311,260 virus samples, nucleotide numbers of the twelve SARS-CoV-2 lineages were obtained (Fig. 2). In all lineages, C numbers are significantly reduced (except Omicron variant), while U numbers are significantly increased. Among them, S clade has the highest variation in both C and U numbers (−13.3 and +14.2), followed by GH clade (−9.5 and +11.2). The C-down and U-up mutations are observed in all lineages, even though the viruses transmitted for only a few months (e.g. L, V and Omicron). G-down and G-up mutations are observed in nine and two lineages respectively, whereas A-down and A-up mutations are observed in eight and three lineages respectively. These data show that mutations of C, G and A into U have occurred in most SARS-CoV-2 lineages.

It is to be noted that, reductions of U number from April to June 2021 in G-series clades are due to exclusion of virus samples with genomic deletions. All current SARS-CoV-2 variants are derived from G-series clades [29], most of which involve nucleotide deletions (Fig. 1). From April to June 2021, more and more G-series viruses contained shortened genomes. Correspondently, less and less G-series viral genomes are eligible for survey (Fig. 2, values in red round frame), because our survey compares difference of nucleotide numbers within specific lengths (Fig. 1). Therefore, G-series viruses have evolved in two directions. One direction is followed by the majority of viruses which underwent fast evolution in elevating U-content and reducing genome-

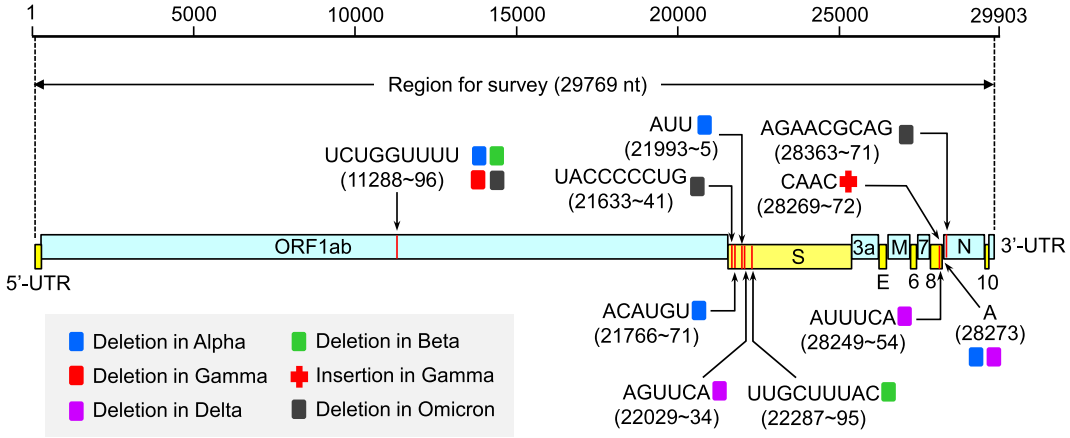

Fig. 1. Genomic region of SARS-CoV-2 for survey. SARS-CoV-2 genomic sequence of 29,903 nucleotides was used as reference [2]. Viral genomic region corresponding to No. 66 to No. 29834 (totalling 29,769 nucleotides) of the reference sequence was taken for survey. Capital letters pointed to the thin red line indicate deleted nucleotides with deletion site below them. Insertion of "CAAC" and deletion of "A" occur in the intergenic region following ORF8. Abbreviations: UTR (untranslated region), ORF (open reading frame), S (spike), E (envelope), M (matrix) and N (nucelocapsid). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

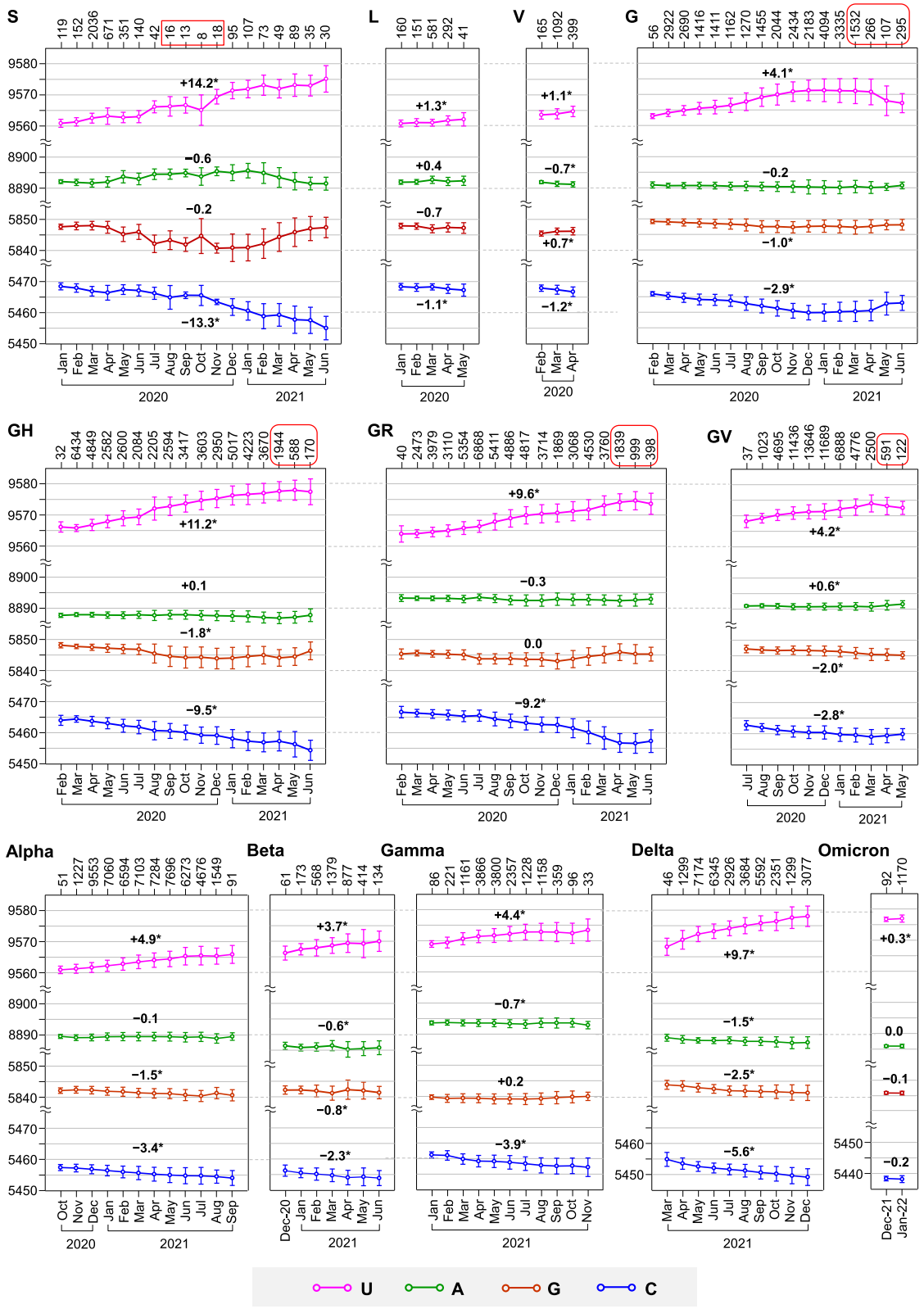

Fig. 2. Variations of nucleotide numbers in SARS-CoV-2 lineages. Nucleotide numbers of each month are presented as mean ± standard deviation with sample numbers on top of the graph. Data from sample number below 30 are excluded for analysis, except those for S clade (in red square frame). Values in red round frame indicate obvious decline in sample numbers compared to previous months. Value above or below each line indicates overall variation between the last and the first month. * indicates significant difference ($p < 0.05$). GK and GRY clades are not surveyed separately, because sequences of these two clades overlap greatly with those of Delta and Alpha variants. Please refer to Table S1 for original data (values in blue colour). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

size, thus becoming ancestors of the new variants. Viruses under this evolutional direction are not considered as of G-series, because they do not have the required genome length (i.e. 29,769 nucleotides). The other direction is followed by the minority of viruses, which evolved relatively slowly. These viruses were slower in elevating U-content and reducing genome-size. Viruses under this evolutional direction are considered as of G-series. Therefore, their U-content in May/June 2021 is generally lower than that in April 2021. Correspondently, their C-content shows a general trend of increase after April 2021, while G- and A-contents are not changed dramatically (Fig. 2).

3.2. Variations in amino acid and codon numbers

Significant variations in nucleotide numbers have resulted in changes of amino acid numbers at various degrees. Numbers of proline, threonine and alanine are significantly reduced in 7–9 lineages, while those of isoleucine, leucine, phenylalanine, valine and tyrosine are significantly increased in 6–9 lineages (Fig. S1). Correspondently, codon numbers for proline (CCA and CCU), threonine (ACA) and alanine (GCU) are significantly reduced in 6–8 lineages, while those for isoleucine (AUU), phenylalanine (UUU), valine (GUU), tyrosine (UAU) and cysteine (UGU) are significantly increased in 6–8 lineages (Fig. S2). In general, codons with a reduced number are C-rich, and those with an increased number are U-rich, being consistent with C-down and U-up mutational trend shown in Fig. 2.

Among the 20 amino acids, threonine was reduced to the greatest extent in GR clade (−3.4), followed by proline in S clade (−2.9). Isoleucine was increased to the greatest extent in S clade (+2.5) followed by tyrosine (+2.4) in S clade (Fig. S1). Furthermore, 16 amino acids have been significantly changed in S clade. It is understandable because it has undergone transmission for over eighteen months. In lineages that have undergone transmission for over twelve months (i.e. G, GH and GR clades), 10–14 amino acids are changed significantly. Compared to these clades, Alpha, Beta, Gamma and Delta variants have relatively higher mutation rate. After transmission for seven to twelve months, 13–15 amino acids are significantly changed already (Fig. S1).

3.3. Mutation patterns of nucleotides

In order to analyse the mutation patterns of nucleotides, we made a statistic on the number of variations among C, G, A and U at different codon positions (Table S3, data within green frame). Based on these data, overall variations of nucleotides at various codon positions were obtained (Table 1). They can be used to infer mutation patterns involved in genome evolution. For example, in S clade, at codon position 1, only U number is increased by 4.7. It is thus considered that 3.5, 0.3 and 0.9 nucleotides of C, G and A have been mutated into U, respectively. Then, at codon position 2, only C number is reduced by 5.0. It is thus considered that 1.3, 0.1 and 3.6 nucleotides of C have been mutated into G, A and U, respectively. As for codon position 3, 4.7 and 1.2 nucleotides of C and G are reduced, and 1.1 and 4.8 nucleotides of A and U are increased. Thus, it is considered that 5.9 (4.7 + 1.2) nucleotides of C/G have been mutated into U/A (4.8 + 1.1).

Based on calculations using data listed in Table 1, mutation patterns of nucleotides in all SARS-CoV-2 lineages were obtained (Fig. 3). It is found that, (i) C to U mutation occurs frequently in all lineages except L and V clades. (ii) C/G to U/A mutation mainly occurs in S, L, GH and Alpha lineages. (iii) C/A to U/G mutation occurs predominantly in V clade and Gamma variant. It also occurs frequently in L, G, GR, Beta and Delta lineages. (iv) G to U mutation occurs frequently in G, GV and Delta lineages. (v) G to A and C to A mutations occur specifically in L and GR lineages, respectively. Mutation patterns of all linages (merged data of twelve lineages) show that C is the major target for mutation, and U is the major product of mutation (Fig. 3).

Table 1
Overall variations of nucleotide numbers at different codon positions.

Lineage	Codon position	C	G	A	U
S	1	−3.5	−0.3	−0.9	+4.7
	2	−5.0	+1.3	+0.1	+3.6
	3	−4.7	−1.2	+1.1	+4.8
L	1	0.0	−0.4	+0.4	0.0
	2	−0.6	−0.4	+0.3	+0.7
	3	−0.4	+0.1	−0.3	+0.6
V	1	−0.6	+0.2	−0.4	+0.7
	2	−0.3	0.0	−0.3	+0.5
	3	−0.3	+0.5	−0.1	−0.1
G	1	−1.2	−0.6	−0.1	+1.9
	2	−1.3	+0.4	−0.1	+1.0
	3	−0.1	−0.3	−0.1	+0.5
GH	1	−1.7	−0.9	−0.2	+2.8
	2	−4.6	+0.4	+0.1	+4.1
	3	−2.2	−1.0	+0.4	+2.8
GR	1	−1.1	+0.5	−1.3	+1.9
	2	−4.3	+0.1	+1.3	+3.0
	3	−2.2	−0.5	−0.3	+2.9
GV	1	−0.8	−1.0	−0.1	+1.8
	2	−1.1	0.0	0.0	+1.2
	3	−1.1	−0.9	0.0	+2.0
Alpha	1	−0.5	−0.7	0.0	+1.2
	2	−1.0	0.0	−0.1	+1.2
	3	−1.8	−0.7	−0.1	+2.5
Beta	1	−0.8	−0.3	−0.2	+1.3
	2	−0.7	+0.1	0.0	+0.6
	3	−1.0	−0.5	−0.4	+1.9
Gamma	1	−0.3	+0.2	−0.5	+0.6
	2	−2.3	0.0	+0.3	+2.0
	3	−1.3	0.0	−0.2	+1.5
Delta	1	−0.2	−2.2	−0.7	+3.2
	2	−2.4	+0.4	−0.6	+2.6
	3	−2.0	+0.5	−0.2	+2.7
Omicron	1	0.0	0.0	0.0	0.0
	2	0.0	0.0	0.0	0.0
	3	−0.2	−0.1	0.0	+0.3

This table lists increased (+) or reduced (−) number of nucleotides at different codon positions over the survey period. Please refer to Table S3 (values within green frame) for detailed data.

3.4. Mutations at different codon positions

In order to understand whether nucleotide mutations occur more frequently at codon position 1, 2 or 3, we made a statistics on monthly variations of C, G, A and U at different codon positions (Table S3, data within purple frame). Average monthly variation rate of nucleotides at different codon positions (Fig. S3) shows that only GR, Gamma and Delta lineages have significant difference in C and G mutations between different codon positions. Mutations of all other nucleotides have no significant difference between codon positions. This means that most mutations of C, G and A take place indiscriminately regardless of which positions they are at. This is true even when the mutation leads to formation of a premature stop codon. As we have identified, codon 254 of ORF3a has been mutated from GGA to UGA in Beta variant. Codons 27 and 68 of ORF8 have been mutated from CAA and AAA to UAA in Alpha variant.

3.5. Genome stability of mutated viruses

Nucleotide composition in an RNA strand has a close relationship with stability of its secondary structure. An RNA strand with high U + A content will form a secondary structure less stable than that with high C + G content, because C-G and U-A base pairs have three and two hydrogen bonds, respectively. Thus, a viral genome with increased U-content will have lower stability. Such genome can be unfolded more easily for replication. In fact, within the region for survey (totalling 29,769 nucleotides), the highest U number has reached 9591 in individual viruses, being 31 higher over the reference sample. As for the five

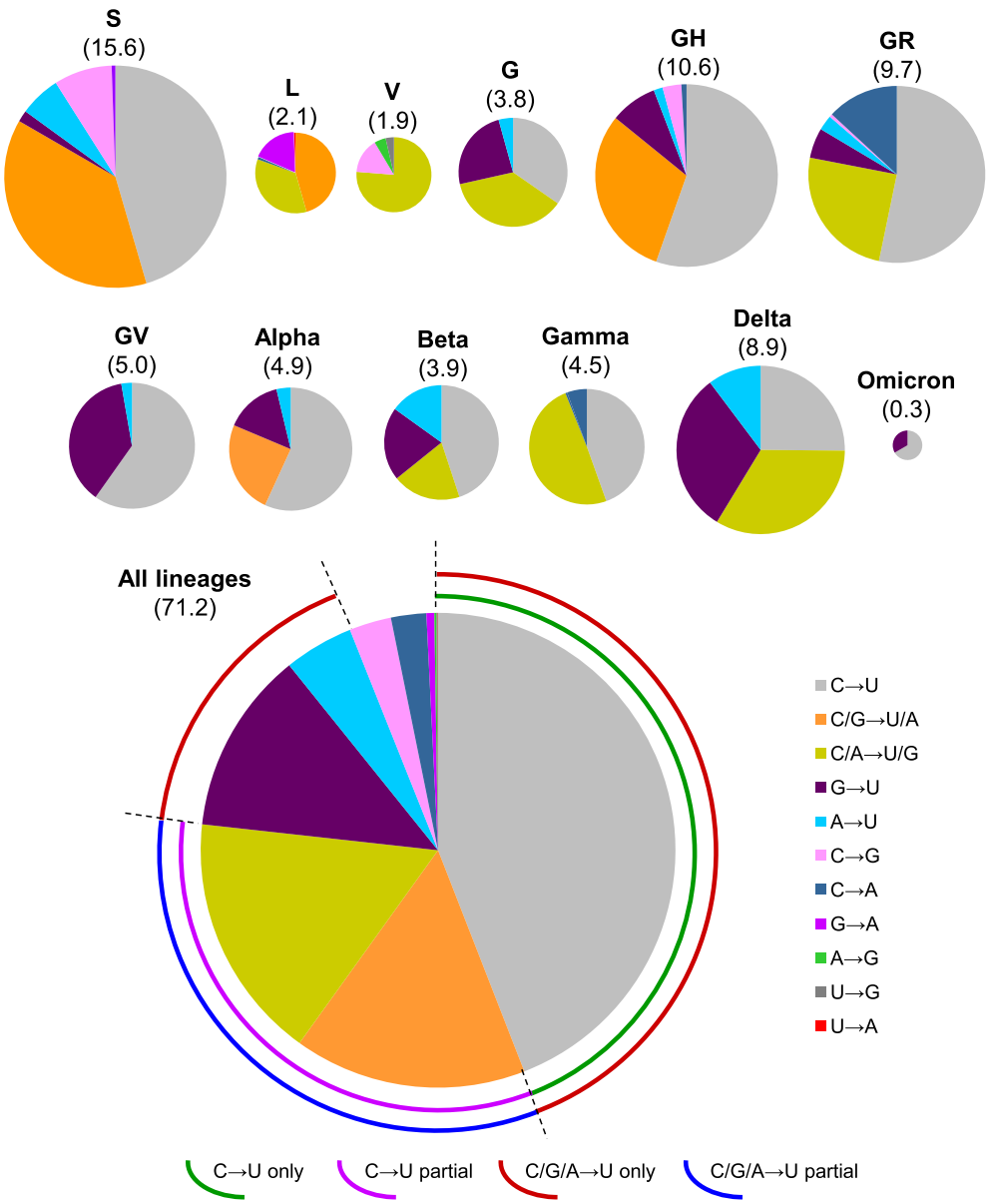

Fig. 3. Mutation patterns involved in genome evolution of SARS-CoV-2. Size of each pie chart is in scale with total number of nucleotide mutations, which is indicated with number below lineage name. Detailed data for each lineage are listed in Table S3 (values in blue colour, within green frame). Pie chart of all lineages is based on merged data of twelve lineages.

variants, if the deleted U numbers (Fig. 1) are compensated, their U numbers could be 29–37 higher than the reference sample (Fig. 4A).

Stability of 5′-UTR (1–200 nt) is considerably reduced in Beta and Delta variants, and in individual samples with the highest U level in S and V clades (Fig. 4B). It is found that, a C to U mutation occurs at site 142 in S clade, a G-to-U mutation occurs at site 109 in V and Beta lineages, and a G to U mutation occurs at site 145 in Delta variant. This single nucleotide mutation has led to a 3–5% reduction of 5′-UTR stability. Stability of TSS-to-end region is reduced slightly in all samples (Fig. 4C). This slight reduction is understandable, because increased U number only accounts for a very low percentage in the whole genome. Yet, the trends of U-content increase and genome-stability reduction are obvious in all lineages.

4. Discussion

4.1. Diversity of nucleotide mutation

Mutation patterns described in Fig. 2 do not include U, G and A to C. This does not mean their absence in genome evolution of SARS-CoV-2. In fact, monthly variations of codon numbers show that any nucleotide may be mutated to any other nucleotide during transmission of the virus (Table S3, data within purple frame). These observations are consistent with presence of all twelve possible mutation patterns in evolution of SARS-CoV-2 genome [27,28] and formation of premature stop codons in coding regions of ORF6 and ORF8 [34,35]. Deletion and insertion are other patterns of nucleotide mutations occurring in SARS-CoV-2 genome [19,36]. A dozen of deletions and an insertion have contributed to the emergence of Alpha, Beta, Gamma, Delta and Omicron variants, which are 19, 18, 5, 13 and 27 nucleotides shorter than the seven surveyed clades, respectively (Fig. 1). Meanwhile, insertion of 1–8 nucleotides occurred in virus samples of GH, GR and GV clades (data not shown). All

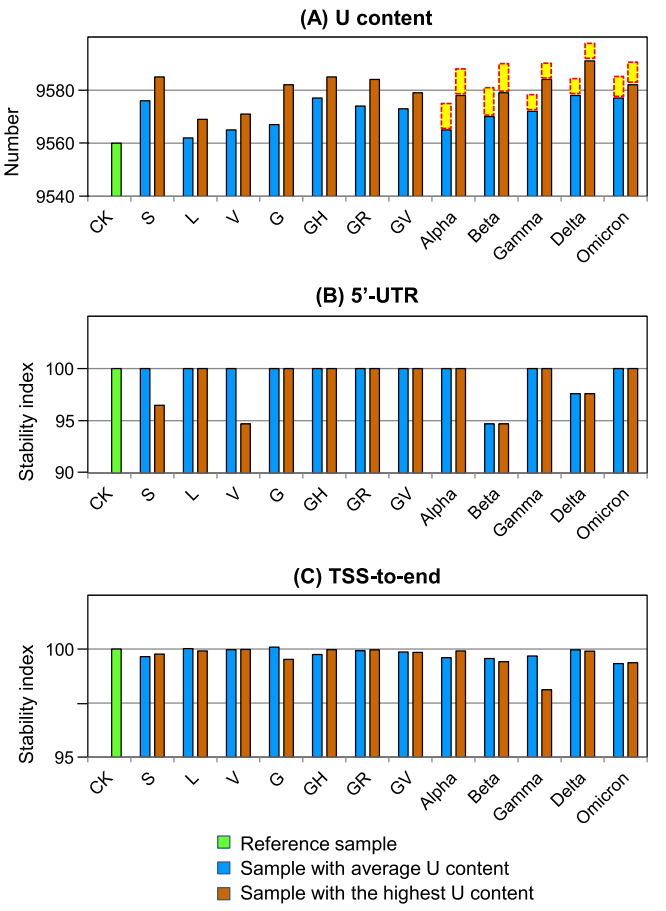

Fig. 4. Genome stability of SARS-CoV-2. For genome stability comparison, two virus samples were taken for calculating free energy of their genomic segments. One of the samples has average U content, while the other has the highest U content. (A) U content in various samples. Yellow block on top of data bar indicates U number lost from deletions (as shown in Fig. 1) in the five variants. (B) and (C) Stability of 5′-UTR (untranslated region) and TSS (translation start site)-to-end region in various samples. Please refer to Table S4 for detailed sample information and free energy values. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

these data demonstrate that nucleotide mutations occurring in SARS-CoV-2 genome are greatly diversified.

4.2. Trend of virus evolution

If nucleotide mutations are greatly diversified in SARS-CoV-2, why do current viruses have higher U-content and smaller genome-size? This could be the consequence of natural selection, because a virus with increased U-content and smaller genome-size can be more successful in replication. On one hand, increased U-content in 5′-UTR reduces stability (Fig. 4B) of its IRES (internal ribosome entry site) structure [31,37]. This makes the virus more cryptic in replication, because less host machinery is recruited to translate viral RNA. On the other hand, higher U-content and smaller genome-size reduce stability of viral RNA (Fig. 4C). This makes the virus more efficient in replication, because less host energy is consumed to disrupt secondary structures of viral RNA. Thus, the virus could become less virulent but more infectious, because more viruses can be replicated from using unit resources.

Mutating C, G and A into U is probably a new trend of SARS-CoV-2 genome evolution. Previously, we reported that C to U and G to A mutations allow SARS-CoV-2 to possess a genome with low C + G content. Thus, a potential C-G base-pair formed in genomic RNA could be replaced by a potential A-U base pair [31]. Our current survey reveals that C, G and A to U mutations all occurred in most SARS-CoV-2 lineages (Fig. 2). These data suggest that SARS-CoV-2 has attempted not only to replace a potential C-G base pair with A-U base pair but also to avoid formation of A-U and/or G-U base pairs, because such mutations could reduce viral genome stability to a greater extent (Fig. 5).

4.3. Mutability of Delta and Omicron variants

Delta variant has a high and unique mutability over other lineages, which has probably enabled it to cause many vaccine-breakthrough infections [38,39]. Its high mutability is reflected in elevating its U

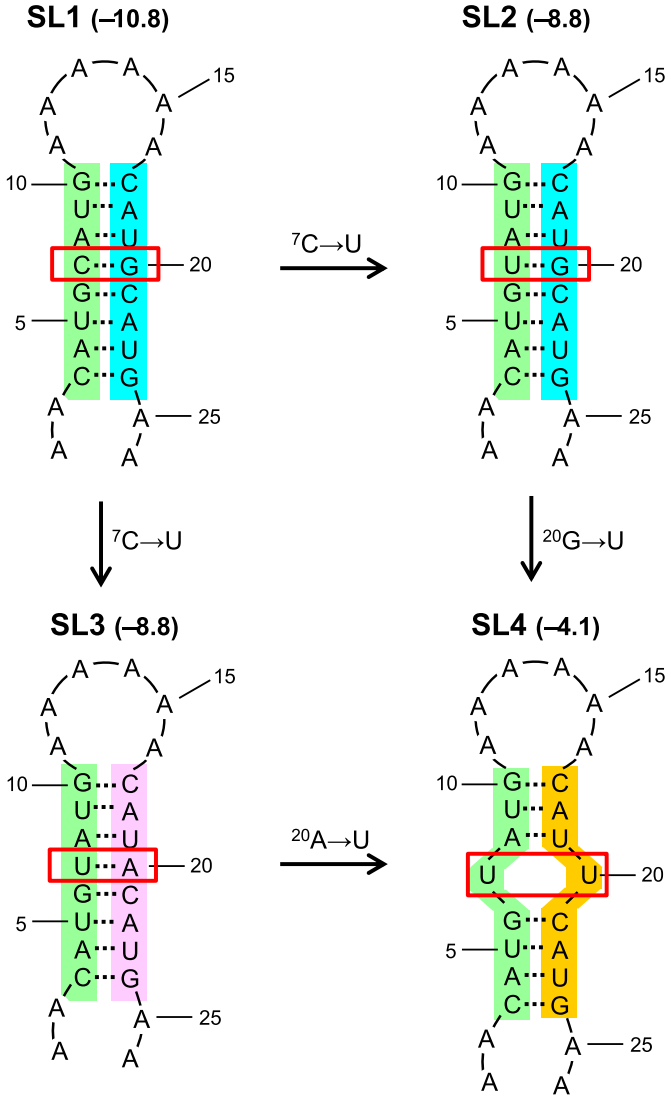

Fig. 5. Stability reduction of a stem-loop structure. Shown here is an example of stability reduction of stem-loop structures formed by hypothetical nucleotide sequences. Firstly, stem-loop 1 (SL1) is formed between the green- and cyan-shaded segments. This structure has a free energy of −10.8 kcal/mol. Then, after C at position 7 (^{7}C) is mutated into U, SL2 is formed because U is able to form a canonical base pair with G [33]. Alternatively, SL3 can be formed because the green-shaded segment may pair with another segment (mauve-shaded). Both SL2 and SL3 are less stable than SL1, because they have a free energy of −8.8 kcal/mol. Finally, after ^{20}G in SL2 and ^{20}A in SL3 are mutated into U, stability of the formed stem-loop (SL4) is further reduced, only having a free energy of −4.1 kcal/mol. Red frame highlights the formed or disrupted base pair. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

number by 9.7 and changing numbers of 14 amino acids significantly within ten months (Figs. 2 and S1). Its unique mutability is reflected in having high number of G and C mutated at codon position 1 and 2 respectively (Fig. S3). Such high and unique mutability could provide a large number of mutated viruses for natural selection against vaccines. This could probably explain why Delta variant caused more and more vaccine-breakthrough infections in many COVID-19 devastated countries [40,41].

Omicron variant emerged in mid-November 2021 by having high number of amino acid substitutions relative to the earliest SARS-CoV-2 virus [42,43]. This high mutability has resulted in rapid transmission of the virus and occurrence of a few vaccine breakthroughs [44,45]. Our survey indicates that Omicron has the smallest genome size among all SARS-CoV-2 lineages. Its genome is 27 nucleotides shorter than the reference one (Fig. 1) and has a considerably lower stability than other SARS-CoV-2 lineages (Fig. 4C). However, because of recent emergence, no sufficient sequence data are available for characterising its nucleotide and amino acid mutations. Thus, whether it has similar mutability like Delta variant awaits future investigation.

In conclusion, human SARS-CoV-2 has evolved to increase U content and reduce genome size. C, G and A to U mutations have all contributed to this U-content increase. Mutations of C, G and A at codon position 1, 2 or 3 have no significant difference in most lineages. Both deletion and insertion are involved in formation of SARS-CoV-2 variants. These results may provide a clue for tracing the origin of SARS-CoV-2, because ancestral viruses should have lower U-content and probably bigger genome-size.

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ijbiomac.2022.02.034.

CRediT authorship contribution statement

Y.W., Q.Y. and K.P.C. conceived the study. Y.W. and X.Y.C. wrote the manuscript. Y.W. compiled the computer program. Y.W., X.Y.C. and L.Y. performed surveys and analyses. All authors reviewed the manuscript.

Declaration of competing interest

The authors declare that they have no conflict of interest.

Data availability

Data will be made available on request.

Acknowledgements

The authors are greatly thankful to all contributors of SARS-CoV-2 genome sequences and to GISAID Initiative team for classifying all the sequences into various lineages. This study was supported by National Key Research and Development Program of China (No. 2018YFE0196600) and National Natural Science Foundation of China (No. 31861143051).

References

[1] WHO, Weekly Epidemiological Update on COVID-19, 2022. www.who.int/publications/m/. (Accessed 2 January 2022).
[2] F. Wu, S. Zhao, B. Yu, Y.M. Chen, W. Wang, Z.G. Song, Y. Hu, Z.W. Tao, J.H. Tian, Y.Y. Pei, M.L. Yuan, Y.L. Zhang, F.H. Dai, Y. Liu, Q.M. Wang, J.J. Zheng, L. Xu, E. C. Holmes, Y.Z. Zhang, A new coronavirus associated with human respiratory disease in China, Nature 579 (7798) (2020) 265–269.
[3] Y.H. Jin, L. Cai, Z.S. Cheng, H. Cheng, T. Deng, Y.P. Fan, C. Fang, D. Huang, L. Q. Huang, Q. Huang, Y. Han, B. Hu, F. Hu, B.H. Li, Y.R. Li, K. Liang, L.K. Lin, L. S. Luo, J. Ma, L.L. Ma, Z.Y. Peng, Y.B. Pan, Z.Y. Pan, X.Q. Ren, H.M. Sun, Y. Wang, Y.Y. Wang, H. Weng, C.J. Wei, D.F. Wu, J. Xia, Y. Xiong, H.B. Xu, X.M. Yao, Y. F. Yuan, T.S. Ye, X.C. Zhang, Y.W. Zhang, Y.G. Zhang, H.M. Zhang, Y. Zhao, M. J. Zhao, H. Zi, X.T. Zeng, Y.Y. Wang, X.H. Wang, A rapid advice guideline for the diagnosis and treatment of 2019 novel coronavirus (2019-nCoV) infected pneumonia (standard version), Mil. Med. Res. 7 (1) (2020) 4.
[4] C.K.C. Lai, W. Lam, Laboratory testing for the diagnosis of COVID-19, Biochem. Biophys. Res. Commun. 538 (2021) 226–230.
[5] J. Gao, Z. Tian, X. Yang, Breakthrough: chloroquine phosphate has shown apparent efficacy in treatment of COVID-19 associated pneumonia in clinical studies, Biosci. Trends 14 (1) (2020) 72–73.
[6] M. Gavriatopoulou, I. Ntanasis-Stathopoulos, E. Korompoki, D. Fotiou, M. Migkou, I.G. Tzanninis, T. Psaltopoulou, E. Kastritis, E. Terpos, M.A. Dimopoulos, Emerging treatment strategies for COVID-19 infection, Clin. Exp. Med. 21 (2) (2021) 167–169.
[7] Y.C. Kim, B. Dema, A. Reyes-Sandoval, COVID-19 vaccines: breaking record times to first-in-human trials, NPJ Vaccines 5 (1) (2020) 34.
[8] A. Awadasseid, Y. Wu, Y. Tanaka, W. Zhang, Current advances in the development of SARS-CoV-2 vaccines, Int. J. Biol. Sci. 17 (1) (2021) 8–19.
[9] R. Güner, I. Hasanoğlu, F. Aktaş, COVID-19: prevention and control measures in community, Turk. J. Med. Sci. 50 (SI-1) (2020) 571–577.
[10] F.J. Peng, L. Tu, Y.S. Yang, P. Hu, R.S. Wang, Q.Y. Hu, F. Cao, T.J. Jiang, J. Sun, G. G. Xu, C. Chang, Management and treatment of COVID-19: the Chinese experience, Can. J. Cardiol. 36 (6) (2020) 915–930.
[11] C. Valle, B. Martin, F. Touret, A. Shannon, B. Canard, J.C. Guillemot, B. Coutard, E. Decroly, Drugs against SARS-CoV-2: what do we know about their mode of action? Rev. Med. Virol. 30 (6) (2020) 1–10.
[12] T. Asselah, D. Durantel, E. Pasmant, G. Lau, R.F. Schinazi, COVID-19: discovery, diagnostics and drug development, J. Hepatol. 74 (1) (2021) 168–184.
[13] F. González-Candelas, M.A. Shaw, T. Phan, U. Kulkarni-Kale, D. Paraskevis, F. Luciani, H. Kimura, M. Sironi, One year into the pandemic: short-term evolution of SARS-CoV-2 and emergence of new lineages, Infect. Genet. Evol. 92 (2021), 104869.
[14] S. Ilmjärv, F. Abdul, S. Acosta-Gutiérrez, C. Estarellas, I. Galdadas, M. Casimir, M. Alessandrini, F.L. Gervasio, K.H. Krause, Concurrent mutations in RNA-dependent RNA polymerase and spike protein emerged as the epidemiologically most successful SARS-CoV-2 variant, Sci. Rep. 11 (1) (2021) 13705.
[15] F. Robson, K.S. Khan, T.K. Le, C. Paris, S. Demirbag, P. Barfuss, P. Rocchi, W.L. Ng, Coronavirus RNA proofreading: molecular basis and therapeutic targeting, Mol. Cell 79 (5) (2020) 710–727.
[16] N.H. Moeller, K. Shi, Ö. Demir, S. Banerjee, L. Yin, C. Belica, C. Durfee, R.E. Amaro, H. Aihara, Structure and dynamics of SARS-CoV-2 proofreading exoribonuclease ExoN, in: bioRxiv, 2021, https://doi.org/10.1101/2021.04.02.438274.
[17] B.S. Hudson, V. Kolte, A. Khan, G. Sharma, Dynamic tracking of variant frequencies depicts the evolution of mutation sites amongst SARS-CoV-2 genomes from India, J. Med. Virol. 93 (4) (2021) 2534–2537.
[18] M. Giovanetti, F. Benedetti, G. Campisi, A. Ciccozzi, S. Fabris, G. Ceccarelli, V. Tambone, A. Caruso, S. Angeletti, D. Zella, M. Ciccozzi, Evolution patterns of SARS-CoV-2: snapshot on its genome variants, Biochem. Biophys. Res. Commun. 538 (2021) 88–91.
[19] T. Koyama, D. Platt, L. Parida, Variant analysis of SARS-CoV-2 genomes, Bull. World Health Organ. 98 (7) (2020) 495–504.
[20] M. Bianchi, A. Borsetti, M. Ciccozzi, S. Pascarella, SARS-Cov-2 ORF3a: mutability and function, Int. J. Biol. Macromol. 170 (2021) 820–826.
[21] S.Q. Wu, C. Tian, P.P. Liu, D.J. Guo, W. Zheng, X.Q. Huang, Y. Zhang, L.J. Liu, Effects of SARS-CoV-2 mutations on protein structures and intraviral protein-protein interactions, J. Med. Virol. 93 (4) (2021) 2132–2140.
[22] B. Korber, W.M. Fischer, S. Gnanakaran, H. Yoon, J. Theiler, W. Abfalterer, N. Hengartner, E.E. Giorgi, T. Bhattacharya, B. Foley, K.M. Hastie, M.D. Parker, D. G. Partridge, C.M. Evans, T.M. Freeman, T.I. de Silva, , Sheffield COVID-19 Genomics Group, C. McDanal, L.G. Perez, H.L. Tang, A. Moon-Walker, S.P. Whelan, C.C. LaBranche, E.O. Saphire, D.C. Montefiori, Tracking changes in SARS-CoV-2 Spike: evidence that D614G increases infectivity of the COVID-19 virus, Cell 182 (4) (2020) 812–827.
[23] C. van Oosterhout, J.F. Stephenson, B. Weimer, H. Ly, N. Hall, K.M. Tyler, COVID-19 adaptive evolution during the pandemic - implications of new SARS-CoV-2 variants on public health policies, Virulence 12 (1) (2021) 2013–2016.
[24] L. Amato, L. Jurisic, I. Puglia, V.D. Lollo, V. Curini, G. Torzi, A.D. Girolamo, I. Mangone, A. Mancinelli, N. Decaro, P. Calistri, F.D. Giallonardo, A. Lorusso, N. D'Alterio, Multiple detection and spread of novel strains of the SARS-CoV-2 B.1.177 (B.1.177.75) lineage that test negative by a commercially available nucleocapsid gene real-time RT-PCR, Emerg. Microbes Infect. 10 (1) (2021) 1148–1155.
[25] A.M. Rice, A.C. Morales, A.T. Ho, C. Mordstein, S. Mühlhausen, S. Watson, L. Cano, B. Young, G. Kudla, L.D. Hurst, Evidence for strong mutation bias toward, and selection against, U content in SARS-CoV-2: implications for vaccine design, Mol. Biol. Evol. 38 (1) (2021) 67–83.
[26] J. Zahradník, S. Marciano, M. Shemesh, E. Zoler, D. Harari, J. Chiaravalli, B. Meyer, C.L. Li, I. Marton, O. Dym, N. Elad, M.G. Lewis, H. Andersen, M. Gagne, R.A. Seder, D.C. Douek, G. Schreiber, Y.N. Rudich, SARS-CoV-2 variant prediction and antiviral drug design are enabled by RBD in vitro evolution, Nat. Microbiol. 6 (9) (2021) 1188–1198.
[27] C. Roy, S.M. Mandal, S.K. Mondal, S. Mukherjee, T. Mapder, W. Ghosh, R. Chakraborty, Trends of mutation accumulation across global SARS-CoV-2 genomes: implications for the evolution of the novel coronavirus, Genomics 112 (6) (2020) 5331–5342.
[28] S.P. Otto, T. Day, J. Arino, C. Colijn, J. Dushoff, M. Li, S. Mechai, G.V. Domselaar, J.H. Wu, D.J.D. Earn, N.H. Ogden, The origins and potential future of SARS-CoV-2 variants of concern in the evolving COVID-19 pandemic, Curr. Biol. 31 (14) (2021) R918–R929.

[29] GISAID, Clade and Lineage Nomenclature Aids in Genomic Epidemiology Studies of Active hCoV-19 Viruses, 2021. www.gisaid.org/references/statements-clarifications/. (Accessed 2 January 2022).
[30] J.H. Chen, R. Wang, M.L. Wang, G.W. Wei, Mutations strengthened SARSCoV-2 infectivity, J. Mol. Biol. 432 (19) (2020) 5212–5226.
[31] Y. Wang, J.M. Mao, G.D. Wang, Z.P. Luo, L. Yang, Q. Yao, K.P. Chen, Human SARS-CoV-2 has evolved to reduce CG dinucleotide in its open reading frames, Sci. Rep. 10 (1) (2020) 12331.
[32] S. Kumar, G. Stecher, M. Li, C. Knyaz, K. Tamura, MEGA X: molecular evolutionary genetics analysis across computing platforms, Mol. Biol. Evol. 35 (2018) 1547–1549.
[33] J.S. Reuter, D.H. Mathews, RNAstructure: software for RNA secondary structure prediction and analysis, BMC Bioinformatics 11 (2010) 129.
[34] S. Delbue, S.D. Alessandro, L. Signorini, M. Dolci, E. Pariani, M. Bianchi, S. Fattori, A. Modenese, C. Galli, I. Eberini, P. Ferrante, Isolation of SARS-CoV-2 strains carrying a nucleotide mutation, leading to a stop codon in the ORF6 protein, Emerg. Microbes Infect. 10 (1) (2021) 252–255.
[35] S. DeRonde, H. Deuling, J. Parker, J. Chen, Identification of a novel SARS-CoV-2 strain with truncated protein in ORF8 gene by next generation sequencing, Res. Sq. (2021), https://doi.org/10.21203/rs.3.rs-413141/v1.
[36] J.Á. Patiño-Galindo, I. Filip, R. Chowdhury, C.D. Maranas, P.K. Sorger, M. AlQuraishi, R. Rabadan, Recombination and lineage-specific mutations linked to the emergence of SARS-CoV-2, Genome Med. 13 (1) (2021) 124.
[37] N. Sonenberg, J. Pelletier, Poliovirus translation: a paradigm for a novel initiation mechanism, Bioessays 11 (5) (1989) 128–132.
[38] M. McCallum, A.C. Walls, K.R. Sprouse, J.E. Bowen, L.E. Rosen, H.V. Dang, A. D. Marco, N. Franko, S.W. Tilles, J. Logue, M.C. Miranda, M. Ahlrichs, L. Carter, G. Snell, M.S. Pizzuto, H.Y. Chu, W.C.Van Voorhis, D. Corti, D. Veesler, Molecular basis of immune evasion by the delta and kappa SARS-CoV-2 variants, in: bioRxiv, 2021, https://doi.org/10.1101/2021.08.11.455956.
[39] T. Farinholt, H. Doddapaneni, X. Qin, V. Menon, Q.C. Meng, G. Metcalf, H. Chao, M.C. Gingras, V. Avadhanula, P. Farinholt, C. Agrawal, D.M. Muzny, P.A. Piedra, R. A. Gibbs, J. Petrosino, Transmission event of SARS-CoV-2 Delta variant reveals multiple vaccine breakthrough infections, BMC Med. 19 (1) (2021) 255.
[40] R. Wang, J.H. Chen, Y. Hozumi, C.C. Yin, G.W. Wei, Emerging vaccine-breakthrough SARS-CoV-2 variants, in: ArXiv, 2021 doi: arXiv:2109.04509v1.
[41] R. Wang, J.H. Chen, G.W. Wei, Mechanisms of SARS-CoV-2 evolution revealing vaccine-resistant mutations in Europe and America, J. Phys. Chem. Lett. 12 (49) (2021) 11850–11857, 2021.
[42] S.K. Saxena, S. Kumar, S. Ansari, J.T. Paweska, V.K. Maurya, A.K. Tripathi, A. S. Abdel-Moneim, Characterization of the novel SARS-CoV-2 omicron (B.1.1.529) variant of concern and its global perspective, J. Med. Virol. (2021), https://doi.org/10.1002/jmv.27524.
[43] X.M. He, W.Q. Hong, X.Y. Pan, G.W. Lu, X.W. Wei, SARS-CoV-2 omicron variant: characteristics and prevention, Med. Comm. 2 (4) (2021) 838–845, https://doi.org/10.1002/mco2.110.
[44] Y.C. Wang, L. Zhang, Q.Q. Li, Z.T. Liang, T. Li, S. Liu, Q.Q. Cui, J.H. Nie, Q. Wu, X. W. Qu, W.J. Huang, The significant immune escape of pseudotyped SARS-CoV-2 variant omicron, Emerg. Microbes Infect. 11 (1) (2022) 1–5, https://doi.org/10.1080/22221751.2021.2017757.
[45] S.R. Kannan, A.N. Spratt, K. Sharma, H.S. Chand, S.N. Byrareddy, K. Singha, Omicron SARS-CoV-2 variant: unique features and their impact on pre-existing antibodies, J. Autoimmun. 126 (2022), 102779.