Analysis of Continuous Mutation and Evolution on Circulating SARS-CoV-2

Jie-Mei Yu, Li-Shu Zhang, Yuan-Hui Fu, Feng-Min Ji, Han-Li Xu, Jia-Qiang Huang, Xiang-Lei Peng, Yan-Peng Zheng, Ying Zhang and Jin-Sheng He

College of Life Sciences and Bioengineering, Beijing Jiaotong University, Beijing, China.

ABSTRACT: Monitoring the mutation and evolution of the virus is important for tracing its ongoing transmission and facilitating effective vaccine development. A total of 342 complete genomic sequences of SARS-CoV-2 were analyzed in this study. Compared to the reference genome reported in December 2019, 465 mutations were found, among which, 347 occurred in only 1 sequence, while 26 occurred in more than 5 sequences. For these 26 further identified as SNPs, 14 were closely linked and were grouped into 5 profiles. Phylogenetic analysis revealed the sequences formed 2 major groups. Most of the sequences in late period (March and April) constituted the Cluster II, while the sequences before March in this study and the reported S/L and A/B/C types in previous studies were all in Cluster I. The distributions of some mutations were specific geographically or temporally, the potential effect of which on the transmission and pathogenicity of SARS-CoV-2 deserves further evaluation and monitoring. Two mutations were found in the receptor-binding domain (RBD) but outside the receptor-binding motif (RBM), indicating that mutations may only have marginal biological effects but merit further attention. The observed novel sequence divergence is of great significance to the study of the transmission, pathogenicity, and development of an effective vaccine for SARS-CoV-2.

KEYWORDS: SARS-CoV-2, mutation dynamics, SNP, evolution, clusters

Introduction

Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) is a highly pathogenic and infectious agent that causes coronavirus disease 2019 (COVID-19). As of May 17, 2020, over 4.53 million cases of COVID-19 have been confirmed, including 312 009 deaths worldwide (https://covid19.who.int/). Although many countermeasures against transmission of SARS-CoV-2 have been taken around the world, there are no signs yet indicating that the pandemic will cease in the near future.

SARS-CoV-2 is a non-segmented positive-sense RNA virus of about 30 kb in the Coronaviridae family and contains crown-shaped peplomers. Excluding the 5’-cap structure and 3’-poly-A tail, the genome consists of 6 major open reading frames (ORFs) common to other coronaviruses and a number of other accessory genes. Four main structural proteins of the virus are encoded by ORFs near the 3’ terminus of the genome, among which a spike (S) protein can induce the host to produce neutralizing antibody and the receptor-binding domain (RBD) determines the histotropism of the virus.

Although mutation is one of the major factors affecting the molecular evolution of RNA viruses, the mutation rate and error rate are low in coronaviruses since the 3’-5’ exonuclease, nonstructural protein 14 (nsp14), has proofreading activity. SARS-CoV-2 is speculated to have originated from a wild animal, such as a bat, with or without an intermediate species. As a newly emerging virus, its transmission, evolution, pathogenesis, and induced immunity are far from clear. Some key questions that need to be addressed are whether, when, and how random mutations generate different variants with functional alterations, such as a higher or lower transmission rate and a likely evasion from the protection induced by developing vaccines. Some hotspots and strong purifying selection have been observed during the study of SARS-CoV-2 evolution. Here, we assess the mutation dynamics of 342 complete SARS-CoV-2 genomes from 20 different countries or regions. No isolate different from the original one was found, but multiple mutations and single nucleotide polymorphisms (SNPs) were detected; some of the mutations were tightly linked in the virus genome and could be grouped into 5 profiles. The distributions of some mutations in the SARS-CoV-2 genome were found to be associated with time or location. The change in mutation frequencies in some of the cluster-determining nucleotide (nt) sites revealed that the prevalent types of the virus transitioned over time.

Results and Discussion

Mutations in SARS-CoV-2 genomes

Despite the identities of the 342 analyzed sequences beingas high as >99%, we identified 465 mutations in the SARS-CoV-2 genome, among which, 347 (74.6%) occurred in only 1 sequence and 26 occurred in more than 5 sequences (ie, occurrence of mutation frequency >1.5%). To avoid interference from false mutations caused by sequencing error, in this study we mainly analyzed those 26 SNPs. They
distribut ed in the genes of 5’UTR, ORF1a, ORF1b, S, ORF3a, ORF7a, ORF8, and N of the virus (Table 1). Interestingly, 14 SNPs linked tightly and formed 5 profiles: #1: C241T/C3037T/C14408T/A23403G, #2: C8782T/T28144C, #3: C17747T/A17858G/C18060T , #4: G1440A/G2891A, and #5: G28881A/G28882A/G28883C. Except for nt site 241, all other 13 nts in the linked SNPs were located in coding regions, including 3 synonymous mutations and ten nonsynonymous ones, leading to 9 amino acid (aa) substitutions in the corresponding SARS-CoV-2 isolates. These sequences increased the diversity of SARS-CoV-2 genome sequences; however, we could not conclude that these SNPs, including the tightly linked SNPs, already conferred a novel trait to these variants.

| POSITION| NT CHANGE | LOCATION | NUMBERSb | AA SUBSTITUTION |
|---------|-----------|----------|-----------|-----------------|
| 241     | C-T       | 5’UTR    | 189       |                 |
| 1059    | C-T       | nsp2     | 60        | T-I             |
| 1397    | G-A       | nsp2     | 11        | V-I             |
| 1440    | G-A       | nsp2     | 8         | G-D             |
| 2891    | G-A       | nsp3     | 8         | A-T             |
| 3037    | C-T       | nsp3     | 186       | no              |
| 4402    | T-C       | nsp3     | 8         |                |
| 4809    | C-T       | nsp3     | 6         | S-F             |
| 1785    | C-T       | nsp3     | 7        |                |
| 11083   | G-T       | nsp6     | 43        | L-F             |
| 14408   | C-T       | nsp12    | 187       | P-L             |
| 14805   | C-T       | nsp12    | 19        | no              |
| 15324   | C-T       | nsp12    | 10        | no              |
| 17747   | C-T       | nsp13    | 22        | P-L             |
| 17858   | A-G       | nsp13    | 22        | Y-C             |
| 18060   | C-T       | nsp14    | 24        | no              |
| 20329   | A-G       | nsp15    | 14        | no              |
| 23403   | A-G       | spike    | 187       | D-G             |
| 25556   | G-T       | ORF3a    | 72        | Q-H             |
| 26144   | G-T       | ORF3a    | 25        | G-V             |
| 27751   | A-T       | ORF7a    | 10        | R-S             |
| 28144   | T-C       | ORF5     | 24        | S-L             |
| 28881-28883 | GGG/AAC | nucleocapsid | 46       | RG/KR          |
| 28888   | T-C       | nucleocapsid | 11        | no              |

"-" not applicable; "no," synonymous mutation without aa substitution. Data marked in the same color shows the tightly related mutations of the virus in the profiles.

*Locations of the SNPs on the genome.

**Total number of mutated sequences.

Phylogeny analysis revealed the increasing divergence of SARS-CoV-2

To trace the potential spreading history of SARS-CoV-2, an unrooted maximum likelihood tree of the 342 genomes was built based on the nearly full genome sequence alignment. Overall, these target SARS-CoV-2 genomes formed 2 major distinct phylogenetic clusters, Clusters I and II, which were determined by the difference in SNP profile#1 (C241T/C3037T/C14408T/A23403G), of which the nt3037 was a synonymous mutation, while nts14408 and 23403 were nonsynonymous mutations. Moreover, Cluster I subclusters (I-1 and I-2) differentiated by a synonymous mutation in nt8782 of T/C and a nonsynonymous mutation in nt 28144 of C/T (S-L).
Subcluster I-2 was further divided into I-2-1 and I-2-2 with a synonymous mutation in nt 26144 of G/T. Also, 2 Cluster II subclusters (II-1 and II-2) were differentiated by a nonsynonymous mutation (Q-H) in nt 25563 of G/T; subcluster II-2 was further clustered into II-2-1 and II-2-2 with variation in SNP profile #5 of GGG22881-28883AAC with 2 nonsynonymous mutations (RG-KR). When the phylogeny tree was made excluding the time parameter, even the sequences reported from the same geographic region could be clustered or subclustered differently displaying the pandemic pattern of mixed lineages rather than a single lineage in each region (Figure 1a). Furthermore, if the time factor was considered, the earlier sequences (before March) were mainly distributed in Cluster I, while the late sequences (from March and April) would be mainly distributed in the newly emerged Cluster II (Figure 1b).

Two earlier studies reported that SARS-CoV-2 evolved into 2 (designated L and S based on T8782C and C28144T) or 3 (designated A, B, and C based on T8782C, C28144T, and G26144T) major types. However, according to our study, all these reported types belonged to Cluster I, with S or A type corresponding to I-1, L or B to I-2, and C to I-2-2. Cluster II, also designated as G clade in a very recent study published in the bioRxiv preprint repository upon increasing genome frequency in the GISAID data, mainly involved the late sampling sequences in March and April. Taken together, we think the random mutation following the ongoing transmission of SARS-CoV-2 was responsible for the appearance of Cluster II or G clade.

Temporal and spatial distributions of the mutations in SARS-CoV-2

The real cutoff point of the collected 342 sequences in this study was 06 April 2020 (updated cutoff point was 15 April 2020). Five samples were collected in December 2019, 50 in January 2020, 60 in February, 216 in March, and 11 in April, respectively (Supplemental files). To further characterize the mutations on SARS-CoV-2 genomes over time, we drew a time-related distribution map for the mutations occurring in more than 5 sequences. Upon the 100% stacked column chart of mutation frequency, it was clearly shown that in late February and early March, the SNPs were more divergent than those observed in the times before and after (Figure 2a). It was also noted that the mutation frequencies at nts 1059, 25563, and SNP profile #1 of 241T/3037T/14408T/23403G increased over time, while frequencies at nts 8782, 28144, and 26144 declined (Figure 2b). Due to the limited sample size, the short time span, and the high susceptibility of the population, it was difficult to conclude that the later mutations and the earlier ones have been shaped by exposure to some selective pressure. Alternatively, we inferred that the fluctuation in mutation frequency indicated that altered predominant variants originated from different founder viruses. We would like to state here that Cluster II is comprised of the sequences with mutations from SNP profile #1 of C241T/C3037T/C14408T/A23403G. Although the variants in Cluster II are currently more prevalent, it does not mean that they are more transmissible evolutionarily. We cannot rule out the founder effect, especially considering the good fitness already displayed in much of the early period and the availability of susceptible hosts. D641G substitution in S protein from mutation of A/G in nt 23403 has previously been reported as a nonsynonymous mutation accompanied by mutations in nts 3037 (synonymous mutation) and 14408 (nonsynonymous). In our study, we also observed that mutation in nt241 was tightly linked to SNPs, and SNP profile #1 ensued.
The sequences in this study were from 20 different countries or regions and covered 6 different continents. Among the 342 records, 105 were from Asia, 86 from Europe, 21 from Africa, 11 from South America, 91 from North America, and 28 from Oceania (supplemental files). The regional distribution of those mutations identified in more than 5 viral sequences was analyzed. Several mutations in the non-cluster-determination site were geography-related, for example, SNPs in nts 1440/A and 2890/A were only detected in Europe and USA, SNPs 4402/C and 4809/T only in Asia, SNP 27751/T only in Europe, and SNP profile #3 of 17747/T/17858/G/18060/T only in the USA and Australia. Due to the small sample pools from Africa and South America and the founder effect, we still cannot infer whether this is related to environmental adaption and this should be further investigated.

Most of the observed SNPs were geographically unrelated. For profile#1, the highest mutation frequency was in Africa (over 95%), followed by Europe (80%), and the lowest mutation frequency was found in Asia (about 30%) (Figure 3). When removing the late sampling records from India and Taiwan, China (March 2020), the frequency in Asia is even lower, only about 10%, suggesting that the mutations were time-related instead of population- or region-specific. It was reported that types A (I-1) and C (I-2-2) were found in significant proportions in Europe and USA, while type B (I-2-1) was predominant in East Asia. However, mutation frequencies in nts 8782, 28144, and 26144 were relatively low in the
current study, suggesting that the status of Cluster I (including A, B, and C type) being prevalent had changed. We speculate that this phenomenon was caused by the expanded and updated data analyzed in this study compared with the data from earlier publications.

**Mutations in S protein and RBD**

The RBD in S1 subunit of SARS-CoV-2 directly binds to the peptidase domain of the host receptor, while S2 is responsible for membrane fusion,\(^{14,15}\) the virus binds to its host receptor through its receptor-binding motif (RBM) composed of 17 residues (K417, G446, Y449, Y453, L455, F456, A475, F486, N487, Y489, Q493, G496, Q498, T500, N501, G502, and Y505).\(^{16}\) Five RBM critical residues were recently reported (L455, F486, Q493, S494, and N501).\(^{15}\)

We identified 27 mutations in the S gene among the 342 sequences analyzed in this study, with the highest mutation frequency (54.7\%) at aa position 614 (Figure 4a). Mutation D614G rapidly became the dominant form when it was introduced to new regions and it has been speculated that structural and immunological factors may enhance its fitness.\(^{13}\) However, SARS-CoV-2 has already well adapted to humans since the beginning and, until now, its infectivity, transmissibility, and pathogenicity are still stable with no solid evidence supporting its functional change. Therefore, our view is that the high frequency of this mutation is caused by the founder effect. Additionally, all mutations in the S protein were nonsynonymous. Notably, 2 RBD mutations (S438F and G476S) were identified in this study: S438F occurred in 2 sequences from India collected in 3rd March 2020 (EPI_ISL_420543 and EPI_ISL_420547) and G476S in a sequence from the US collected in 14th March 2020 (EPI_ISL_417380) (Figure 4b). The changed residues were located outside RBM, suggesting a limited direct effect on viral infection at present, more detailed experimental and modeling studies are needed to elucidate the effect of these mutations because the substantial selective advantage is likely produced by some trivial differences over time. Additionally, further significant mutations that may occur in the RBD region also warrant continuous monitoring.

In summary, through the analyses of 342 full-length genome sequences of SARS-CoV-2, we characterized their mutation and evolution dynamics and found more sequence divergence than that found in earlier studies. Although some mutation frequencies did display geographical and temporal fluctuation, we cannot conclude their biological effect on the transmission and pathogenicity of SARS-CoV-2. It is both fascinating and challenging to know the real-time evolution of SARS-CoV-2 and predict its effect on the transmission and pathogenesis of the virus, as well as on the development of a vaccine.

**Materials and Methods**

**Genome sequence dataset retrieval**

Full-length sequence of the prototype SARS-CoV-2 (NC_045512.2) from Wuhan, China in 2019, bat SARS-like coronavirus (NC_004718.3), and SARS-CoV (MN996532.1) were downloaded from the NCBI GenBank Database (https://www.ncbi.nlm.nih.gov/genbank). The set of 342 nearly complete genome sequences was downloaded from Global Initiative on Sharing All Influenza Data (GISAID, https://www.gisaid.org). The genomic sequences of all SARS-CoV-2 were aligned by MAFFT 7 (online version, https://mafft.cbrc.jp/alignment/software/).

**Phylogenetic analysis**

The obtained alignments were used for phylogeny analysis. The phylogenetic tree was performed in MEGA 7.0.26 software using the maximum likelihood method with Tamura-Nei substitution model.
Acknowledgements

We gratefully thank the researchers and laboratories who generated and shared the sequences from GISAID’s EpiFlu (TM) on which our study is based.

Author Contributions

JMY and JSH designed the study. JMY, LSZ, YHF, FMJ, HLX, YPZ and YZ analyzed the data. JMY wrote the manuscript. JSH, LSZ and JQH reviewed the manuscript. All authors have read and approved the manuscript.

ORCID iD

Jie-Mei Yu https://orcid.org/0000-0003-2988-384X

Supplemental Material

Supplemental material for this article is available online.

REFERENCES

1. Cui J, Li F, Shi ZL. Origin and evolution of pathogenic coronaviruses. Nat Rev Microbiol. 2019;17:181-192.
2. Chen Y, Liu Q, Guo D. Emerging coronaviruses: genome structure, replication, and pathogenesis. J Med Virol. 2020;92:418-423.
3. Shang J, Ye G, Shi K, et al. Structural basis of receptor recognition by SARS-CoV-2. Nature. 2020;581:221-224.
4. Sanjuan R, Nebot MR, Chirico N, Mansky LM, Belshaw R. Viral mutation rates. J Virol. 2010;84:9713-9748.
5. Ferro M, Subisi L, Silva De Morais AT, et al. Structural and molecular basis of mismatch correction and ribavirin excision from coronavirus RNA. Proc Natl Acad Sci USA. 2018;115:E162-E171.
6. Andersen KG, Rambaut A, Lipkin WI, Holmes EC, Garry RF. The proximal origin of SARS-CoV-2. Nat Med. 2020;26:450-452.
7. Zhou P, Yang XL, Wang XG, et al. A pneumonia outbreak associated with a new coronavirus of probable bat origin. Nature. 2020;579:270-273.
8. Zhou H, Chen X, Hu T, Hughes AC, Bi Y, Shi W. A close relative of SARS-CoV-2 found in bats offers more evidence it evolved naturally. Curr Biol. 2020;30:2196-2203.
9. Li X, Giorgi EE, Marichann MH, et al. Emergence of SARS-CoV-2 through recombination and strong purifying selection. Sci Adv. 2020;6:27.
10. Pachetti M, Marini B, Benedetti F, et al. Emerging SARS-CoV-2 mutation hotspots include a novel RNA-dependent-RNA polymerase variant. J Transl Med. 2020;18:179.
11. Forster P, Forster L, Renfrew C, Forster M. Phylogenetic network analysis of SARS-CoV-2 genomes. Proc Natl Acad Sci USA. 2020;117:9241-9243.
12. Tang X, Wu C, Li X, et al. On the origin and continuing evolution of SARS-CoV-2. Nat Rev. 2020;6:6.
13. Keber R, Fischer W, Gnanakaran S, et al. Spike mutation pipeline reveals the emergence of a more transmissible form of SARS-CoV-2. bioRxiv preprint. 2020.
14. Li F. Receptor recognition mechanisms of coronaviruses: a decade of structural studies. J Viral. 2015;89:1954-1964.
15. Wan Y, Shang J, Graham R, Basig RS, Li F. Receptor recognition by the novel coronavirus from Wuhan: an analysis based on decade-long structural studies of SARS Coronavirus. J Virol. 2020;94:e00127-20.
16. Lan J, Ge J, Yu J, et al. Structure of the SARS-CoV-2 spike receptor-binding domain bound to the ACE2 receptor. Nature. 2020;581:215-220.

Figure 4. (a) Distribution of mutations in the S protein. All mutations in the S protein were nonsynonymous, with position 614 aa having the highest frequency. Mutations in RBD. Two sequences from India and one from the USA collected in March exhibited nonsynonymous mutations in RBD. *" stands for the conserved aa. All the sequences were compared with the SARS-CoV-2 prototype; SNPs mentioned in this study were marked by red columns; residues involved in RBM were marked by triangles (blue and red), while those residues critical for host range determination were marked by red triangles.