Genomic Evolution and Variation of SARS-CoV-2 in the Early Phase of COVID-19 Pandemic in Guangdong Province, China

Bai-sheng LI1, 2†, Zhen-cui LI1, 2†, Yao HU1, 2†, Li-jun LIANG1, 2, Li-rong ZOU1, 2, Qian-fang GUO1, 2, Zhong-hua ZHENG1, 2, Jian-xiang YU1, 2, Tie SONG1, 2, Jie WU1, 2, 3#
1Guangdong Provincial Center for Disease Control and Prevention, Guangzhou 510000, China
2Guangdong Workstation for Emerging Infectious Disease Control and Prevention, Chinese Academy of Medical Sciences, Guangzhou 510000, China
3Southern Medical University, Guangzhou 510000, China

© Huazhong University of Science and Technology 2021

Summary: Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) with unknown origin spread rapidly to 222 countries, areas or territories. To investigate the genomic evolution and variation in the early phase of COVID-19 pandemic in Guangdong, 60 specimens of SARS-CoV-2 were used to perform whole genome sequencing, and genomics, amino acid variation and Spike protein structure modeling analyses. Phylogenetic analysis suggested that the early variation in the SARS-CoV-2 genome was still intra-species, with no evolution to other coronaviruses. There were one to seven nucleotide variations (SNVs) in each genome and all SNVs were distributed in various fragments of the genome. The Spike protein bound with human receptor, an amino acid salt bridge and a potential furin cleavage site were found in the SARS-CoV-2 using molecular modeling. Our study clarified the characteristics of SARS-CoV-2 genomic evolution, variation and Spike protein structure in the early phase of local cases in Guangdong, which provided reference for generating prevention and control strategies and tracing the source of new outbreaks.

Key words: severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2); spike protein; genomic evolution

Among the several coronaviruses that are pathogenic to humans, most are associated with mild clinical symptoms[1], with two notable exceptions: severe acute respiratory syndrome coronavirus (SARS-CoV)[2, 3] and Middle East respiratory syndrome coronavirus (MERS-CoV)[4], which have caused more than 10 000 cases, with mortality rates of 10% for SARS-CoV and 37% for MERS-CoV[5, 6]. These facts suggest that there is always a threat of coronavirus infection to human beings, especially novel coronavirus from animal origin.

In December, 2019, a series of cases with clinical manifestations of viral pneumonia of unknown cause emerged. Deep sequencing analysis from lower respiratory tract samples indicated a novel coronavirus, which was named severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2)[7]. Meanwhile, World Health Organization (WHO) named novel coronavirus pneumonia as Corona Virus Disease 2019 (COVID-19). According to the current data of WHO, the number of infected and dead cases has increased to 73 575 202 and 1 656 317, respectively (Dec. 2019 to Dec. 19, 2020).

Guangdong Province has the largest number of reported cases in China except for Hubei Province. The first confirmed COVID-19 cases in Guangdong appeared on January 14, 2020. It is an important content to monitor and analyze the genome variation of the virus, which is helpful to predict the epidemic trend of the disease. Here we describe the genomic characterization of 60 genomes of SARS-CoV-2 from patients in Guangdong as well as publicly available genomes, providing important information on the genomic variation of this new virus in the early phase of COVID-19 pandemic in Guangdong.

1 MATERIALS AND METHODS

1.1 Sample Sources
Specimens were collected between January 14, 2020 and January 31, 2020, through Guangdong provincial COVID-19 laboratory monitoring network with Guangdong Provincial Center for Disease Control and Prevention (CDC) as the central laboratory (table 1). Samples collected from suspected patients were sent to the local CDC to detect viral RNA of SARS-CoV-2

---

Bai-sheng LI, E-mail: libsn@126.com; Zhen-cui LI, E-mail: lizhencui@126.com; Yao HU, E-mail: 550126772@qq.com
†The authors contributed equally to this work.
#Corresponding author, E-mail: 771276998@qq.com
Table 1 Specimens were collected between January 14, 2020 and January 31, 2020

| Sequence ID | Sample ID | Gender | Age | Location | Type          | Onset date | Sampling date |
|-------------|-----------|--------|-----|----------|---------------|------------|---------------|
| Shenzhen/20SF012/2020 | 20SF012 | Male | 66 | Shenzhen | Tracheal_aspirate | 1/1/2020 | 1/14/2020 |
| Shenzhen/20SF014/2020 | 20SF014 | Male | 63 | Shenzhen | Blaf | 1/8/2020 | 1/14/2020 |
| Shenzhen/20SF025/2020 | 20SF025 | Female | 68 | Shenzhen | Throat_swab | 1/8/2020 | 1/15/2020 |
| Zhuhai/20SF028/2020 | 20SF028 | Male | 76 | Shenzhen | Nasal_swab | 1/12/2020 | 1/18/2020 |
| Zhuhai/20SF025/2020 | 20SF025 | Female | 69 | Shenzhen | Throat_swab | 1/12/2020 | 1/21/2020 |
| Shenzhen/20SF118/2020 | 20SF118 | Male | 68 | Shenzhen | Throat_swab | 1/11/2020 | 1/23/2020 |
| Shenzhen/20SF117/2020 | 20SF117 | Female | 69 | Shenzhen | Nasal_swab | 1/11/2020 | 1/23/2020 |
| Shenzhen/20SF243/2020 | 20SF243 | Male | 77 | Shenzhen | Throat_swab | 1/12/2020 | 1/22/2020 |
| Zhuhai/20SF200/2020 | 20SF200 | Female | 57 | Shenzhen | Throat_swab | 1/12/2020 | 1/23/2020 |
| Guangzhou/20SF206/2020 | 20SF206 | Male | 73 | Shenzhen | Throat_swab | 1/12/2020 | 1/23/2020 |
| Foshan/20SF207/2020 | 20SF207 | Male | 57 | Shenzhen | Throat_swab | 1/12/2020 | 1/23/2020 |
| Foshan/20SF210/2020 | 20SF210 | Female | 57 | Shenzhen | Throat_swab | 1/12/2020 | 1/23/2020 |
| Foshan/20SF211/2020 | 20SF211 | Male | 57 | Shenzhen | Throat_swab | 1/12/2020 | 1/23/2020 |
| Shenzhen/20SF243/2020 | 20SF243 | Male | 63 | Shenzhen | Throat_swab | 1/12/2020 | 1/23/2020 |
| Huizhou/20SF198/2020 | 20SF198 | Male | 50 | Shenzhen | Throat_swab | 1/12/2020 | 1/23/2020 |
| Guangzhou/20SF115/2020 | 20SF115 | Female | 63 | Shenzhen | Throat_swab | 1/12/2020 | 1/23/2020 |
| Zhanjiang/20SF123/2020 | 20SF123 | Female | 77 | Shenzhen | Throat_swab | 1/12/2020 | 1/23/2020 |
| Zhanjiang/20SF190/2020 | 20SF190 | Male | 60 | Shenzhen | Throat_swab | 1/12/2020 | 1/23/2020 |
| Shenzhen/20SF118/2020 | 20SF118 | Male | 71 | Shenzhen | Throat_swab | 1/12/2020 | 1/23/2020 |
| Shenzhen/20SF117/2020 | 20SF117 | Female | 71 | Shenzhen | Throat_swab | 1/12/2020 | 1/23/2020 |
| Zhuhai/20SF028/2020 | 20SF028 | Male | 68 | Shenzhen | Throat_swab | 1/12/2020 | 1/23/2020 |
| Zhuhai/20SF040/2020 | 20SF040 | Female | 49 | Shenzhen | Throat_swab | 1/12/2020 | 1/23/2020 |
| Zhuhai/20SF043/2020 | 20SF043 | Male | 76 | Shenzhen | Throat_swab | 1/12/2020 | 1/23/2020 |
| Zhuhai/20SF043/2020 | 20SF043 | Female | 63 | Shenzhen | Throat_swab | 1/12/2020 | 1/23/2020 |
| Zhuhai/20SF040/2020 | 20SF040 | Female | 49 | Shenzhen | Throat_swab | 1/12/2020 | 1/23/2020 |
| Zhuhai/20SF043/2020 | 20SF043 | Male | 76 | Shenzhen | Throat_swab | 1/12/2020 | 1/23/2020 |
| Zhuhai/20SF040/2020 | 20SF040 | Female | 49 | Shenzhen | Throat_swab | 1/12/2020 | 1/23/2020 |
| Zhuhai/20SF043/2020 | 20SF043 | Male | 76 | Shenzhen | Throat_swab | 1/12/2020 | 1/23/2020 |
| Zhuhai/20SF040/2020 | 20SF040 | Female | 49 | Shenzhen | Throat_swab | 1/12/2020 | 1/23/2020 |
| Zhuhai/20SF043/2020 | 20SF043 | Male | 76 | Shenzhen | Throat_swab | 1/12/2020 | 1/23/2020 |
| Zhuhai/20SF040/2020 | 20SF040 | Female | 49 | Shenzhen | Throat_swab | 1/12/2020 | 1/23/2020 |
| Zhuhai/20SF043/2020 | 20SF043 | Male | 76 | Shenzhen | Throat_swab | 1/12/2020 | 1/23/2020 |
| Zhuhai/20SF040/2020 | 20SF040 | Female | 49 | Shenzhen | Throat_swab | 1/12/2020 | 1/23/2020 |
| Zhuhai/20SF043/2020 | 20SF043 | Male | 76 | Shenzhen | Throat_swab | 1/12/2020 | 1/23/2020 |
| Zhuhai/20SF040/2020 | 20SF040 | Female | 49 | Shenzhen | Throat_swab | 1/12/2020 | 1/23/2020 |
| Zhuhai/20SF043/2020 | 20SF043 | Male | 76 | Shenzhen | Throat_swab | 1/12/2020 | 1/23/2020 |
| Zhuhai/20SF040/2020 | 20SF040 | Female | 49 | Shenzhen | Throat_swab | 1/12/2020 | 1/23/2020 |
| Zhuhai/20SF043/2020 | 20SF043 | Male | 76 | Shenzhen | Throat_swab | 1/12/2020 | 1/23/2020 |
| Zhuhai/20SF040/2020 | 20SF040 | Female | 49 | Shenzhen | Throat_swab | 1/12/2020 | 1/23/2020 |
| Zhuhai/20SF043/2020 | 20SF043 | Male | 76 | Shenzhen | Throat_swab | 1/12/2020 | 1/23/2020 |
| Zhuhai/20SF040/2020 | 20SF040 | Female | 49 | Shenzhen | Throat_swab | 1/12/2020 | 1/23/2020 |

Note: The table continues with similar entries for the remaining specimens collected between January 14, 2020 and January 31, 2020.
by real-time reverse transcription PCR (RT-PCR) method. The positive samples were sent to Guangdong Provincial CDC for reexamination and identification through nucleic acid testing.

1.2 Whole RNA Meta-transcriptomic Sequencing

We collected clinical samples including bronchoalveolar lavage fluid (BALF), endotracheal aspirates, throat swabs, and nasal swabs from patients and performed meta-transcriptomic sequencing. Total RNA was extracted from 200 μL sputum fluid with the human rRNA Depletion Kit (NEB #E6350). A meta-transcriptomic library was constructed for single-end (75 bp) sequencing using an Illumina NextSeq 550Dx, and the sequencing data were analyzed with the rapid pathogen detection system (RPD-seq, Guangzhou Sagene Biotech Co., Ltd.). Sequence reads were de novo assembled and screened for the whole genome with potential mutations.

1.3 Phylogenetic Analysis of the Virus RNA Sequence and the Spike Protein Structure Modeling

We retrieved the coronavirus sequences including the SARS, bat SARS-like and available genomes from NCBI viral genome database (https://www.ncbi.nlm.nih.gov/) and the GISAID (https://www.gisaid.org/). Multiple sequence alignment of all coronavirus genomes was performed by using MUSCLE software[9]. Out of coronavirus representative genomes of all category were used for phylogenetic tree development using MEGAX software based on neighbor joining method[9]. The phylogenetic tree bootstrap value was 1000 to evaluate reliability. The glycoprotein region of the SARS CoV, Bat-SL-RaTG13 CoV and SARS-CoV-2 were aligned and visualized using Multalin software[11]. The conservation of the amino acid motifs in clinical variants of SARS-CoV-2 genome was presented by performing multiple sequence alignment using MEGAX software. The three dimensional structure of SARS-CoV-2 envelope (spike, S) glycoprotein was generated by using SWISS-MODEL online server[12] and the structure was marked and visualized by using RasMol (www.openrasmol.org). The model of the SARS-RBD combined with receptor complex (PDB code 6acg) was used to predict possibility of the SARS-CoV-2 RBD binding with the potential human receptor (ACE2)[13].

2 RESULTS

2.1 Virus Genome Sequence Assembly

The original macro-transcriptome data of 60 samples were obtained by Next-Generation Sequencing (NGS). After removing the interference of host data, the complete genome sequence of SARS-CoV-2 was successfully assembled (table 2). Large fragment sequences of viruses can be assembled when the sequencing genome has almost complete coverage and the average sequencing depth is more than 10 times. Using reference sequence to correct the assembly sequence can improve the assembly quality.

2.2 Genomic Evolution

The first 6 confirmed cases in Guangdong province and the first 5 confirmed cases in Wuhan were selected to construct a phylogenetic tree with other Beta coronaviruses (fig. 1). The 11 cases of SARS-CoV-2 represent highly homologous sequences, which are obviously clustered. The evolutionary relationship within the cluster is not obvious, which is significantly different from other coronavirus genomes. This suggests that the current variation in the SARS-CoV-2 genome is still intraspecies, with no evolution to other

---

**Fig. 1** Phylogenetic analysis of the 11 cases with infection of SARS-CoV-2 and other Beta coronaviruses

The green branches are related species and the purple branches are distant evolutionary species.
### Table 2 60 samples of SARS-CoV-2 assembly information

| Sequence ID | Sample ID | Reads count | Genome coverage | Avg. coverage | Assembly length |
|-------------|-----------|-------------|-----------------|---------------|-----------------|
| Shenzhen/20SF012/2020 | 20SF012 | 74 680 | 100.00% | 168.43 | 29 915 |
| Shenzhen/20SF014/2020 | 20SF014 | 5 114 906 | 100.00% | 11680.12 | 29 919 |
| Shenzhen/20SF025/2020 | 20SF025 | 112 794 | 100.00% | 267.18 | 29 925 |
| Zhuhai/20SF028/2020 | 20SF028 | 190 912 | 100.00% | 452.61 | 29 918 |
| Zhuhai/20SF040/2020 | 20SF040 | 1 932 | 100.00% | 10.92 | 29 903 |
| Shenzhen/20SF118/2020 | 20SF118 | 7 833 | 99.83% | 5.07 | 29 904 |
| Shenzhen/20SF117/2020 | 20SF117 | 9 364 | 99.67% | 6.08 | 29 905 |
| Shenzhen/20SF117/2020 | 20SF117 | 9 364 | 99.67% | 6.08 | 29 905 |
| Shenzhen/20SF118/2020 | 20SF118 | 7 833 | 99.83% | 5.07 | 29 904 |
| Shenzhen/20SF117/2020 | 20SF117 | 9 364 | 99.67% | 6.08 | 29 905 |
| Shenzhen/20SF117/2020 | 20SF117 | 9 364 | 99.67% | 6.08 | 29 905 |
| Shenzhen/20SF118/2020 | 20SF118 | 7 833 | 99.83% | 5.07 | 29 904 |
2.3 Variation among SARS-CoV-2 Genomes

The four groups of family cluster cases were clustered into clusters (fig. 2), among which only one SNP existed between the internal strains in group 1, and the other three groups had no variation during transmission. The 60 complete genomes were nearly identical across the whole genome, with sequence identity being above 99.99%, indicating the genome is stable in the process of virus transmission. Notably, the sequence identity between the virus genomes from family clustering cases was more than 99.99%. There were 179 nucleotide and 107 amino acid variations in 60 genomes (fig. 2). The number of nucleotide variations in each genome varies from one to seven. There is no highly variable region, and all the single nucleotide variations (SNVs) are distributed in various fragments of the genome.

2.4 A Salt Bridge Between Lys417 and Asp12 Was Found as a Strong Interaction Force Between the SARS-CoV-2 Spike-RPD and the Receptor ACE2

As shown in the Spike receptor binding domain (RBD) sequences, GD-Pangolin-CoV was found to be the closest Beta-coronavirus to the three COVID-19 genomes, showing as high as 97.22% identity and similarity between them. The coronavirus isolates (Bat-CoV-RaTG) came right after the beta-coronavirus, it showed an overall similarity 88.89% with the SARS-CoV-2 Spike-RPD sequences. Surprisingly, the SARS Spike-RPD displayed a much far distance, of which similarity was 72.63% (fig. 3).

The predicted crystal structures of SARS Spike trimmer with the receptor angiotensin-converting enzyme II (ACE2) were applied as the modeling template due to the availability, and relatively high identity and similarity between their RBD sequences[14–16] (fig. 3). Unlike SARS-CoV, a salt bridge between Lys417 and Asp12 was found as a strong interaction force between the SARS-CoV-2 Spike-RPD and the receptor ACE2 in structural prediction model (fig. 4). By checking back to their primary protein sequences, we found the Lys417 of SARS-CoV-2 Spike-RBD was an amino acid replacement. On the SARS-CoV, it was a neutral amino acid Valine there shown on the sequence alignment above and marked by a rectangle with red color (fig. 3A).

2.5 The PRRA-insert Brought a Furin Cleavage Site in Spike Protein

The Spike protein belongs to the Class I viral fusion protein, including SARS Spike protein (S), HIV envelop glycoprotein (Env), flu Hemagglutinin (HA) and Ebolavirus glycoproteins (GP). For further elucidation of the host-virus interactions, we checked the Spike fusion ability with the host membrane through scanning possible S1/S2 cleavage site of SARS-CoV-2 Spike protein. Compared with those beta-coronavirus Spike sequences, there was a four amino acid PRRA-insert in SARS-CoV-2, but not in any others such as SARS-CoV, Bat-CoV or the most closely-related GD-Pangolin-CoV (fig. 5). With another Arginine (R685) right after it, the PRRA-insert resulted in a typical protease furin cleavage site RRAR685 on SARS-CoV-2, in which the S1/S2 boundary was highly assumed.

3 DISCUSSION

Through multiple sequence comparison and evolution analysis of those assembly sequences obtained from 60 specimens, results showed that all the specimens had very few mutations, which was highly consistent with the whole genome sequence of SARS-CoV-2 in the early outbreak. Those variations tend to be randomly dispersed due to no selective pressure. By analyzing the SNPs in 60 specimens, the results showed that the mutation frequency was low in high depth regions (DP-30). A branch with ORF8: L84S variation was commonly observed, indicating that the available virus specimen was not under much selective pressure with a very slow mutation rate. However, there were also some strains that did not have a significant epidemiological association but also had a 100% sequence identity, suggesting that there may be some epidemiological associations not observed from the spatial and temporal distribution.

Through establishing a molecular model of interaction between Spike and human receptor, an amino acid salt bridge was found in SARS-CoV-2, but not in SARS-CoV. Furthermore, a potential furin cleavage site right located on the S1/S2 boundary of Spike protein greatly enhanced the virus invasion and pathogenicity. The receptor-binding analysis may identify important factors in the infection and invasion of COVID-19 much stronger than those of SARS in 2003. More than increasing its invasion, this potential furin cleavage site in SARS-CoV-2 might also result from the difference of the virus package mechanism from the SRAS-CoV[17]. It is not in the case of SRAS-CoV, although its invasion is still required. The SLLR667 on the SARS-CoV spike glycoprotein enhances cell-cell fusion but does not affect virion entry, which has not been considered as a typical furin cleavage site. The SLLR667 on the SARS-CoV spike glycoprotein enhances cell-cell fusion but does not affect virion entry, which has not been considered as a typical furin cleavage site between S1 and S2 boundary of their Spike proteins. With a different packaging mechanism,
Fig. 2 Variation among SARS-CoV-2 genomes

The color of the branch node on the left is red for cases in Wuhan and blue for cases in Guangdong. The four groups of family cluster cases are highlighted by different color on strain number. The nucleotide variations (SNVs) are represented by vertical lines.
**Fig. 4** Structural prediction for Spike-RPD bound with the receptor ACE2 in SARS-CoV-2

In summary, we selected 60 specimens of SARS-CoV-2 whole genome sequences in the early phase of COVID-19 pandemic in Guangdong, and systematically analyzed the characteristics of the SARS-CoV-2 genomics evolution, amino acid variation and Spike protein structure. This study provided reference for generating prevention and control strategies and the tracing the source of new outbreaks.

this novel coronavirus acquired even higher cell transmission efficiency, and this might be one of the important reasons for the spreading of COVID-19 much wider than the SARS. With this typical furin cleavage of Spike protein, the immune treatment and drug use might be extended into anti-virus categories, especially the application of furin-protease inhibitors for SARS-CoV-2.
Conflict of Interest Statement
The authors declare that there is no conflict of interest with any financial organization or corporation or individual that can inappropriately influence this work.

REFERENCES
1 Su S, Wong G, Shi W, et al. Epidemiology, Genetic Recombination, and Pathogenesis of Coronaviruses. Trends Microbiol, 2016,24(6):490-502
2 Kisiazek T G, Erdman D, Goldsmith C S, et al. A novel coronavirus associated with severe acute respiratory syndrome. N Engl J Med, 2003,348(20):1953-1966
3 Kuiken T, Fouchier R A, Schutten M, et al. Newly discovered coronavirus as the primary cause of severe acute respiratory syndrome. Lancet, 2003,362(9380):263-270
4 Zaki AM, van Boheemen S, Bestebroer TM, et al. Isolation of a novel coronavirus from a man with pneumonia in Saudi Arabia. N Engl J Med, 2012,367(19):1814-1820
5 WHO. Summary of probable SARS cases with onset of illness from 1 November 2002 to 31 July 2003. Dec 31, 2003. https://www.who.int/csr/sars/country/table2004_04_21/en/ (accessed Jan 19, 2020).
6 WHO. Middle East respiratory syndrome coronavirus (MERS-CoV). November, 2019. http://www.who.int/emergencies/mers-cov/en/ (accessed Jan 19, 2020).
7 Lu R, Zhao X, Li J, et al. Genomic characterisation and epidemiology of 2019-nCoV pneumonia in China, 2019. N Engl J Med, 2020,382(8):727-733
8 Wrapp D, Wang N, Corbett K S, et al. Cryo-EM structure of the SARS coronavirus spike glycoprotein in complex with its host cell receptor ACE2. PLoS Pathog, 2020,16(2):e1007615
9 Belouzard S, Chu VC, Whittaker GR. Activation of the SARS coronavirus spike protein via sequential proteolytic cleavage at two distinct sites. Proc Natl Acad Sci U S A, 2009,106(14):5871-5876

Fig. 5 A very potential cleavage site RRAR685 near the S1/S2 boundary, compared with furin cleavage sites for other viruses. RRAR685 in SARS-CoV-2 is marked with a red rectangle, and the corresponding short sequence in SARS is SLLR667.