Current Medical Science
DOI https://doi.org/10.1007/s11596-021-2340-3
41(2):228-235,2021

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

Bai-sheng LI[1, 2†], Zhen-cui LI[1, 2†], Yao HU[1, 2†], Li-jun LIANG[1, 2], Li-rong ZOU[1, 2], Qian-fang GUO[1, 2], Zhong-hua ZHENG[1, 2], Jian-xiang YU[1, 2], Tie SONG[1, 2], Jie WU[1, 2, 3#]
[1]*Guangdong Provincial Center for Disease Control and Prevention, Guangzhou 510000, China*
[2]*Guangdong Workstation for Emerging Infectious Disease Control and Prevention, Chinese Academy of Medical Sciences, Guangzhou 510000, China*
[3]*Southern Medical University, Guangzhou 510000, China*

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	63	Shenzhen	Throat_swab	1/8/2020	1/15/2020
Zhuhai/20SF028/2020	20SF028	Male	68	Zhuhai	Throat_swab	1/11/2020	1/17/2020
Zhuhai/20SF040/2020	20SF040	Female	49	Zhuhai	Nasal_swab	1/17/2020	1/18/2020
Zhuhai/20SF043/2020	20SF043	Female	76	Zhuhai	Nasal_swab	1/12/2020	1/18/2020
Shenzhen/20SF118/2020	20SF118	Male	69	Shenzhen	Throat_swab	1/12/2020	1/21/2020
Shenzhen/20SF117/2020	20SF117	Female	71	Shenzhen	Throat_swab	1/15/2020	1/21/2020
Shaoguan/20SF190/2020	20SF190	Male	60	Shaoguan	Throat_swab	1/22/2020	1/23/2020
Huizhou/20SF195/2020	20SF195	Female	58	Huizhou	Throat_swab	1/17/2020	1/22/2020
Yangjiang/20SF200/2020	20SF200	Female	57	Yangjiang	Throat_swab	1/22/2020	1/23/2020
Yangjiang/20SF201/2020	20SF201	Male	25	Yangjiang	Throat_swab	1/22/2020	1/23/2020
Guangzhou/20SF206/2020	20SF206	Female	73	Guangzhou	Throat_swab	1/20/2020	1/22/2020
Foshan/20SF207/2020	20SF207	Male	57	Foshan	Throat_swab	1/19/2020	1/22/2020
Foshan/20SF210/2020	20SF210	Female	57	Foshan	Throat_swab	1/21/2020	1/22/2020
Foshan/20SF211/2020	20SF211	Male	57	Foshan	Throat_swab	1/22/2020	1/22/2020
Shenzhen/20SF243/2020	20SF243	Male	63	Shenzhen	Throat_swab	1/21/2020	1/22/2020
Huizhou/20SF198/2020	20SF198	Female	50	Huizhou	Throat_swab	1/18/2020	1/22/2020
Guangzhou/20SF115/2020	20SF115	Female	63	Guangzhou	Throat_swab	1/14/2020	1/21/2020
Zhanjiang/20SF123/2020	20SF123	Female	77	Zhanjiang	Throat_swab	1/15/2020	1/20/2020
Zhuhai/20SF134/2020	20SF134	Female	49	Zhuhai	Nasal_swab	1/17/2020	1/21/2020
Zhuhai/20SF136/2020	20SF136	Male	77	Zhuhai	Nasal_swab	1/11/2020	1/21/2020
Guangzhou/20SF156/2020	20SF156	Female	52	Guangzhou	Nasal_swab	1/21/2020	1/22/2020
Zhuhai/20SF167/2020	20SF167	Male	77	Zhuhai	Throat_swab	1/11/2020	1/22/2020
Zhuhai/20SF174/2020	20SF174	Male	77	Zhuhai	Nasal_swab	1/11/2020	1/22/2020
Zhuhai/20SF179/2020	20SF179	Male	36	Zhuhai	Throat_swab	1/17/2020	1/22/2020
Qingyuan/20SF252/2020	20SF252	Male	19	Qingyuan	Throat_swab	1/21/2020	1/21/2020
Huizhou/20SF253/2020	20SF253	Male	56	Huizhou	Throat_swab	1/17/2020	1/23/2020
Huizhou/20SF254/2020	20SF254	Female	38	Huizhou	Throat_swab	1/20/2020	1/23/2020
Shenzhen/20SF262/2020	20SF262	Female	78	Shenzhen	Throat_swab	1/23/2020	1/23/2020
Shenzhen/20SF263/2020	20SF263	Female	54	Shenzhen	Throat_swab	1/20/2020	1/23/2020
Shenzhen/20SF265/2020	20SF265	Female	64	Shenzhen	Throat_swab	1/22/2020	1/23/2020
Guangzhou/20SF273/2020	20SF273	Male	64	Guangzhou	Throat_swab	1/21/2020	1/23/2020
Huizhou/20SF316/2020	20SF316	Female	66	Huizhou	Throat_swab	1/19/2020	1/24/2020
Zhuhai/20SF326/2020	20SF326	Male	26	Zhuhai	Nasal_swab	1/19/2020	1/24/2020
Guangzhou/20SF374/2020	20SF374	Male	50	Guangzhou	Throat_swab	1/23/2020	1/26/2020
Meizhou/20SF440/2020	20SF440	Male	27	Meizhou	Throat_swab	1/24/2020	1/25/2020
Shenzhen/20SF616/2020	20SF616	Female	55	Shenzhen	Throat_swab	1/22/2020	1/24/2020
Dongguan/20SF629/2020	20SF629	Male	26	Dongguan	Throat_swab	1/23/2020	1/26/2020
Dongguan/20SF630/2020	20SF630	Male	30	Dongguan	Throat_swab	1/23/2020	1/26/2020
Dongguan/20SF632/2020	20SF632	Female	35	Dongguan	Throat_swab	1/24/2020	1/26/2020
Zhongshan/20SF665/2020	20SF665	Male	55	Zhongshan	Throat_swab	1/23/2020	1/27/2020
Shantou/20SF684/2020	20SF684	Female	39	Shantou	Throat_swab	1/27/2020	1/27/2020
Zhuhai/20SF753/2020	20SF753	Female	28	Zhuhai	Nasal_swab	1/26/2020	1/28/2020
Zhuhai/20SF758/2020	20SF758	Male	44	Zhuhai	Throat_swab	1/24/2020	1/27/2020
Dongguan/20SF840/2020	20SF840	Female	7	Dongguan	Throat_swab	1/29/2020	1/29/2020
Huizhou/20SF1152/2020	20SF1152	Male	40	Huizhou	Throat_swab	1/29/2020	1/30/2020
Huizhou/20SF1153/2020	20SF1153	Female	53	Huizhou	Throat_swab	1/30/2020	1/30/2020
Zhuhai/20SF1159/2020	20SF1159	Female	58	Zhuhai	Throat_swab	1/22/2020	1/31/2020
Zhanjiang/20SF602/2020	20SF602	Female	40	Zhanjiang	Throat_swab	1/25/2020	1/25/2020
Shantou/20SF685/2020	20SF685	Female	17	Shantou	Throat_swab	1/27/2020	1/27/2020
Huizhou/20SF808/2020	20SF808	Male	38	Huizhou	Nasal_swab	1/27/2020	1/28/2020
Huizhou/20SF812/2020	20SF812	Female	56	Huizhou	Nasal_swab	1/21/2020	1/28/2020
Foshan/20SF822/2020	20SF822	Female	68	Foshan	Sputum	1/21/2020	1/28/2020
Huizhou/20SF813/2020	20SF813	Male	41	Huizhou	Throat_swab	1/27/2020	1/28/2020
Huizhou/20SF1107/2020	20SF1107	Female	57	Huizhou	Throat_swab	1/28/2020	1/29/2020
Shanwei/20SF1136/2020	20SF1136	Female	45	Shanwei	Throat_swab	1/30/2020	1/30/2020
Guangzhou/20SF2546/2020	20SF2546	Male	90	Guangzhou	Throat_swab	2/9/2020	2/6/2020
Guangzhou/20SF4047/2020	20SF4047	Female	71	Guangzhou	Throat_swab	1/27/2020	2/17/2020
Guangzhou/20SF4051/2020	20SF4051	Female	41	Guangzhou	Throat_swab	2/1/2020	2/17/2020

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[10]. 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 identified amino acids were aligned with whole viral genome database using BLASTp. 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

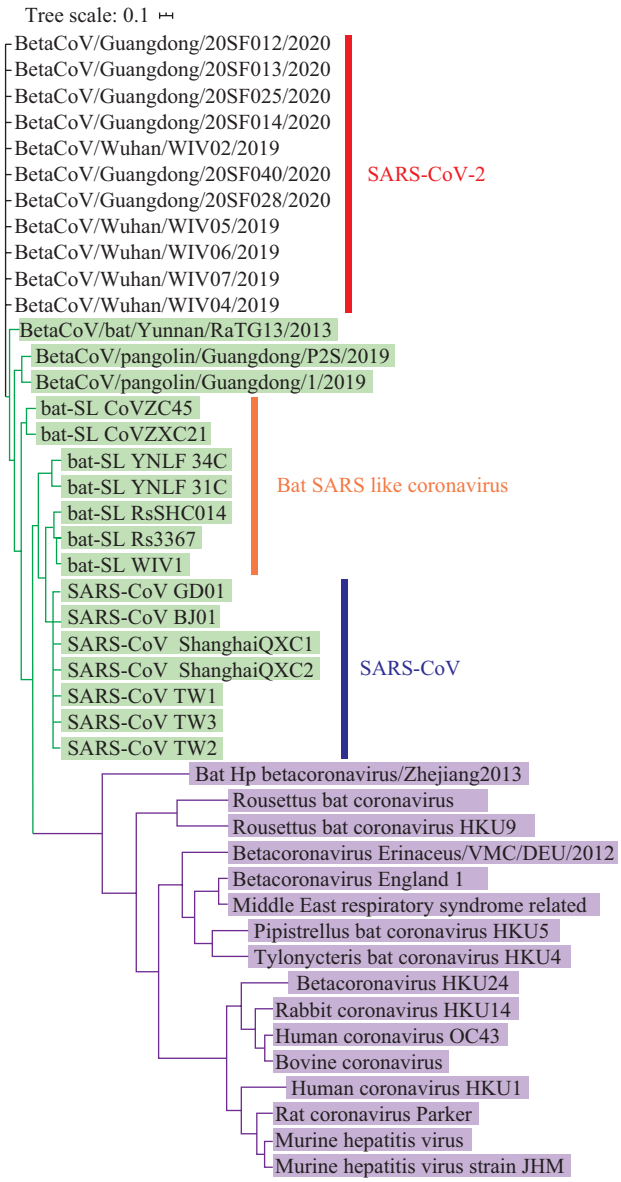


Fig. 1 Phylogenetic analysis of the 11 cases with infection of SARS-CoV-2 and other Beta coronaviruses
The green branches are related species and the purple branches are distant evolutionary species.

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
Zhuhai/20SF043/2020	20SF043	404	93.97%	2.30	29 903
Shenzhen/20SF118/2020	20SF118	7 833	99.83%	5.07	29 903
Shenzhen/20SF117/2020	20SF117	9 364	99.67%	6.08	29 903
Shaoguan/20SF190/2020	20SF190	4 786	99.21%	3.12	29 904
Huizhou/20SF195/2020	20SF195	8 755	99.62%	5.75	29 903
Yangjiang/20SF200/2020	20SF200	10 033	99.52%	6.57	29 903
Yangjiang/20SF201/2020	20SF201	136 370	99.92%	88.84	29 911
Guangzhou/20SF206/2020	20SF206	54 393	99.89%	35.49	29 909
Foshan/20SF207/2020	20SF207	55 677	99.86%	36.27	29 903
Foshan/20SF210/2020	20SF210	46 158	99.84%	30.11	29 903
Foshan/20SF211/2020	20SF211	21 516	99.72%	14.09	29 903
Shenzhen/20SF243/2020	20SF243	5 396	99.60%	3.51	29 903
Huizhou/20SF198/2020	20SF198	9 584	99.78%	6.30	29 903
Guangzhou/20SF115/2020	20SF115	13 628	99.61%	8.91	29 903
Zhanjiang/20SF123/2020	20SF123	9 835	99.70%	6.44	29 909
Zhuhai/20SF134/2020	20SF134	4 154	99.26%	2.71	29 904
Zhuhai/20SF136/2020	20SF136	12 568	99.58%	8.23	29 903
Guangzhou/20SF156/2020	20SF156	3 376	98.82%	2.20	29 902
Zhuhai/20SF167/2020	20SF167	8 774	99.70%	5.75	29 902
Zhuhai/20SF174/2020	20SF174	50 475	99.70%	33.09	29 903
Zhuhai/20SF179/2020	20SF179	4 193	98.72%	2.73	29 902
Qingyuan/20SF252/2020	20SF252	7 335	95.52%	4.69	29 903
Huizhou/20SF253/2020	20SF253	7 028	94.52%	4.52	29 899
Huizhou/20SF254/2020	20SF254	64 671	99.93%	41.81	29 907
Shenzhen/20SF262/2020	20SF262	10 324	96.20%	6.63	29 903
Shenzhen/20SF263/2020	20SF263	3 485	89.85%	2.24	29 903
Shenzhen/20SF265/2020	20SF265	8 626	97.24%	5.48	29 903
Guangzhou/20SF273/2020	20SF273	91 595	99.94%	58.90	29 903
Huizhou/20SF316/2020	20SF316	25 638	99.95%	16.00	29 904
Zhuhai/20SF326/2020	20SF326	2 677	83.05%	1.68	29 901
Guangzhou/20SF374/2020	20SF374	4 804	90.55%	3.01	29 904
Meizhou/20SF440/2020	20SF440	6 797	97.34%	4.31	29 903
Shenzhen/20SF616/2020	20SF616	9 524	99.85%	6.24	29 903
Dongguan/20SF629/2020	20SF629	24 406	99.91%	15.98	29 907
Dongguan/20SF630/2020	20SF630	6 948	99.80%	4.56	29 907
Dongguan/20SF632/2020	20SF632	23 353	99.88%	15.34	29 903
Zhongshan/20SF665/2020	20SF665	17 855	99.85%	11.74	29 903
Shantou/20SF684/2020	20SF684	11 660	99.84%	7.56	29 903
Zhuhai/20SF753/2020	20SF753	38 047	99.86%	24.86	29 903
Zhuhai/20SF758/2020	20SF758	1 111 321	99.91%	725.57	29 875
Dongguan/20SF840/2020	20SF840	10 457	99.92%	6.72	29 903
Huizhou/20SF1152/2020	20SF1152	9 352	99.92%	6.13	29 903
Huizhou/20SF1153/2020	20SF1153	10 865	99.85%	7.08	29 903
Zhuhai/20SF1159/2020	20SF1159	4 432	99.74%	2.89	29 903
Zhanjiang/20SF602/2020	20SF602	7 665	99.34%	5.05	29 903
Shantou/20SF685/2020	20SF685	9 623	99.83%	6.31	29 903
Huizhou/20SF808/2020	20SF808	26 316	99.86%	17.00	29 903
Huizhou/20SF812/2020	20SF812	93 242	99.87%	60.81	29 925
Foshan/20SF822/2020	20SF822	5 483	99.74%	3.58	29 903
Huizhou/20SF813/2020	20SF813	66 634	99.85%	43.58	29 903
Huizhou/20SF1107/2020	20SF1107	10 525	99.84%	6.84	29 900
Shanwei/20SF1136/2020	20SF1136	5 916	99.86%	3.86	29 903
Guangzhou/20SF2546/2020	20SF2546	44 458	99.85%	65.98	29 859
Guangzhou/20SF4047/2020	20SF4047	83 538	99.85%	54.45	29 851
Guangzhou/20SF4051/2020	20SF4051	8 455	99.42%	5.55	29 465

coronaviruses.

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 99.9%, 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-conronavirus, 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 and has not been considered as a typical furin cleavage site[18–20]. In contrast, the SARS-CoV-2 might use the package way similar to the mouse hepatitis virus (MHV), human immunodeficiency virus (HIV) or Ebolavirus, since most other beta-coronavirus did not display a typical furin cleavage site between S1 and S2 boundary of their Spike proteins. With a different packaging mechanism,

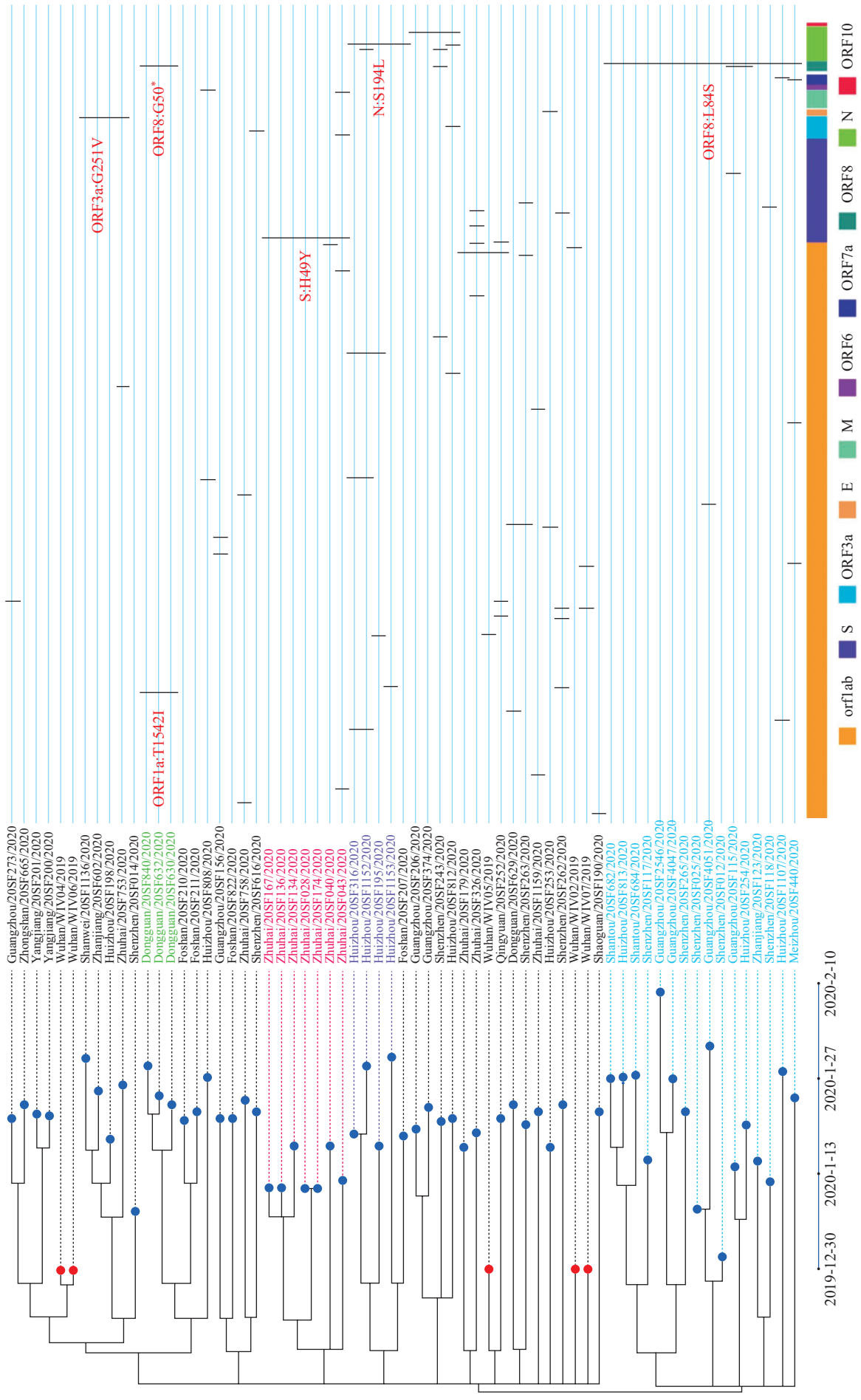


Fig. 2 Variation among SARS-CoV-2 genomes
The color of the branch node on the left is red for cases in Wuhan and blue for cases in Guangdong. The four groups of family cluster cases are highlighted by different color on strain number. The nucleotide variations (SNVs) are represented by vertical lines.

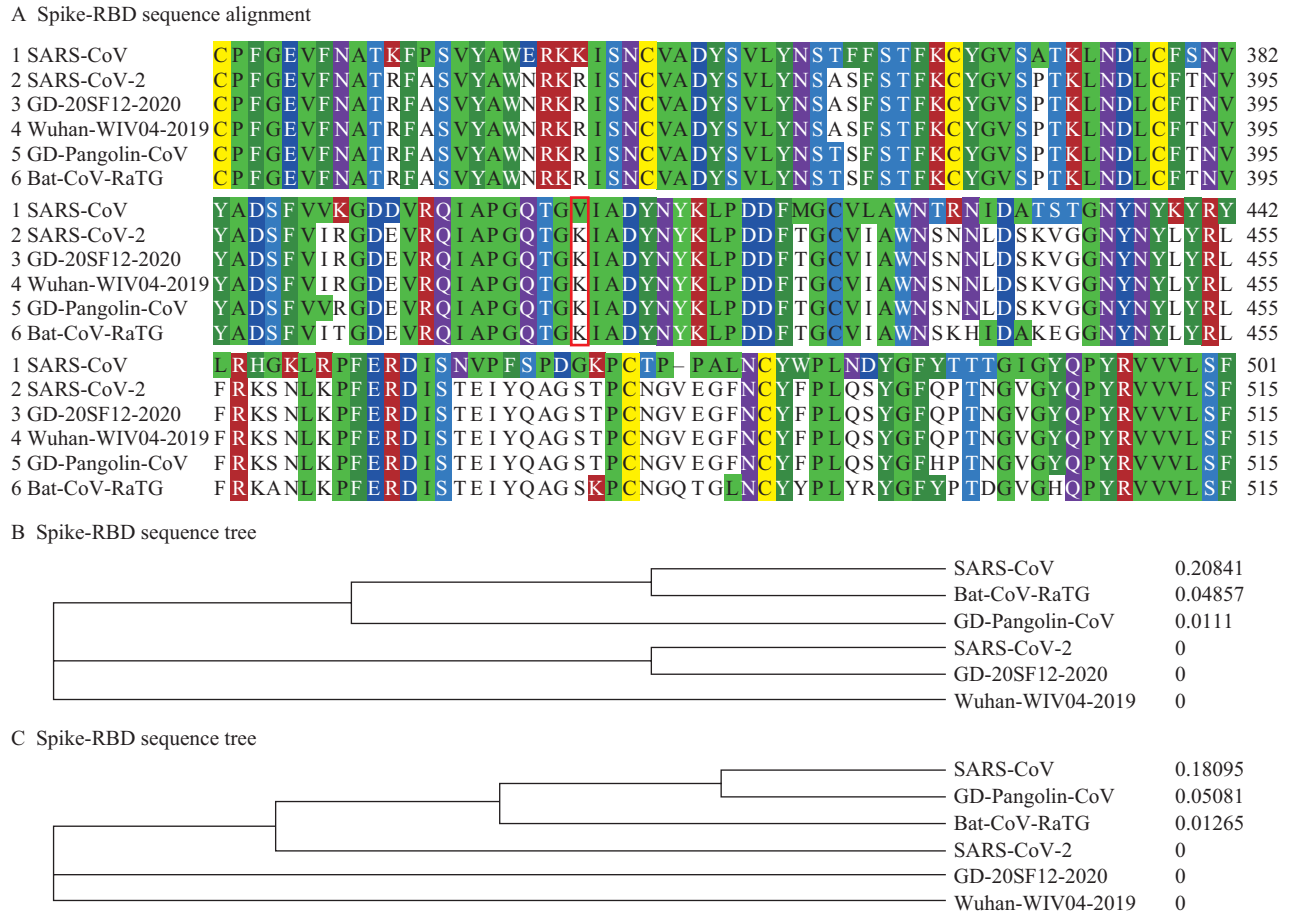


Fig. 3 The phylogenetic analysis for Spike receptor binding domain (RBD) sequences of beta-coronaviruses closest to the novel coronary pneumonia COVID-19

The numbers on the right in figures B and C indicated the evolutionary distance. SARS-CoV denoted for the NC_004718.3; SARS-CoV-2 for the NC_045512.2 as the seafood market pneumonia virus isolate Wuhan-Hu-1; GD-20SF12-2020 for the first case of SARS-CoV-2 in Guangdong CDC; Wuhan-WIV04-2019 for the MN996528.1 SARS-CoV-2 isolate WIV04; GD-Pangolin-CoV for the BetaCoV/pangolin Guandong/1/2019|EPI_ISL_410721; Bat-CoV-RaTG for the MN996532.1 Bat coronavirus isolate RaTG13.

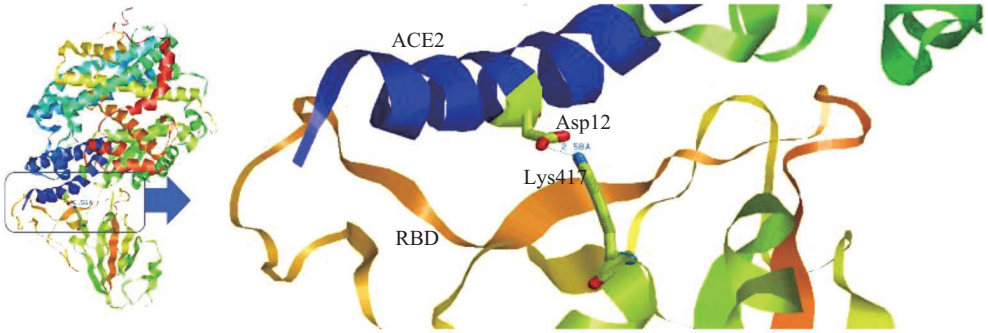
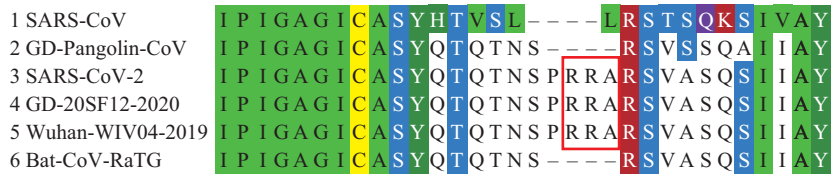


Fig. 4 Structural prediction for Spike-RPD bound with the receptor ACE2 in SARS-CoV-2

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.

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.



Virus	Site Seq	Genome	Host
SARS-CoV	$SLLR_{667}$	NC_004718.3	Human
SARS-CoV-2	$RRAR_{685}$	NC_045512.2	Human
MHV	$RAHR_{628}$	NC_001846.1	Mouse
HIV	$REKR_{515}$	BAF31430.1	Human
Ebolavirus	$RKIR_{302}$	AAD14585.1	Human

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.

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 Ksiazek 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 novel coronavirus: implications for virus origins and receptor binding. Lancet, 2020, 395(10224):565-574

8 Bankevich A, Nurk S, Antipov D, *et al*. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. J Comput Biol, 2012,19(5):455-477

9 Elbe S, Buckland-Merrett G. Data, disease and diplomacy: GISAID's innovative contribution to global health. Glob Chall, 2017,1(1):33-46

10 Kumar S, Stecher G, Li M, *et al*. MEGA X: Molecular Evolutionary Genetics Analysis across Computing Platforms. Mol Biol Evol, 2018,35(6):1547-1549

11 Corpet F. Multiple sequence alignment with hierarchical clustering. Nucleic Acids Res, 1988,16(22):10881-10890

12 Biasini M, Bienert S, Waterhouse A, *et al*. SWISS-MODEL: modelling protein tertiary and quaternary structure using evolutionary information. Nucleic Acids Res, 2014,42(Web Server issue):W252-W258

13 Song W, Gui M, Wang X, *et al*. Cryo-EM structure of the SARS coronavirus spike glycoprotein in complex with its host cell receptor ACE2. PLoS Pathog, 2018,14(8):e1007236

14 Zhu N, Zhang D, Wang W, *et al*. A Novel Coronavirus from Patients with Pneumonia in China, 2019. N Engl J Med, 2020,382(8):727-733

15 Wrapp D, Wang N, Corbett K S, *et al*. Cryo-EM structure of the 2019-nCoV spike in the prefusion conformation. Science, 2020,367(6483):1260-1263

16 Alam N, Goldstein O, Xia B, *et al*. High-resolution global peptide-protein docking using fragments-based PIPER-FlexPepDock. PLoS Comput Biol, 2017,13(12): e1005905

17 Song W, Gui M, Wang X, *et al*. Cryo-EM structure of the SARS coronavirus spike glycoprotein in complex with its host cell receptor ACE2. PLoS Pathog, 2018,14(8):e1007236

18 Pillay T S. Gene of the month: the 2019-nCoV/SARS-CoV-2 novel coronavirus spike protein. J Clin Pathol, 2020,73(7):366-369

19 Follis KE, York J, Nunberg JH. Furin cleavage of the SARS coronavirus spike glycoprotein enhances cell-cell fusion but does not affect virion entry. Virology, 2006,350(2):358-369

20 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

(Received Jan. 9, 2021; accepted Mar. 12, 2021)