Genomic surveillance of a resurgence of COVID-19 in Guangzhou, China

Zhou-Bin Zhang

Guangzhou Center for Disease Control and Prevention, Guangzhou 510440, China.

Peng Li

Center for Disease Control and Prevention of PLA, Beijing, China

Hong-Jie Liu

Beijing Institute of Radiation Medicine, Beijing, China

Jia-Yong Zhong

State Key Laboratory of Ophthalmology, Zhongshan Ophthalmic Center, Sun Yat-sen University, Guangzhou, China

Yingfeng Zheng

State Key Laboratory of Ophthalmology, Zhongshan Ophthalmic Center, Sun Yat-sen University, Guangzhou, China

Kui-Biao Li

Guangzhou Center for Disease Control and Prevention, Guangzhou 510440, China.

Peng-Zhe Qin

Guangzhou Center for Disease Control and Prevention, Guangzhou 510440, China.

Qing Zeng

Berry Genomics Co. Ltd, Beijing, China

Jin-Hui Li

Center for Disease Control and Prevention of PLA, Beijing, China

Li-Zhong Li

Center for Disease Control and Prevention of PLA, Beijing, China

Lan Cao

Guangzhou Center for Disease Control and Prevention, Guangzhou 510440, CGuangzhou Center for Disease Control and Prevention, Guangzhou 510440, China.hina.

Ji-Bin Wu

Guangzhou Center for Disease Control and Prevention, Guangzhou 510440, China.

Yi-Yun Chen

Guangzhou Center for Disease Control and Prevention, Guangzhou 510440, China.

Lei-Li Jia

Center for Disease Control and Prevention of PLA, Beijing, China

Hong-Bin Song

Center for Disease Control and Prevention of PLA, Beijing, China

Hua-Dong Wang

Department of Pathophysiology, Key Laboratory of State Administration of Traditional Chinese Medicine of the People's Republic of China—School of Medicine, Jinan University, Guangzhou, China

Qi-Wei Zhang

Guangdong Provincial Key Laboratory of Virology, Institute of Medical Microbiology, Jinan University, Guangzhou, China

Jun Fan

Department of Medical Biochemistry and Molecular Biology, Jinan University, Guangzhou, China

Biao Di

Guangzhou Center for Disease Control and Prevention, Guangzhou 510440, China.

Hong-Ling Jia

Department of Medical Biochemistry and Molecular Biology, Jinan University, Guangzhou, China

Ming Ni

Beijing Institute of Radiation Medicine, Beijing, China

Chuan-Le Xiao (✉ xiaochuanle@126.com)

State Key Laboratory of Ophthalmology, Zhongshan Ophthalmic Center, Sun Yat-sen University, Guangzhou, China <https://orcid.org/0000-0002-4680-0682>

Article

Keywords: COVID-19, corona virus, genomic surveillance, transmission patterns, Guangzhou, China

Posted Date: June 23rd, 2020

DOI: <https://doi.org/10.21203/rs.3.rs-35869/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Abstract

In the middle of March, the World Health Organization declared the outbreak of COVID-19 caused by SARS-CoV-2 infection a global pandemic. While China experienced a dramatic decline in daily growth rate of COVID-19, multiple importations of new cases from other countries and their related local infections caused a rapid rise. Between March 12 and April 15, we collected nasopharyngeal samples from 109 imported cases from 25 countries and 69 local cases in Guangzhou, China. In order to characterize the transmission patterns and genetic evolution of this virus among different populations, we sequenced the genome of SARS-CoV-2. The imported viral strains were assigned to lineages distributed in Europe (33.0%), America (17.4%), Africa (25.7%), or Southeast/West Asia (23.9%). Importantly, 10 imported cases from Africa formed two novel sub-lineages not identified in global tree previously. A detailed analysis showed that the imported viral strains from Philippines and Pakistan were closely related and within the same sub-lineage, whereas Ethiopia had varied lineages in the African phylogenetic tree. In spite of the diversity of imported SARS-CoV-2, 60 of 69 local infections could be traced back to two specific small lineages imported from Africa. A combined genetic and epidemiological analysis revealed a high-resolution transmission network of the imported SARS-CoV-2 in local communities, which might help inform the public health response and genomic surveillance in other cities and regions. Finally, we observed in-frame deletions on seven loci of SARS-CoV-2 genome, some of which were intra-host mutations, and they exhibited no enrichment on the S protein. Our findings provide new insight into the viral phylodynamics of SARS-CoV-2 and beta coronavirus.

Introduction

In the middle of May 2020, coronavirus disease 2019 (COVID-19) pandemic caused by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) has resulted in 6 million confirmed cases (till 31-May) and killed over 0.36 million¹.

Genomic surveillance of SARS-CoV-2 has played a critical role in containing this pandemic. This is exemplified by the whole genome sequence of the first identified SARS-CoV-2 strain². Thereafter, related coronavirus strains in bat and pangolin have been identified³⁻⁷. Over 18,700 genome sequences of SARS-CoV-2 strains have been shared by researchers worldwide in public databases such as GISAID³ and GenBank. The Nextstrain website⁸ allows us to analyze and visualize these global viral genomes in a real-time manner. As a result, lineages of SARS-CoV-2 in global and specific regions can be rapidly identified, and the introduction and transmission SARS-CoV-2 in local areas can be discussed⁹⁻¹⁵.

In late January, China rapidly implemented travel restrictions and transmission control measures, together with great medical assistance to the epidemic area. In February and March, there is a remarkable drop of COVID-19 infections in Wuhan and other cities in China, and in late of March to April, many areas including Hubei had no local new infections¹⁶. However, as COVID-19 pandemic in many countries continued, importations of SARS-CoV-2 infections from other countries into China were reported, and there were spikes in provinces or cities such as Heilongjiang Province, Shanghai, Guangdong Province,

and Beijing. Genomic surveillance of imported cases and related cases is crucial, owing to its capacity to elucidate the phylogenetics and transmission of SARS-CoV-2. Such information will help facilitate decision making about COVID-19 control and prevention. Genomic analysis of international passengers also enables us to explore the genomic information in infected individuals from countries where the research data are currently not available.

Results

Enhanced surveillance in Guangzhou

Guangzhou, the capital city of Guangdong Province, is a transport hub in South China with over 15 million inhabitants. From February 27 to April 11, there were 443 inbound flights, with more than 56,000 passengers arriving at Baiyun Airport, Guangzhou. As local infections associated with overseas arrivals continued to increase, Guangzhou took stringent measures to trace and control the cases. The city has been implementing a “five-shield” system to prevent the spread of SARS-CoV-2 from passengers arriving in Guangzhou. The system includes (1) status check at checkpoints; (2) medical observation and isolation at designated venues; (3) community screening; (4) close contact screening; and (5) fever clinics and treatment at hospitals.

Since March 21, all travelers entering Guangzhou have been required to undergo central quarantine by taking nucleic acid tests for SARS-CoV-2 and undergoing a 14-day quarantine at designated venues. The city has also conducted screening that involved people who arrived before the central quarantine measure was put into place, as well as people from high-risk regions and via high-risk flights.

As of April 5, Guangzhou has screened 6,353 people. Between March 15 to 12 April, 119 imported cases have been confirmed. 25 of them were of foreign nationality and 94 were of Chinese nationals. This was coincident with an increase in SARS-CoV-2 infection in the local communities (Figure 1a).

Patients' characteristics and contact history

We obtained nasopharyngeal samples between March 12 and April 15 in this study. These included 109 (109/119, 91.6%) imported cases, 69 local cases and 5 information missing cases. The imported cases were from 25 countries, including Asia ($n = 26$), Africa ($n = 28$), Europe ($n = 36$), North and South Americas ($n = 19$) (Stab 1). Most of the local cases had been in close contact with imported SARS-CoV-2 cases or related infections¹⁷. Additionally, we included 17 samples, collected between January 23 and February 22, from local cases who were confirmed to have contracted the virus from Hubei cases. All the 200 cases included in this study were classified for the severity of the disease according to accepted criteria. The majority (98%) were classified as mild, and only three cases were critical or severe (Stab 1).

Viral genome sequencing

We conducted multiplex SARS-CoV-2 specific amplification followed by next generation sequencing (NGS) to obtain viral genome sequences. We reported the results of a high coverage genome analysis (at least 10X sequence coverage for more than 90% of the genome nucleotides) for 73% (n = 146) of cases, each sequenced to a mean depth of > 12,000-fold (SD = 5,483). The rest of the cases (27%) were sequenced to a median genome coverage of 67% (Figure 1b and Stab 1). The genome coverage was correlated with viral loads, which were quantified by C_t values of real-time reverse transcription-polymerase chain reaction (qRT-PCR) assay¹⁸ (Figure 1c). Using genome sequence of the Wuhan-hu-1 strain as the reference, we found that the density of single nucleotide polymorphisms (SNPs) of the virus was ~0.2 nucleotide (nt) per 1,000 nts, which showed no significant correlation with C_t values (Fig 1c). C-to-U substitutions dominated the variations. In addition, 56.2% of the SNPs were amino acid-changing (non-synonymous, Figure S1).

Lineages of imported strains within the global phylogeny

To trace the foreign imported SARS-CoV-2 strains, we constructed a maximum likelihood (ML) phylogenetic tree of high-coverage genomes from 77 foreign imported strains, in combination with the data from 6,453 genomes shared by researchers worldwide (till Apr 20, 2020) (Figure 1e). These viral strains were widely distributed in the global SARS-Cov-2 phylogeny and were highly diverse. With the GISAID nomenclature, 49 of the imported cases could be identified as G clade (characterized by D614G in S protein, A23403G in genome), followed by 6 in S clade (L845S in ORF8, U28144C) and 4 in V clade (G251V in ORF3a, G26144U). The rest were assigned to other clades, including 17 from Asia and 1 from Europe. With another nomenclature suggested by ref.¹⁰, 3 of 77 imported strains (3.9%) could be classified as lineage A that shared nucleotides at position 8782 (U) and 28144 (C) with the closest known bat virus RaTG13⁶, and 71 of 77 were from lineage B with different nucleotides (CU). Hereafter we adopted the lineage A/B nomenclature.

Closed related strains of the imported SARS-CoV-2 could be identified and assigned to detailed lineages (SFig2-3 and Stab3), as numerous public genomes from European and North America are now available. 41 of 77 imported infections were assigned to two sub-lineages mainly sampled in Europe/North America (B.1) and Southeast/West Asia (B.6), in concordant with countries/regions where they were traveling from. A further analysis showed that the B.1 viral strains, imported from Europe (n = 18), North America (9) and South America (1), shared the SNPs C241U, C3037U, C14408U, and A23403G.

Our viral genomic data in the imported cases have broaden the phylogeny of the SARS-CoV-2, and expanded the global coverage in areas where the viral genomic surveillance data are currently limited. We found that many genomes were closely clustered within country, including the ones from Philippines (n = 11), Pakistan (5) and Thailand (2); whereas the three genomes from Dubai in United Arab Emirates diverged into two sub-clusters (Figure 2f). On the other hand, the strain from Ethiopia were heterogeneous, possibly due to the fact that Ethiopia has a major hub airport in Africa. In particular, five

of these Ethiopia strains could be split into four sub-clusters in the phylogenetic tree of Africa, compared to the fewer genetic branches from Nigeria strains (Figure 1g).

The phylogenetic analyses also offer a unique opportunity to investigate how the virus is transmitted among passengers on the same flight and among family members (Stab3). For instances, among the ten imported cases from five countries arriving at Guangzhou on the same airplane, the strains could be assigned to two haplotypes ($n = 5$ and $n = 4$) of B.1 lineage and one haplotype of A lineage ($n = 1$). A couple travelled together were infected by different viral strains; one of them lived in Ethiopia and the other in The Republic of Congo.

Phylogenetic analysis of imported and local cases

Next, we investigated the relationships between foreign imported and local spreading SARS-CoV-2. With a cutoff of $>90\%$ genome coverage, we included 77 imported and 52 local cases (March - April) in the ML phylogenetic analysis.

Despite the diversity of foreign viral strains, most of the local infections (50 of 52) were related to two specific lineages imported from Africa countries (Figure 2). The first lineage, denoted as L1, included 38 local cases and 5 imported cases from Nigeria or Ethiopia. L1 was characterized by C15324U (Asn5020Asn of ORF1b) and descended from the B.1 lineage. About 4% (260 of 6453) of the publicly available SARS-CoV-2 genomes share the same haplotype with L1 at these positions. These genomic data were reported in five continents (including Europe, Africa, Oceania, North and South America), but had not been previously identified in Asia (Stab 2).

The second lineage circulating in local cases was a descendant from L1 (denoted by L2) that harbored an additional C19524U (Leu6420Leu of ORF1b). Twelve local cases and three imported cases, from Uganda, Tanzania, and Ivory Coast, respectively, could be assigned to L2, as well as one case from Nigeria with a low genome coverage. Up to this writing, the L2 has not been reported in any public genome datasets.

Intriguingly, two local cases belonged to a new lineage characterized by G25563U (denoted as L3). Although many imported cases from Europe ($n = 7$), North America ($n = 7$), Africa ($n = 5$) also harbored the G25563U, they shared no other SNPs (such as C2416U or C1059U) with the two local cases (Figure 2). Thus, there was a lack of evidence that the L3 virus passed from international travelers to the local patients. This was consistent with their exposure and contact history (See Methods).

Notably, we found that four imported strains from Nigeria ($n = 2$), Ethiopia ($n = 1$), and Angola ($n = 1$) with G25563U were descended from the lineage with G25563U and C2416U. These strains need to be considered as novel ones in the current global phylogeny, as they contained a haplotype of two rare SNPs (C5654U and C16846U) (Figure 2). A careful examination of the global phylogeny suggested that this

haplotype is likely to be the result of a recombination between strains sampled in European (GISAID accession, EPI_ISL_428358 and EPI_ISL_420045) and Asian (EPI_ISL_420084) (SFig 4).

To explore the genetic characteristics of viruses among different waves of COVID-19 outbreak in China, we conducted a separate analysis to include 12 strains sampled (high genome coverage) from January to February 2020, seven of which were importation from Hubei province. Comparison of the phylogenetic information showed that the viral strains obtained in January were distinct from the imported strains identified in March and April (Figure 2).

The transmission in local community

To elucidate the spread of imported strains in local communities from March to April, we conducted a detailed analysis by including both viral genomic data and contact tracing information.

Besides the 51 local cases with high-coverage viral genome, 13 other local cases, in despite of the < 90% viral genome coverage, could be assigned to each of the L1 to L3 lineage based on characteristic variants.

Finally, 47 of 64 local infections were predominantly assigned to L1, followed by 13 to L2 and 4 to L3. It should be noted that the sample collection date for the first L3 case was prior to the ones for their corresponding imported cases, and it is likely that the genome sequencing data was not available for a few imported cases (Figure 3a-b).

38 of these 64 local infections were visitors to each of the three locations, including a restaurant, a tavern, and a trading market (Figure 3c). Before the onset of the disease, L1 local cases ($n = 5$) had a history of close contact with the L1 imported cases ($n = 3$) in the restaurant, and L1 imported cases (Trading market: $n = 2$, Tavern: $n = 1$) were visitors to all the above-mentioned locations. L3 local cases did not have a direct contact history with any of the imported cases, but they had previous contacts with other visitors to these locations.

One household and two close contacts of visitors to these locations were infected by L1 or L2 viral strains. However, a total of 20 local cases had neither history of visiting these locations, nor any contact with visitors in these areas. They were assigned to lineages L1 ($n = 13$) or L2 ($n = 7$), of which there is one L1 lineage household ($n = 3$) and one L2 lineage household ($n = 2$).

In sum, imported viral strains from all the three lineages circulated in local communities, but with distinguished scale. Of all the local cases, we did not identify any other SNPs that could be utilized to form new sub-lineage (Fig 2), suggesting that the circulation of imported SARS-CoV-2 in

local community might have been limited.

Genomic deletions and intra-host variations

Among the genomes that we studied, 10 harbored 12 events of viral in-frame deletions on SARS-CoV-2 genome, many of which were intra-host variations. As deep sequencing was error-prone for deletions and insertions, we filtered and selected deletions with a >30% mutated allele frequencies (MuAFs), and validated the results by Sanger sequencing (SFig 5a). These deletions were located at seven loci of the viral genome, with a length ranged from one to 14 amino acids (AAs, median 3). At the 82–86 AAs of the ORF1a protein, four cases, including importations from different countries, had 3–5 AA deletions. Other in-frame deletions located in the S, M and N proteins. In the S protein, one in-frame deletion P589del (MuAF 34.28%) was identified in a Europe-imported case. This deletion was not close to the junction site of S1 and S2 sub-units, and was predicted functionally conservative. No enrichment of these events was found in the S protein. We detected shared frame-shift deletions in the ORF8 for three cases, who were households (SFig 5b).

To further investigate the intra-host selection of SARS-CoV-2, we analyzed the intra-host single nucleotide variants (iSNVs) with a MuAF ranged from 5% to 95%. Given the noise in iSNV when viral loads are low, we included only the samples with a $<29 C_t$ value that exhibited no-significant correlation between iSNV density and C_t value (SFig 6 a-b). The results showed a strong purifying selection of the intra-host SARS-CoV-2. However, we found that iSNVs, whose loci were shared among individuals, were under a stronger purifying selection than those sporadic singleton iSNVs. The MuAFs of shared non-synonymous iSNVs were significantly lower than the synonymous shared iSNVs (two-side Wilcoxon rank-sum test, $P = 0.037$), as with non-synonymous singleton iSNVs ($P = 0.0163$, Figure 4c). The MuAF spectrum of shared non-synonymous exhibited the largest deviation from the expectation under neutral selection (Figure 4d-e).

We next examined the substitutions of iSNVs, and found that the distribution of the shared iSNVs was not concordant with that of SNPs (Figure 4f-g and SFig 1). Both the singleton iSNVs and SNPs had many G-to-U. Shared SNPs were dominated by C-to-U while shared iSNVs showed a G-to-U and A-to-G pattern. Interestingly, singleton iSNVs were dissonant with shared ones in terms of MuAFs distribution (Figure 4h-i). For the shared iSNVs, both C-to-U and U-to-C had significantly higher MuAFs than the other substitutions (two-side Wilcoxon rank-sum test, $P < 0.001$). These findings support a potential selection advantage of C-to-U during viral transmission.

Conclusion And Discussion

SARS-CoV-2 has spread rapidly in large scale to over 200 countries and territories. However, data are limited on how SARS-CoV-2 enters and spreads in a population and how it continues to mutate via international travel. As a metropolitan city with an international airport connecting around 200 cities in 50 different countries, Guangzhou is in a unique position to track the international transmission of SARS-CoV-2. As of February 2020, efforts to contain the internal spread of the virus had been very effective, with extremely low incidence of newly diagnosed cases in Guangzhou. In March, amid the global surge of

COVID–19, Guangzhou has reported a fresh wave of SARS-CoV–2 mostly linked to travelers who were infected from outside China. This was accompanied by newly clusters of local infections in April.

Genomic testing from infected travelers returning to Guangzhou suggests that the outbreak between March and April was not mainly caused by local chain of transmission, but instead by multiple new imported infections from over 20 countries, and these strains were genetically diverse in the global lineages. The circulation of import strains predominately originated from two lineages was largely responsible for the local transmission. The diversity of imported infections justified the need to step up measures for travelers arriving from abroad (including countries lacking viral genomic information) during the global COVID–19 pandemic, instead of a program that only applied to those coming from heavily affected countries.

The viral genome sequencing also provides insight into the phylogenetics of the pandemic in countries or regions lacking viral genomic information. For example, the strains from Ethiopia showed a high level of genetic diversity with various uncertain lineages (Fig 1g). It is unclear whether these lineages formed and circulated in Ethiopia or were imported from inbound travelers, suggesting a need for genomic surveillance in the country.

Our genomic surveillance in combination with epidemiological information show that the imported viruses were not spreading at high speed through the local society. More importantly, by the time (May 15) when we completed the genomic sequencing, no newly identified local infections were reported¹⁷. Such a rapid rise and fall of the local transmission may be due to the tighten quarantine measure in Guangzhou, where the control strategy was shifting away from self-quarantine to the one in which all arriving international passengers were required to stay in government-designated facilities for 14 days¹⁹. This measure has proved effective and can be applied to COVID–19 outbreaks in other cities.

Our work also provides new information on the evolution of SARS-CoV–2. It is widely accepted that the SARS-CoV–2 harbors a four-amino-acid insertion at the junction of S1 and S2. Little is known about the history and function of this insertion, which appears to occur naturally in animal betacoronaviruses²⁰. Zhou et al also identified a bat-derived coronavirus RmYN02 harboring a similar insertion with SARS-CoV–2 at the junction site⁶. In addition to this, we identified novel deletion events in the intra-host viral population, but there was no apparent enrichment in the S protein. Liu et al have also reported that common deletions on the junction and flank sites emerged through cell passage of SARS-CoV–2²¹. Furthermore, we documented the intra-host dynamics of MuAFs, presumably due to a larger selection pressure for the shared iSNVs. The selective advantage might also contribute to the predominant C-to-U substitution in the lineage characteristic variants. However, as with all phylogenetic analyses, our findings should be interpreted with caution, as the numbers of genetic changes or mutations were small and thus the analyses were statistically underpowered. This highlights a need for global alliance to foster a pooled phylogenetic analysis of the SARS-CoV–2 viruses.

Methods

Ethics

This study was approved by the ethics committee of the Center for Disease Control and Prevention (CDC) of Guangzhou (GZCDC-ECHR-2020P0002). Written informed consent was obtained from patients about the surveillance and data related to disease control and further analysis. All information regarding individual persons has been anonymized in this study.

Surveillance of imported COVID-19 cases in Guangzhou

In response to the COVID-19 outbreak outside China, Guangzhou has strengthened and adjusted the screening strategy on inbound passengers. From February 27 to March 20, any passengers came or transferred from high-risk regions were required to undergo quarantine at 3 designated places and receive nucleic acid test for SARS-CoV-2. Other passengers were required to self-quarantine. However, starting from 0 am March 21, 2020, Guangzhou has tightened the controls. Any passengers entering into Guangzhou from Hong Kong, Macau, Taiwan and overseas via ports located in Guangzhou were required to quarantine in designated facilities in Guangzhou for medical inspection for 14 days and receive tests for SARS-CoV-2¹⁹. Individuals with fever or respiratory symptoms were delivered to 11 designated hospitals in Guangzhou for further confirmation.

Samples were collected from the quarantine facilities or hospitals. Those tested positive for SARS-CoV-2 by laboratories in hospitals and third-party institutions were subsequently sent to central laboratory of CDC of Guangzhou for double-check and final confirmation. Between January 3 and April 15, samples from 200 confirmed cases were included in this study. These cases were classified based on severity as mild, moderate, severe, or critical cases according to the Diagnosis and Treatment Protocol for Novel Coronavirus Pneumonia (the 7th edition) issued by the National Health Commission of China²².

Epidemiological investigation of imported and local cases

Epidemiological investigation has been conducted by the municipal CDC of Guangzhou and 11 districts in Guangzhou. The data included patient's general personal information, care-seeking behavior, residence, travel history, potential exposures, and source of infection¹².

Epidemiological data on both foreign imported and local cases have clearly indicated a small outbreak of SARS-CoV-2 infections in local communities, due to inbound passengers from non-high-risk regions before March 27.

Laboratory nucleic acid test

Nucleic acid testing for SARS-CoV-2 was performed using real-time reverse transcription-PCR (qRT-PCR) assays recommended by the Chinese Center for Disease Control and Prevention following the guidelines of World Health Organization¹². Briefly, total nucleic acids were isolated from 200 µL viral transport media containing oropharyngeal swabs specimen through magnetic bead-based viral RNA nucleic acid extraction system (BioPerfectus Technologies, Jiansu, China). The qRT-PCR was performed with 2019-nCoV detection kit (Cat No. DA0931, Daan Gene, Guangzhou, China) according to the manufacturer's instructions on ABI QDx real-time PCR platform (Thermo Fisher Scientific).

Amplicon-based SARS-CoV-2 virus sequencing

The remnant nucleic acid samples for qRT-PCR assay were used for viral genome sequencing through a multiplex PCR approach, similar to a protocol reported previously¹⁸. If multiple samples were available for the same individual, only the sample with the highest viral load (the lowest qRT-PCR C_t value) was used. Cases were excluded, if samples were tested negative by qRT-PCR. First strand cDNA was reverse-transcribed from 13 µL of RNA using NEBNext Ultra II RNA First Strand Synthesis Module (New England Biolabs) before SARS-CoV-2 specific amplification.

Two primer sets targeting SARS-CoV-2 were used in this study. The first was the version 3 of the ARTICCOVID-19 multiplex PCR primers (https://github.com/artic-network/artic-ncov2019/tree/master/primer_schemes/nCoV-2019/V3), which has 98 pair of primers and an average amplicon length of ~400 bp (Stab 6). To complement the sequencing depth bias among the ARTIC primers, another 36 pairs of primer with a ~1k bp amplicon length (Stab 7) were designed using the Primal Scheme website (primal.zibreproject.org). The implementation of the two sets of pairs was listed in Stab 6–7. Multiplex PCR amplification for both primer sets was performed with NEB Q5 DNA High-fidelity Polymerase (New England Biolabs). Regimen of thermal cycling for the ARTIC primer set followed the ARTIC protocol: 30s at 98°C; 35 cycles of 15s at 98°C and 5 min at 65°C; then hold at 4°C.

The PCR products were pooled and cleaned using GeneJET Gel Extraction Kit (Thermo Fisher Scientific) according to the manufacturer's instructions. The products of the ARTIC primers, except seven samples (Stab 6), were prepared with the multiplex NGS library, without fragmentation, using TruSeq DNA PCR-Free High Throughput Library Prep Kit (Illumina). The Illumina NovaSeq 6000 platform was used to generate 2 x 250 bp paired-end reads. Seven samples were fragmented into a desired size of 250 bp and their libraries were prepared using NEBNext Ultra II DNA Library Prep Kit for Illumina (New England Biolabs). Sequencing was performed on Illumina MiSeq platform (Illumina) to generate 2 x 150 bp paired-end reads.

The 1k bp PCR products were fragmented, and multiplex NGS library was prepared with 5x WGS Fragmentation Mix (Enzymatics, #Y9410), 10X Fragmentation Buffer (Enzymatics, #B0330), High Concentration T4 DNA Ligase (Enzymatics, #L6030-600000) and 5X Rapid Ligation Buffer (Enzymatics,

#B9020). The NovalSeq 6000 sequencing system was then used to generated 2 x 150 bp paired-end reads.

Assembly and variants calling

A reference-based assembly of the NGS sequencing reads was conducted according to ARTIC bioinformatics pipeline

(https://github.com/CDCgov/SARS-CoV-2_Sequencing/tree/master/protocols/BFX-UT_ARTIC_Illumina).

Briefly, raw NGS reads were aligned to the genome of SARS-CoV-2 Wuhan-hu-1⁵ (GenBank accession MN908947.3, GISAID accession EPI_ISL_402125) by using BWA v0.7.17²³. The SARS-CoV-2 specific PCR primer regions of reads were removed based on the alignment results (BAM formatted files) by using 'trim' of iVar v1.0²⁴. Mpileup files were generated by using the 'mpileup' program of SAMtools v1.9²⁵ with parameters of "-A -d 20000 -B -Q 0". Consensus sequences were then generated by using 'consensus' program of iVar with the default parameter.

The single nucleotide variations (SNVs) and their mutated allele frequencies (MuAFs) were identified as described in ref²⁶. Based on the alignment by using BWA v0.7.17²³, SAMtools v1.9²⁵ was used to generate 'mpileup' files with a 20000 limit for the maximum site depth. The bioinformatics workflow for SNV calling is available at <http://github.com/generality/iSNV-calling>, which uses mpileup files as input. A Q20 quality filtering was first conducted for the sequencing bases. The MuAFs of SNVs were obtained for sites with a ≥ 100 -fold depth. Intra-host SNVs were defined as those with a 0.05–0.95 MuAF and ≥ 5 support reads for the substitutions. Mutations of small insertion and deletion (delins) were identified by using VarScan2 v2.4.4²⁷, which uses mpileup files as input. The parameters for delin calling were as following: "--min-coverage 50--min-reads 25--min-var-freq 0.3". The visualization tools IGV v2.4.17²⁸ was used for manual examination and filtering delins. Both the SNVs and small delins were annotated by using SnpEff v4.3t²⁹ with a reference SARS-CoV-2 Wuhan-hu-1 with default parameters.

Public datasets of SARS-CoV-2 genomes

All available genome sequences of SARS-CoV2 were downloaded from GISAID¹⁸ (n = 10867, gisaid.org), NCBI GenBank (n = 1207, www.ncbi.nlm.nih.gov/genbank/sars-cov-2-seqs/) and BIGD³⁰ (n = 1182, bigd.big.ac.cn/ncov) on April 23, 2020. The sequences were filtered to remove duplicated individuals, incomplete genome sequences (<29000bp) and those with low quality ($\geq 0.1\%$ N and $\geq 0.05\%$ unique amino acid mutation). Finally, we obtained a total of 6453 nCoV-2019 sequences (Stab 2).

Phylogenetic analysis

Multiple alignments of genome sequences were performed by using MAFFT v7.458³¹ and manually inspected by using MEGA v10.1.8³². Given the bias of genome coverage of public genome and sequences in this study, part of the 5' and 3' untranslated region was removed and the aligned genome length was 29697 nucleotides. We explored the phylogenetic structure with maximum likelihood (ML) method. ML Phylogenies of large alignment (>6000 genomes) were inferred by using IQ-Tree2³³ (rc2) with the best-fitting substitution model parameters (GTR+F+R2) estimated by Model Finder and 1000 rapid bootstrapping replicates. Phylogenetic analyses of <200 viral genomes were performed by using RAxML v8.2.12³⁴ with 1000 bootstrap replicates and employing the GTRGAMMA+I model. The generated phylogenetic trees were visualized with the R package ggtree v1.14.6³⁵.

Sanger sequencing

PCR primers were designed by using Primer3-0.4.0³⁶ (<http://bioinfo.ut.ee/primer3-0.4.0/>) for a validation of the identified SNVs and delins of interest. As the remaining samples were highly degraded when we performed Sanger sequencing, the products of SARS-CoV-2 specific amplification were utilized for Sanger sequencing validation. The primer sequences and the mutations for validation are listed in Stab 8. Template DNA was amplified using 2×Taq PCR Mix (TIANGEN, #KT201) on T100 Thermal Cycler (BIO-RAD, US). PCR conditions was set as follows: 3 min at 94°C; then for 30 cycles of 30 s at 94°C, 30 s at 53°C, and 1 min at 72°C; followed by 5 min at 72°C. PCR products were sent for Sanger sequencing (Tianyi Huiyuan Life Science and Technology Company, Beijing, China).

Declarations

Acknowledgements

This study was supported by National Natural Science Foundation of China (31870079, 91953122, 31871326), National Science and Technology Major Project of the Ministry of Science and Technology of China (2017ZX10103011, 2018ZX10305410, 2018ZX10201001), Guangdong Provincial Novel Coronavirus Scientific and Technological Project (2020111107001), Guangdong Basic and Applied Basic Research Foundation (2020A1515010776 and 2020B1515020057) and the Beijing Nova Program (Z181100006218114 and Z181100006218110) to M.N. and P.L..

Contributions

C.L.X., M.N., Z.B.Z., B.D. and J.F. supervised the overall project design and execution. M.N., H.J.L. and J.Y.Z. participated in data analysis. Z.B.Z., B.D., H.L.J., K.B.L., P.Z.Q., L.C., J.B.W. and Y.Y.C., X.P.T. collected the samples and clinical information. H.L.J., K.B.L. and P.Z.Q. performed the virus detection, P.L. J.H.L., L.Z.L., L.L.J., H.B.S. and Q.Z. performed the virus sequencing and the Sanger sequencing validation. M.N.,

H.J.L., J.Y.Z. and C.L.X. wrote the manuscript and created diagrams. Y.F.Z., N.M., P.L. and J.F. revised the paper.

Data availability

The SARS-CoV-2 sequences reported in this paper have been deposited in the China National Microbiology Data Center (<http://www.nmdc.cn/>) with accession numbers from NMDC60013143 to NMDC60013288. Code for all figures, tree files, haplotype files and raw data for Figures 1, 2, 3, and 4 are available at https://github.com/zhengdafangyuan/Genomic_surveillance_of_imported_CoVID-19_Guangzhou.

References

1. *Coronavirus disease 2019*, <<https://www.who.int/emergencies/diseases/novel-coronavirus-2019>> (2020).
2. Wu, *et al.* A new coronavirus associated with human respiratory disease in China. *Nature* **579**, 265-269, doi:10.1038/s41586-020-2008-3 (2020).
3. Shu, & McCauley, J. GISAID: Global initiative on sharing all influenza data - from vision to reality. *Euro Surveill* **22**, 30494, doi:10.2807/1560-7917.ES.2017.22.13.30494 (2017).
4. Lam, T.-Y. *et al.* Identifying SARS-CoV-2 related coronaviruses in Malayan pangolins. *Nature*, doi:10.1038/s41586-020-2169-0 (2020).
5. Xiao, K. *et al.* Isolation of SARS-CoV-2-related coronavirus from Malayan pangolins. *Nature*, doi:10.1038/s41586-020-2313-x (2020).
6. Zhou, H. *et al.* A Novel Bat Coronavirus Closely Related to SARS-CoV-2 Contains Natural Insertions at the S1/S2 Cleavage Site of the Spike Protein. *Current Biology*, doi:<https://doi.org/10.1016/j.cub.2020.05.023> (2020).
7. Zhou, *et al.* A pneumonia outbreak associated with a new coronavirus of probable bat origin. *Nature* **579**, 270-273, doi:10.1038/s41586-020-2012-7 (2020).
8. Hadfield, J. *et al.* Nextstrain: real-time tracking of pathogen evolution. *Bioinformatics* **34**, 4121-4123, doi:10.1093/bioinformatics/bty407 (2018).
9. Bedford, *et al.* Cryptic transmission of SARS-CoV-2 in Washington State. *medRxiv*, 2020.2004.2002.20051417, doi:10.1101/2020.04.02.20051417 (2020).
10. Rambaut, A. *et al.* A dynamic nomenclature proposal for SARS-CoV-2 to assist genomic epidemiology. *bioRxiv*, 2020.2004.2017.046086, doi:10.1101/2020.04.17.046086 (2020).
11. Eden, -S. *et al.* An emergent clade of SARS-CoV-2 linked to returned travellers from Iran. *Virus evolution* **6**, veaa027-veaa027, doi:10.1093/ve/veaa027 (2020).
12. Lu, J. *et al.* Genomic Epidemiology of SARS-CoV-2 in Guangdong Province, China. *Cell*, S0092-8674(0020)30486-30484, doi:10.1016/j.cell.2020.04.023 (2020).

13. Gámbaro, *et al.* Introductions and early spread of SARS-CoV-2 in France. *bioRxiv*, 2020.2004.2024.059576, doi:10.1101/2020.04.24.059576 (2020).
14. Tang, *et al.* On the origin and continuing evolution of SARS-CoV-2. *National Science Review*, doi:10.1093/nsr/nwaa036 (2020).
15. Oude Munnink, B. B. *et al.* Rapid SARS-CoV-2 whole genome sequencing for informed public health decision making in the Netherlands. *bioRxiv*, 2020.2004.2021.050633, doi:10.1101/2020.04.21.050633 (2020).
16. Tian, *et al.* An investigation of transmission control measures during the first 50 days of the COVID-19 epidemic in China. *Science* **368**, 638-642, doi:10.1126/science.abb6105 (2020).
17. Municipality, P. s. G. o. G. *Notice of COVID-19 in Guangzhou city*, <<http://www.gz.gov.cn/zt/qlyfdyyqfkyz/qktb/>> (2020).
18. Quick, J. *et al.* Multiplex PCR method for MinION and Illumina sequencing of Zika and other virus genomes directly from clinical samples. *Nat Protoc* **12**, 1261-1276, doi:10.1038/nprot.2017.066 (2017).
19. Municipality, P. s. G. o. G. *Notice for travelers entering Guangdong from abroad*, <http://www.gz.gov.cn/guangzhouinternational/home/bulletin/content/post_5739204.html> (2020).
20. Andersen, K. G., Rambaut, A., Lipkin, I., Holmes, E. C. & Garry, R. F. The proximal origin of SARS-CoV-2. *Nature Medicine* **26**, 450-452, doi:10.1038/s41591-020-0820-9 (2020).
21. Liu, Z. *et al.* Identification of a common deletion in the spike protein of SARS-CoV-2. *bioRxiv*, 2020.2003.2031.015941, doi:10.1101/2020.03.31.015941 (2020).
22. Municipality, P. s. G. o. G. *Notification on Further Strengthening COVID-19 Prevention and Control Work, Guangzhou Command Center for COVID-19 Control and Prevention (No.7)*, <http://www.gz.gov.cn/guangzhouinternational/home/bulletin/content/post_5742313.html> (2020).
23. Heng, Li, Richard & Durbin. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* (2009).
24. Grubaugh, N. *et al.* An amplicon-based sequencing framework for accurately measuring intrahost virus diversity using PrimalSeq and iVar. *Genome Biol* **20**, 8-8, doi:10.1186/s13059-018-1618-7 (2019).
25. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078-2079, doi:10.1093/bioinformatics/btp352 (2009).
26. Ni, M. *et al.* Intra-host dynamics of Ebola virus during 2014. *Nature Microbiology* **1**, 16151 (2016).
27. Koboldt, C. *et al.* VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Res* **22**, 568-576, doi:10.1101/gr.129684.111 (2012).
28. Thorvaldsdóttir, H., Robinson, J. & Mesirov, J. P. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Brief Bioinform* **14**, 178-192, doi:10.1093/bib/bbs017

- (2013).
29. Cingolani, *et al.* A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain w1118; iso-2; iso-3. *Fly (Austin)* **6**, 80-92, doi:10.4161/fly.19695 (2012).
 30. Zhao, M. *et al.* The 2019 novel coronavirus resource. *Yi chuan = Hereditas* **42**, 212-221, doi:10.16288/j.yczzz.20-030 (2020).
 31. Katoh, K., Misawa, K., Kuma, K.-i. & Miyata, M. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res* **30**, 3059-3066, doi:10.1093/nar/gkf436 (2002).
 32. Sudhir, , Glen, S., Li, M., Christina, K. & Koichiro, T. MEGA X: Molecular Evolutionary Genetics Analysis across Computing Platforms. *Molecular Biology & Evolution*, 6 (2018).
 33. Minh, B. Q. *et al.* IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. *Molecular biology and evolution* **37**, 1530-1534, doi:10.1093/molbev/msaa015 (2020).
 34. Stamatakis, A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**, 1312-1313, doi:10.1093/bioinformatics/btu033 (2014).
 35. Yu, , Smith, D. K., Zhu, H., Guan, Y. & Lam, T. Y. ggtree: an r package for visualization and annotation of phylogenetic trees with their covariates and other associated data. *Methods in Ecology & Evolution* (2017).
 36. Untergasser, *et al.* Primer3–new capabilities and interfaces. *Nucleic Acids Res* **40**, e115-e115, doi:10.1093/nar/gks596 (2012).

Figures

Figure 1

The sequencing of SRARS-CoV-2 and phylogenetic analysis. a, The daily new cases of total COVID-19 cases and foreign imported cases through late of Jan 2020 to middle of Apr 2020. b, The depth profile by amplicon-sequencing over SARS-CoV-2 genome. The maximum depth was limited to 20,000X. Bold black line denotes median depth, and grey shadow indicates the 25th and 75th quantiles of depth (n=200). c-d, The genome coverage (>10X, c) and SNP density per kilo-nucleotide (d) along with the C_t values of qRT-PCR test for SARS-CoV-2 (n=200). e, the maximum likelihood phylogenetic tree with 1000 bootstrap replicates, based on foreign importation to Guangzhou (n=77, imported cases with high-coverage genome) and public resources. The branches of tree are colored by continent, and the main lineages the imported strain belong to are indicated by arrow. f-g, The phylogenetic tree of Asia (n=22, f) and Africa

(n=20, g). To illustrate the B.6 lineage in Asia, the data from China, Japan, and South Korea are excluded due to limited space. Bar-plots along the corresponding branches of tree provide the counts of imported cases (solid bar) in this study and cases previously published (hollow bar) by countries.

Figure 2

The relationship between foreign imported and local case. A RaxML phylogenetic tree at the left panel presents imported (n=77), local cases (March-April n=52, Jan-Feb n=12) and NA cases (information not available, n=5) with high genome coverage. The continents where foreign imported cases were from

present side correspondingly. Distributions of SNPs in these cases are illustrated in the right panel, aligned with the tree branches, with frequent variants indicated. The cases related to local transmission are highlighted in light yellow (with C15324U) or blue (with both 15324U and C19525U). In the bottom, the counts of SNPs at loci along the viral genome are shown.

Figure 3

The local transmission. a-b, Identification of foreign imported (n=109, a) and local cases (March-April n=69, b) harboring variants of specific lineages (see text) versus sampling date. Note that the sampling

date is ahead of the confirmed date (Figure 1a). Light grey bars in a indicate cases without the given variants or failed in sequencing. Dark grey bars in c mean the failure of genotyping the characteristic variants. NA, unavailable. c, The network of transmission. The counts of infected cases belonging to L1, L2, and L3 (see text) are indicated at the right side of the clusters. The numbers with star prefix indicate the case(s) with a history of visiting two or all the three locations. Arrow lines are drawn based on the reports of epidemiology investigations, and dashed arrow lines denote the imported cases visited the place but did not contact with the local infected cases. The distance between the restaurant and the trading market is 300 meters. The distance between the restaurant and the tavern is 4.2 kilometers. The other local case assigned to L3 (ID070) is not shown as she/he did not visit the locations or live in the communities. She/he contracted the virus from a Europe-imported case (ID014).

Figure 4

The intra-host genomics deletions and substitutions. a, Distributions of in-frame deletions along the genomes. The loci were numbered and indicated by dashed arrow lines. b, the in-frame deletions are displayed as numbers in a. The strains from cases was aligned to the Wuhan-hu-1 strain, with genes and genome positions. The identities of foreign imported cases are displayed as brown, and cases related this cluster infection are indicated by shadow. The mutated allele frequencies (MuAF) of each deletion are

shown. c, Box and scatter plots of the MuAFs of non-synonymous (nonsyn) and synonymous (syn) iSNVs. The singleton and shared iSNVs are demonstrated separately, with the P values of two-side Wilcoxon rank-sum test. Star indicates a significant P value<0.05. d-e, The MuAF spectra of non-synonymous and synonymous iSNVs for the singletons (d) and shared (e). Expectation under neutral selection (neutral exp.) is shown for comparison. f-g, The counts of singleton iSNVs (f) and shared iSNVs (g) for each substitution. h-i, The distribution of MuAFs of singleton (h) and shared iSNVs (i) in each substitution. Red arrows highlight the substitution with a relatively higher MuAFs. For box-plot, the boxes represent the interquartile range between the first and third quartiles. Horizontal blond lines inside the boxes indicate median values, and the lines outside represent values within 1.5 times the IRQ.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [SupplementaryFigures.pdf](#)