# Genomic characterization of SARS-CoV-2 identified in a reemerging COVID-19 outbreak in Beijing's Xinfadi market in 2020

Yong Zhang [a,1], Yang Pan [b,1], Xiang Zhao [a], Weifeng Shi [c,d], Zhixiao Chen [a], Sheng Zhang [a], Peipei Liu [a], Jinbo Xiao [a], Wenjie Tan [a], Dayan Wang [a], William J. Liu [a], Wenbo Xu [a,*], Quanyi Wang [b,*], Guizhen Wu [a,*]

[a] NHC Key Laboratory of Biosafety, National Institute for Viral Disease Control and Prevention, Chinese Center for Disease Control and Prevention, Beijing 102206, China
[b] Beijing Center for Disease Control and Prevention, Beijing 100013, China
[c] Key Laboratory of Etiology and Epidemiology of Emerging Infectious Diseases in Universities of Shandong, Shandong First Medical University and Shandong Academy of Medical Sciences, Tai'an 271000, China
[d] The First Affiliated Hospital of Shandong First Medical University (Shandong Provincial Qianfoshan Hospital), Jinan 271099, China

## ARTICLE INFO

## ABSTRACT

After 56 days without coronavirus disease 2019 (COVID-19) cases, reemergent cases were reported in Beijing, China on June 11, 2020. Here, we report the genetic characteristics of severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) sequenced from the clinical specimens of 4 human cases and 2 environmental samples. The nucleotide similarity among six SARS-CoV-2 genomes ranged from 99.98% to 99.99%. Compared with the reference strain of SARS-CoV-2 (GenBank No. NC_045512), all six genome sequences shared the same substitutions at nt241 (C → T), nt3037 (C → T), nt14408 (C → T), nt23403 (A → G), nt28881 (G → A), nt28882 (G → A), and nt28883 (G → C), which are the characteristic nucleotide substitutions of L-lineage European branch I. This was also proved by the maximum likelihood phylogenetic tree based on the full-length genome of SARS-CoV-2. They also have a unique shared nucleotide substitution, nt6026 (C → T), which is the characteristic nucleotide substitution of SARS-CoV-2 in Beijing's Xinfadi outbreak. It is noteworthy that there is an amino acid D614G mutation caused by nt23403 substitution in all six genomes, which may enhance the virus's infectivity in humans and help it become the leading strain of the virus to spread around the world today. It is necessary to continuously monitor the genetic variation of SARS-CoV-2, focusing on the influence of key mutation sites of SARS-CoV-2 on viral transmission, clinical manifestations, severity, and course of disease.

After 56 days without COVID-19 cases, reemergent cases were reported in Beijing, China on June 11, 2020. From June 11 to July 4, 334 COVID-19 cases were reported in Beijing. Of the patients affected, 47% were Xinfadi market staff, while others were people who have had contact with the market [1].

Throat swab samples were taken from the patients, and environmental swab samples were taken from the environment in Xinfadi market, including the body surface of seafood, the channel at the gate of the stall, trash, outer surface of fish tanks, inner surface of refrigerators, etc. Total RNA was extracted from the supernatant using the Viral RNA Extraction Kit (Tianlong, Xi'an, China). The rRNA was removed using the TransNGS rRNA Depletion (Human/Mouse/Rat) kit (TransGen, Beijing, China), while the remaining RNA was incorporated for library preparation with the Illumina TruSeq DNA Preparation Protocol and sequenced on the Illumina NextSeq 550 platform (Illumina, San Diego, CA, USA) with 150 bp paired-end reads. The full-length genome sequences were assembled using QIAGEN CLC Genomics Workbench (Qiagen, Hilden, Germany). Three full-le5ngth genome sequences of SARS-CoV-2 from the Xinfadi market were deposited at the National Genomics Data Center (two from patients and one from environmental sample) and three were deposited at GISAID (two from patients and one from environmental sample).

At present, the widely accepted molecular typing method of SARS-CoV-2 is based on the difference of nucleotide at nt8782 and nt28144 in its full-length genome sequence. According to this molecular typing method, SARS-CoV-2 is divided into S-lineage (nt8782T and nt28144C) and L-lineage (nt8782C and nt28144T). It is speculated that S-lineage may be an older type. In Wuhan City, Hubei Province, S-lineage was the

* Corresponding authors: NHC Key Laboratory of Biosafety, National Institute for Viral Disease Control and Prevention, Chinese Center for Disease Control and Prevention, Beijing 102206, China.
*E-mail addresses:* xuwb@ivdc.chinacdc.cn (Wenbo Xu), bjcdcxm@126.com (Quanyi Wang), wugz@ivdc.chinacdc.cn (Guizhen Wu).
[1] Yong Zhang and Yang Pan contributed equally to this manuscript.

main epidemic type in the early stages of the outbreak, while L-lineage is currently the most prevalent lineage in the world. Therefore, most of the imported cases in China belong to L-lineage. Due to the global pandemic

of SARS-CoV-2, imported cases have been frequently reported in China since March 2020. In order to better identify the SARS-CoV-2 types prevalent in various countries, according to the full-length genome sequence of
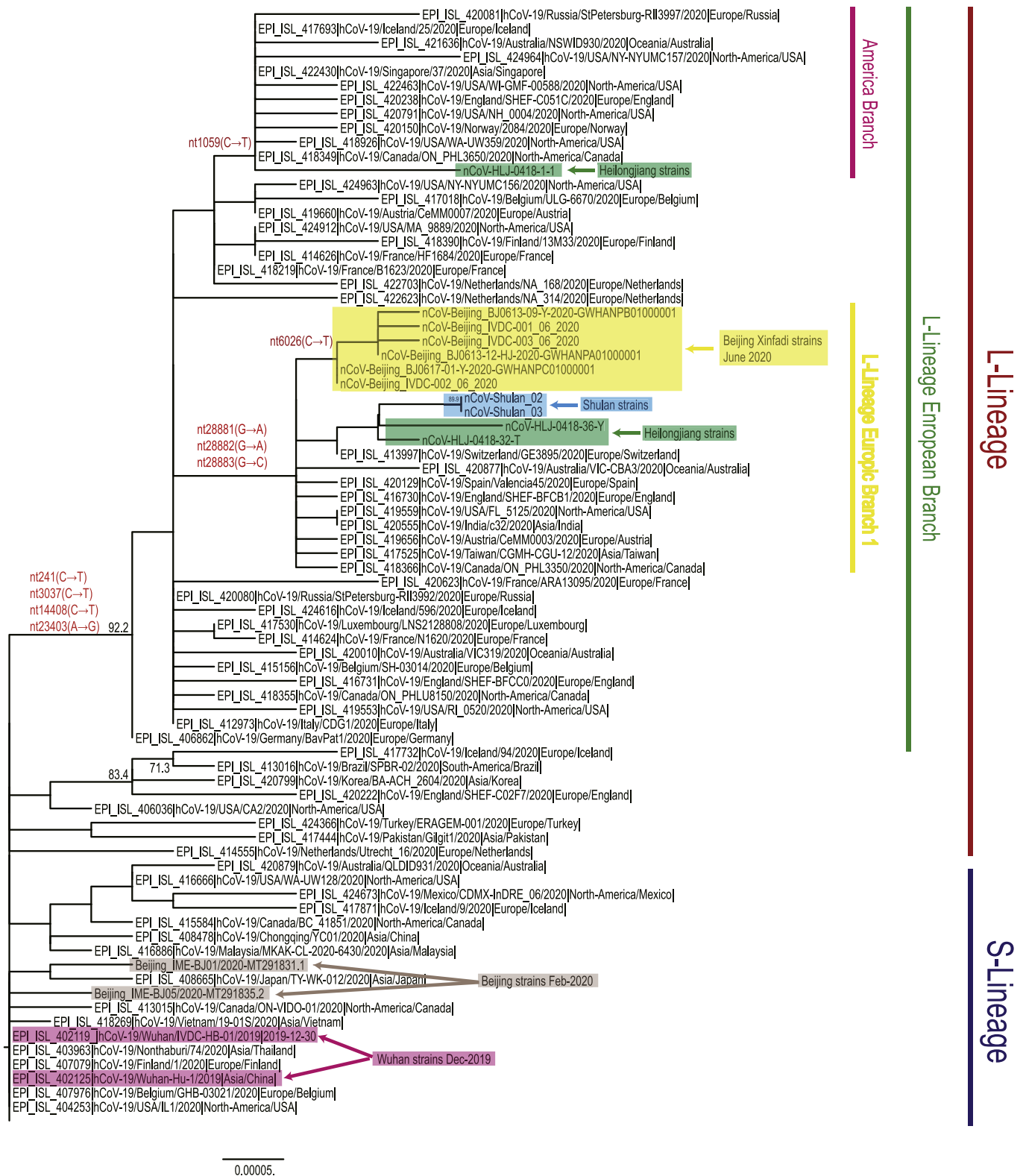


**Fig. 1.** Maximum likelihood phylogenetic tree based on the full-length genome sequences of the SARS-CoV-2. The tree was constructed using MEGA (v7.0) software with 1,000 bootstrap replicates. The genomes of the SARS-CoV-2 from Beijing's Xinfadi market are highlighted in shades of yellow. The genomes of the SARS-CoV-2 from Beijing (Feb-2020) and Wuhan (Dec-2019) are highlighted in shades of grey and pink, respectively. The recent reemergence of SARS-CoV-2 in Northeastern China (Shulan and Heilongjiang) that was associated with imported cases are highlighted in shades of blue and green, respectively. S- or L-lineage of the SARS-CoV-2 are marked and colored on the right.

the virus detected in imported cases in China, L-lineage were further divided into L-lineage European branch where the characteristic substitution sites were nt241(C → T), nt3037(C → T), nt14408(C → T), and nt23403 (A → G) and L-lineage global branch with substitution sites other than those mentioned above. L-lineage European branch can be further divided into L-lineage European branch I (the characteristic substitution sites were nt28881(G → A), nt28882(G → A), and nt28883(G → C)) and L-lineage European branch II (the characteristic substitution site was nt25563 (G → T)). The L-lineage European branch II can be further divided into L-lineage European branch II.1 (the characteristic substitution site was nt1059(C → T)) and L-lineage European branch II.2 (the characteristic substitution site was nt2416(C → T)). Because L-lineage European branch II.1 is a prevalent branch in America, it is also called the American branch.

The nucleotide similarity among six SARS-CoV-2 genomes ranged from 99.98% to 99.99%. Compared with the reference strain of SARS-CoV-2 (GenBank No. NC_045512), the nucleotide similarity was 99.96%–99.97%. A maximum likelihood tree based on the full-length genome sequences of SARS-CoV-2 was constructed using MEGA software (v7.0) [2] with 1,000 bootstrap replicates. According to this phylogenetic tree, all the six genome sequences belong to L-lineage European branch I (Fig. 1). SARS-CoV-2 in L-lineage European branch I are prevalent in Asia, South America, Africa, and North America, but mainly in Europe. This further confirmed that the outbreak in Xinfadi may be caused by the introduction of an infectious source. It is different from the virus that was prevalent in Wuhan in December 2019 and Beijing in February 2020, which belonged to S-lineage; therefore, it is suggested that the outbreak in Xinfadi did not involve the continued transmission of native virus.

Compared with the reference strain of SARS-CoV-2 (GenBank No. NC_045512), all six genome sequences shared the same substitutions at nt241 (C → T), nt3037(C → T), nt14408(C → T), nt23403(A → G), nt28881 (G → A), nt28882(G → A), and nt28883(G → C), among which nt14408, nt23403, and nt28881–nt28883 were nonsynonymous substitutions in ORF 1ab gene, S gene, and N gene, respectively. All the seven nucleotide substitutions are the characteristic nucleotide substitutions of L-lineage European branch I (Fig. 2). All six genome sequences also have a unique shared nucleotide substitution, nt6026(C → T), which is the characteristic nucleotide substitution of SARS-CoV-2 from Xinfadi market (Fig. 2).

It is worth noting that there is an amino acid D614G mutation caused by nt23403 substitution in all six genomes of SARS-CoV-2 found in Xinfadi market. The mutation from 614-D to 614-G occurred first in Europe, then in North America and Oceania, and then in Asia, and now this mutation is shared by all viral genomes in L-lineage European branch I.

In a recent study, researchers found that a D614G mutation in the SARS-CoV-2 genome enhances the virus's ability to infect human cells, helping it to become the leading strain of the virus spreading around the world today [3]. The research team also proposed that D614G mutation pseudovirus was associated with higher infectivity. Quantitative analysis showed that viral particles carrying 614-G mutation had significantly higher infectivity titer than the corresponding 614-D mutation, which increased by 2.6 to 9.3 times, and was confirmed in a variety of cell types [3]. However, there is no evidence that it will lead to a more serious condition [4] because there is a big difference between the laboratory results and the changes in the actual transmission of the virus, which scientists have monitored.

The complete genome sequence analysis of the SARS-CoV-2 from Xinfadi further confirmed that the source of the epidemic was not a new "overflow" from the host or intermediate host. The Xinfadi market has become the SARS-CoV-2 transmission "hub", the most important reason being that the market environment is relatively wet and cold. One of the main characteristics of the virus is that it is sensitive to heat and not to cold. In such an environment, the virus can survive for a long time. Moreover, it is also closed and poorly ventilated, which also contributes to the spread of the virus. For example, if a person infected with SARS-CoV-2 sneezes, it is difficult to spread out, and the droplets may settle on the ground and pollute other places after flushing and sweeping. These conditions help to explain the occurrence of a large number of cases in a short period of time. However, this does not mean that the market itself is the source of the virus.

The prevention and control of COVID-19 epidemic in big cities should be based on the principle of precise prevention and control. For example, areas with new confirmed cases can be classified as medium-risk areas. However, it is not necessary to freeze all the communities in medium-risk areas. Precise prevention and control involves the freezing of only the communities that have confirmed cases. Beijing's rapid, effective, and accurate epidemic prevention and control measures will draw on experience from other places and the future normalization of epidemic prevention.

At present, with the rapid development of genomics technology, bioinformatics, and big data management technology, the new generation of genomic sequencing technology can achieve more detailed analysis of



| | 241 | 3237 | 6026 | 8782 | 11910 | 12085 | 14408 | 23282 | 23403 | 28144 | 28881 | 28882 | 28883 | 29694 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| NC_045512 | C | C | C | C | A | C | C | G | A | T | G | G | G | A |
| BJ-IVDC-001 | T | T | T | • | • | T | T | • | G | • | A | A | C | C |
| BJ-IVDC-001 | T | T | T | • | • | • | T | • | G | • | A | A | C | • |
| BJ-IVDC-003 | T | T | T | • | G | • | T | • | G | • | A | A | C | C |
| BJ-0613-12-HJ | T | T | T | • | • | • | T | • | G | • | A | A | C | C |
| BJ-0613-09-Y | T | T | T | • | • | • | T | T | G | • | A | A | C | C |
| BJ-0617-01-Y | T | T | T | • | • | • | T | • | G | • | A | A | C | • |

Synonymous mutation — P1291S ORF 1ab — P4715L ORF 1ab — D614G S protein — R203K N protein — G204R N protein
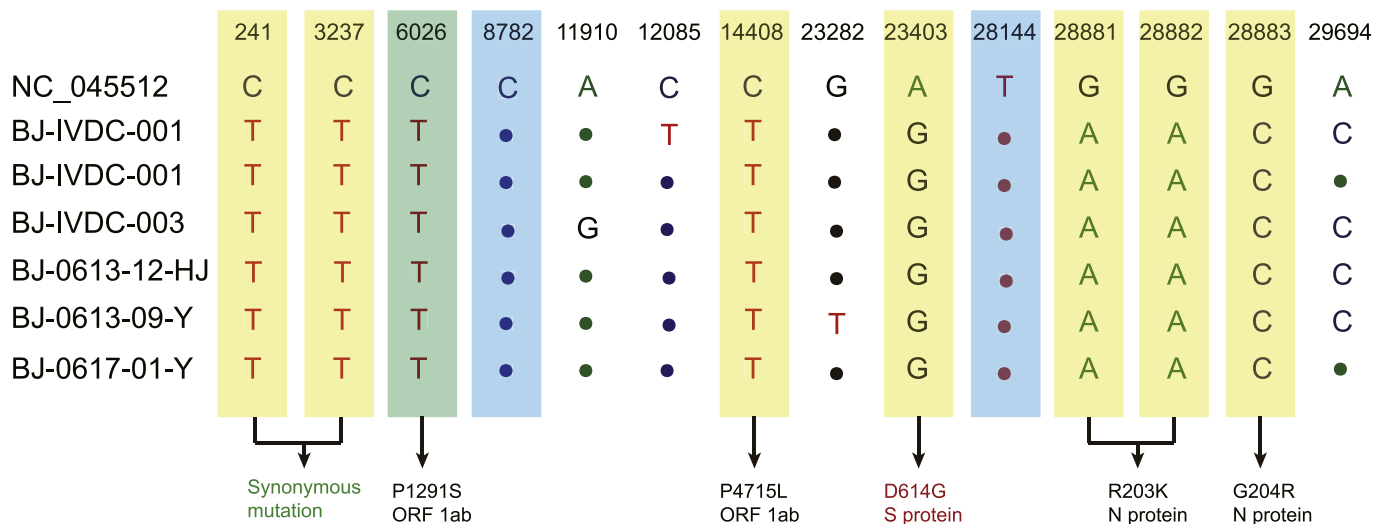
Fig. 2. Nucleotide substitutions of six SARS-CoV-2 genome sequences identified from Beijing's Xinfadi market compared with the reference strain of SARS-CoV-2 (GenBank No. NC_045512). The characteristic nucleotide substitutions of L-lineage, nt8782C and nt28144T, are highlighted in shades of blue. The characteristic nucleotide substitutions of L-lineage European branch I, nt241(C → T), nt3037(C → T), nt14408(C → T), nt23403(A → G), nt28881(G → A), nt28882(G → A), and nt28883 (G → C), are highlighted in shades of yellow. The characteristic nucleotide substitution of SARS-CoV-2 from the Xinfadi market, nt6026(C → T), is highlighted in shades of green. nt23403(A → G) is a nonsynonymous substitution in S gene, which leads to D614G mutation of S protein (indicated in red).

pathogenic microorganisms in specimens. Through bioinformatics analysis and high-speed operation of large-scale servers, massive whole genome sequence data can be processed and sorted, which can accurately trace the source of viruses and gain insight into their distribution. In the process of pathogen evolution, the virulence, pathogenic factors, and drug resistance genes are found, so as to grasp the pathogenic and drug-resistant mechanism of pathogens and provide a theoretical basis for the precise prevention and control of viral diseases.

The field of pathogen genomics has been developing rapidly, and various molecular technologies are being integrated into the diagnosis, prevention, and control of viral infectious diseases. The integration of genomic technology with bioinformatics and epidemiology has improved the public health surveillance, research, and control of infectious diseases, and has played an increasingly important role in the world [5]. For example, the traceability analysis of SARS virus in 2003, the transmission of imported wild poliovirus in Xinjiang, China in 2011, and the traceability of Ebola virus epidemic transmission in West Africa from 2014 to 2016 have been widely used [6].

With the support of the National Science and Technology Major Project, the National Institute for Viral Disease Control and Prevention has taken the lead in establishing a set of surveillance and traceability technology systems based on genome typing and clustering analysis algorithm of difference viruses including SARS-CoV-2, which can be used for early warning and prediction of viral infectious diseases. These systems were introduced in disease prevention and control centers at all levels in China, as well as scientific research and hospital systems. This will greatly improve the prevention and control of viral infectious diseases and play an important role in the prevention and control of major infectious diseases and the construction of biosafety system in China.

Finally, we also need to note whether the epidemic of D614G mutant SARS-CoV-2 is random or naturally selected, as the virus is circulating globally at present. Although there is no evidence from available epidemiological and clinical data that the mutation of S protein D614G leads to increased pathogenicity or virulence of the virus, whether the transmission of the virus is enhanced still needs to be examined by systematic global assessment. Therefore, we also need to conduct more research to understand the characteristics of this virus to help end the global outbreak as soon as possible. It is necessary to continuously monitor the genetic variation of SARS-CoV-2, focusing on the influence of key mutation sites of SARS-CoV-2 on viral transmission, clinical manifestations, severity, and course of disease.

## Acknowledgements

## Conflict of interest statement

The authors declare that there are no conflicts of interest. Given their roles as Editorial Board Member, Y. Zhang, W. Tan, D. Wang, W.J. Liu, W. Xu, and G. Wu had no involvement in the peer-review of this article and had no access to information regarding its peer-review. Full responsibility for the editorial process for this article was delegated to the editor Chuan Qin.

## Author contributions

**Yong Zhang:** Conceptualization, Data curation, Writing – original draft, Writing – review & editing. **Yang Pan:** Investigation. **Xiang Zhao:** Data curation, Formal analysis. **Weifeng Shi:** Data curation, Formal analysis. **Zhixiao Chen:** Formal analysis. **Sheng Zhang:** Formal analysis. **Peipei Liu:** Formal analysis. **Jinbo Xiao:** Formal analysis. **Wenjie Tan:** Formal analysis. **Dayan Wang:** Formal analysis. **William J. Liu:** Formal analysis. **Wenbo Xu:** Conceptualization, Funding acquisition. **Quanyi Wang:** Conceptualization, Funding acquisition. **Guizhen Wu:** Conceptualization, Funding acquisition.

## References

[1] W. Tan, P. Niu, X. Zhao, Y. Pan, Y. Zhang, L. Chen, L. Zhao, Y. Wang, D. Wang, J. Han, G.F. Gao, C. Huang, W. Xu, G. Wu, Reemergent cases of COVID-19 — Xinfadi wholesales market, Beijing municipality, China, June 11, 2020, China CDC Weekly 2 (2020) 502–504, https://doi.org/10.46234/ccdcw2020.132.

[2] S. Kumar, G. Stecher, K. Tamura, MEGA7: molecular evolutionary genetics analysis version 7.0 for bigger datasets, Mol. Biol. Evol. 33 (2016) 1870–1874, https://doi.org/10.1093/molbev/msw054.

[3] B. Korber, W.M. Fischer, S. Gnanakaran, H. Yoon, J. Theiler, W. Abfalterer, N. Hengartner, E.E. Giorgi, T. Bhattacharya, B. Foley, K.M. Hastie, M.D. Parker, D.G. Partridge, C.M. Evans, T.M. Freeman, T.I.D. Silva, C. McDanal, L.G. Perez, H. Tang, A. Moon-Walker, S.P. Whelan, C.C. LaBranche, E.O. Saphire, D.C. Montefiori, Tracking changes in SARS-CoV-2 Spike: evidence that D614G increases infectivity of the COVID-19 virus, Cell 182 (2020) 812–827, https://doi.org/10.1016/j.cell.2020.06.043.

[4] M. Eaaswarkhanth, A. Al Madhoun, F. Al-Mulla, Could the D614G substitution in the SARS-CoV-2 spike (S) protein be associated with higher COVID-19 mortality? Int. J. Infect. Dis. 96 (2020) 459–460, https://doi.org/10.1016/j.ijid.2020.05.071.

[5] J. Lu, L. du Plessis, Z. Liu, V. Hill, M. Kang, H. Lin, J. Sun, S. Francois, M.U.G. Kraemer, N.R. Faria, J.T. McCrone, J. Peng, Q. Xiong, R. Yuan, L. Zeng, P. Zhou, C. Liang, L. Yi, J. Liu, J. Xiao, J. Hu, T. Liu, W. Ma, W. Li, J. Su, H. Zheng, B. Peng, S. Fang, W. Su, K. Li, R. Sun, R. Bai, X. Tang, M. Liang, J. Quick, T. Song, A. Rambaut, N. Loman, J. Raghwani, O.G. Pybus, C. Ke, Genomic epidemiology of SARS-CoV-2 in Guangdong Province, China, Cell 181 (2020) 997–1003, https://doi.org/10.1016/j.cell.2020.04.023.

[6] S.K. Gire, A. Goba, K.G. Andersen, R.S. Sealfon, D.J. Park, L. Kanneh, S. Jalloh, M. Momoh, M. Fullah, G. Dudas, S. Wohl, L.M. Moses, N.L. Yozwiak, S. Winnicki, C.B. Matranga, C.M. Malboeuf, J. Qu, A.D. Gladden, S.F. Schaffner, X. Yang, P.P. Jiang, M. Nekoui, A. Colubri, M.R. Coomber, M. Fonnie, A. Moigboi, M. Gbakie, F.K. Kamara, V. Tucker, E. Konuwa, S. Saffa, J. Sellu, A.A. Jalloh, A. Kovoma, J. Koninga, I. Mustapha, K. Kargbo, M. Foday, M. Yillah, F. Kanneh, W. Robert, J.L. Massally, S.B. Chapman, J. Bochicchio, C. Murphy, C. Nusbaum, S. Young, B.W. Birren, D.S. Grant, J.S. Scheiffelin, E.S. Lander, C. Happi, S.M. Gevao, A. Gnirke, A. Rambaut, R.F. Garry, S.H. Khan, P.C. Sabeti, Genomic surveillance elucidates Ebola virus origin and transmission during the 2014 outbreak, Science 345 (2014) 1369–1372, https://doi.org/10.1126/science.1259657.