Please cite this article in press as: Zhang and Holmes, A Genomic Perspective on the Origin and Emergence of SARS-CoV-2, Cell (2020), https://doi.org/10.1016/j.cell.2020.03.035

Cell

A Genomic Perspective on the Origin and Emergence of SARS-CoV-2

Yong-Zhen Zhang[1] and Edward C. Holmes[1,2,*]
[1]Shanghai Public Health Clinical Center and School of Life Science, Fudan University, Shanghai, China
[2]Marie Bashir Institute for Infectious Diseases and Biosecurity, School of Life and Environmental Sciences and School of Medical Sciences, The University of Sydney, Sydney, Australia
*Correspondence: edward.holmes@sydney.edu.au
https://doi.org/10.1016/j.cell.2020.03.035

The ongoing pandemic of a new human coronavirus, SARS-CoV-2, has generated enormous global concern. We and others in China were involved in the initial genome sequencing of the virus. Herein, we describe what genomic data reveal about the emergence SARS-CoV-2 and discuss the gaps in our understanding of its origins.

A New Human Coronavirus

The first reports of a novel pneumonia (COVID-19) in Wuhan city, Hubei province, China, occurred in late December 2019, although retrospective analyses have identified a patient with symptom onset as early as December 1st. Because the number of SARS-CoV-2 cases is growing rapidly and spreading globally, we will refrain from citing the number of confirmed infections. However, it is likely that the true number of cases will be substantially greater than reported because very mild or asymptomatic infections will often be excluded from counts. Any under-reporting of case numbers obviously means that the case fatality rate (CFR) associated with COVID-19 in the worst-hit regions will be lower than that currently cited. CFRs will also vary geographically, between age groups and temporally. Although these uncertainties will likely not be resolved without large-scale serological surveys, from current data it is clear that the CFR for COVID-19 is substantially higher than that of seasonal influenza but lower than that of two closely related coronaviruses that have similarly recently emerged in humans: SARS-CoV, responsible for the SARS outbreak of 2002–2003, and MERS-CoV that since 2015 has been responsible for the ongoing outbreak of MERS largely centered on the Arabian peninsula. However, it is also evident that SARS-CoV-2 is more infectious than both SARS-CoV and MERS-CoV and that individuals can transmit the virus when asymptomatic or presymptomatic, although how frequently remains uncertain.

An important early association was observed between the first reported cases of COVID-19 and the Huanan seafood and wildlife market in Wuhan city (which we both visited several years ago) where a variety of mammalian species were available for purchase at the time of the outbreak (Figure 1). Given that SARS-CoV-2 undoubtedly has a zoonotic origin, the link to such a "wet" market should come as no surprise. However, as not all of the early cases were market associated, it is possible that the emergence story is more complicated than first suspected. Genome sequences of "environmental samples"—likely surfaces—from the market have now been obtained, and phylogenetic analysis reveals that they are very closely related to viruses sampled from the earliest Wuhan patients. While this again suggests that the market played an important role in virus emergence, it is not clear whether the samples were derived from people who inadvertently deposited infectious material or from animals or animal matter present at that location. Unfortunately, the apparent lack of direct animal sampling in the market may mean that it will be difficult, perhaps even impossible, to accurately identify any animal reservoir at this location.

After clinical cases began to appear, our research team, along with a number of others, attempted to determine the genome sequence of the causative pathogen (Lu et al., 2020; Wu et al., 2020; Zhou et al., 2020; Zhu et al., 2020). We focused on a patient admitted to the Central Hospital of Wuhan on December 26, 2019, six days after the onset of symptoms (Wu et al., 2020). This patient was experiencing fever, chest tightness, cough, pain, and weakness, along with lung abnormalities indicative of pneumonia that appear to be commonplace in COVID-19 (Huang et al., 2020). Fortunately, next-generation meta-transcriptomic sequencing enabled us to obtain a complete viral genome from this patient on January 5, 2020. Initial analysis revealed that the virus was closely related to those of SARS-like viruses (family *Coronaviridae*). This result was immediately reported to the relevant authorities, and an annotated version of the genome sequence (strain Wuhan-Hu-1) was submitted to NCBI/GenBank on the same day. Although the GenBank sequence (GenBank: MN908947) was the first of SARS-CoV-2 available, it was subsequently corrected to ensure its accuracy. With the help of Dr. Andrew Rambaut (University of Edinburgh), we released the genome sequence of the virus on the open access Virological website (http://virological.org/) early on January 11, 2020. Afterwards, the China CDC similarly released SARS-CoV-2 genome sequences (with associated epidemiological data) on the public access GISAID database (https://www.gisaid.org/). At the time of writing, almost 200 SARS-CoV-2 genomes are publicly available, representing the genomic diversity of the virus in China and beyond and providing a freely accessible global resource. Importantly, the release of the SARS-CoV-2 genome sequence data facilitated the rapid development of diagnostic tests (Corman et al., 2020) and now an

Please cite this article in press as: Zhang and Holmes, A Genomic Perspective on the Origin and Emergence of SARS-CoV-2, Cell (2020), https://doi.org/10.1016/j.cell.2020.03.035

Figure 1. The Huanan Seafood and Wildlife Market in Wuhan, China

The photographs (credit: E.C.H.) were taken when both authors visited the market together in October 2014 and highlight some of the wide variety of wildlife on sale, providing a potent mechanism for zoonotic transmission. Importantly, although many of the early COVID-19 cases were linked to this market, its role in the initial emergence of SARS-CoV-2 remains uncertain.

infectious clone (Thao et al., 2020). The race to develop an effective vaccine and antivirals is ongoing, with trails of the latter underway (Wang et al., 2020).

Comparisons between SARS-CoV-2 and Other Coronaviruses

The earliest genomic genome sequence data made it clear that SARS-CoV-2 was a member of the genus *Betacoronavirus* and fell within a subgenus (*Sarbecovirus*) that includes SARS-CoV (MERS-CoV falls in a separate subgenus, *Merbecovirus*) (Lu et al., 2020; Wu et al., 2020; Zhou et al., 2020; Zhu et al., 2020). Indeed, initial comparisons revealed that SARS-CoV-2 was approximately 79% similar to SARS-CoV at the nucleotide level. Of course, patterns of similarity vary greatly between genes, and SARS-CoV and SARS-CoV-2 exhibit only ~72% nucleotide sequence similarity in the spike (S) protein, the key surface glycoprotein that interacts with host cell receptors.

Given these close evolutionary relationships, it is unsurprising that the genome structure of SARS-CoV-2 resembles those of other betacoronaviruses, with the gene order 5′-replicase ORF1ab-S-envelope(E)-membrane(M)-N-3′. The long replicase ORF1ab gene of SARS-CoV-2 is over 21 kb in length and contains 16 predicted non-structural proteins and a number of downstream open reading frames (ORFs) likely of similar function to those of SARS-CoV. Comparative genomic analysis has been greatly assisted by the availability of a related virus from a *Rhinolophus affinis* (i.e., horseshoe) bat sampled in Yunnan province, China, in 2013 (Zhou et al., 2020). This virus, denoted RaTG13, is ~96% similar to SARS-CoV-2 at the nucleotide sequence level. Despite this sequence similarity, SARS-CoV-2 and RaTG13 differ in a number of key genomic features, arguably the most important of which is that SARS-CoV-2 contains a polybasic (furin) cleavage site insertion (residues PRRA) at the junction of the S1 and S2 subunits of the S protein (Coutard et al., 2020). This insertion, which may increase the infectivity of the virus, is not present in related betacoronaviruses, although similar polybasic insertions are present in other human coronaviruses, including HCoV-HKU1, as well as in highly pathogenic strains of avian influenza virus. In addition, the receptor binding domain (RBD) of SARS-CoV-2 and RaTG13 are only ~85% similar and share just one of six critical amino acid residues. Both sequence and structural comparisons suggest that the SARS-CoV-2 RBD is well suited for binding to the human ACE2 receptor that was also utilized by SARS-CoV (Wrapp et al., 2020). Importantly, an independent insertion(s) of the amino acids PAA at the S1/S2 cleavage site was recently observed in a virus (RmYN02) sampled in mid-2019 from another *Rhinolophus* bat in Yunnan province, indicating that these insertion events reflect a natural part of ongoing coronavirus evolution (Zhou et al., 2020). While RmYN02 is relatively divergent from SARS-CoV-2 in the S protein (~72% sequence similarity), it is the closest relative (~97% nucleotide sequence similarity) of the human virus in the long replicase gene.

Although SARS-CoV and MERS-CoV are both closely related to SARS-CoV-2

and have bat reservoirs, the biological differences between these viruses are striking. As noted above, SARS-CoV-2 is markedly more infectious, resulting in very different epidemiological dynamics to those of SARS-CoV and MERS-CoV. In these latter two viruses, there was a relatively slow rise in case numbers, and MERS-CoV has never been able to fully adapt to human transmission: the majority of the cases are due to spillover from camels on the Arabian peninsula with only sporadic human-to-human transmission (Sabir et al., 2016). In contrast, the remarkable local and global spread of SARS-CoV-2 caught most by surprise. Determining the virological characteristics that underpin such transmissibility is clearly a priority.

The Zoonotic Origins of SARS-CoV-2

The emergence and rapid spread of COVID-19 signifies a perfect epidemiological storm. A respiratory pathogen of relatively high virulence from a virus family that has an unusual knack of jumping species boundaries, that emerged in a major population center and travel hub shortly before the biggest travel period of the year: the Chinese Spring Festival. Indeed, it is no surprise that epidemiological modeling suggests that SARS-CoV-2 had already spread widely in China before the city of Wuhan was placed under strict quarantine (Chinazzi et al., 2020).

It was also no surprise that early genomic comparisons revealed that the most closely related viruses to SARS-CoV-2 came from bats (Zhou et al., 2020). Sampling in recent years has identified an impressive array of bat coronaviruses, including RaTG13 and RmYN02 (Hu et al., 2017; Yang et al., 2015). Hence, bats are undoubtedly important reservoir species for a diverse range of coronaviruses (Cui et al., 2019). Despite this, the exact role played by bats in the zoonotic origin of SARS-CoV-2 is not established. In particular, the bat viruses most closely related to SARS-CoV-2 were sampled from animals in Yunnan province, over 1,500 km from Wuhan. There are relatively few bat coronaviruses from Hubei province, and those that have been sequenced are relatively distant to SARS-CoV-2 in phylogenetic trees (Lin et al., 2017). The simple inference from this is that our sampling of bat viruses is strongly biased toward some geographical locations. This will need to be rectified in future studies. In addition, although sequence similarity values of 96%–97% make it sound like the available bat viruses are very closely related to SARS-CoV-2, in reality this likely represents more than 20 years of sequence evolution (although the underlying molecular clock may tick at an uncertain rate if there was strong adaptive evolution of the virus in humans). It is therefore almost a certainty that more sampling will identify additional bat viruses that are even closer relatives of SARS-CoV-2. A key issue is whether these viruses, or those from any other animal species, contain the key RBD mutations and the same furin-like cleavage site insertion as found in SARS-CoV-2.

Although bats are likely the reservoir hosts for this virus, their general ecological separation from humans makes it probable that other mammalian species act as "intermediate" or "amplifying" hosts, within which SARS-CoV-2 was able to acquire some or all of the mutations needed for efficient human transmission. In the case of SARS and MERS, civets and camels, respectively, played the role of intermediate hosts, although as MERS-CoV was likely present in camels for some decades before it emerged in humans during multiple cross-species events, these animals may be better thought of as true reservoir hosts (Sabir et al., 2016). To determine what these intermediate host species might be, it is imperative to perform a far wider sampling of animals from wet markets or that live close to human populations. This is highlighted by the recent discovery of viruses closely related to SARS-CoV-2 in Malayan pangolins (*Manis javanica*) illegally imported into southern China (Guangdong and Guangxi provinces). The Guangdong pangolin viruses are particularly closely related to SARS-CoV-2 in the RBD, containing all six of the six key mutations thought to shape binding to the ACE2 receptor and exhibiting 97% amino acid sequence similarity (although they are more divergent from SARS-CoV-2 in the remainder of the genome). Although pangolins are of great interest because of how frequently they are involved in illegal trafficking and their endangered status, that they carry a virus related to SARS-CoV-2 strongly suggests that a far greater diversity of related betacoronaviruses exists in a variety of mammalian species but has yet to be sampled.

While our past experience with coronaviruses suggests that evolution in animal hosts, both reservoirs and intermediates, is needed to explain the emergence of SARS-CoV-2 in humans, it cannot be excluded that the virus acquired some of its key mutations during a period of "cryptic" spread in humans prior to its first detection in December 2019. Specifically, it is possible that the virus emerged earlier in human populations than envisaged (perhaps not even in Wuhan) but was not detected because asymptomatic infections, those with mild respiratory symptoms, and even sporadic cases of pneumonia were not visible to the standard systems used for surveillance and pathogen identification. During this period of cryptic transmission, the virus could have gradually acquired the key mutations, perhaps including the RBD and furin cleavage site insertions, that enabled it to adapt fully to humans. It wasn't until a cluster of pneumonia cases occurred that we were able to detect COVID-19 via the routine surveillance system. Obviously, retrospective serological or metagenomic studies of respiratory infection will go a long way to determining whether this scenario is correct, although such early cases may never be detected.

Another issue that has received considerable attention is whether SARS-CoV-2 is a recombinant virus, and whether such recombination might have facilitated its emergence (Lu et al., 2020; Wu et al., 2020). The complicating factor here is that sarbeviruses, and coronaviruses more broadly, experience widespread recombination, so that distinguishing recombination that assisted virus emergence from "background" recombination events is not trivial. Recombination is visible at multiple locations across the sarbevirus genome, including in the S protein, and in bat viruses closely related to SARS-CoV-2. For example, there is some evidence for recombination among SARS-CoV-2, RaTG13, and the Guangdong pangolin CoVs (Lam et al., 2020), and the genome of RmYN02 has similarly been widely impacted by recombination (Zhou et al., 2020). However, trying to

determine the exact pattern and genomic ancestry of recombination events is difficult, particularly as many of the recombinant regions may be small and are likely to change as we sample more viruses related to SARS-CoV-2. To resolve these issues, it will again be necessary to perform a far wider sampling of viral diversity in animal populations.

Ongoing Genomic Evolution of SARS-CoV-2

As the COVID-19 epidemic has progressed, so more viral genomes have been sequenced. As expected given their recent common ancestry, the earliest samples from Wuhan contained relatively little genetic diversity. While this can prevent detailed phylogenetic and phylogeographic inferences, it does show that the public health authorities in Wuhan did a remarkable job in detecting the first cluster of pneumonia cases. However, this seemingly recent common ancestry does not exclude a pre-outbreak period of cryptic transmission in humans. Although accumulating genetic diversity means that it is now possible to detect distinct phylogenetic clusters of SARS-CoV-2 sequences, it is difficult to determine using genomic comparisons alone whether the virus is fixing phenotypically important mutations as it spreads through the global population, and any such claims require careful experimental verification.

Given the high mutation rates that characterize RNA viruses, it is obvious that many more mutations will appear in the viral genome and that these will help us to track the spread of SARS-CoV-2 (Grubaugh et al., 2019). However, as the epidemic grows, our sample size of sequences will likely be so small relative to the total number of cases that it will be very difficult, if not impossible, to detect individual transmission chains. Caution must therefore always be exercised when attempting to infer exact transmission events. As an aside, although coronaviruses likely have lower mutation rates than other RNA viruses because of an inherent capacity for some proof-reading activity due to a 3′-to-5′ exoribonuclease (Minskaia et al., 2006), their long-term rates of nucleotide substitution (i.e., of molecular evolution) fall within the distribution of those seen in other RNA viruses (Holmes et al., 2016). This suggests that lower mutation rates are to some extent compensated by high rates of virus replication within hosts. Although there is no evidence that this capacity to mutate (common to RNA viruses) will result in any radical changes in phenotype—such as in transmissibility and virulence—as these only rarely change at the scale of individual disease outbreaks (Grubaugh et al., 2020), it is obviously important to monitor any changes in phenotype as the virus spreads. In all likelihood, any drop in the number of cases and/or CFR of COVID-19 will likely be due to rising immunity in the human population and epidemiological context rather than mutational changes in the virus.

Conclusions

It seems inevitable that SARS-CoV-2 will become the fifth endemic coronavirus in the human population (along with HKU1, NL63, OC43, and 229E) and one that is currently spreading in a totally susceptible population. Coronaviruses clearly have the capacity to jump species boundaries and adapt to new hosts, making it straightforward to predict that more will emerge in the future, although quite why coronaviruses possess this capacity in comparison to some other RNA viruses is unclear. Critically, the surveillance of animal coronaviruses should include animals other than bats, as the role of intermediate hosts is likely of major importance, providing a more direct pathway for the virus to emerge in humans. Given the enormous diversity of viruses in wildlife and their ongoing evolution, arguably the simplest and most cost-effective way to reduce the risk of future outbreaks is to limit our exposure to animal pathogens as much as possible. While our intimate relationship with the animal world means we cannot build impregnable barriers, stronger action against the illegal wildlife trade and removing all mammalian (and perhaps avian) wildlife from wet markets will provide an important buffer.

ACKNOWLEDGMENTS

This work was funded by the National Natural Science Foundation of China (grants 81861138003 and 31930001), the Special National Project on investigation of basic resources of China (grant 2019FY101500), and the Australian Research Council (grant FL170100022).

WEB RESOURCES

GISAID, https://www.gisaid.org/
Virological, http://virological.org/

REFERENCES

Chinazzi, M., Davis, J.T., Ajelli, M., Gioannini, C., Litvinova, M., Merler, S., Pastore Y Piontti, A., Mu, K., Rossi, L., Sun, K., et al. (2020). The effect of travel restrictions on the spread of the 2019 novel coronavirus (COVID-19) outbreak. Science. eaba9757. Published online March 6, 2020. https://doi.org/10.1126/science.aba9757.

Corman, V.M., Landt, O., Kaiser, M., Molenkamp, R., Meijer, A., Chu, D.K.W., Bleicker, T., Brünink, S., Schneider, J., Schmidt, M.L., et al. (2020). Detection of 2019 novel coronavirus (2019-nCoV) by real-time RT-PCR. Euro Surveill. *25*, 2000045.

Coutard, B., Valle, C., de Lamballerie, X., Canard, B., Seidah, N.G., and Decroly, E. (2020). The spike glycoprotein of the new coronavirus 2019-nCoV contains a furin-like cleavage site absent in CoV of the same clade. Antiviral Res. *176*, 104742.

Cui, J., Li, F., and Shi, Z.-L. (2019). Origin and evolution of pathogenic coronaviruses. Nat. Rev. Microbiol. *17*, 181–192.

Grubaugh, N.D., Ladner, J.T., Lemey, P., Pybus, O.G., Rambaut, A., Holmes, E.C., and Andersen, K.G. (2019). Tracking virus outbreaks in the twenty-first century. Nat. Microbiol. *4*, 10–19.

Grubaugh, N.D., Petrone, M.E., and Holmes, E.C. (2020). We shouldn't worry when a virus mutates during disease outbreaks. Nat Microbiol. Published online February 18, 2020. https://doi.org/10.1038/s41564-020-0690-4.

Holmes, E.C., Dudas, G., Rambaut, A., and Andersen, K.G. (2016). The evolution of Ebola virus: Insights from the 2013-2016 epidemic. Nature *538*, 193–200.

Hu, B., Zeng, L.P., Yang, X.L., Ge, X.Y., Zhang, W., Li, B., Xie, J.Z., Shen, X.R., Zhang, Y.Z., Wang, N., et al. (2017). Discovery of a rich gene pool of bat SARS-related coronaviruses provides new insights into the origin of SARS coronavirus. PLoS Pathog. *13*, e1006698.

Huang, C., Wang, Y., Li, X., Ren, L., Zhao, J., Hu, Y., Zhang, L., Fan, G., Xu, J., Gu, X., et al. (2020). Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. Lancet *395*, 497–506.

Lam, T.T.-Y., Shum, M.H.-H., Zhu, H.-C., Tong, Y.-G., Ni, X.-B., Liao, Y.-S., Wei, W., Cheung, W.Y.-M., Li, W.-J., Li, L.-F., et al. (2020). Identification of 2019-nCoV related coronaviruses in Malayan pangolins in southern China. bioRxiv. https://doi.org/10.1101/2020.02.13.945485.

Lin, X.-D., Wang, W., Hao, Z.-Y., Wang, Z.-X., Guo, W.-P., Guan, X.-Q., Wang, M.-R., Wang, H.-W., Zhou, R.-H., Li, M.-H., et al. (2017). Extensive diversity of coronaviruses in bats from China. Virology *507*, 1–10.

Lu, R., Zhao, X., Li, J., Niu, P., Yang, B., Wu, H., Wang, W., Song, H., Huang, B., Zhu, N., et al. (2020). Genomic characterisation and epidemiology of 2019 novel coronavirus: implications for virus origins and receptor binding. Lancet *395*, 565–574.

Minskaia, E., Hertzig, T., Gorbalenya, A.E., Campanacci, V., Cambillau, C., Canard, B., and Ziebuhr, J. (2006). Discovery of an RNA virus 3′->5′ exoribonuclease that is critically involved in coronavirus RNA synthesis. Proc. Natl. Acad. Sci. USA *103*, 5108–5113.

Sabir, J.S.M., Lam, T.T.-Y., Ahmed, M.M.A., Li, L., Shen, Y., Abo-Aba, S.E.M., Qureshi, M.I., Abu-Zeid, M., Zhang, Y., Khiyami, M.A., et al. (2016). Co-circulation of three camel coronavirus species and recombination of MERS-CoVs in Saudi Arabia. Science *351*, 81–84.

Thao, T.T.N., Labroussaa, F., Ebert, N., V'kovski, P., Stalder, H., Portmann, J., Kelly, J., Steiner, S., Holwerda, M., Kratzel, A., et al. (2020). Rapid reconstruction of SARS-CoV-2 using a synthetic genomics platform. bioRxiv. https://doi.org/10.1101/2020.02.21.959817.

Wang, Y., Zhou, F., Zhang, D., Zhao, J., Du, R., Hu, Y., Cheng, Z., Gao, L., Jin, Y., Luo, G., et al. (2020). Evaluation of the Efficacy and Safety of Intravenous Remdesivir in Adult Patients with Severe Pneumonia caused by COVID-19 virus infection: study protocol for a Phase 3 Randomized, Double-blind, Placebo-controlled, Multicentre trial. BMC Trials. https://doi.org/10.21203/rs.2.24058/v1.

Wrapp, D., Wang, N., Corbett, K.S., Goldsmith, J.A., Hsieh, C.-L., Abiona, O., Graham, B.S., and McLellan, J.S. (2020). Cryo-EM structure of the 2019-nCoV spike in the prefusion conformation. Science *367*, 1260–1263.

Wu, F., Zhao, S., Yu, B., Chen, Y.-M., Wang, W., Song, Z.G., Hu, Y., Tao, Z.W., Tian, J.H., Pei, Y.Y., et al. (2020). A new coronavirus associated with human respiratory disease in China. Nature *579*, 265–269.

Yang, X.L., Hu, B., Wang, B., Wang, M.N., Zhang, Q., Zhang, W., Wu, L.J., Ge, X.Y., Zhang, Y.Z., Daszak, P., et al. (2015). Isolation and characterization of a novel bat coronavirus closely related to the direct progenitor of Severe Acute Respiratory Syndrome coronavirus. J. Virol. *90*, 3253–3256.

Zhou, P., Yang, X.L., Wang, X.G., Hu, B., Zhang, L., Zhang, W., Si, H.R., Zhu, Y., Li, B., Huang, C.L., et al. (2020). A pneumonia outbreak associated with a new coronavirus of probable bat origin. Nature *579*, 270–273.

Zhu, N., Zhang, D., Wang, W., Li, X., Yang, B., Song, J., Zhao, X., Huang, B., Shi, W., Lu, R., et al.; China Novel Coronavirus Investigating and Research Team (2020). A novel coronavirus from patients with pneumonia in China, 2019. N. Engl. J. Med. *382*, 727–733.