Extended ORF8 Gene Region Is Valuable in the Epidemiological Investigation of Severe Acute Respiratory Syndrome–Similar Coronavirus

Shuaiyin Chen,1, a Xin Zheng,2, a Jingyuan Zhu,1, a Ronghua Ding,1 Yuefei Jin,1 Weiguo Zhang,1 Haiyan Yang,1 Yingjuan Zheng,2 Xin Li,5 and Guangcai Duan1

1Zhengzhou University College of Public Health, Zhengzhou, China, 2Taoharmony Biotech Ltd, Hangzhou, China, 3Duke University Medical Center, Duke University, Durham, North Carolina, USA, 4First Affiliated Hospital, Zhengzhou University, Zhengzhou, China, and 5Beijing Ditan Hospital, Capital Medical University, Beijing, China

Severe acute respiratory syndrome coronavirus (SARS-CoV) was discovered as a novel pathogen in the 2002–2003 SARS epidemic. The emergence and disappearance of this pathogen have brought questions regarding its source and evolution. Within the genome sequences of 281 SARS-CoVs, severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), and SARS-related CoVs (SARSr-CoVs), a ~430 bp genomic region (from 27 701 bp to 28 131 bp in AY390556.1) with regular variations was investigated. This ~430 bp region overlaps with the ORF8 gene and is prone to deletions and nucleotide substitutions. Its complexity suggested the need for a new genotyping method for coronaviruses related to SARS-similar coronaviruses (SARS-CoV, SARSr-CoV, and SARS-CoV-2). Bat SARS-CoV presented 3 genotypes, of which type 0 is only seen in bat SARSr-CoV, type I is present in SARS in the early phase, and type II is found in all SARS-CoV-2. This genotyping also shows potential usage in distinguishing the SARS-similar coronaviruses from different hosts and geographic areas. This genomic region has important implications for predicting the epidemic trend and studying the evolution of coronavirus.

Keywords. SARS-CoV; SARS-CoV-2; ORF8 gene; genotyping method.
term “SARS-similar coronavirus” was used to refer to all the 3 subtypes of coronaviruses.

METHODS

A total of 154 genome sequences of SARS-CoVs and bat SARSr-CoVs were collected from the National Center for Biotechnology Information (http://www.ncbi.nlm.nih.gov). Among these, 115 SARS-CoV-2 isolates were obtained from humans, 4 SARS-CoV-2 isolates were collected from the environment, and 8 SARSr-CoVs from pangolin were downloaded from GISAID. Isolate information is listed in Supplementary Tables 1 and 2. Sequences were aligned with Clustal Omega (http://www.clustal.org, version 1.2.4) [13, 14]. Jalview was adopted to view the alignment (http://www.jalview.org, version 2.11.0) [15]. Phylogeny analysis was performed using the maximum likelihood algorithm under the Jukes-Cantor model with bootstrap values determined by 1000 replicates with Mega (http://www.megasoftware.net, version 10.1) [16]. AY390556.1 (GZ02), which was isolated at the early stage of the SARS epidemic, was used as the reference isolate.

RESULTS

Phylogeny of SARS-CoV and SARSr-CoV Based on the S Gene

The spike(s) gene is directly involved in ACE2 binding, which is highly variable in the SARS-CoV and SARSr-CoV genomes. The phylogenetic distances among the 154 isolates of SARS-CoVs and SARSr-CoVs on the basis of their similarity to the S gene are shown in Figure 1.

SARS-CoVs from humans were clearly separated from the SARSr-CoVs from bats or civets (Figure 1). Viruses from the 3 epidemic phases also showed within-group variations in the tree. By contrast, the bat viruses from the same cave in Yunnan, Hong Kong, and other areas in China did not show any clear distinction. The viruses from the cave were crossed with those collected from other areas, such as Jiangxi and Hubei, China at different times. Some bat viruses from Hong Kong and Yunnan caves formed subgroups in the tree. The SARS-CoVs from civets were located between human SARS-CoVs and SARSr-CoVs and were close to 2 human SARS-CoVs (AY568539.1 and AY613947.1) from a localized breakout in Tong De Li Restaurant (TDLR) in Guangzhou, China. Most of the civet SARS-CoVs were collected at the end of 2003 and showed close relationship with the 2 civet SARS-CoVs collected in early 2003 (AY304486.1 and AY304488.1). The nucleotide substitution in each type also demonstrated consistency to some extent. This regular deletion was related to the epidemic stage of the disease. Data on civet SARS-CoV (Figure 3A) revealed 3 deletion types in the region. One happened around 27 769 bp in the reference genome (AY390556.1) with sizes around 39 nt (39 nt spot), one happened around 27 881 bp in the ~430 bp region. Meanwhile, the ~430 bp region in bat SARSr-CoVs presented different types. First type displayed deletions of 2, 6, and 7 nt compared with the consensus sequence. The second type displayed deletions of 1, 2, 3, and 9 nt. The third type displayed deletions of 3 and 9 nt. Three other deletions (5, 29, and 336 nt) happened in a single isolate. In addition, nucleotide substitution in each type also demonstrated consistency to some extent. This regular deletion was related to the epidemic stage of the disease. Data on civet SARS-CoV (Figure 3C) showed limited substitutions and minor deletions of 1 or 2 nt. The detailed alignment of all collected genomes is accessible in the Supplementary Materials.

Figure 4A shows the comparison of the SARS-CoV of early phase and the SARSr-CoVs of the first type above. Among the 12 SARSr-CoVs, 8 showed high identity with SARSr-CoVs.

Distinct Phylogeny Revealed by the Extended Region Spanning ORF8

ORF8 gene is another hypervariable region in the genomes of SARS-CoV and SARSr-CoV. In the genome of reference isolate SARS-CoV AY390556.1 (GZ02), ORF8 is located from 27 779 bp to 28 147 bp. The variations in the ORF8 gene region revealed an indistinct phylogeny in our study (Supplementary Figure 1). But when we explored the phylogeny in a gene-spanning way, it was found that extending to upstream of ORF8 brought such a region that all main variations around ORF8 were included and the phylogenetic tree became clearer in epidemic phases. Therefore, we extended the gene region and figured out the longer fragment, from 27 701 bp to 28 131 bp, which turned out to establish a distinct phylogeny among the virus isolates in Figure 2.

Compared with the phylogeny analysis based on the S gene (Figure 1), the viruses from the cave of Yunnan were closer (Figure 2). The SARS-CoVs from the early, mid, and late phases were grouped in the tree, which also encompassed all civet SARS-CoVs. Among the 18 civet SARS-CoVs, 16 isolates remained close to the 2 SARS-CoV viruses from patients in the TDLR breakout. However, another 2 civet SARS-CoVs (AY304486.1 and AY304488.1) collected in early 2003 were separated from the others due to some nucleotide (nt) substitutions (Figure 3C). One isolate from the late phase, AY345988.1, was grouped with early isolates.

Sequence alignments further exhibited the underlying genetic variations. Data on human SARS-CoV (Figure 3A) revealed 3 deletion types in the region. One happened around 27 769 bp in the reference genome (AY390556.1) with sizes around 39 nt (39 nt spot), one happened around 27 881 bp with sizes around 29 nt (29 nt spot), and the other was a deletion of 415 nt. Deletions in the 29 nt spot happened in 4 of 13 early-phase isolates, and in all 16 mid-phase isolates. No 39 nt deletion was found in known early/mid-phase isolates. Among the 60 late-phase isolates, 56 carried the 29 nt deletion, 4 of which also showed deletions in the 39 nt spot. The remaining 4 of the 60 isolates had 415 nt deletion in the ~430 bp region. Meanwhile, the ~430 bp region in bat SARSr-CoVs presented different types. First type displayed deletions of 2, 6, and 7 nt compared with the consensus sequence. The second type displayed deletions of 1, 2, 3, and 9 nt. The third type displayed deletions of 3 and 9 nt. Three other deletions (5, 29, and 336 nt) happened in a single isolate. In addition, nucleotide substitution in each type also demonstrated consistency to some extent. This regular deletion was related to the epidemic stage of the disease. Data on civet SARS-CoV (Figure 3C) showed limited substitutions and minor deletions of 1 or 2 nt. The detailed alignment of all collected genomes is accessible in the Supplementary Materials.

Sequence Alignment of ~430 bp Region in SARS-CoV-2

This ~430 bp region was also found in 119 isolates of SARS-CoV-2 and 8 isolates of SARSr-CoVs from pangolins. Its type
in these isolates was compared with that in the 154 isolates of SARS-CoVs and SARSr-CoVs. In Figure 4B, the SARS-CoV-2 isolates presented high similarity with 3 bat SARSr-CoVs (MG772933.1, MG772934.1, and RaTG13), which are of the second type in Fib.3b. These 3 bat SARSr-CoVs had highly conserved (90.64%–97.49%) nucleotide sequence identities for SARS-CoV-2. The pangolin isolates appeared close to SARS-CoV-2 with 93 nt substitutions.
Coronaviruses belong to a large diverse family of Coronaviridae. SARS-CoV was discovered as a novel pathogen causing SARS in 2002–2003. The epidemic continued for approximately 9 months and never reappeared, leaving unresolved epidemiological questions about the virus origin, how it disappeared, and whether it will return. Previous studies strongly suggested that SARS-CoV originated from SARSr-CoV in bats [17–19].

DISCUSSION

Coronaviruses belong to a large diverse family of Coronaviridae. SARS-CoV was discovered as a novel pathogen causing SARS in 2002–2003. The epidemic continued for approximately 9 months and never reappeared, leaving unresolved epidemiological questions about the virus origin, how it disappeared, and whether it will return. Previous studies strongly suggested that SARS-CoV originated from SARSr-CoV in bats [17–19].
Genomic analysis confirmed the existence of highly similar genes in both viruses. In late 2002, SARS-CoV evolved in bats, took civet as the intermediate transmitting host, and finally appeared in humans as a new pathogen. SARS-CoV is suggested to have emerged from the recombination of bat SARSr-CoV viruses. The most variable genomic regions in SARS-CoV and SARSr-CoV are the S and ORF8 genes [20, 21]. The S gene is essential for receptor binding during SARS-CoV infection and is highly variable due to the deletions of 5, 12, or 13 amino acids in the S protein [18, 22–24]. The S protein displays 78.2%–97.2% amino acid identity between SARS-CoV and SARSr-CoV [23]. Published studies also indicated that polymorphism in the S gene is critical to the virus’s affinity for the human ACE2 receptor and consequently to its transmission in human hosts [25]. The ORF8 gene presents multiple genotypes in the SARS 2002–2003 outbreak including 82 nt deletion, 29 nt deletion, and whole ORF8 loss [3]. This gene displays 47.7%–100% identity between SARS-CoV and SARSr-CoV. ORF8 is split into ORF8a, which enhances SARS-CoV replication and induces caspase-dependent apoptosis, and ORF8b, which affects DNA synthesis and degradation of E protein [26, 27]. The functions of these proteins in SARS-CoV are trivial, and their deletions do not affect the virus survival. Additional work is needed to fully elucidate the influence of the deletions with substantially long size. Therefore, we explored the association of ORF8 variations with disease transmission. The genomic variations in ORF8 differ among the various phases in the 2002–2003 epidemic [3]. In the present study, we aimed to track the genomic variations of the ORF8 gene in SARS-CoVs and SARSr-CoVs in different hosts and epidemic phases in 2002–2003 to clarify the potential role of the ORF8 polymorphism in the epidemic. AY390556.1 (GZ02) was one of the earliest isolated strains in the 2002–2003 epidemic, and hence was adopted as the reference strain to reflect the evolution in the aspect of epidemiology.

Figure 3. A, Sequence alignment of ~430 bp region in selected severe acute respiratory syndrome (SARS) coronaviruses. The wrapped alignment is divided into 3 parts by 2 horizontal lines. The upper part is of early-phase strains; the middle part is of mid-phase strains; the lower part is of late-phase strains. The 29 nt and 39 nt spots are labeled. B and C, Sequence alignment of ~430 bp region in bat (B) and civet (C) SARS-related coronaviruses.
Figure 3. Continued.
Evolution was first examined in the S gene as a reference of how the isolates change in the genome. The phylogeny tree clearly divided the viruses from bats and humans without any apparent similarities across different hosts. Although in the transient breakout at the end of 2003 in Guangzhou, China, the human SARS-CoV carried an almost identical S gene as the virus from civets [28]. This finding is consistent with other studies [23].

In the analysis of ORF8 genomic variations, the division among hosts and epidemic phases was not as clear as that in the S gene phylogeny, thereby suggesting that ORF8 variations were not relevant to the epidemic phases. In observing the variations in ORF8 extending to its upstream, we adopted the region toward the upstream of the 5’ end of the ORF8 gene until 27 701 bp in the reference genome to ensure that the region contains the most variable fragments for thorough distinction among these viruses. The phylogenetic tree classified the hosts and phases better than the previous result. The SARS-CoVs isolated from the early, mid, and late phases were divided clearly in the tree established from the extended region.

With regard to the reference genome, the main genomic variations are deletions as shown in Figure 3A. Two hot spots were found in this region. The isolates in the early and mid phases had deletions of 2 different sizes at the 29 nt spot: one with 29 nt deletion (1 of 13 early isolates and 15 of 16 mid isolates), and the other with 82 nt deletion (3 of 13 early isolates and 1 of 16 mid isolates). Among the 60 late isolates, 56 carried deletions on the 29 nt spot, 4 of which also had deletions of varying sizes at the 39 nt spot. Among 60 late isolates, 4 had the deletion of 415 nt spanning both spots. Although the phases were defined roughly according to the key events in the disease breakout, the genomic deletions continued progressing and removed additional nucleotides. Given that ORF8 could play a role in the pathogenesis [6, 26, 27], its deletion might lead to the loss or reduced function of the ORF8 protein. However, the virus survival appeared unaffected by the deletions according to epidemiological evidence [3].

The bat isolates were collected from different areas in China. One main collection site was Hong Kong, and the other collection area was a single cave in Yunnan. The ~430 bp region in these isolates presented 3 types of sequence aligning as shown in Figure 3B. One (type I) resembled the ~430 bp region in 8 isolates in the early phase. The second (type II) appeared similar
to the ~430 bp region in known SARS-CoV-2 isolates and in the coronaviruses from pangolins. However, the pangolin coronaviruses display distinct substitutions within the region. The third type differed from either SARS-CoV or SARS-CoV-2, and only appeared in SARSr-CoV from wild bats. This isolate type could be thought as the wild type (type 0). These suggested that the ~430 bp region might be an ideal region for genotyping SARS-similar coronaviruses. The sequence alignment (Figure 3B) also presented many substitutions scattered in this region, which could be used for finer genotyping.

The civet SARS-CoVs were grouped together beside the human viruses in the phylogeny tree (Figure 3C). Most of the civet SARS-CoVs (16 of 18) were collected at the end of 2003 in Guangzhou where a breakout happened in TDLR and appeared almost identical in the region with the 2 isolates from infected persons in the restaurant. Another 2 civet SARS-CoVs were collected in the late phase before July 2003. Only limited nucleotide substitutions were found among the viruses collected from different time points, thereby suggesting that the viral genome did not change in the intermediate transmission host—that is, palm civet (Figure 3C). However, the distinction was still apparent between the 2 viruses and TDLR viruses.

Figure 3B displays 3 different sequence types of the ~430 bp region, Figure 4A further presents the similarity between type II of bat SARS-CoVs and type II of bat SARS-CoVs. Green bar maps to the 3 SARS-CoVs of type II. Blue bar maps to SARS-CoV-2 strains (not all SARS-CoV-2 strains are put here for space limitations). Pink bar maps to SARS-CoVs from pangolins.
Figure 4. Continued.

ORF8 Extension in SARS-CoV-2 Genotyping • JID 2020:222 (15 July) • 231
I and early-phase SARS-CoVs, and Figure 4B shows the detailed comparison among type II, SARS-CoV-2, and SARSr-CoVs isolated from pangolins, which were suggested to be the transmitting host of SARS-CoV-2. The SARS-CoV-2 isolates and type II SARSr-CoVs from bats were highly similar, showing a 2 nt deletion compared with 5 of 6 SARSr-CoVs from pangolins (Figure 4B). One pangolin isolate was more similar to SARS-CoV-2 with the 2 nt deletion than to other pangolin isolates. According to the alignment in Figure 4B, the bat isolates from Yunnan and Nanjing and the pangolin isolate (EPI-ISL_410721) were most similar to the known SARS-CoV-2. This finding is a clue for further study in tracing the transmission of SARS-CoV-2.

In the work of Tang et al, a single-nucleotide polymorphism was found at 28,144 bp of NC_045512 and was located at site ~330 bp in this ~430 bp region [29]. The C/T polymorphism suggested 2 genotypes. Among 115 human-isolated strains, 95.5% (21 of 22) of those from in Wuhan carried T, compared with the 63.4% (59 of 93) isolated outside Wuhan. Although the sampling was not randomized and the impact of polymorphism was unclear, the mutation in the ~430 bp region suggests its value in the fight against the ongoing pandemic. The sequence alignment of collected 119 SARS-CoV-2 isolates is shown in Supplementary Figure 2.

In conclusion, the various changes exhibited by the extended ORF8 region in SARS-CoV, bat SARSr-CoVs, and SARS-CoV-2 isolates indicated its value for tracing the evolution of SARS-like coronaviruses. We suggest that this region could be used to monitor emerging pathogenic coronaviruses in wild animals and analyze the epidemic trend of ongoing SARS-CoV-2 infection. In addition, the biological function involved in the region may be worth further investigation.

Supplementary Data

Supplementary materials are available at The Journal of Infectious Diseases online (http://jid.oxfordjournals.org/). Consisting of data provided by the authors to benefit the reader, the posted materials are not copyedited and are the sole responsibility of the authors, so questions or comments should be addressed to the corresponding author.

Notes

Author contributions. S.C. analyzed the data, contributed to data acquisition, and wrote parts of the manuscript. X. Z. and J. Z. contrived data analysis methods, analyzed data, contributed to research design, and wrote parts of the manuscript. R. D. and Y. J. contributed to some aspects of the analysis. W. Z. and H. Y. provided advice and technical assistance. Y. Z. and X. L. contributed to data collection and analysis. G. D. contributed to research design and conceived the overall project. All authors commented on the manuscript.

Acknowledgments. We thank all of the scientists who kindly shared their genomic sequences of the 279 coronaviruses used in this study.

Financial support. This work was supported by the National Science and Technology Major Projects of China (grant number 2018ZX103010407).

Potential conflicts of interest. The authors: No reported conflicts of interest. All authors have submitted the ICMJE Form for Disclosure of Potential Conflicts of Interest.

References

1. Peiris JS, Guan Y, Yuen KY. Severe acute respiratory syndrome. Nat Med 2004; 10:888–97.
2. Smith RD. Responding to global infectious disease outbreaks: lessons from SARS on the role of risk perception, communication and management. Soc Sci Med 2006; 63:3113–23.
3. Chinese SARS Molecular Epidemiology Consortium. Molecular evolution of the SARS coronavirus during the course of the SARS epidemic in China. Science 2004; 303:1666–9.
4. Liang G, Chen Q, Xu J, et al. Laboratory diagnosis of SARS. J Microbiol Immunol Infect 2004; 37:177–81.
5. Drexler JF, Corman VM, Drosten C. Ecology, evolution and classification of bat coronaviruses in the aftermath of SARS. Antiviral Res 2014; 101:45–56.
6. Hu B, Zeng LP, Yang XL, et al. Discovery of a rich gene pool for SARS-related coronaviruses provides new insights into the origin of SARS coronavirus. PLoS Pathog 2017; 13:e1006698.
7. Guan Y, Zheng BJ, He YQ, et al. Isolation and characterization of viruses related to the SARS coronavirus from animals in southern China. Science 2003; 302:276–8.
8. Song HD, Tu CC, Zhang GW, et al. Cross-host evolution of severe acute respiratory syndrome coronavirus in palm civet and human. Proc Natl Acad Sci U S A 2005; 102:2430–5.
9. Zhu N, Zhang D, Wang W, et al. A novel coronavirus from patients with pneumonia in China, 2019. N Engl J Med 2020; 382:727–33.
10. Wu A, Peng Y, Huang B, et al. Genome composition and divergence of the novel coronavirus (2019-nCoV) originating in China. Cell Host Microbe 2020; 27:325–8.
11. Zhou P, Yang XL, Wang XG, et al. A pneumonia outbreak associated with a new coronavirus of probable bat origin. Nature 2020; 579:270–3.
12. Wu F, Zhao S, Yu B, et al. A new coronavirus associated with human respiratory disease in China. Nature 2020; 579:265–9.
13. Sievers F, Higgins DG. Clustal Omega for making accurate alignments of many protein sequences. Protein Sci 2018; 27:135–45.
14. Sievers F, Wilm A, Dineen D, et al. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. Mol Syst Biol 2011; 7:539.
15. Waterhouse AM, Procter JB, Martin DM, Clamp M, Barton GJ. Jalview Version 2—a multiple sequence alignment editor and analysis workbench. Bioinformatics 2009; 25:1189–91.
16. Kumar S, Stecher G, Li M, Knyaz C, Tamura K. MEGA X: Molecular Evolutionary Genetics Analysis across computing platforms. Mol Biol Evol 2018; 35:1547–9.
17. Li W, Shi Z, Yu M, et al. Bats are natural reservoirs of SARS-like coronaviruses. Science 2005; 310:676–9.
18. Lau SK, Woo PC, Li KS, et al. Severe acute respiratory syndrome coronavirus-like virus in Chinese horseshoe bats. Proc Natl Acad Sci U S A 2005; 102:14040–5.
19. Ge XY, Li JL, Yang XL, et al. Isolation and characterization of a bat SARS-like coronavirus that uses the ACE2 receptor. Nature 2013; 503:535–8.
20. Li W, Moore MJ, Vasilieva N, et al. Angiotensin-converting enzyme 2 is a functional receptor for the SARS coronavirus. Nature 2003; 426:450–4.
21. Wu Z, Yang L, Ren X, et al. ORF8-related genetic evidence for Chinese horseshoe bats as the source of human severe acute respiratory syndrome coronavirus. J Infect Dis 2016; 213:579–83.
22. Babcock GJ, Eshshaki DJ, Thomas WD Jr, Ambrosino DM. Amino acids 270 to 510 of the severe acute respiratory syndrome coronavirus spike protein are required for interaction with receptor. J Virol 2004; 78:4552–60.
23. Cui J, Li F, Shi ZL. Origin and evolution of pathogenic coronaviruses. Nat Rev Microbiol 2019; 17:181–92.
24. Ren W, Li W, Yu M, et al. Full-length genome sequences of two SARS-like coronaviruses in horseshoe bats and genetic variation analysis. J Gen Virol 2006; 87:3355–9.
25. Liu L, Fang Q, Deng F, et al. Natural mutations in the receptor binding domain of spike glycoprotein determine the reactivity of cross-neutralization between palm civet coronavirus and severe acute respiratory syndrome coronavirus. J Virol 2007; 81:4694–700.
26. Chen CY, Ping YH, Lee HC, et al. Open reading frame 8a of the human severe acute respiratory syndrome coronavirus not only promotes viral replication but also induces apoptosis. J Infect Dis 2007; 196:405–15.
27. Keng CT, Choi YW, Welkers MR, et al. The human severe acute respiratory syndrome coronavirus (SARS-CoV) 8b protein is distinct from its counterpart in animal SARS-CoV and down-regulates the expression of the envelope protein in infected cells. Virology 2006; 354:132–42.
28. Wang M, Yan M, Xu H, et al. SARS-CoV infection in a restaurant from palm civet. Emerg Infect Dis 2005; 11:1860–5.
29. Tang X, Wu C, Li X, et al. On the origin and continuing evolution of SARS-CoV-2 [manuscript published online ahead of print 3 March 2020]. Natl Sci Rev 2020. doi:10.1093/nsr/nwaa036.