Current understanding of the origin, molecular biology and continuing evolution of severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2)

Mohamad S. Hakim¹, Luthvia Annisa¹, Endah Supriyati², Edwin W. Daniwijaya¹, Rakhmat A. Wibowo³, Eggi Arguni¹, Titik Nuryastuti¹

¹Department of Microbiology, ²Centre of Tropical Medicine, ³Department of Physiology, ⁴Department of Child Health, Faculty of Medicine, Public Health and Nursing, Universitas Gadjah Mada, Yogyakarta, Indonesia

ABSTRACT

Recent outbreaks of human coronaviruses, officially named severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), have put health authorities worldwide on a high alert. Firstly emerged in the city of Wuhan, China, SARS-CoV-2 infection is rapidly escalating into a global pandemic. It is first thought as the result of a zoonotic transmission event, similar to the previous epidemic of coronaviruses. However, a continuously increasing number of confirmed cases indicates that the virus gains capacity of efficient human-to-human transmission. Soon after the pandemic is arising, many efforts are focused on identifying the origin of SARS-CoV-2 infection in the human population. Current evidence suggests that the virus is probably derived from bat or pangolin coronaviruses as the natural host. Whether intermediate host(s) exist in the transmission cascade from bat or pangolin to humans is, to a great extent, elusive. This information is essential as the basis for infection prevention and control measures. In this review, we discuss our recent understanding of SARS-CoV-2 biology, highlighting its origin and molecular evolution.

Keywords: bats; evolution; origin; pangolin; SARS-CoV-2;
INTRODUCTION

Coronaviruses (CoVs) are a large family of RNA viruses infecting humans and many different animal species, including bats, swine, camels, cats, dogs, and other wildlife animals. In animals, CoVs may cause a variety of respiratory, gastrointestinal, hepatic, and neurologic infections with different clinical severity. For a long period, it has been recognized four strains of human CoVs (HCoV-NL63, HCoV-229E, HCoV-OC43, and HKU1), which cause only mild upper respiratory infections in immunocompetent individuals with unknown fatalities. Therefore, they are first considered as low pathogenic viruses in humans. However, the emergence of two novel human CoV strains responsible for worldwide outbreaks, i.e., severe acute respiratory syndrome CoV (SARS-CoV) (2002) and Middle East respiratory syndrome CoV (MERS-CoV) (2012), highlights the characteristics of CoV as highly pathogenic viruses that potentially cause lethality in the human population. More than 8,000 cases were identified during the SARS-CoV outbreak firstly emerged in Guangdong province, China, with about 774 fatalities (±10% case fatality rate, CFR). Since then, SARS-CoV has been successfully contained and eradicated. In 2012, MERS-CoV outbreak firstly emerged in Jeddah, Saudi Arabia, more than 2,400 cases were detected with ±34% CFR. In contrast to SARS-CoV, MERS-CoV continuously re-emerges through sporadic cases both in the community and in the hospital.

Less than 20 years after the first deadly SARS-CoV outbreak, another crisis of viral pneumonia was recently emerged in the city of Wuhan, China. The third highly pathogenic human CoV, officially named as SARS-CoV-2, has been quickly identified as the causative agent of this rapidly spreading coronavirus disease 2019 (COVID-19). SARS-CoV-2 has ±80% genomic similarity with SARS-CoV, hence its name. Both viruses have also been suggested to evolve from SARS-related CoV (SARSr-CoV) circulating in bats and employ the similar cell entry receptor (angiotensin-converting enzyme II, ACE2). Despite these similarities, the pandemic profile is different, in terms of its level of adaptation, virulence and transmissibility. It is currently estimated that the average basic reproduction number ($R_0$) of SARS-CoV-2 is 3.28 (median 2.79), indicating that the confirmed cases will continuously increase. About four months since the beginning of the outbreak, by 26 April 2020, COVID-19 has been confirmed in more than 2.8 million cases worldwide. Moreover, 193,710 deaths were reported in more than 200 countries and territories. We here discussed our recent understanding of SARS-CoV-2 biology, highlighting its origin and molecular evolution.

DISCUSSION

The molecular biology of SARS-CoV-2

CoVs are enveloped, single-strand, positive-sense RNA viruses of approximately 26-32 kilobases (kb) in length, and thus it is known as the largest RNA virus. In addition, 5'-capped and poly (A) tail are present on the 5'- and 3'-end of the genome, respectively. CoVs form spherical viral particles of 100-160 nm in diameter. The viral membrane is composed of the spike (S) glycoprotein, the membrane (M) glycoprotein, and the envelope (E) protein, surrounding a flexible nucleocapsid (N) (FIGURE 1A). CoVs belong to the family of Coronaviridae, within the subfamily of Coronavirinae. It is then further divided into Alpha-(α), Beta-(β), Gamma-(γ), and Delta-(δ) coronavirus based on serologic, genomic structures, and phylogenetic relationships. The four strains of “low pathogenic” coronavirus belong to the genus of Alphacoronavirus subgroup.
1b (HCoV-NL63 and HCoV-229E) and Betacoronavirus subgroup 2a (HCoV-OC43 and HCoV-HKU1). Interestingly, all three strains of highly pathogenic CoVs (SARS-CoV, MERS-CoV, and SARS-CoV-2) cluster together within the subgroup 2b of Betacoronavirus genus.14

FIGURE 1. A. Coronavirus viral particles consist of four structural proteins, i.e., spike (S), envelope (E), membrane (M), and nucleocapsid (N). The S protein forms a layer of glycoproteins that protrude from the envelope. B. A large portion of the genome of SARS-CoV, MERS-CoV, and SARS-CoV-2 is made up of a single open reading frame 1ab (orf1ab). This gene encodes a “giant” polyprotein (pp), pp1ab, which is further cut off by two viral-derived proteases to release 15 non-structural proteins (nsp, nsp1-nsp10 and nsp12-nsp16).15 Another orf (orf1a) is also generated by a programmed ribosomal frameshifting during translation process in the ribosome. A single nsp (nsp11) is missing in the genome of SARS-CoV-2. The remaining 3’-end of the genome encodes the structural protein S, E, M, and N. Eight (for SARS-CoV and SARS-CoV-2) or five (for MERS-CoV) open reading frames (ORF) encoding so-called accessory proteins are interspersed between and within these structural protein-coding regions.

The genomic structure of SARS-CoV-2

The basic genomic structure of SARS-CoV-2 is similar with the other CoVs within the subgroup 2b of Betacoronavirus genus (FIGURE 1B).14 A large portion of the genome starting from its 5’-end is made up by a single open reading frame 1ab (orf1ab). This gene encodes a “giant” polyprotein, pp1ab, which is further cut off by two viral-derived proteases to release 15 non-structural proteins (nsp, nsp1-nsp10 and nsp12-nsp16).15 Another orf (orf1a) is also generated by a programmed ribosomal frameshifting, encoding pp1a protein which is further cleaved into ten non-structural proteins (nsp1-nsp10).15 Those proteins are mainly involved in replication and transcription of the genome in the early phase of the viral life cycle in the infected cells.16 The remaining 3’-end of the genome encodes the structural protein S, E, M, and N. Eight open reading frames (ORF) encoding so-
called accessory proteins (3a, 3b, p6, 7a, 7b, 8b, 9b, and orf14) are interspersed between and within these structural protein-coding regions. These genes are dispensable for virus replication, yet they are pivotal for viral-host interactions, including counteracting the host innate immunity. Analysis of SARS-CoV-2 published sequences revealed that its genome shares 79.5% similarity to former SARS-CoV.

**The host receptor usage of SARS-CoV-2**

It has been shown that SARS-CoV-2 employs ACE2 as the cell entry receptor, similar to SARS-CoV. SARS-CoV-2 initial attachment to target cells is mediated by binding S protein to ACE2. Thus, the S glycoprotein plays an essential role in determining viral tropism, cross-species transmission, and infectivity. The S glycoprotein is structurally composed of two subunits, S1 and S2. The S1 subunit contains a receptor-binding domain (RBD) responsible for engagement to ACE2 located on the upper part of S protein. Indeed, RBD is the most variable part of the CoV genome. This phenomenon tends to be common for the virus in general since it is the region that obtains continuous evolutionary pressure from the host immune system. At the lower part of S protein, there is a relatively conserved S2 sub-unit that function as fusion machinery with the cellular membrane, and thus, it is important for entry process. Noteworthy, SARS-CoV-2 S protein-mediated cell entry was potently blocked by murine-derived SARS-CoV S polyclonal antibodies. This mechanism indicates that targeting the conserved region of S protein by cross-neutralizing antibodies is one of the potential strategies to develop vaccines against SARS-CoV-2.

As previously mentioned, engagement of S protein to ACE2 is crucial for entry process, like a lock and key mechanism. However, this engagement is not sufficient for viral entry. Another second feature is also needed, i.e., proteolytic cleavage event executed by cellular protease, transmembrane protease serine 2 (TMPRSS2). This protease cleaves the RBD of S protein to separate from the fusion domain. This event drives fusion of viral membrane with the host's cellular membrane. It was reported that constitutive expression TMPRSS2 enhanced susceptibility of VeroE6 cell line to SARS-CoV-2 infection. Indeed, the S protein of SARS-CoV-2 is a classic class I fusion protein, similar to hemagglutinin (HA) protein of influenza virus. Previous mutational analysis demonstrated that six amino acid residues are critical for binding to human ACE2. Surprisingly, five substitutions were identified among those six residues in SARS-CoV-2 genome compared with SARS-CoV. Structural studies and biochemical experiments showed that the RBD of SARS-CoV-2 binds with high affinity to human ACE2, although this interaction is not optimal. These studies suggest that S protein of SARS-CoV-2 is most likely naturally selected on humans or other species with high homology to human ACE2. Other features of S protein of SARS-CoV-2 that result in more optimal binding to human ACE2 remain to be elucidated to give insight into its capacity of efficient human-to-human transmission.

Another unique feature SARS-CoV-2 spike glycoprotein is the acquisition of a polybasic cleavage site (PRRAR), which allows more effective cleavage by another unidentified cellular protease(s). This prominent feature was not found in SARS-CoV and SARSr-CoV. This cleavage site is located at the junction of S1 and S2 subunits. Previous studies in influenza virus found that acquisition of this feature converts low-pathogenic into high-pathogenic influenza viruses. Current pandemic situations suggest that
SARS-CoV-2 is less pathogenic yet more transmissible than SARS-CoV and MERS-CoV. Therefore, further experiments are required to investigate the impact of the polybasic cleavage site on SARS-CoV-2 pathogenicity and transmissibility.

**SARS-CoV-induced ACE2 downregulation**

ACE2 is widely expressed in the gastrointestinal tract, heart, kidney, and lung, especially in type II alveolar cells. It is also found on the oral cavity mucosa, particularly in the epithelial cells of the tongue. Its counterpart, ACE, is widely expressed on the surface of endothelial cells, particularly in the lungs. In cooperation with renin produced by the kidney, ACE and ACE2 are involved in the renin-angiotensin system (RAS). ACE mediates the conversion of angiotensin I to angiotensin II, which then induces the vasoconstriction of the blood flow. In contrast, ACE2 generates angiotensin from angiotensin II with a vasodilation effect. Thus, ACE and ACE2 regulate the homeostatic balance of vasoconstriction and vasodilation of the vascular system.

It has been suggested that patients who are treated with ACE inhibitors (ACEIs) and angiotensin receptor blockers (ARBs) may be at increased risk of severe SARS-CoV-2 infection. This is because both ACEIs and ARBs increase the expression of ACE2 in the cardiopulmonary circulation. However, it has been shown that ACE2 and angiotensin play a protective role from lung injury. Downregulation of pulmonary ACE2 was associated with acute lung injury, as characterized by increased vascular permeability, lung edema, inflammation, and reduced lung function. These phenomena reflect the dysfunction of the RAS system due to over accumulation of angiotensin II.

The S protein of SARS-CoV downregulates ACE2 expression following initial attachment and fusion of the viral membrane, and thereby contributes to severe lung injury. Therefore, ACE2-associated lung injury has been suggested to be involved in SARS-CoV pathogenesis. This SARS-CoV-induced downregulation of ACE2 can be reversed by angiotensin receptor blocker (ARB) treatment that is commonly used in hypertension patients. It is also suggested that administration of soluble form of ACE2 may competitively bind to S protein to neutralize the virus and also rescue the cellular ACE2 levels. However, limited clinical data are available about the efficacy of ARBs and ACEIs in the treatment of SARS-CoV-2-induced lung injury. Since SARS-CoV-2 employs ACE2 as a host cell entry receptor, it is possible that ACE2 is involved in the inflammatory response of SARS-CoV-2. However, the role of ACE2 in the pathogenesis of COVID-19 is poorly understood. Due to these conflicting mechanisms, careful pharmacoepidemiologic studies are urgently needed to investigate whether RAS inhibition improves or worsens the outcomes of COVID-19 patients.

**Evolution of SARS-CoV-2 and cross-species transmission**

**Bat coronaviruses as the origin of SARS-CoV and MERS-CoV**

Bats are well-known to harbor highly pathogenic, emerging, or re-emerging RNA viruses posing a great threat to human health, including bat lyssaviruses (Rabies virus) and henipaviruses (Hendra and Nipah viruses). Therefore, it is not surprising that bats have also been recognized as the natural reservoir of most, if not all, mammalian CoVs, including SARS-CoV and MERS-CoV. Earlier investigation of SARS outbreak (2003) identified SARS-CoV and anti-SARS-CoV antibodies in the masked palm civet (*Paguma larvata*). However, a large scale identification of both farmed and wild civets found that other animals were the source of SARS-CoV.
transmission to civets.\textsuperscript{40,41} Besides, civets are not known as the natural host of CoVs. Thereafter, SARSr-CoV was identified in horseshoe bats (genus \textit{Rhinolophus}), indicating a natural origin of human SARS-CoV and that civets are the most likely intermediate host for SARS-CoV.\textsuperscript{42} It has been hypothesized that SARS-CoV in civets acquires new mutations before cross-transmission to humans.\textsuperscript{4} SARS-CoV eradication from the intermediate animal reservoir (civets) is among one measure for a successful SARS-CoV containment and eradication during SARS-CoV pandemic in 2003.\textsuperscript{43}

Epidemiologic investigations of the early cases of MERS-CoV revealed that most patients had contact histories with dromedary camels, suggesting zoonotic events.\textsuperscript{2} Phylogenetic analysis of MERS-CoV derived from human cases and dromedary camels showed that they were almost identical in sequence, indicating that humans and camels were infected with the same source of MERS-CoV.\textsuperscript{2} Interestingly, anti-MERS-CoV antibodies were detected in archived serum samples of camels collected in 1983.\textsuperscript{44} This suggests that the spillover event to camel already occurred at least 30 years before the MERS outbreak. This study also suggests that MERS-CoV had circulated and adapted in camels for 30 years before it acquired unique characteristics that enable an efficient transmission to humans. More than ten bat species from two bat families, \textit{Vespertilionidae} and \textit{Nycteridae}, were found to harbor MERS-related CoVs (MERSr-CoVs), suggesting the original ancestor of human MERS-CoV.\textsuperscript{2}

### Bat or pangolin CoVs as the origin of SARS-CoV-2

Phylogenetic analysis of SARS-CoV-2 and various CoVs from different hosts demonstrated that SARS-CoV-2 belong to the \textit{Betacoronavirus} genus.\textsuperscript{45} Thus, it is classified in the subgenus \textit{Sarbecovirus} of the \textit{Betacoronavirus} genus.\textsuperscript{6,14} Based on molecular clock analysis, its most recent common ancestor emerged around at the end of November 2019.\textsuperscript{46} At the amino acid levels, the similarity between SARS-CoV-2 and former SARS-CoV is 76.47%.\textsuperscript{45}

It is logical to suggest that bats are the natural hosts of SARS-CoV-2 based on its similarity with former SARS-CoV. Indeed, initial sequence analyses demonstrated that SARS-CoV-2 was closely related to bat-derived SARSr-CoVs (bat-SL-CoV-ZC45 and bat-SL-CoV-ZXC21) originally identified in the city of Nanjing, China.\textsuperscript{6,47,48} However, SARS-CoV-2 formed a monophyletic cluster different from these two CoV strains, suggesting that they are not the direct ancestors of SARS-CoV-2.\textsuperscript{6,47,48} In addition, it has been suggested that SARS-CoV-2 is a recombinant virus between bat CoVs and unknown CoV. This recombination event probably occurs in S protein recognizing the cell-surface receptor.\textsuperscript{47} All these findings suggest that SARS-CoV-2 share similar genetic information with bat CoVs and that bat SARS-like CoVs as the probable origin of SARS-CoV-2.\textsuperscript{46}

It was subsequently identified that bat CoV strain RaTG13 had 96.2% overall genome sequence identity with SARS-CoV-2.\textsuperscript{49} Bat CoV RaTG13 was found from \textit{Rhinolophus affinis} bat
sampled from Yunnan Province, China. However, bat CoV RaTG13 differs in the RBD with SARS-CoV-2. It shares 89% identity at amino acid level with SARS-CoV-2. Importantly, four out of five key amino acid residues in the RBD of bat CoV RaTG13 were different with SARS-CoV-2. These suggest that S protein of bat CoV RaTG13 may not bind efficiently to human ACE2. Thus indicating that they are not the direct ancestors of SARS-CoV-2.

The second parental virus predicted to be the origin of SARS-CoV-2 is CoVs identified in Malayan pangolins (Manis javanica) obtained during anti-smuggling operations in Guangdong and Guangxi, named as pangolin-CoVs. Pangolins are among the sources of meat in a certain region in China. At the whole genome level, pangolin-CoV shares 91.02% sequence similarity with that of SARS-CoV-2. However, its RBD was predicted to interact with human ACE2. Surprisingly, five key amino acids in RBD are similar between pangolin-CoV and SARS-CoV-2, while RaTG13 only has one identical residue. Since the sequence divergence at the whole genome level and also the absence of polybasic cleavage sites in pangolin-CoV, it is not likely that pangolin-CoVs are the intermediate host of SARS-CoV-2. However, considering the fact that CoV infection in Manis javanica results in poor conditions, in contrast to bats, it is also possible that pangolins serve as the intermediate and not the natural host for SARS-CoV-2. A deeper study is urgently needed to clarify this issue.

Altogether, it is most likely that bats or pangolins serve as natural hosts of the newly identified SARS-CoV-2. It is possible that the ancestral of SARS-CoV-2 directly cross to humans and acquires further adaptation (i.e., acquisition of polybasic cleavage site) in human population. There is also a possibility of the presence of unknown and unidentified intermediate hosts in the transmission cascade from bats or pangolins to humans. Epidemiologic investigations of pneumonia outbreaks in China showed that most of the initial patients (66%) were directly exposed to wildlife animals at the Huanan seafood market. This market sells fish and wildlife animals, including marmots and snakes. However, bats CoVs are not thought to infect humans directly since most bats in Wuhan are in a hibernating period in the late December and no bats were sold at Huanan market. Thus, intermediate host(s) may exist between bats or pangolin and humans (FIGURE 2). An early study employing relative synonymous codon usage (RSCU) analysis showed that SARS-CoV-2 has similar synonymous codon usage with snakes, suggesting its potential as the intermediate host. However, this conclusion was disputed by another study. Besides, SARSr-CoVs have not identified yet in snakes. Therefore, there is no solid evidence up to this moment that snakes or other wildlife animals are responsible as intermediate hosts for SARS-CoV-2. Therefore, the identification of animal reservoirs serves as intermediate host(s) of SARS-CoV-2 is essential to understand its spill over to human populations and to formulate proper containment measures. Possibilities of another source of infection are likely since a number of initial patients had no contact history with Huanan market.
FIGURE 2. Model of evolution and cross-species transmission of coronaviruses from bats. Five out of seven human coronaviruses are known to derive from bat coronaviruses. Bat-derived recombinant viruses are shown in mixed colors. Each color represents the ancestral virus circulating in the natural reservoir (bats). These recombinant viruses potentially lead to the cross-species transmission to intermediate hosts and finally to humans as end hosts. For SARS-CoV-2, whether pangolin serves as the intermediate host is unclear. Two other human coronaviruses (HCoV-OC43 and HCoV-HKU1) are likely originated from rodents (not shown). Severe acute diarrhea syndrome CoV (SADS-CoV) causes large outbreaks and fatal diseases in pigs, with no known transmission to the human population.

The evolution of SARS-CoV-2 in human populations

A chronological set of sequence of SARS-CoV strains isolated during the early, middle, and late phases of epidemic demonstrated continuous and dynamic mutational changes of SARS-CoV following epidemic expansion in 2003. It is generally estimated that mutation rates for RNA virus are about $10^{-3}$ to $10^{-5}$. Surprisingly, in contrast to other RNA viruses, SARS-CoV possesses 3'-5' exonuclease activity encoded by nsp14. This nsp14-encoded protein serves as RNA proof-reading activity and consequently, lowering the error possibility during replication of RNA genome. Therefore, it is generally estimated that the mutation rates of SARS-CoV and other CoVs were about $2 \times 10^{-6}$. However, recent analysis suggests that the evolutionary rate of SARS-CoV-2 is $6.58 \times 10^{-3}$ substitution site per year. This suggests that although the virus does not frequently mutate, the high number of human infections has provided sufficient opportunity for such events.

Indeed, sequence analysis of different SARS-CoV-2 strains collected from various countries showed the presence of many mutations and deletions in the genome, indicating a rapid evolution of the virus during this current pandemic. Phylogenetic study revealed that SARS-CoV-2 across the world have evolved into three variants (A, B and C variants) based on amino acid differences, suggesting a dynamic evolution of SARS-CoV-2 following introduction to human population. All these results suggest that SARS-CoV-2 continuously evolves following its first introduction to the human population. Consequently, this event may influence its virulence, infectivity, and transmissibility.
CONCLUSIONS AND FUTURE PERSPECTIVES

The emergence of SARS-CoV-2 infection, the responsible agent for COVID-19 pandemic, has put countries worldwide, on a high alert. Studies have been published to delineate the origin as well as the molecular characteristics of this newly emerging virus. Much information has been gained, yet more details need to be clarified. Sequence analysis of SARS-CoV-2 genome has shown that it is likely originated from bat or pangolin coronaviruses. However, how this cross-species transmission occurs, and further adaptation in humans requires further studies. Close and continuous monitoring of the virus during this pandemic crisis is highly required to keep track of the virus’ mutation and adaptation in humans as well as its infectivity, pathogenicity, and transmissibility. Finally, studies of animal coronaviruses, especially bat coronaviruses, need to be continued in parallel for early detections and warning signs for the possibilities of future outbreaks.

ACKNOWLEDGMENTS

The authors thank Talitha Rahmawati and Risky Oktriani for critical reading of this manuscript. The authors also thank Adam Ash Shidiq for preparing the figures.

REFERENCES


27. Patel AB, Verma A. COVID-19 and Angiotensin-converting enzyme inhibitors and angiotensin receptor...
https://doi.org/10.1001/jama.2020.4812

https://doi.org/10.1093/jtm/taaa041

https://doi.org/10.1038/nature03712

https://doi.org/10.1038/ncomms4594

https://doi.org/10.1113/expphysiol.2007.040048

https://doi.org/10.1038/nm1267

https://doi.org/10.1093/cid/ciaa329

https://doi.org/10.1007/s12250-020-00207-4

https://doi.org/10.12688/f1000research.22211.2

https://doi.org/10.1042/CS20200163

https://doi.org/10.1080/22221751.2020.1746200

https://doi.org/10.3390/v11030210

https://doi.org/10.1126/science.1087139

https://doi.org/10.1128/JVI.79.18.11892-11900.2005

https://doi.org/10.3201/eid1012.040520


