Molecular Evolution of the Novel Coronavirus SARS-CoV-2

Emily C. Vook[1], Jeffrey O. Henderson[1,*]

[1]Department of Science and Mathematics, Judson University, Elgin, IL 60123, USA

Abstract

The coronavirus infectious disease (20)19 (COVID-19) pandemic is caused by a newly identified virus (2019) SARS-CoV-2, a beta coronavirus that shares similarities with other human-infecting coronaviruses. Genomic analysis suggests that SARS-CoV-2 is closely related to SARS-CoV, a bat-related coronavirus, RaTG13, and to other pangolin-associated coronaviruses. The spike protein of coronaviruses are glycoproteins and are responsible for attaching the virus to the host cell and entering. Amino acid changes within the spike protein-encoding gene from SARS-CoV to SARS-CoV-2 enable SARS-CoV-2 to form a stable spike protein, to form a stable complex between the S protein and the receptor ACE2, to increase binding points between the S protein and ACE2, and to survive at higher temperatures. SARS-CoV-2 is zoonotic, with genomic analysis implicating bats as the original host and pangolins as the most likely intermediate host to infect humans. As SARS-CoV-2 infects humans, viral point mutations will continually occur and cause the emergence of new competitive SARS-CoV-2 strains. Two major strains include D614G and N501Y and have increased infectivity and transmission, further complicating the scope of the current COVID-19 pandemic. Vigilant monitoring of viral development and evolution is necessary for developing proper treatment methods and vaccine targets.

Introduction

The coronavirus infectious disease (20)19 (COVID-19) pandemic has afflicted over 120 million and killed nearly 3 million people worldwide. COVID-19 is caused by a newly identified virus (2019) SARS-CoV-2. SARS-CoV-2 is a beta coronavirus that shares similarities with other human-infecting coronaviruses that have resulted in significant pandemics and endemics. SARS-COV-2 shows similarity to nonhuman animal-associated coronaviruses, such as coronaviruses associated with bats and pangolins, suggesting the zoonotic origin and cross-species transmission of SARS-CoV-2 from an original host to a probable intermediate host and finally to humans. Due to the significance of the COVID-19 pandemic, SARS-CoV-2 has been studied extensively. The general background of coronaviruses along with comparative analysis between SARS-CoV-2 and other pertinent coronaviruses implicates evolutionary relationships among coronaviruses, cross-species transmission capabilities, and the evolution of SARS-CoV-2 strains within the human population.

Coronaviruses

Viruses are disease-causing agents known to infect a wide variety of living organisms. Among viruses are coronaviruses which cause diseases in humans as well as other mammalian species and some avian species. Within the category of coronaviruses, SARS-CoV-2 falls in the realm of *Riboiviria,* order *Nidovirales*, suborder *Coronavirinae,* family *Coronaviridae,* and subfamily Orthocoronavirinae[1]. Within Orthocoronavirinae, there are four genera: alpha, beta, gamma, and delta coronaviruses. Gamma and delta coronaviruses primarily infect birds, however, few have been associated with mammalian infections. Alpha and beta coronaviruses are associated with human infections and can also infect other mammals, such as civets, camels, bats, and pangolins. Alpha coronaviruses include HCoV-NL63 and HCoV-229E which cause mild respiratory infections in humans [2]. Beta coronaviruses include SARS-CoV, MERS-CoV, HCoV-HKU1, HCoV-OC43, and the newest emerging SARS-CoV-2. HCoV-HKU1 and HCoV-OC43 also cause mild respiratory infections with SARS-CoV, MERS-CoV, and SARS-CoV-2 causing more severe

respiratory infections [3]. The exact species name of the virus of interest, SARS-CoV-2, is *Severe acute respiratory syndrome-related coronavirus* [1].

Coronaviruses are positive-strand RNA viruses characterized by their outer envelope which gives the virus its "crown-like" appearance, thus contributing to the name "coronavirus." The visible spikes on the envelope are due to glycoproteins which allow the virus to recognize and attach to a specific receptor on a host target and enter the cell. The coronavirus genome is also unique in that it has the largest RNA genome amongst all RNA viruses, averaging 27-32kb [3]. The average overall length of the SARS-CoV-2 genome specifically is reported at 30kb [4]. Coronaviruses additionally encode a 3`-to-5`-exoribonuclease which promotes high-fidelity replication and facilitates fast transmission and proofreading capabilities [5].

The coronavirus genome codes four to five structural proteins including spike (S), membrane (M), envelope (E), nucleocapsid (N), and hemagglutinin-esterase protein (HE) with the SARS-CoV-2 genome specifically expressing S, M, E, and N [3]. The S protein of coronaviruses are glycoproteins and are responsible for attaching the virus to the host cell and enter, therefore determining viral infectivity [6]. S proteins come together to form homotrimers which form the spikes on the envelope. Cleavage occurs at the S1 and S2 subunit boundary by host proteases which induce irreversible conformational changes allowing the virus to fuse to the host cell [7]. This cleavage is initiated by cell surface-associated TMPRSS2 and cathepsin, separating the protein into two domains, S1 and S2 [8]. Unique to SARS-CoV-2 is a furin-like cleavage site at this boundary which, after initial cleavage, is cleaved again by the furin enzyme [9]. The furin site is thought to contribute to the especially high infectivity of SARS-CoV-2 by facilitating priming of the spike protein overall which would improve receptor binding [7, 10]. S1 is responsible for the specific binding of the virus to the host and S2 is responsible for structural support and mediates membrane fusion [3, 11]. The S1 domain includes an N-terminal domain and three C-terminal domains [6]. Within the S1 domain, there is a receptor-binding domain (RBD). The RBD contains a receptor-binding motif, RBM which is the main functional motif in RBD composed of two regions. The

RBM, specifically, makes contact with its appropriate receptor. The RBD of the S1 subunit undergoes hinge-like conformational movements that will shield or expose the RBM from binding to the receptor and thus controls the overall binding that can occur. The binding of the virus to the receptor of the host cell causes the pre-fusion trimer to destabilize and the S1 subunit will be shed and the S2 subunit will transition to a stable post-fusion conformation. After this membrane fusion, the RNA viral genome will be released into the cytoplasm in which important proteins necessary for replication, including pp1a and pp1ab encoding for non-structural proteins and forming the replication-transcription complex, will be translated and allow the viral particle buds to form [8]. The identities of the RBM and the RBD are crucial in determining the transmission and infectivity of a virus. In the case of SARS-CoV-2 as well as SARS-CoV and other bat and pangolin-associated coronaviruses, the receptor that is recognized has been determined to be angiotensin-converting enzyme 2, ACE2. ACE2 and ACE2 orthologs are found in many mammalian species, with SARS-CoV-2 related virus infections found in pangolins, bats, pigs, civets, and humans. ACE2 is mainly found in lung epithelial cells and the small intestine and renal tubules, however, it is also found in heart cells, arterial smooth muscle cells, and within the gastrointestinal and nervous system. ACE2 is typically rare in the circulatory system of the body but is readily expressed within specific organs, especially in the lungs, kidneys, and gastrointestinal tract [8]. The M protein of coronaviruses is abundant and provides the virus with its shape, forming a dimer, and is specifically responsible for the curvature of the membrane, thus further assisting in binding to the nucleocapsid. The E protein of coronaviruses is scarcer, yet importantly contributes to virus assembly and release. Additional implications of the E protein suggest that it has ion channel activity and facilitates pathogenesis. The N protein of coronaviruses is a component of the nucleocapsid and helps package the viral genome as well as facilitating interactions accomplished by the M protein. The HE protein of coronaviruses is only present in HCoV-OC43 and HCoV-HKU1 and directly binds to sialic acid found on the S proteins facilitating viral entry. HE proteins also play a role in the transmission of the virus in mucosa [3].

Generally, coronaviruses are transmitted via respiratory droplets. Direct transmission of the virus occurs through coughing, sneezing, and sputum, but physical contact between droplets and mucous membranes can also transmit it [3, 4, 12, 13]. In humans, the general organ system targeted in SARS-CoV-2 is the respiratory system which contains epithelium cells with ACE2. Additional symptoms associated with organ systems, including the nervous, digestive, and urinary systems, are due to the expression of ACE-2 in their associated cells. The binding of SARS-CoV-2 to ACE2 leads to the downregulation of ACE2 expression and this could disrupt the protective effect ACE2 possesses [8].

General symptoms reported in SARS-CoV-2 infectious cases include fever, cough, myalgia, and malaise[3, 8, 9, 13]. Additional reports of common symptoms include headache, nausea, ageusia (loss of taste), and anosmia (loss of smell). Symptom presence and severity vary widely amongst infected individuals. Symptom severity tends to increase with an individual's age and with the presence of comorbidities (e.g. obesity, diabetes, immunocompromised state, etc.). More severe symptoms include pneumonia, comas, seizures, and organ failures, which may lead to death. Although severity tends to increase in age, there is evidence of younger people without comorbidities developing these more severe symptoms. The expression of ACE2 has a great factor in SARS-CoV-2 transmission and symptom severity. Animal model studies show that the expression of ACE2 tends to increase with age, and therefore, humans can be assumed to follow a similar trend of morbidity and mortality [14]. This increase in ACE2 expression with age reflects the rates of infectivity across age groups that is seen across infected individuals. In addition to this, certain populations are genetically predisposed to expressing more or less ACE2. The Chinese population, for example, express higher levels of ACE2 in tissues, and such populations may be more susceptible to SARS-CoV-2 infection [8, 12]. A final complexity to SARS-CoV-2 cases is that patients may be asymptomatic and possess none of these symptoms.

SARS-CoV-2 and other Coronaviruses

SARS-CoV-2 is a newly discovered virus placed taxonomically as a coronavirus. The similarities between SARS-CoV-2 and other coronaviruses have helped shed

light on the pathogenicity, the efficiency of transmission, and overall characteristics of SARS-CoV-2. The exact evolutionary relationship between SARS-CoV-2 and other coronaviruses remains unknown as genetic distances between these viruses depend on factors such as nucleotide sequence, amino acid sequence, and domain region sequences; however, similarities between SARS-CoV-2 and beta coronaviruses classify it as a beta coronavirus. This classification of SARS-Cov-2 will help researchers and physicians understand its infectivity, transmission, variants, and make breakthroughs in treatment and control of the virus. The Coronaviridae Study Group of the International Committee on Taxonomy of Viruses published a maximum likelihood tree of several coronaviruses, including representative species within the genus *Betacoronavirus* and *Alphacoronavirus*. Of note, SARS-CoV-2 shows the closest relationship to SARSr-CoV RaTG13, a bat-associated coronavirus, and shares a common ancestor with SARS-CoV [1].

There are several coronaviruses known to infect mammals, both human and nonhuman, which include bats, pangolins, and civets as examples. Relevant cases have seen coronaviruses crossing the species barrier from the original host, of a nonhuman mammal, typically a bat for coronaviruses, to an intermediate host, such as civets or camels, and then to humans. The evolutionary relationship between nonhuman animal coronaviruses and human infecting coronaviruses shows that coronaviruses have zoonotic origins [6].

Coronaviruses can be highly pathogenic, and several are associated with human endemics and pandemics with varying degrees of severity [4]. The human endemic coronaviruses include HCoV-229E, HCoV-OC43, HCoV-NL63, and HCoV-HKU1 which contribute to 15-30% of all respiratory infections in humans every year, causing the common cold with mild symptoms [3]. The human pandemic coronaviruses with more significant symptom severity are SARS-CoV, MERS-CoV, and SARS-CoV-2. SARS-CoV appeared in 2002 in the Guangdong province of China, infecting around 8000 and killing almost 800 [4]. MERS-CoV appeared in 2012 in the Arabian Peninsula, infecting around 2500 people and killing over 800 [6, 7, 8]. More similar to SARS-CoV, SARS-CoV-2 was detected in Wuhan, China in December

2019. The current statistics as of March 2021 are over 150 million confirmed infected and over 3 million dead worldwide [15]. From raw data alone, SARS-CoV-2 shows greater transmission, higher infectivity, and increased symptom severity. This current pandemic is more significant because of its more rampant spread throughout countries and because it has lasted over a year. The spread and duration of SARS-CoV-2 could be due to its lower mortality rate at 3% compared to around 10% for SARS-CoV and 35% for MERS-CoV [4, 16]. SARS-CoV-2 is highly similar to other coronaviruses, especially SARS-CoV, RaTG13, and some pangolin CoVs, discussed below; however, the uniqueness of SARS-CoV-2 is demonstrated specifically in the RBD, RBM, spike protein, and the ACE2 complex formed.

SARS-CoV-2 and SARS-CoV genomic analysis demonstrates 79.5% nucleotide identity [6]. SARS-CoV-2 lies in a different subgenus than MERS-CoV (SARS-CoV-2 in *Sarbecovirus* and MERS-CoV in *Merbecovirus*) and has less overall similarity at ~50% [3, 8]. The overall similarity between SARS-CoV-2 and RaTG13, a beta coronavirus that infects bats, is approximately 96% which is significantly higher, however, the 4% difference is at critical genomic features suggesting a non-linear evolutionary path from RaTG13 to SARS-CoV-2 [10, 17]. Phylogenetic analysis reveals that SARS-CoV-2 and RaTG13 are the most closely related [1]. The overall similarity between SARS-CoV-2 and the GD (Guangdong) Pangolin-CoV strain is near 91% [8].

The receptor in which the virus binds to and enters the cell has important implications on infectivity, symptoms, and host susceptibility. It has been demonstrated that SARS-CoV-2 and SARS-CoV bind to the same receptor, ACE2 for host cell entrance [4]. RaTG13 and GD Pangolin-CoV also bind ACE2 and are predicted to recognize human ACE2 [8, 18]. Conversely, MERS-CoV uses the receptor dipeptidyl peptidase 4 (DDP4) which is commonly found on T cells [4]. HCoV-OC43 and HCoV-HKU1 bind to 5-N-acetyl-9-O-acetylsialosides that are on glycoproteins and glycolipids [7]. The similarity in the receptor of SARS-CoV-2, SARS-CoV, and the bat and pangolin coronaviruses can be explained by the similarity of the RBM and RBD. The slight differences that are found here, however, contribute heavily to differences in infectivity and

transmission of SARS-CoV-2.

The spike protein, the protein responsible for interactions between the virus and the host, has been analyzed and compared between coronaviruses. Differences (mutations) in the spike protein are typically found within this region because this protein contributes heavily to the virus' infectivity and transmissibility. Therefore, positive selection would favor beneficial mutations that would persist, leading to new strains or new species such as SARS-CoV-2. Therefore, most divergence arises in the S gene as the high mutation rate of RNA viruses and natural selection will favor nonsynonymous substitutions that confer greater infectivity and therefore survivability of the virus [17]. As mentioned earlier, the spike protein consists of two subunits, S1 and S2, with the S1 subunit containing the receptor-binding domain, RBD. The RBD is the sequence of the virus spike protein that is responsible for forming the complex between the spike protein and ACE2 with this domain consisting of approximately 193 amino acid residues[3, 6, 8, 9, 17, 18, 19]. The interaction occurring between the RBD of the spike protein and ACE2, therefore, is a determining factor of functionality.

Overall, the spike protein of SARS-CoV-2 shares 72-78% sequence identity with SARS-CoV [7, 19]. The RBD of S1 was 64% similar and the S2 fusion domain was 90% similar [8]. The receptor-binding motif shows a 50-53% similarity from SARS-CoV-2 to SARS-CoV [3]. Although some high similarity is seen in sequence identity the differences are great and impact functionality.

The RBD domain of S1 proteins of SARS-CoV-2 and SARS-CoV exhibit an overall sequence similarity of 89.2% with a sequence identity of 73.7%. The binding sites on the RBD proteins are highly conserved with corresponding residues having a sequence similarity of 83.3%. Both SARS-COV-2 and SARS-CoV have 13 hydrophobic residues at the binding site which allow for important protein-protein interactions [6]. The differences seen between SARS-CoV-2 and SARS-CoV spike protein affect the overall binding that takes place with ACE2 and thus affects the complex that is formed. Although the structure of the complex is structurally similar, there are specific contact points that differ. Subtle changes in the amino acids present may benefit SARS-CoV-2 and allow

for improved attachment and host entry. Lan and coworkers (2020) [9] elucidated amino acid conservation and the interaction between the two homologous proteins of SARS-CoV-2 and SARS-CoV and the ACE2 receptor. They found several amino acid mutations that did not affect biochemical properties within the spike protein. In comparing the RBD of SARS-CoV and SARS-CoV-2, there is a change in the positions of six amino acids. In SARS-CoV, four of these amino acid positions are critical in forming the bond between the spike protein and ACE2. At these four positions in SARS-CoV-2, three changes confer improved binding. The first change results in additional interaction points between the spike protein and several ACE2 amino acid positions. The second change is a single point mutation allowing for the formation of an extra hydrogen bond between the spike protein and ACE2 that is not present in SARS-CoV. The third change is in the RBM and SARS-CoV-2 has a unique residue that forms an additional salt-bridge with ACE2. Collectively, specific amino acid changes from SARS-CoV to SARS-CoV-2 most likely contribute to a greater binding affinity of SARS-CoV-2 to the ACE2 receptor with the increase in contact points [9].

He and colleagues (2020) [6] calculated the binding free energies of SARS-CoV-2 and SARS-CoV. The binding free energy (ΔG) between the interaction of SARS-CoV-2 RBD and human ACE2 was demonstrated to be -50kcal/mol, and the binding free energy between SARS-CoV RBD and human ACE2 was -36.75 kcal/mol. The lower value found with the RBD of SARS-CoV-2 suggests that SARS-CoV-2 has a higher binding affinity to ACE2 than SARS-CoV. The solvation energy contribution, ΔG_{solv}, has an impact on the overall binding affinity with SARS-CoV-2 calculated to be 674.97 kcal/mol with SARS-CoV at 696.56 kcal/mol. The binding free energy in a vacuum, ΔG_{gas}, adds to the total value with SARS-CoV-2 having a higher value at -725.41 and SARS-CoV at -733.31. These values indicate that in water, SARS-CoV-2 will bind to ACE2 more efficiently than SARS-CoV, and in gas, SARS-CoV will bind to ACE2 more efficiently than SARS-CoV-2 [6].

Biochemical analysis of spike protein stability was carried out [6] in SARS-CoV-2 and SARS-CoV, a factor that would determine the efficiency of attachment to

host cells and membrane fusion. For the spike protein overall, the stability was compared between SARS-CoV-2 and SARS-CoV. The G-total of the spike protein of SARS-CoV-2 was calculated to be-67,303.28 kcal/mol in contrast to the G-total of SARS-CoV at -63,139.96 kcal/mol. These results suggest that the spike protein of SARS-CoV-2 is generally more stable and would be able to persist at a higher temperature than the spike protein of SARS-CoV. This may imply that evolution and adaptation favor lower free energies as this would allow SARS-CoV-2 to survive in a variety of hosts with varying internal temperatures. This also may support the proposal of SARS-CoV-2 originating in bats that have a higher body temperature than humans, thus explaining why SARS-CoV-2 can survive at these higher temperatures. The free energy in a vacuum for the spike protein was calculated to be -36,405.44 kcal/mol for SARS-CoV-2 and -32,053.43 kcal/mol for SARS-CoV, with solvation energy showing similar values with SARS-CoV-2 at -30,897.84 kcal/mol and SARS-CoV at -31,086.53 kcal/mol. Amino acid changes account for the differences in inter-residue interactions. He and coworkers (2020) further suggest that it is more favorable from an evolutionary and adaptation standpoint to select for beneficial internal reactions rather than selecting for solvation energy [6].

Examination of the differences in free energies between the RBD protein of the spike protein in SARS-CoV-2 and SARS-CoV. Similar to the trend seen across these analyses, SARS-CoV-2 shows lower free energy at -4,090.04 kcal/mol compared to SARS-CoV with higher free energy at -3,617.73 kcal/mol. The free energy for SARS-CoV-2 in a vacuum is -2,104.37 kcal/mol and for SARS-CoV is -1,703.66 kcal/mol. Solvation energy is 1,985.68 kcal/mol for SARS-CoV-2 and at -1,914.07 kcal/mol for SARS-CoV. The lower solvation energy in SARS-CoV-2 can be explained by the binding of the RBD to ACE2 in humans. RBD moves away from the spike protein itself and into water to bind human ACE2 so it being more soluble in water would allow for easier movement and therefore, higher infection [6].

In the comparison of SARS-CoV-2 to bat-related coronaviruses, the spike protein amino acid identity ranges from 75-97%. Bat coronavirus, RaTG13,

specifically shows the highest overall similarity at 97% with the S1 subunit at 96% similar and S2 at 100%. Sequence identity studies specifically indicate that the spike protein overall has a 93.1 % similarity between SARS-CoV-2 and RaTG13 [9]. The RBD of SARS-CoV-2 and RaTG13 are approximately 85-89% similar and share one of the six critical amino acid residues influencing ACE2 binding [8, 10].

In the comparison of SARS-CoV-2 and pangolin-related coronaviruses, the spike protein shows a similarity of 92% for GX Pangolin-CoV (Guanxi) and 89% for GD Pangolin-CoV (Guangdong) with the most divergence found in the S1 subunit. The RBD of SARS-CoV-2 and GD Panoglin-CoV show a higher similarity at 97% at the amino acid level and share all six critical amino acid residues influencing ACE2 binding [10, 19]. GX Pangolin-CoV shows more difference at only 87% identity similarity [8].

With overall high genetic similarity between SARS-CoV-2 and these other coronaviruses, the subtle difference in sequences has caused an apparent difference in affinity, solubility, and binding of the spike protein to ACE2, all related to increasing infectivity and transmission. Genomic similarity further indicates the zoonotic nature of SARS-CoV-2, a characteristic that adds to the complexity of this pandemic.

Zoonotic Nature of SARS-CoV-2

Coronaviruses are known to cross species barriers because RNA viruses have a high mutation rate [17] allowing adaptation to new animal niches. Coronaviruses, however, do not have as high of a mutation rate as retroviruses due to their 3`-to-5` exoribonuclease proof-reading activity. Coronaviruses compensate for the low mutation rate by having a very high rate of virion replication within hosts which results in clonal expansion of multiple mutational variants [10]. Mutations within the RBD of the spike protein, which contribute heavily to transmission and infectivity, will be conserved. SARS-CoV-2 has already been demonstrated to have a more stable spike protein which forms a stronger complex with ACE2, a lower free binding energy, and other beneficial characteristics which have all been due to subtle nucleotide and amino acid changes. There is also an added furin site which may contribute to increased infectivity as well. In reflecting

on this, such subtle mutations may allow for survivability and efficient binding within a variety of different hosts [6]. Along with mutations, viral recombination additionally increases diversity by reassorting genomic segments which could lead to the subtle, beneficial changes mentioned above. Recombination and mutations often persist within the S gene encoding the S protein as changes here could lead to recognition of different receptors and may allow for different host susceptibility [2, 8]. Mutations can also occur during these recombination events between two different viruses with the phenotypically beneficial mutations lasting [12]. A second reason is due to the virus binding to a receptor that is found in several animal species. SARS-CoV-2, as well as SARS-CoV and bat and pangolin-associated coronaviruses, bind to ACE2 which is found in several vertebrate species including humans, bats, pangolins, civets, and snakes. Therefore, the coronaviruses that utilize ACE2, or other widespread receptors, would most likely be able to bind to ACE2 found in several different species. This ability, however, is contingent on the virus's ability to survive in the host body as factors like internal temperature and environment can make transmission more or less difficult. For example, SARS-CoV-2, as mentioned above, can withstand higher temperatures allowing it to more readily bind to ACE2 in a variety of host species, even hosts with higher internal body temperatures.

Human-infecting coronaviruses are associated with having zoonotic origins. SARS-CoV is one of these coronaviruses that most likely emerged through genomic recombination of bat SARS-related coronaviruses, particularly SARSr-CoVs. The primary reservoir and original host species are suspected to be bats which is a common reservoir for human coronaviruses [12]. Bats have been found to have sequences of SARS-related CoVs and evidence points to bats being infected with such viruses, additionally implicating bats as the original reservoir. Two novel bat SARS-related CoVs show high similarity with SARS-CoV and utilize the same ACE2 receptor, further reinforcing that SARS-CoV originated in bats. Civets were additionally found to have SARS-CoV- antibodies, and genome analysis allowed for the conclusion of bats as the original host and civets as the intermediate host before moving to humans [2, 4, 8]. MERS-CoV, similar to SARS-CoV, is closely related to two

bat coronaviruses which imply that bats are the original host reservoirs. A MERS-CoV-related bat coronavirus, HKU4, is most likely the original strain [7]. MERS-CoV has additionally shown identical strains in humans and camels, implicating camels as the intermediate host between bats and humans [3].

Two mammal-associated coronaviruses, in particular, show great similarity to SARS-CoV-2 and have been suggested as original or intermediate host possibilities. SARS-CoV-2 is most closely related to the nonhuman animal coronaviruses including bat RaTG13 and GD and GX Pangolin-CoVs, discussed above [17]. To begin with bat-related coronaviruses, in terms of host environment, the ability of SARS-CoV-2 to withstand higher temperatures could imply that it can survive in bats as they possess a higher internal temperature [6]. SARS-CoV-2 shows the greatest genome similarity to RaTG13 which implicates bats as the original host [3]. It has been demonstrated that overall genome similarity between bat coronaviruses and SARS-CoV-2 to generally be around 96% [17, 19]. In vitro, RaTG13 interacts with human ACE2; therefore, RaTG13 has the potential to cross over into humans. Amino acid substitutions, however, have been demonstrated to allow SARS-CoV-2 to recognize human ACE2 more efficiently than RaTG13, suggesting that such changes allowed for the cross-species transmission to humans in addition to the more apparent stable characteristics, such as the formation of a stable complex between the spike protein and ACE2. RaTG13 and SARS-CoV-2 also have similar residues within the ACE2 binding ridge [18]. Within the RBD, SARS-CoV-2 differs from RaTG13 with only having one identical amino acid out of the six critical amino acids which are essential in binding to ACE2 [17]. The RBD shows an overall amino acid similarity between SARS-CoV-2 and RaTG13 of 89%. The S protein amino acid identity shows a range from 75-97% among bat-related coronaviruses with RaTG13 showing the highest. Despite this high similarity, these few differences lead to the implication of bats acting as the original host and not the direct host into humans [8, 10].

The pangolin CoV demonstrates an overall genomic similarity of 91% to SARS-CoV-2 [19]. The spike protein shows a great similarity between SARS-CoV-2 and pangolin CoVs. Six out of the six critical amino acids remain the same within this region, however, there are

enough differences at the nucleotide level and in other amino acid positions to suggest these similarities arising through convergent evolution [10, 17]. Such differences result in a similarity within the S protein of around 92% for GX Pangolin-CoV and 89% for GD Pangolin-CoV with more divergence within the S1 subunit specifically. The RBD of GD Pangolin CoV shows an overall similarity of 97%, much greater in comparison to the GX Pangolin CoV showing 87% and even RaTG13 showing 89% [8]. The GD Pangolin CoV has been demonstrated to utilize ACE2 and GX Pangolin-CoV does not; therefore, the GD pangolin is more likely to be the intermediate host than the GX pangolin [18].

In reflecting on how a great number of viruses are zoonotic and can infect a wide variety of different host species, understanding the pathways that are expected and known to happen is essential in further understanding the virus itself as well as preventing cross-species transmission in the future. SARS-CoV-2 undoubtedly has zoonotic origins. Recognizing this and performing comparative analyses can help in determining treatment targets and developing preventative measures [8]. Even though many factors influence cross-species transmission, this can still happen readily and has been proven to do so, especially with SARS-CoV, MERS-CoV, and SARS-CoV-2. The current scientific consensus is that bats are the original host of SARS-CoV-2 while pangolins most likely serve as the intermediate host that infects humans. Interestingly, two research groups, Cui (2018) and Ge (2013), even predicted that novel variants would arise from SARS-CoV specifically and forewarns of the potential rise of new viruses that can again infect humans just as SARS-CoV and MERS-CoV did, a prediction that was indeed fulfilled [2, 20].

Evolution of SARS-CoV-2 within the Human Population

As long as SARS-CoV-2 infects and circulates in humans, viral point mutations will continually occur and cause the emergence of new competitive SARS-CoV-2 strains. Mutations will accumulate within the spike protein-encoding gene which alters transmission, pathogenesis, receptor usage, and infectivity [21]. Although coronaviruses do contain a proofreading mechanism, mutations in the spike protein especially will persist as the spike protein-encoding gene dictates

effective interactions with ACE2 and would therefore experience both positive and purifying selection [11, 22]. To add to this complexity, environments will positively or negatively select for these new strains [12]. The following two mutations, positively selected for, have resulted in two widespread strains with a point mutation of the spike protein or of the RBD.

One strain that arose early in the emergence of SARS-CoV-2 is the D614G mutation in which an aspartate (D) in the 614th amino acid position is mutated to glycine (G) [21, 23]. The G614 strain with the S protein mutation dominates over the original D614 strain. Plante and colleagues (2020) demonstrated that the G614 virus results in enhanced viral replication within cells [22]. Additionally, these researchers demonstrated that G614 has an increased viral load in the upper respiratory tract than the D614 virus but with similar viral loads in the lungs [22]. These increased efficiencies will allow for further transmission and infection of the G614 virus and are confirmed to rapidly outcompete the D614 virus. Infectivity has been found to additionally be increased during entry into the host cell [5]. Plante and colleagues (2020) compared the stability and infectivity of G614 to D614 at different temperatures of 33°C, 37°C, and 42°C, respectively and found that G614 displayed higher infectivity than D614 at all of these temperatures, which could suggest that the G614 mutation confers greater overall stability of SARS-CoV-2 virions [22]. There are two conformations of the RBD within the spike protein, the open and closed conformation [21]. The open conformation is necessary for binding to ACE2 with the closed conformation partially shielding this site, thus resulting in less binding in the closed conformation. In analyzing the S1 subunit specifically, Yurkovetskiy and colleagues (2020) found that the RBDs of the G614 virus contain a greater percentage of RBD in the open conformation, with 58% in the open compared to 18% open of the D614, implying that this improves binding to ACE2 [21]. The actual binding affinity, however, remains unchanged with this structural change mentioned, only allowing for better interaction rather than binding more tightly to ACE2 [5]. The suggestion of improved interaction infers greater transmission of this mutant strain. With increased infectivity, the prevalence of the G614 virus will more readily spread to individuals and reach a

greater number of people but not necessarily causing significantly increased symptomology severity. These improved characteristics of increased survivability, infectivity, transmissibility, and stability have been seemingly conferred by a single amino acid change within the crucial region of the S protein. This rapid rise to prominence is significant and the inevitable rise of more mutations in this protein-encoding gene should be taken note of.

Controlled environment lab studies of animal and human cells and epidemiologic data, at this point, appear to conflict with each other in the analysis of disease severity of G614. Several lab studies did not find any evidence of an association with G614 and increased disease severity [11, 21, 22]. Epidemiology data, however, suggests that there may be an association with G614 and a higher case fatality rate. This increased higher rate of observed fatality cases may be due to skewed statistical data or as a result of immunologic rather than virologic mechanisms [23]. Regardless, further study is warranted in determining the symptom severity conferred by mutant strains.

Other mutations of significance are within the B117 lineage which descends from D614G. This lineage has risen to significant prevalence especially in Britain and South Africa and includes 17 mutations with 8 in the gene that encodes the spike protein [24]. The most prominent mutation is the N501Y mutation in which an asparagine (N) in the 501 position is mutated to a tyrosine (Y) within the RBD of the spike protein. Analysis showed that the Y501 residue forms a hydrogen bond with K353 of the ACE2 receptor as well as interacts with three additional amino acids of ACE2. Additionally, a change in the spacing of amino acid residues in the spike protein was observed, increasing the interaction between Y501 and several residues at the spike-ACE2 interface. This change increases the spike-ACE2 interaction force and confers greater transmissibility [25].

The dominant SARS-CoV-2 strains are positively selected for increased transmission and infectivity as opposed to strains having the severity to kill the host and decrease transmission. This is observed in the strains mentioned above in which transmission and infectivity have been improved in comparison to the original SARS-CoV-2 strain, leading to their dominance in

COVID19 cases throughout the world. The persistence of a strain, however, is not always due to a single mutation as the survivability of the strain within a population is also attributed to a variety of factors such as environment, contact with individuals, and the virus continually circulating through a population. Therefore, it is essential to limit the spread of such severe viruses to limit the ability of more efficient strains in functionality to freely run through a population. As of now, the rise of a strain that confers a greater increase in severe symptoms or an increased death rate has not been detected, however, as seen with these two particular mutations, the development of such a strain may be able to develop and persist in the future. Therefore, heightened vigilance in monitoring and eliminating the virus in a short amount of time is essential in preventing such an occurrence. Studies such as the ones described here, however, do have limitations as they may not be entirely representative of undetected and currently circulating mutant strains due to the small sample size that is analyzed. Increasing sample size in addition to obtaining samples from various locations could result in a more representative study. Regardless, particular mutant strains are persisting and dominating, indicating that improved functionality differences are indeed selected for and that more will potentially arise in the future [5].

This brings up the question; are these viruses, D614G and N501Y of the B lineage, able to be treated the same way as the original or novel strain of SARS-CoV-2? Korber and colleagues (2020) state that vaccines and treatments tend to target the envelope of viruses containing the proteins that bind to receptors. For the case of SARS-CoV-2, the target is the spike protein which allows for host entry [11]. The two mutants analyzed have mutations within this spike protein and the RBD. Hou and coworkers (2020) however, suggest that current vaccine targets should have a similar effect on the D614G virus [21]. Nonetheless, the accumulation of mutations as SARS-CoV-2 continues to circulate through populations could result in the persistence of a strain that changes the interacting points of the spike protein enough to warrant new treatment methods or even new vaccines. Therefore, continual surveillance of circulating strains is necessary for maintaining or adjusting current strategies.

Conclusion and Prospectus

In conclusion, the general background of coronaviruses along with comparative analysis between SARS-CoV-2 and other pertinent coronaviruses implicates evolutionary relationships among coronaviruses, cross-species transmission capabilities, and the evolution of variant SARS-CoV-2 strains within the human population. SARS-CoV-2 is a beta coronavirus that has resulted in a pandemic greatly impacting the functioning of societies. Understanding the evolutionary relationship between SARS-CoV-2 and other coronaviruses helps in understanding its emergence and rise to prevalence while also aiding in determining the most effective treatment and vaccine targets.

SARS-CoV-2 crosses the species boundary and infects several similar mammalian species. The zoonotic nature of SARS-CoV-2 indicates that transmission from nonhuman animals to humans has a significant impact on the health of human populations. Bats, most likely, are the original host species of SARS-CoV-2. SARS-Co-V-2 then moved to an intermediate species, probably pangolins, to begin infecting humans. This pathway to human infection, which is accepted as of now, still needs to be further investigated. Further genomic research of the suspected intermediate host, including pangolins, is necessary for developing targeted immunotherapy and limiting cross-species transmission into the human population. The reason for cross-species transmission is typically due to the improper handling of wildlife in places like wet markets, which are known to transmit zoonotic viruses to human communities. Therefore, to prevent the spread of these viruses within the human population, it will be important to impose strict regulations when handling wildlife in general but also in improving sanitary conditions and limiting contact with known susceptible mammals.

The persistence of SARS-CoV-2 in the human population has resulted in the emergence of significantly more infectious mutant strains. It is expected that mutations that lead to such strains will continue to accrue, and selective pressures will lead to more competition between strains. The characteristics of these new strains will most likely retain mutations that increase transmission and infectivity. This ensures that the virus will continue to survive in host species. New strains will most likely not confer increased symptomology or disease severity because it is typically not favored by natural selection. The host environment that arises due to severe symptomology will compromise the host, maybe kill the host, and thus decrease the transmission of the virus to the next host. Overall, viral transmission will be decreased and may lead to limited viral spread. Therefore, emergent strains will most likely continue to become more infectious which results in the further spread and further complication of overall progress. Current safety precautions and recommendations will continue indefinitely due to its high level of infectivity. The complexities that can arise with mutant strains may make vaccine efficacy difficult to achieve. Vaccines that are being distributed currently will be monitored as time passes due to the inevitable emergence of strains that may no longer be inactivated by the vaccines or by the immunologic response.

Acknowledgements

Competing Interests

The authors declare no competing interests.

References

1. Coronaviridae Study Group of the International Committee on Taxonomy of Viruses (2020). The species Severe acute respiratory syndrome-related coronavirus: classifying 2019-nCoV and naming it SARS-CoV-2. *Nature Microbiology*, *5*(4), 536–544. https://doi.org/10.1038/s41564-020-0695-z

2. Cui, J., Li, F., & Shi, Z.-L. (2018). Origin and evolution of pathogenic coronaviruses. *Nature Reviews Microbiology*, *17*(3), 181–192. https://doi.org/10.1038/s41579-018-0118-9

3. Rabaan, A., Al-Ahmed, S., Haque, S., Sah, R., Tiwari, R., Singh Malik, Y., Dhama, K., Yatoo, M., Katterine Bonilla-Aldana, D., Rodriguez-Morales, A., & And Kashmir, J. (n.d.). *SARS-CoV-2, SARS-CoV, and MERS-CoV: a comparative overview*. https://www.infezmed.it/media/journal/Vol_28_2_2020_7.pdf

4. Das, S., Das, S., & Ghangrekar, M. M. (2020). The COVID-19 pandemic: biological evolution, treatment options and consequences. *Innovative Infrastructure Solutions*, *5*(3). https://doi.org/10.1007/s41062-020-00325-8

5. Yurkovetskiy, L., Wang, X., Pascal, K. E., Tomkins-Tinch, C., Nyalile, T. P., Wang, Y., Baum, A., Diehl, W. E., Dauphin, A., Carbone, C., Veinotte, K., Egri, S. B., Schaffner, S. F., Lemieux, J. E., Munro, J. B., Rafique, A., Barve, A., Sabeti, P. C., Kyratsous, C. A., & Dudkina, N. V. (2020). Structural and Functional Analysis of the D614G SARS-CoV-2 Spike Protein Variant. *Cell*, *183*(3), 739-751.e8. https://doi.org/10.1016/j.cell.2020.09.032

6. He, J., Tao, H., Yan, Y., Huang, S.-Y., & Xiao, Y. (2020). Molecular Mechanism of Evolution and Human Infection with SARS-CoV-2. *Viruses*, *12*(4), 428. https://doi.org/10.3390/v12040428

7. Walls, A. C., Park, Y.-J., Tortorici, M. A., Wall, A., McGuire, A. T., & Veesler, D. (2020). Structure, Function, and Antigenicity of the SARS-CoV-2 Spike Glycoprotein. *Cell*, *181*(2), 281-292.e6. https://doi.org/10.1016/j.cell.2020.02.058

8. Tizaoui, K., Zidi, I., Lee, K. H., Ghayda, R. A., Hong, S. H., Li, H., Smith, L., Koyanagi, A., Jacob, L., Kronbichler, A., & Shin, J. I. (2020). Update of the current knowledge on genetics, evolution, immunopathogenesis, and transmission for coronavirus disease 19 (COVID-19). *International Journal of Biological Sciences*, *16*(15), 2906–2923. https://doi.org/10.7150/ijbs.48812

9. Lan, J., Ge, J., Yu, J., Shan, S., Zhou, H., Fan, S., Zhang, Q., Shi, X., Wang, Q., Zhang, L., & Wang, X. (2020). Structure of the SARS-CoV-2 spike receptor-binding domain bound to the ACE2 receptor. *Nature*, *581*(7807), 215–220. https://doi.org/10.1038/s41586-020-2180-5

10. Zhang, Y.-Z., & Holmes, E. C. (2020). A Genomic Perspective on the Origin and Emergence of SARS-CoV-2. *Cell*, *181*(2), 223–227. https://doi.org/10.1016/j.cell.2020.03.035

11. Korber, Bette, et al. "Tracking Changes in SARS-CoV-2 Spike: Evidence That D614G Increases Infectivity of the COVID-19 Virus." *Cell*, vol. 182, no. 4, Aug. 2020, pp. 812-827.e19, www.ncbi.nlm.nih.gov/pmc/articles/PMC7332439/, 10.1016/j.cell.2020.06.043

12. Das, P., & Choudhuri, T. (2020). Decoding the global outbreak of COVID-19: the nature is behind the scene. *VirusDisease*, *31*(2), 106–112. https://doi.org/10.1007/s13337-020-00605-y

13. Yan, Y., Chang, L., & Wang, L. (2020). Laboratory testing of SARS-CoV, MERS-CoV, and SARS-CoV-2 (2019-nCoV): Current status, challenges, and countermeasures. *Reviews in Medical Virology*, *30*(3). https://doi.org/10.1002/rmv.2106

14. Yachou, Y., El Idrissi, A., Belapasov, V., & Ait Benali, S. (2020). Neuroinvasion, neurotropic, and neuroinflammatory events of SARS-CoV-2: understanding the neurological manifestations in COVID-19 patients. *Neurological Sciences*, *41*(10), 2657–2669. https://doi.org/10.1007/s10072-020-04575-3

15. *WHO Coronavirus (COVID-19) Dashboard*. (2021). Who.int. https://covid19.who.int/

16. Lu, L., Zhong, W., Bian, Z., Li, Z., Zhang, K., Liang, B., Zhong, Y., Hu, M., Lin, L., Liu, J., Lin, X., Huang, Y., Jiang, J., Yang, X., Zhang, X., & Huang, Z. (2020). A comparison of mortality-related risk factors of COVID-19, SARS, and MERS: A systematic review and meta-analysis. *Journal of Infection*, *81*(4), e18–e25. https://doi.org/10.1016/j.jinf.2020.07.002

17. Tang, X., Wu, C., Li, X., Song, Y., Yao, X., Wu, X., Duan, Y., Zhang, H., Wang, Y., Qian, Z., Cui, J., & Lu, J. (2020). On the origin and continuing evolution of SARS-CoV-2. *National Science Review*, *7*(6), 1012–1023. https://doi.org/10.1093/nsr/nwaa036

18. Shang, J., Ye, G., Shi, K., Wan, Y., Luo, C., Aihara, H., Geng, Q., Auerbach, A., & Li, F. (2020). Structural basis of receptor recognition by SARS-CoV-2. *Nature*, *581*(7807), 221–224. https://doi.org/10.1038/s41586-020-2179-y

19. Zhang, T., Wu, Q., & Zhang, Z. (2020). Probable Pangolin Origin of SARS-CoV-2 Associated with the COVID-19 Outbreak. *Current Biology*, *30*(7),

1346-1351.e2. https://doi.org/10.1016/j.cub.2020.03.022

20. Ge, X.-Y., Li, J.-L., Yang, X.-L., Chmura, A. A., Zhu, G., Epstein, J. H., Mazet, J. K., Hu, B., Zhang, W., Peng, C., Zhang, Y.-J., Luo, C.-M., Tan, B., Wang, N., Zhu, Y., Crameri, G., Zhang, S.-Y., Wang, L.-F., Daszak, P., & Shi, Z.-L. (2013). Isolation and characterization of a bat SARS-like coronavirus that uses the ACE2 receptor. *Nature*, *503*(7477), 535–538. https://doi.org/10.1038/nature12711

21. Hou, Y. J., Chiba, S., Halfmann, P., Ehre, C., Kuroda, M., Dinnon, K. H., Leist, S. R., Schäfer, A., Nakajima, N., Takahashi, K., Lee, R. E., Mascenik, T. M., Graham, R., Edwards, C. E., Tse, L. V., Okuda, K., Markmann, A. J., Bartelt, L., de Silva, A., & Margolis, D. M. (2020). SARS-CoV-2 D614G variant exhibits efficient replication ex vivo and transmission in vivo. *Science*, eabe8499. https://doi.org/10.1126/science.abe8499

22. Plante, J. A., Liu, Y., Liu, J., Xia, H., Johnson, B. A., Lokugamage, K. G., Zhang, X., Muruato, A. E., Zou, J., Fontes-Garfias, C. R., Mirchandani, D., Scharton, D., Bilello, J. P., Ku, Z., An, Z., Kalveram, B., Freiberg, A. N., Menachery, V. D., Xie, X., & Plante, K. S. (2020). Spike mutation D614G alters SARS-CoV-2 fitness. *Nature*. https://doi.org/10.1038/s41586-020-2895-3

23. Becerra-Flores, M., & Cardozo, T. (2020). SARS-CoV-2 viral spike G614 mutation exhibits higher case fatality rate. *International Journal of Clinical Practice*, *74*(8). https://doi.org/10.1111/ijcp.13525

24. Kupferschmidt, K. (2021). Viral evolution may herald new pandemic phase. *Science*, *371*(6525), 108–109. https://doi.org/10.1126/science.371.6525.108

25. Santos, J. C., & Passos, G. A. (2021). *The high infectivity of SARS-CoV-2 B.1.1.7 is associated with increased interaction force between Spike-ACE2 caused by the viral N501Y mutation*. https://doi.org/10.1101/2020.12.29.424708