REVIEW ARTICLE

# Molecular epidemiology of SARS-CoV-2: a review of current data on genetic variability of the virus

Miłosz Parczewski[1], Andrzej Ciechanowicz[2]

1 Department of Infectious, Tropical Diseases and Immune Deficiency, Pomeranian Medical University in Szczecin, Szczecin, Poland
2 Department of Clinical and Molecular Biochemistry, Pomeranian Medical University in Szczecin, Szczecin, Poland

**KEY WORDS**

coronavirus, coronavirus disease 2019 genomics, molecular epidemiology, severe acute respiratory syndrome coronavirus 2

**ABSTRACT**

Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), associated with coronavirus disease 2019 (COVID-19), is a novel pathogen recently introduced to the human population. It is characterized by rapid epidemic transmissions due to lack of herd immunity as well as by notable mortality which increases with age and in patients with comorbidities. Outbreak forecasting and modelling suggest that the number of infected people will continue to rise globally in the forthcoming months. Upon investigation of the disease patterns, differences in mortality between south-European and north-European countries became striking with mortality of more than 10% in Italy and Spain and less than 5% in Germany and Poland so far. It is unknown if this difference is associated with a higher virulence of viral strains, differences in host genomics, access to medical resources, or other unknown variables. Little is also known about SARS-CoV-2 evolutionary and transmission patterns as a limited number of large-scale sequence and phylogenetic analyses have been performed so far. In this review, we aimed to provide concise data on the SARS-CoV-2 genomics, molecular evolution, and variability with special consideration of the disease course.

Correspondence to:
Prof. Miłosz Parczewski, MD, PhD, Department of Infectious, Tropical Diseases and Immune Deficiency, Pomeranian Medical University, ul. Arkońska 4, 71-455 Szczecin, Poland, phone: +48 91 813 94 56, email: mparczewski@yahoo.co.uk
**Received:** July 13, 2020.
**Revision accepted:** August 10, 2020.
**Published online:** August 11, 2020.
Pol Arch Intern Med. 2021; 131 (1): 63-69
doi:10.20452/pamw.15550


**Introduction** Coronaviruses are positive-sense, single-stranded, spherical, enveloped RNA viruses, well known to cause mild flu-like symptoms in humans, which also affect an array of mammals. In general, coronaviruses cause infections of the respiratory or gastrointestinal tracts by fusion with macrophages and epithelial cells. These viruses have long been known to be of high potential for a zoonotic cross-species transmission to humans. From the perspective of emerging infectious diseases, transmissions of RNA, as opposed to DNA viruses, from animals have been relatively frequent with a high mutation rate in these viruses allowing for a rapid adaptation to the novel hosts.[1]

In December 2019, the Wuhan Municipal Health Committee (Wuhan, China) identified an outbreak of viral pneumonia of unknown cause. Novel coronavirus designated as severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), was found to be genetically similar to coronaviruses found in bats, which are so far the most likely host of the virus. Except for the bats, it has been suggested that Malayan pangolins (*Manis javanica*) may also be a reservoir of SARS-CoV-2 not only due to homologic coronaviruses circulating in these animals but also because of similarity of the binding site of the angiotensin-converting enzyme 2 (ACE2) receptor.[2]

Most theories link the introduction of the virus with the Huanan seafood market in Wuhan, Hubei Province of China; however, recently published molecular data have indicated possible initial expansion of the infected populations between December 11, 2019 and January 22, 2020, coinciding with the Chinese New year, or even earlier—between 13 November 2019 and 26 December 2019, when only a single case of COVID-19 was reported. It is therefore possible that the virus had been already widely circulating in Wuhan in November 2019.[3] The Polish index case was diagnosed on March 4, 2020 with subsequent spread reaching approximately 35 000 cases and approximately 4.5% mortality as of the day of the manuscript submission (June 30, 2020). To compare,

in the neighboring Germany, despite a significantly larger epidemic (>190 000) mortality is similar (approximately 4.5%), while in the southern European countries with the progressive epidemic, namely Italy and Spain, the case count exceeded 200 000 cases with mortality of 14.5% and 11.5%, respectively. Little is known about the reason for this difference. It is likely associated with a distinct demographic profile of the populations and higher percentage of the population aged older than 65 years in the south; however, these clear differences in the mortality remain not fully elucidated, and may also be linked to the genetic differences among the host populations or divergent molecular characteristics of the virus per se.

**Taxonomy of coronaviruses and SARS-CoV-2** Within the realm Riboviria, order Nidovirales, suborder Cornidovirineae, family Coronaviridae, subfamily Orthocoronavirinae, 4 genera have been identified, namely *Alpha-*, *Beta-*, *Delta-*, and *Gammacoronaviridae*. So far, almost 50 species that belong to this family of viruses have been discovered.[4] Coronaviruses include mammalian *Alphacoronaviruses* and *Betacoronaviruses*, as well as *Gammacoronaviruses* and *Deltacoronaviruses* which generally cause infections in birds. Within the *Alphacoronavirus* genus, various species infecting a vast array of animals have been identified, including, but not limited to, human coronaviruses 229E and NL63, miniopterus bat coronaviruses 1 and HKU8, porcine epidemic diarrhea virus, rhinolophus bat coronavirus HKU2, scotophilus bat coronavirus 512. Genus *Betacoronavirus* includes murine and bovine coronaviruses, clinically mild human OC43 and HKU1 coronaviruses, several bat infecting species (pipistrellus bat coronavirus HKU5, rousettus bat coronavirus HKU9, tylonycteris bat coronavirus) as well as severe acute respiratory syndrome-related coronavirus (SARS-CoV) and SARS-CoV-2, Middle East respiratory syndrome-related coronavirus (MERS-CoV), and hedgehog coronaviruses. Two other genera, *Gamma-* and *Deltacoronaviridae*, include beluga whale coronavirus SW1, infectious bronchitis virus, and bulbul coronavirus HKU11, porcine coronavirus HKU15, respectively with no human transmissions noted so far.[5]

The novel coronavirus, responsible for the COVID-19 epidemic and associated with severe acute respiratory syndrome, has only recently been classified by phylogeny and taxonomy to belong to *Betacoronaviridae* based on the sequence similarity to the sister SARS-CoV.[4] Other genetically similar SARS coronaviruses have also been previously identified, for example, civet SARS-CoV_PC4-227 and SARSr-CoV-btKY72.[6] It should be noted that classification of the RNA viruses is not easy—many exist as a swarm of genetically interrelated, co-evolving quasispecies. Moreover, coronaviruses are ubiquitous among vertebrates, with the current COVID-19 epidemic representing the third major zoonotic transmission of the novel pathogenic coronaviruses capable of causing life-threatening disease in humans in the recent history.[7] Famously, these epidemics were caused by SARS-CoV in 2002 to 2003 and MERS-CoV ongoing since 2012. It should be emphasized that SARS-CoV-2 is not descending from SARS-CoV and has a separate history of introduction into the human species, lower pathogenicity, and higher infectivity rate compared with SARS-CoV and MERS-CoV.[8,9] Two hypotheses have been proposed for the origin of this virus, namely natural selection in humans following zoonotic introduction or evolution in humans after the transmission. Of note, mildly symptomatic infections with other *Alpha-* (human coronaviruses 229E, NL63) and *Betacoronaviruses* (human coronaviruses OC4, HKU1), common in both adults and children, are also highly likely to originate from the bat or rodent reservoir.[10]

**Viral structure and the replicative cycle** The spherical structure of the virus contains the core with ribonucleoprotein of the helical structure enclosed by nucleocapsid (N) proteins. Within the viral membrane, envelope (E) proteins are anchored, with the crown-shaped spikes formed by the spike (S) protein protruding from the virion membrane.[8]

Receptor binding for coronaviruses is dependent on the S protein, which is equipped with an extracellular, transmembrane anchor and intracellular tail domains,[11] and has been previously identified as a likely vaccine target.[12] The SARS-CoV-2 S protein is prone to accumulate mutations compared with SARS-CoV, especially at the interface with the ACE2 receptor and therefore shows lower sequence homology and higher genetic variation (81% and 19%, respectively).[13,14]

Key for the binding with the target host cell is an extracellular part with 2 subunits involved in receptor binding (S1 domain) and membrane fusion (S2 domain).[15] Spike proteins vary across coronavirus species, with differences in structure correlating with cellular tropism and virulence.[10,16] Typically, S proteins consist of approximately 1300 amino acids which form trimeric structures anchored in the virus membrane.[17] Of note, at the junction of the S1 and S2 domains, polybasic furin cleavage site with the RRAR (arginine-arginine--alanine-arginine) motif is located. Polybasic motifs are well known to increase pathogenicity of viruses,[18] and in this case, the furin cleavage site enhances the virus-cell fusion.[2] In SARS-CoV-2, similarly to SARS-CoV, receptor binding domain is complexing via a form of the hydrophobic tunnel with salt bridges within the ACE2 receptors on the human cells.[19]

After the ACE2 complex binds to the S1 part of the S protein, the ACE-virus complex translocates to the endosomes. Subsequently, the S1/S2 protein is cleaved by the endosomal proteases (eg, cathepsin L), which unmasks the S2 fusion peptides activating integration between the viral and host membranes within the endosome (FIGURE 1). In this process, coronavirus receptor

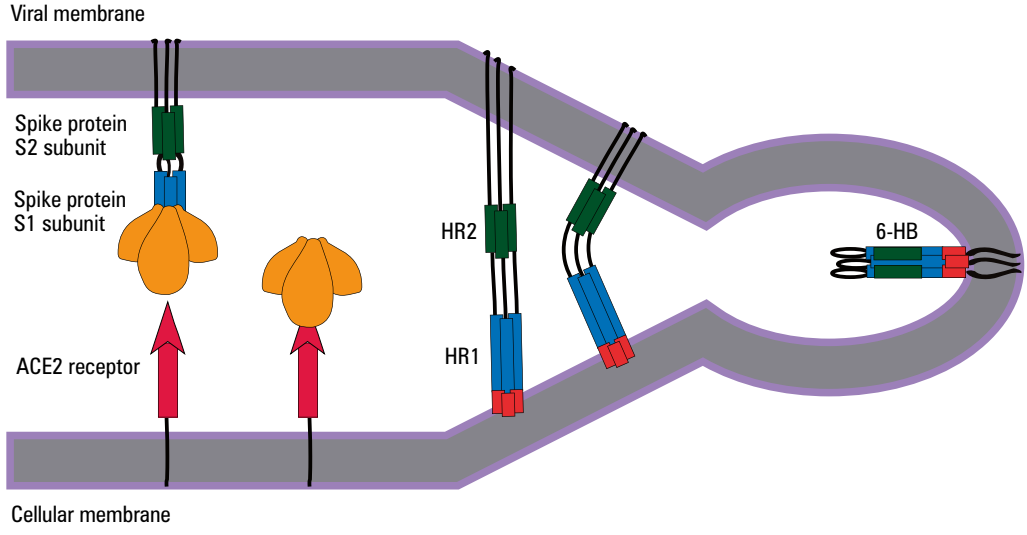


**FIGURE 1** Simplified outline of the SARS-CoV-2 integration with the host cell.
Abbreviations: ACE2, angiotensin-converting enzyme 2; HR, heptad repeat structure; 6-HB, folded 6-helix heptad structure

binding domains link to the hypothesized "virus binding hotspot" of the ACE2 receptor with mutation shifts allowing for adaptation across various species including ferret, bat, pig, civet cat, and other animals.[17] The C terminal portion of the S protein contains 2 trimeric helical heptad repeat structures (HR1 and HR2). These structures are of primary importance for the virus-host cell fusion, folding into a stable protease resistant 6-helix (6-HB) structure. These folded forms are observed post fusion.[15] It should be emphasized here that 6-HB structures have been previously identified to be similar to influenza hemagglutinin, Ebola glycoprotein, or HIV glycoprotein 41.[20]

Interestingly, to ensure efficient replication, in SARS-CoV-2, not 1 but 2 RNA-dependent polymerases are involved: the first is primer dependent and the second has primase activity, therefore with the capacity to initiate replication. The viral genome is released and translated by the viral replicase complex and cut by proteinases. The full-length negative template serves as a basis for mRNA synthesis. Viral nucleocapsids are assembled from genomic RNA and bound to the N protein in the cytoplasm. A release from the infected cell through exocytosis follows budding from the endoplasmic reticulum-Golgi compartment, completing the life cycle of the virus.

**Clinical course of coronavirus disease 2019** In most cases (approximately 80%), COVID-19 presents as a mild-to-moderate self-limited acute respiratory illness with fever, cough, and shortness of breath, but infection may also progress to interstitial pneumonia, severe acute respiratory syndrome, kidney failure, and death.[21] Clinical stages of the disease have been well established and divided into asymptomatic or mild type presenting only with mild upper or genitourinary symptoms, stable patients with respiratory symptoms and radiological confirmation of pneumonia, clinically unstable patients with respiratory failure defined as impaired gas exchange capacity (tachypnea, dyspnea, decreased $SpO_2$ <90%) and acute respiratory distress syndrome, which may include shock, multiorgan failure, and impaired consciousness.[22,23] Established risk factors for severe COVID-19 infections and mortality include older age (>65 years), chronic lung or cardiovascular diseases, diabetes, male sex, as well as cancers (including hematological), obesity, and renal and liver diseases.[24]

Notably, in severe COVID-19, increased activity of the inflammatory parameters, including interleukin 1 (IL-1), IL-6, or tumor necrosis factor α (TNF-α) levels, reflect the cytokine storm and may be a predictor of disease severity.[25] Of these, IL-6 has become a key laboratory parameter predicting disease severity in COVID-19. Physiologically, IL-6 promotes expansion and activation of T cell populations, B cell differentiation, regulates acute phase response, and to a certain extent affects the hormone-like properties of vascular disease, lipid metabolism, insulin resistance, mitochondrial activity, neuroendocrine system, and neuropsychological behavior.[26] In SARS-CoV-2 infections, high expression of IL-6 is a result of a hyperactive humoral response from the cytotoxic T lymphocytes and is a marker of respiratory failure, shock, and multiorgan failure. However, it is unknown if increases in IL-6 and other acute phase parameters are associated with the differences in the virulence of the infecting strains reflected by the molecular variability. COVID-19 infections have also been associated with immune exhaustion of the NK and CD8 T lymphocytes.[27]

**The coronavirus disease 2019 genome and sequence variability** As noted above, the virus was first identified from samples of a seller from a seafood market in Wuhan with diagnosed severe pneumonia. After confirming that bronchoalveolar lavage samples contained the coronavirus genetic material, the next generation sequencing of the viral RNA was performed identifying a virus with 96% bat RaTG13 (sampled from *Rhinolophus affinis)* viral sequence homology, 89% nucleotide identity with bat SARS-like-CoVZXC2, 82% to 87% similarity to human SARS-CoV and 79.6% to SARS-CoV BJ01.[8,28] In similarity plots of this

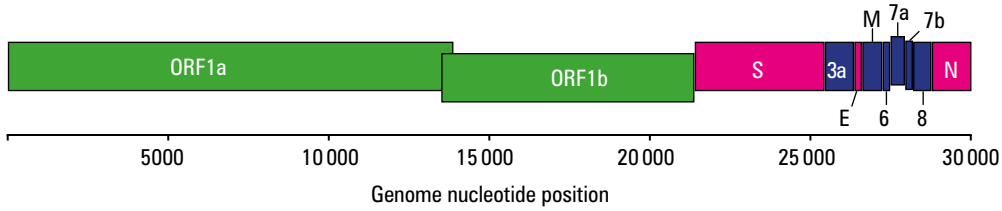


**FIGURE 2** SARS-CoV-2 genome organization
Abbreviations:
E, envelope protein;
N, nucleocapsid protein;
ORF, open reading frame;
S, spike protein

novel virus, the highest sequence similarity (closest ancestry) with the bat RaTG13 has further been confirmed with SARS-CoV-2 lineage clearly distinct from the SARS-CoV.[3,14] Additionally, the S protein notably differs from other coronaviruses, with the highest similarity to the bat RaTG13 mentioned above, indicating separate origin and strongly suggesting zoonotic transmission of the virus.[29] As a result, bat coronaviruses are frequently used as an outgroup in the phylogenetic studies.[3,30]

The SARS-CoV-2 genome encodes for 8 open reading frames (ORFs), which is typical of coronaviruses. The genome of 29 903 nucleotides contains genes encoding for 3C-like proteinase, RNA-dependent RNA polymerase (RdRp), 2'-O-ribose methyltransferase, S protein, E protein, N phosphoprotein, membrane (M) protein, and several unknown proteins (FIGURE 2).[31,32] Within ORF1a, replicase polyproteins are encoded, as well as papain-like proteinase (nonstructural protein 3) involved in the cleavage of the nonstructural proteins and blockage of the immune response and cytokine expression by inhibition of the interferon-stimulated genes. Furthermore, this ORF encodes the nonstructural protein 4 involved in the formation of the double membrane vesicles and the conserved 3CLPro protease involved in RNA replication.[33]

The M protein of coronaviruses is known to induce neutralizing antibody response which is well recognized by CD8 lymphocytes.[34] The RdRp polymerase is directly involved in the transcription of the viral RNA because it is coupled with a nonstructural protein 14 exonuclease which has a proofreading function. Of note, antiviral nucleotide analogues including remdesivir or favipiravir inhibit RdRp.[35] Over the course of the epidemic, RdRp tends to accumulate mutations, diverging from the ancestral viral clades. Mutational patters within the frames coding for this enzyme differ between regions, which may result in differences in the viral replication rates and therefore infectivity. It is possible that RdRp replication complexes from some European strains have lesser proofreading activity and therefore are linked with decreased virulence.[36] The N protein is not only a structural protein but is also crucial for the viral transcription and assembly, sharing approximately 90% to 93% amino acid sequence identity with SARS, which confirms conserved nature of this protein. It contains 2 RNA binding domains—one at the N- and the other at the C-terminus of the protein linked by the serine / arginine rich domain which improves oligomerization and as a whole is positively charged to facilitate nucleic acid binding.[37] Nucleocapsid is also highly immunogenic, involved in the deregulation of the host cell cycle (arrest), inhibition of interferon production by blockage of the IRF3 and NFkB activity, up-regulation of the proinflammatory cyclooxygenase-2 protein.[38] Importantly, the N protein is abundantly expressed during infection.[39]

**Molecular evolution of SARS-CoV-2** Genetic diversity among coronaviruses results from the RdRp-generated errors as well as recombination, both within host and heterologous, which is a well-known mechanism involved in the viral evolution.[40] Sequence data collected so far guide phylogenetic investigation to inform molecular epidemiology, analyze transmission patterns and infection hotspots, and investigate the lineages of COVID-19. Virus variability, leading to the development of quasispecies, provides the background for virus evolution and adaptation to new hosts. It has been suggested that analyses of both amino acid and nucleotide sequences may indicate the nature of transmission and evolution of the virus.[41] A report analyzing 2666 S proteins from China, including 507 of human origin, has predicted risk of cross-species transmissions based on the amino acid sequence of the S protein, which highlights the importance of the molecular models for the prediction of infectivity.[29] Additionally, it has been demonstrated that changes in the methylation patterns in the S1 and S2 segments of the S protein may affect the binding forces on the host cells and therefore disease course.[42] Another study suggested the differences in the S protein cleavage site sequence may be associated with differences in the tissue tropism of the virus,[43] with in silico analyses predicting changes in the S protein affinity to the ACE2 receptor associated with the genetic variability and mutations in this region.[16] From the treatment perspective, molecular variability of the virus has been associated with mechanisms of chloroquine action, therefore, knowledge on the amino acid composition of SARS-CoV-2, including the S region is highly relevant for the development of vaccines and novel therapeutic targets.[11,44]

Data on the phylogenetic networks indicate that SARS-CoV-2 evolved into at least 58 haplotypes and 2 clades (ancestral, closely related to bat RaTg13 coronavirus clade I with 19 haplotypes and clade II with 39 haplotypes). It is possible that distinct haplotypes acquired adaptive mutations allowing for higher infectivity rate.[3] Analysis of the phylogenetic networks showed that

differences in the mutation patterns at various genomic positions (such as T8782C and C28144T) allowed to clearly distinguish viral clades originating from East Asia with mostly local spread from the non-Asia transmitted variants.[30]

Phylogenetic analyses using next-generation sequence data have been used to track the clustering of COVID-19 infections and identify index cases in introduction to the specific spot.[45,46] For this purpose, metagenomic sequencing technologies optimized for the identification of the viral pathogens from upper respiratory samples have been implemented, with novel clusters and possibility of the intra-host evolution of the virus being identified.[47,48] It was noted, based on the substitutions in the ORF3a region, that mutations in the COVID-19 genome form phylogenetic cluster with a common origin and new clades (clade V) based on the G251V substitution in this reading frame have been defined.[48]

Beside clade V, Guan et al[49] in their recent report defined 4 other major clades of SARS-CoV-2. Similarly to clade V (ORF3a; codon position, G251V), these 4 clades were named I, D, G, and S due to missense mutations in: ORFab (positions V378I and G392D), S protein (position D614G), and ORF8 (position L84S), respectively. In addition, the authors identified 9 minor clades which were named either after the amino acid mutation: H (ORFab, Q676H), H2 (M, D209H), L2 (N, S194L), S2 (N, P344S), Y (S, H49Y), I2 (ORF1ab, T6136I), and K (Orf1ab, T2016K) or after the following nucleotide substitutions: G11410A or C17373A in ORFab. The major 5 clades representing 85.7% of 2058 analyzed sequences (minor clades represent 3.2% of all sequences) were classified using only 10 single nucleotide polymorphisms (SNPs) in the viral genome. Using the same SNP-based approach Guan et al[49] were also able to successfully classify 95.6% of 4000 additional viral genomes deposited in GISAID between March 31 and April 15. Guan et al[49] reported that clade G represents 46.2% of all viral sequences, followed by S (25.4%), V (9.4%), I (2.6%), and D (2.1%). The remaining 14.3% were not assigned to a major clade. Clade G has been found to be widely distributed in Africa, Europe, West Asia, and South America, whereas clade S represented 63% of North American sampled genomes, and nearly a quarter of those from Oceania. Clade I has been identified in approximately one-third of genomes derived from South and West Asia, and Oceania, while Southeast Asia and South Asia have had the greatest number of unassigned genomes (56.9%). In addition, increasing prevalence of 1 or 2 clades in each geographic region was found. For example, the Asian and Oceanian genomes were largely clade I, whereas clade S predominated in the cases from North America, and European genomes were predominantly classified as clade G.[49] Korber et al[50] reported that the earliest D614G mutation of SARS-CoV-2 in Europe was identified in Germany (EPI_ISL_406862, sampled January 28, 2020). The D614G mutation began to spread rapidly first in Europe, and then in other parts of the world, and has become the dominant pandemic variant in many countries. The authors concluded that the alarming rate of the D614G frequency increase indicates a relative fitness advantage to the original Wuhan strain that enables more rapid spread. Recently, Zhang et al[51] have observed that retroviruses pseudotyped with G614 S variant infected ACE2-expressing cells more efficiently than those with D614 ones. This greater infectivity was correlated with less S1 shedding and greater incorporation of the S protein into the pseudovirion. Of note, G614 S variant did not bind ACE2 more efficiently than D614 S, and the pseudoviruses containing these S proteins were neutralized with comparable efficiencies by convalescent plasma. These results show the D614 S variant is less stable than G614 ones, which is consistent with epidemiological data suggesting that viruses with latter variant transmit more efficiently. Furthermore, apart from the clades described above, Van Dorp et al[52] revealed 198 recurrent mutations (about 80% representing nonsynonymous changes) in SARS-CoV-2 by analysis of a set of 7666 complete viral genome sequences acquired from the GISAID. The authors focused on the mutations which have emerged independently multiple times (homoplasies) and found that 3 sites in ORF1ab in the regions encoding Nsp6, Nsp11, or Nsp13 (nucleotide positions, G11083T, T13402G, or C16887T, respectively) and one in the S protein (nucleotide position, C21575T) accumulated particularly large number of recurrent mutations (>15 events). On the other hand, in a set of 2058 SARS-CoV-2 sequences Guan et al[49] identified 1221 SNPs with 753 missense, 452 silent, 12 nonsense, and 4 intergenic substitutions. The authors also observed that the genes S, N, and ORF3a accumulated markedly more mutations than expected solely by random drift. For example, the D614G mutation (clade G–defining mutation) is located in subdomain 1, and the substitution of aspartic acid by glycine would entail losing these stabilizing electrostatic interactions and increase the dynamics in this region.[49] It is noteworthy that the D614G mutation was also the most common SNP detected by van Dorp et al[52] in a set of SARS-CoV-2 genomes from GISAID included in their homoplasy analysis. Guan et al[49] suggested that the nonsynonymous mutations in the N protein, which play a key role in viral assembly, might also have functional implications. The hotspot mutations in the S202N, R203K, and G204R positions all cluster in a linker region where they might potentially enhance RNA binding and alter the response to serine phosphorylation events.[49] In addition, both R203K and G204R variants were detected in more than one-fifth of sequences analyzed by van Dorp et al.[52] In contrast, Guan et al[49] also indicated that several nonstructural proteins showed a lower-than-expected mutation rate. They also suggested that, similarly to the other *Betacoronavirus*

analogues, might be involved in evading host immune defenses, enhancing viral expression and cleavage of the replicase polyprotein.

**Conclusion** In the review of the molecular evolution of the virus described above, we briefly summarized evolutionary history of the SARS-CoV-2. Research on genetic variability and mutation characteristics of SARS-CoV-2 is crucial to understand the transmission patterns and course of the viral disease spread among people. Further genetic evolution of the virus is certain—possible changes in the affinity to human receptors such as ACE2,[13] escape from immunologic pressure, or other genetic changes may be observed in the future. For RNA viruses, high mutation rate is expected and adaptations in the SARS-CoV-2 sequence may result in an increased efficacy of transmissions and boost in virulence.[53] It is also possible that COVID-19 will become less virulent through human-to-human transmissions because genetic bottlenecks for RNA viruses often occur during respiratory droplet transmission.

Additionally, it was suggested that in vivo *Betacoronaviruses* may evolve into complex and dynamic distributions of closely related variants. Analyses of sequence variability support the presence of viral quasispecies in the longitudinal clinical samples.[48]

In the opinion of the authors it is more likely that the propagating viral species will tend to become less virulent which allows for a prolonged infectious period and higher number of exposures. Additionally, it has been hypothesized that observed differences in the population frequency and dynamics between the regions may arise from the previous immunization with the bacille Calmette-Guérin vaccine; however, the mechanism for such protection remains unclear.[54] Also, it should be considered that in North Europe, infections with non-SARS-CoV-2 coronaviruses are common and cross-immunity, and therefore selective pressure from the host to the viral species, may be an additional attenuating factor for COVID-19, as suggested by several recent studies on the T-cell reactivity to SARS-CoV-2 proteins, especially the S protein.[55,56] Of note, these hypotheses require further confirmation by high-quality scientific studies, as the nature of the host cross-reactive or vaccine-derived selective pressure on the viral genetic structure remains unknown.

To sum up, sequences generated so far may be used to model the amino acid and protein composition and potentially inform the development of the therapeutic targets, link sequence variability to differences in inflammation and disease severity as well as predict the virulence of COVID-19.

## ARTICLE INFORMATION

**CONFLICT OF INTEREST** None declared.



**HOW TO CITE** Parczewski M, Ciechanowicz A. Molecular epidemiology of SARS-CoV-2: a review of current data on genetic variability of the virus. Pol Arch Intern Med. 2021; 131: 63-69. doi:10.20452/pamw.15550

## REFERENCES

1 Ka-Wai Hui E. Reasons for the increase in emerging and re-emerging viral infectious diseases. Microbes Infect. 2006; 8: 905-916.

2 Andersen KG, Rambaut A, Lipkin WI, et al. The proximal origin of SARS--CoV-2. Nat Med. 2020; 26: 450-452.

3 Yu WB, Tang GD, Zhang L, et al. Decoding the evolution and transmissions of the novel pneumonia coronavirus (SARS-CoV-2/HCoV-19) using whole genomic data. Zool Res. 2020; 41: 247-257.

4 Gorbalenya AE, Baker SC, Baric RS, et al. The species Severe acute respiratory syndrome-related coronavirus: classifying 2019-nCoV and naming it SARS-CoV-2. Nat Microbiol. 2020; 5: 536-544.

5 Wertheim JO, Chu DKW, Peiris JSM, et al. A case for the ancient origin of coronaviruses. J Virol. 2013; 87: 7039-7045.

6 Wu Y, Ho W, Huang Y, et al. SARS-CoV-2 is an appropriate name for the new coronavirus. Lancet. 2020; 395: 949-950.

7 Prasad A, Prasad M. SARS-CoV-2: the emergence of a viral pathogen causing havoc on human existence. J Genet. 2020; 99.

8 Brüssow H. The novel coronavirus - a snapshot of current knowledge. Microb Biotechnol. 2020; 13: 607-612.

9 Prajapati S, Sharma M, Kumar A, et al. An update on novel COVID-19 pandemic: a battle between humans and virus. Eur Rev Med Pharmacol Sci. 2020; 24: 5819-5829.

10 Corman VM, Muth D, Niemeyer D, et al. Hosts and Sources of endemic human coronaviruses. Adv Virus Res. 2018; 100: 163-188.

11 Du L, He Y, Zhou Y, et al. The spike protein of SARS-CoV - a target for vaccine and therapeutic development. Nat Rev Microbiol. 2009; 7: 226-236.

12 Amanat F, Krammer F. SARS-CoV-2 vaccines: status report. Immunity. 2020; 52: 583-589.

13 Ou X, Liu Y, Lei X, et al. Characterization of spike glycoprotein of SARS--CoV-2 on virus entry and its immune cross-reactivity with SARS-CoV. Nat Commun. 2020; 11: 1620-1620.

14 Rehman SU, Shafique L, Ihsan A, et al. Evolutionary trajectory for the emergence of novel coronavirus SARS-CoV-2. Pathogens. 2020; 9: 240.

15 Heald-Sargent T, Gallagher T. Ready, set, fuse! The coronavirus spike protein and acquisition of fusion competence. Viruses. 2012; 4: 557-580.

16 Ortega JT, Serrano ML, Pujol FH, et al. Role of changes in SARS-CoV-2 spike protein in the interaction with the human ACE2 receptor: an in silico analysis. Excli J. 2020; 19: 410-417.

17 Walls AC, Park Y-J, Tortorici MA, et al. Structure, function and antigenicity of the SARS-CoV-2 spike glycoprotein. Cell. 2020; 181: 281-292.

18 Nao N, Yamagishi J, Miyamoto H, et al. Genetic predisposition to acquire a polybasic cleavage site for highly pathogenic avian influenza virus hemagglutinin. mBio. 2017; 8: e02298-e02316.

19 Wu K, Chen L, Peng G, et al. A virus-binding hot spot on human angiotensin-converting enzyme 2 is critical for binding of two different coronaviruses. J Virol. 2011; 85: 5331-5337.

20 Lamb RA, Jardetzky TS. Structural basis of viral invasion: lessons from paramyxovirus F. Curr Opin Struct Biol. 2007; 17: 427-436.

21 Guan W-j, Ni Z-y, Hu Y, et al. Clinical characteristics of coronavirus disease 2019 in China. N Eng J Med. 2020; 382: 1708-1720.

22 Flisiak R, Horban A, Jaroszewicz J, et al. Management of SARS-CoV-2 infection: recommendations of the Polish Association of Epidemiologists and Infectiologists as of March 31, 2020. Pol Arch Intern Med. 2020; 130: 352-357.

23 Flisiak R, Horban A, Jaroszewicz J, et al. Management of SARS-CoV-2 infection: recommendations of the Polish Association of Epidemiologists and Infectiologists. Annex no. 1 as of June 8, 2020. Pol Arch Intern Med. 2020; 130: 557-558.

24 Gandhi RT, Lynch JB, del Rio C. Mild or moderate covid-19. N Eng J Med. 2020; 383: 1157-1166.

25 Ulhaq ZS, Soraya GV. Interleukin-6 as a potential biomarker of COVID-19 progression. Med Mal Infect. 2020; 50: 382-383.

26 Zhang C, Wu Z, Li J-W, et al. Cytokine release syndrome in severe COVID-19: interleukin-6 receptor antagonist tocilizumab may be the key to reduce mortality. Int J Antimicrob Agents. 2020: 105954.

27 Zheng M, Gao Y, Wang G, et al. Functional exhaustion of antiviral lymphocytes in COVID-19 patients. Cell Mol Immunol. 2020; 17: 533-535.

28 Zhou P, Yang X-L, Wang X-G, et al. A pneumonia outbreak associated with a new coronavirus of probable bat origin. Nature. 2020; 579: 270-273.

29 Qiang X-L, Xu P, Fang G, et al. Using the spike protein feature to predict infection risk and monitor the evolutionary dynamic of coronavirus. Infect Dis Poverty. 2020; 9: 33.

30 Forster P, Forster L, Renfrew C, Forster M. Phylogenetic network analysis of SARS-CoV-2 genomes. Proc Natl Acad Sci USA. 2020; 117: 9241-9243.

31 Beck BR, Shin B, Choi Y, et al. Predicting commercially available antiviral drugs that may act on the novel coronavirus (SARS-CoV-2) through a drug-target interaction deep learning model. Comput Struct Biotechnol J. 2020; 18: 784-790.

32 Mousavizadeh L, Ghasemi S. Genotype and phenotype of COVID-19: their roles in pathogenesis. J Microbiol Immunol Infect. 2020 Mar 31. [Epub ahead of print].

33 Tahir Ul Qamar M, Alqahtani SM, Alamri MA, Chen L-L. Structural basis of SARS-CoV-2 3CLpro and anti-COVID-19 drug discovery from medicinal plants. J Pharm Anal. 2020; 10: 313-319.

34 Liu J, Sun Y, Qi J, et al. The membrane protein of severe acute respiratory syndrome coronavirus acts as a dominant immunogen revealed by a clustering region of novel functionally and structurally defined cytotoxic T-lymphocyte epitopes. J Infect Dis. 2010; 202: 1171-1180.

35 Wang M, Cao R, Zhang L, et al. Remdesivir and chloroquine effectively inhibit the recently emerged novel coronavirus (2019-nCoV) in vitro. Cell Res. 2020; 30: 269-271.

36 Pachetti M, Marini B, Benedetti F, et al. Emerging SARS-CoV-2 mutation hot spots include a novel RNA-dependent-RNA polymerase variant. J Transl Med. 2020; 18: 179.

37 Zeng W, Liu G, Ma H, et al. Biochemical characterization of SARS--CoV-2 nucleocapsid protein. Biochem Biophys Res Commun. 2020; 527: 618-623.

38 Surjit M, Lal SK. The SARS-CoV nucleocapsid protein: a protein with multifarious activities. Infect Genet Evol. 2008; 8: 397-405.

39 Kang S, Yang M, Hong Z, et al. Crystal structure of SARS-CoV-2 nucleocapsid protein RNA binding domain reveals potential unique drug targeting sites. Acta Pharm Sin B. 2020; 10: 1228-1238.

40 Menachery V, Graham R, Baric R. Jumping species-a mechanism for coronavirus persistence and survival. Curr Opin Virol. 2017; 23: 1-7.

41 Zhang C, Zheng W, Huang X, et al. Protein structure and sequence reanalysis of 2019-nCoV genome refutes snakes as its intermediate host and the unique similarity between its spike protein insertions and HIV-1. J Proteome Res. 2020; 19: 1351-1360.

42 Jin X, Lian J-S, Hu J-H, et al. Epidemiological, clinical and virological characteristics of 74 cases of coronavirus-infected disease 2019 (COVID-19) with gastrointestinal symptoms. Gut. 2020; 69: 1002-1009.

43 Wang Q, Qiu Y, Li J-Y, et al. A unique protease cleavage site predicted in the spike protein of the novel pneumonia coronavirus (2019-nCoV) potentially related to viral transmissibility. Virol Sin. 2020; 35: 337-339.

44 Xia S, Liu M, Wang C, et al. Inhibition of SARS-CoV-2 (previously 2019-nCoV) infection by a highly potent pan-coronavirus fusion inhibitor targeting its spike protein that harbors a high capacity to mediate membrane fusion. Cell Res. 2020; 30: 343-355.

45 Bal A, Destras G, Gaymard A, et al. Molecular characterization of SARS-CoV-2 in the first COVID-19 cluster in France reveals an amino-acid deletion in nsp2 (Asp268Del). Clin Microbiol Infect. 2020; 26: 960-962.

46 Haveri A, Smura T, Kuivanen S, et al. Serological and molecular findings during SARS-CoV-2 infection: the first case study in Finland, January to February 2020. Euro Surveill. 2020; 25: 2000266.

47 Bal A, Pichon M, Picard C, et al. Quality control implementation for universal characterization of DNA and RNA viruses in clinical respiratory samples using single metagenomic next-generation sequencing workflow. BMC Infect Dis. 2018; 18: 537.

48 Capobianchi MR, Rueca M, Messina F, et al. Molecular characterization of SARS-CoV-2 from the first case of COVID-19 in Italy. Clin Microbiol Infect. 2020; 26: 954-956.

49 Guan Q, Sadykov M, Nugmanova R, et al. A genetic barcode of SARS-CoV-2 for monitoring global distribution of different clades during the COVID-19 pandemic. Int J Infect Dis. 2020; 100: 216-223.

50 Korber B, Fischer W, Gnanakaran S, et al. Tracking Changes in SARS--CoV-2 Spike: Evidence that D614G Increases Infectivity of the COVID-19 Virus. Cell. 2020; 182: 812-827.e19.

51 Zhang L, Jackson CB, Mou H, et al. SARS-CoV-2 spike-protein D614G mutation increases virion spike density and infectivity. Nat Commun. 2020; 11: 6013.

52 van Dorp L, Acman M, Richard D, et al. Emergence of genomic diversity and recurrent mutations in SARS-CoV-2. Infect Genet Evol. 2020; 83: 104351.

53 Prabakaran P, Xiao X, Dimitrov DS. A model of the ACE2 structure and function as a SARS-CoV receptor. Biochem Biophys Res Commun. 2004; 314: 235-241.

54 Escobar LE, Molina-Cruz A, Barillas-Mury C. BCG vaccine protection from severe coronavirus disease 2019 (COVID-19). Proc Natl Acad Sci USA. 2020; 117: 17720-17726.

55 Grifoni A, Weiskopf D, Ramirez SI, et al. Targets of T cell responses to SARS-CoV-2 coronavirus in humans with COVID-19 disease and unexposed individuals. Cell. 2020; 181: 1489-1501.e1415.

56 Sette A, Crotty S. Pre-existing immunity to SARS-CoV-2: the knowns and unknowns. Nat Rev Immunol. 2020; 20: 457-458.