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

Authors: Miłosz Parczewski, Andrzej Ciechanowicz

Article type: Review article

Received: July 13, 2020.

Accepted: August 10, 2020.

Published online: August 11, 2020.

ISSN: 1897-9483

This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License (CC BY-NC-SA 4.0), allowing third parties to copy and redistribute the material in any medium or format and to remix, transform, and build upon the material, provided the original work is properly cited, distributed under the same license, and used for noncommercial purposes only. For commercial use, please contact the journal office at pamw@mp.pl.
Title: Molecular epidemiology of SARS CoV-2: a review of current data on genetic variability of the virus

Miłosz Parczewski¹, Andrzej Ciechanowicz²

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.

*Correspondence to:
Prof. Miłosz Parczewski MD, Ph.D.
Pomeranian Medical University, Department of Infectious, Tropical Diseases and Immune Deficiency, Arkońska 4 71-455 Szczecin, Poland
Tel. 0048918139456, fax 0049918139449
e-mail: mparczewski@yahoo.co.uk

Conflict of interest: none declared
Abstract:

Severe acute respiratory syndrome coronavirus 2 (SARS CoV-2), associated with coronavirus disease (COVID-19) is a novel pathogen recently introduced to the human population. It may be characterised by the rapid epidemic transmissions due to lack of the herd immunity and notable mortality, increasing with age and among patients with comorbidities. Outbreak forecasting and modelling suggest that infection numbers 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 in Italy and Spain exceeding 10% and <5% in Germany and Poland so far. It is unknown if this difference is associated with the higher virulence of the 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 limited number of the large-scale sequence and phylogenetic analyses have been performed so far. In this review, we aim to provide concise data on the SARS CoV-2 genomics, molecular evolution and variability with special consideration of the disease course.

Key words: coronavirus, COVID-19 genomics, molecular epidemiology, SARS CoV-2.
**Introduction**

Coronaviruses (CoVs) are positive-sense, single stranded, spherical, enveloped RNA viruses, well known to cause mild flu-like symptoms in humans, but are also affecting an array of mammals. In general, coronaviruses cause respiratory or gastrointestinal tract infections by fusion with macrophages and epithelial cells. These viruses have long been known to be of high potential for zoonotic cross-species transmissions to humans. From the emerging infectious diseases perspective, transmissions of RNA, as opposed to DNA viruses from animals have been relatively frequent with high mutation rate in these viruses allowing for the 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 the bat CoVs, which is 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 reservoir of the SARS CoV-2, due to not only homologic coronaviruses circulating in these animals, but also similarity of the ACE-2 receptor binding site [2].

Most theories link the introduction of the virus with the Huanan seafood market in the Wuhan Province, however, recently published molecular data have indicated possible initial expansion of the infected populations between 11 December and 2019 and 22 January 2020 coinciding with the Chinese New year or even earlier – between 13 November 2019 and 26 December 2019, when only single case of COVID-19 was reported. It is therefore possible, that virus had been already widely circulating in Wuhan province of China in November 2019 [3]. Polish index case was diagnosed on the 4th of March, with subsequent spread, reaching ~35 000 cases and ~4.5% mortality as of the day of the manuscript submission (30.06.2020). To compare, in the neighboring Germany, despite significantly larger epidemics (>190 000 )
mortality is similar (~4.5%), while in South-European countries with the progressive epidemics, namely Italy and Spain the case count exceeded 200 thousand cases with mortality of 14.5% and 11.5% respectively. Little is known about the reason for this difference, it is likely associated with different demographic profile of the populations and higher percentage of the population >65 years of age 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 suborder *Coronavirinae*, order *Nidovirales*, realm *Riboviria*, family *Orthocoronaviridae* four genera have been identified, namely *alpha-*, *beta-*, *delta-*, and *gammacoronaviridae*. So far, >forty-nine almost fifty 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 coronavirus 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*, *Roussettus bat coronavirus HKU9*, *Tylonycteris bat coronavirus*) as well as severe acute respiratory syndrome-related coronaviruses (*SARS-CoV, SARS-CoV-2*), *Middle East respiratory syndrome-related coronavirus* (*MERS-CoV*) and *Hedgehog coronaviruses*. Two remaining 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 epidemics 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 virus [4]. Other genetically similar SARS coronaviruses have also been previously identified, e.g. 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 epidemics representing the third major zoonotic transmission of the novel pathogenic CoVs capable of causing life-threatening disease in humans in the recent history [7]. Famously, these epidemics were caused by severe acute respiratory syndrome coronavirus (SARS-CoV) in 2002-2003 and Middle East respiratory syndrome coronavirus (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 to SARS and MERS [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 (229E, NL63) and betacoronaviruses (OC4, HKU1), common in both adults and children, are also highly likely to originate from bat or rodent reservoir [10].

Viral structure and the replicative cycle

Spherical structure of the virus contains the core with ribonucleoprotein of the helical structure enclosed by the nucleocapsid proteins. Within the viral membrane envelope proteins are anchored, with the crown-shaped spikes formed by the S protein protruding from the virion membrane [8].
Receptor binding for coronaviruses is dependent on the spike protein, which is equipped with an extracellular, transmembrane anchor and intracellular tail domains [11], and have been previously identified as a likely vaccine target [12]. SARS CoV-2 spike protein is prone to accumulate mutations compared to the SARS, especially at the interface with the ACE-2 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 two 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, the spike proteins consist of approximately 1300 amino acids which form virus membrane anchored trimeric structures [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 furin cleavage site enhances virus-cell fusion [2]. In SARS CoV-2, similarly to SARS-CoV receptor binding domain (RBD) is complexing via a form of the hydrophobic tunnel with salt bridges within the ACE2 receptors on the human cells [19].

Following binding of the ACE2 complex to the S1 part of the spike protein translocation of the ACE-virus complex to the endosomes occurs. Subsequently, S1/S2 protein is cleaved by the endosomal proteases (eg. Cathepsin L), which unmask the S2 fusion peptides activating integration between the viral and host membranes within the endosome (Figure 1). In this process, CoVs receptor 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 ACE-2 [17]. The C terminal portion of the S protein contains two trimeric helical heptad repeat structures (HR1 and HR2). These
structures are of the primary importance for the virus-host cell fusion, folding into a stable protease resistant six-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 gp41 [20].

Interestingly, to ensure efficient replication SARS CoV-2 not one but two RNA dependent polymerases are involved, one primer dependent and the second with primase activity, therefore with the capacity to initiate replication. Viral genome is released and translated by viral replicase complex and cut by proteinases. 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. 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 COVID-19

In most (~80%) cases, 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/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 SpO2 <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 gender, as well as cancers (including haematological), obesity, renal and liver diseases [24].
Notably, in severe COVID-19, increased activity of the inflammatory parameters, including interleukin 1 (IL-1), 6 (IL-6) or tumor necrosis factor (TNF α) levels reflect the cytokine release 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 the 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 the IL-6 is a result of a hyperactive humoral response from the cytotoxic T lymphocytes and is a marker of respiratory failure, shock and multi-organ failure. It is unknown however, if the IL-6 and other acute phase parameter increases 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].

**COVID-19 genome and sequence variability**

As noted above, the virus was first identified from samples of seller from the seafood market in Wuhan with diagnosed severe pneumonia. After confirming that bronchoalveolar lavage samples contained the coronavirus genetic material 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, 82-87.% similarity to human SARS CoV and 79.6% with SARS-CoV BJ0 1 [8, 28]. In similarity plots of this 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 spike 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 outgroup in the phylogenetic studies [3, 30].

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

Membrane (M) protein of coronaviruses in known to induce neutralizing antibody response being well recognized by the CD8 lymphocytes [34]. RdRp polymerase is directly involved in the transcription of the viral RNA being coupled with non-structural protein 14 exonuclease which possesses the proofreading function. Of note, antiviral nucleotide analogues including remdesivir or favipiravir inhibit RdRp [35]. Over the course of epidemics, 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 possess lesser proofreading activity and therefore are linked with decreased virulence [36]. Nucleocapsid (N) protein is not only a structural protein, but is also crucial for the viral transcription and assembly, sharing approximately 90-93% amino acid sequence identity with SARS, which confirms conserved nature of this protein. It contains two 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 proinflammatory cyclooxygenase-2 protein [38]. Importantly, nucleocapsid protein is abundantly expressed during infection [39].

**Molecular evolution of SARS CoV-2**

Genetic diversity among CoVs results from the RdRp generated errors as well as recombination, both within host and heterologous – a well-known mechanism involved in the viral evolution [40]. Sequence data collected so far guide phylogenetic investigation to inform molecular epidemiology, analyse transmission patterns, infection hotspots and to 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 both aminoacid and nucleotide sequence analyses may indicate the character of transmission and evolution of the virus [41]. In the report analyzing the 2666 spike proteins from China, including 507 of human origin have predicted risk of cross-species transmissions based on the aminoacid sequence of the spike protein underscoring 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 spike protein may affect the binding forces on the host cells and therefore disease course [42]. Further research suggests 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 aminoacid composition of SARS CoV-2, including the S (spike) region is highly relevant for the development of vaccines and novel therapeutic targets [11, 44].

Data on the phylogenetic networks have indicated that SARS CoV-2 evolved into at least 58 haplotypes and two 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]. In analyses of the phylogenetic networks, 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 has been 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. in their recent report defined four other major clades of SARS-CoV-2 [49]. Similarly to the clade V (ORF3a, codon position - G251V), these four clades were named as I, D, G and S due to missense mutations in: ORFab (position: V378I and G392D), S protein (position D614G) and ORF8 (position L84S), respectively. In addition, authors identified nine minor clades which were named either for 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 for the following nucleotide substitutions: G11410A or C17373A in ORFab. The major five 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. were also able to successfully classify 95.6% of 4000 additional viral genomes deposited in GISAID between March 31st and April 15th [47]. Guan et al. 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 one or two clades in each geographic region was found. For example, the Asian and Oceanian genomes were largely clade I whereas clade S predominated in North America cases while European genomes were predominantly classified as clade G [49]. Korber et al. reported that the earliest D614G mutation of SARS-CoV-2 in Europe was identified in Germany (EPI_ISL_406862, sampled 1/28/2020). 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 D614G frequency increase at such an alarming rate indicates a relative fitness advantage to the original Wuhan strain that enables more rapid spread [50]. Recently, Zhang et al., have observed that retroviruses pseudotyped with spike G614 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 Spike protein into the pseudovirion. Of note, spike G614 variant did not bind ACE2 more efficiently than spike D614, and the pseudoviruses containing these spike proteins were neutralized with comparable efficiencies by convalescent plasma. These results
show spike D614 variant is less stable than G614 ones, which is consistent with epidemiological data suggesting that viruses with latter variant transmit more efficiently [51]. Furthermore, apart from the clades described above, Van Dorp et al., revealed 198 recurrent mutations (ca. 80% representing non-synonymous changes) in the SARS-CoV-2 by analysis of a set of 7666 complete viral genome sequences acquired from the GISAID [52]. The authors focused on the mutations which have emerged independently multiple times (homoplasies) and found that three 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. 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 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 D614G mutation was also most common SNP detected by van Dorp et al. in a set of SARS-CoV-2 genomes from GISAID included in their homoplasy analysis [52]. Guan et al. suggested that the nonsynonymous mutations in the N protein, which have key roles 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, the 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. 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 [49].

**Conclusion**

Review of the molecular evolution of the virus described above summarizes brief evolutionary history of the SARS CoV-2. To understand the transmission patterns and course of the viral disease spread among people, it is crucial to investigate genetic variability and mutation characteristics of the COVID-19. Further genetic evolution of the virus is certain – possible changes in the affinity to human receptors such as ACE-2 [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 the 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 due to genetic bottlenecks for RNA viruses often occur during respiratory droplet transmission. Additionally, it has already been 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 the 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 BCG vaccine, however the mechanism for such protection remains unclear [54]. Also, it should be considered that in North-European countries 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 the COVID-19, as suggested by several recent studies on the T-cell reactivity to the SARS-CoV-2 proteins, especially spike
[55, 56]. Of note, these hypotheses require further confirmation by rigorous 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, disease severity and predict the virulence of the COVID-19.
Figures:

Figure 1. Simplified outline of the SARS CoV-2 integration with the host cell. HR1 and HR2 - heptad repeat structure 1 and 2, 6-HB – folded 6 helix heptad structure.

Figure 2. SARS CoV-2 genome organization. Abbreviations for the coding regions: ORF – open reading frame, S – spike protein, E – envelope, N – nucleocapsid protein.
References

1. Ka-Wai Hui E. Reasons for the increase in emerging and re-emerging viral infectious diseases. Microbes and Infection. 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. Nature Microbiology. 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.

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: e02298-02216.

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. New England Journal of Medicine. 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. New England Journal of Medicine. 2020 Apr 24. [Epub ahead of print].

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. The cytokine release syndrome (CRS) of severe COVID-19 and Interleukin-6 receptor (IL-6R) antagonist Tocilizumab may be the key to reduce the mortality. Int J Antimicrob Agents. 2020: 105954-105954.

27. Zheng M, Gao Y, Wang G, et al. Functional exhaustion of antiviral lymphocytes in COVID-19 patients. Cellular and Molecular Immunology. 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: 33-33.

30. Forster P, Forster L, Renfrew C, et al. Phylogenetic network analysis of SARS-CoV-2 genomes. Proceedings of the National Academy of Sciences. 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, et al. Structural basis of SARS-CoV-2 3CLpro and anti-COVID-19 drug discovery from medicinal plants. Journal of Pharmaceutical Analysis. 2020 Mar 26. [Epub ahead of print].

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 Pharmaceutica Sinica B. 2020 Apr 20. [Epub ahead of print].

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. The genomic variation landscape of globally-circulating clades of SARS-CoV-2 defines a genetic barcoding scheme. bioRxiv. 2020 Apr 23. [Epub ahead of print].
50. Korber B, Fischer W, Gnanakaran S, et al. Spike mutation pipeline reveals the emergence of a more transmissible form of SARS-CoV-2. bioRxiv. 2020 Apr 30. [Epub ahead of print].

51. Zhang L, Jackson CB, Mou H, et al. The D614G mutation in the SARS-CoV-2 spike protein reduces S1 shedding and increases infectivity. bioRxiv. 2020 Jun 12. [Epub ahead of print].

52. van Dorp L, Acman M, Richard D, et al. Emergence of genomic diversity and recurrent mutations in SARS-CoV-2. Infection, Genetics and Evolution. 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). Proceedings of the National Academy of Sciences. 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. Nature Reviews Immunology. 2020; 20: 457-458.