**Introduction**

In December 2019, an outbreak of pneumonia of unknown origin was reported in Wuhan, Hubei Province, China. Most of these cases were epidemiologically linked to the Huanan Seafood Wholesale Market. Inoculation of bronchoalveolar lavage fluid obtained from patients with pneumonia of unknown origin into human airway epithelial cells and Vero E6 and Huh7 cell lines led to the isolation of a novel coronavirus, SARS-CoV-2, previously named 2019-nCoV [1].

Coronaviruses belong to the family Coronaviridae and are positive single-stranded RNA viruses surrounded by an envelope. They are divided into four genera:...
such as vomiting, diarrhea, and abdominal pain [13, 14]. Patients with COVID-19 had gastrointestinal symptoms failure [13].

syndrome may worsen rapidly and die of multiple organ failure [1, 13]. Patients that develop acute respiratory distress syndrome (ARDS) with tachypnea and ground-glass opacity are described on chest X-ray [1, 13]. Those patients who develop pneumonia, multiple mot-
ting and ground-glass opacity are described on chest X-ray [1, 13]. Patients that develop acute respiratory distress syndrome may worsen rapidly and die of multiple organ failure [13].

It has also been reported that about 2–10% of the pa-
tients with COVID-19 had gastrointestinal symptoms such as vomiting, diarrhea, and abdominal pain [13, 14]. Diarrhea and nausea preceded the development of fever and respiratory symptoms in 10% of patients [13].

**Viral Transmission**

At present, the exact mechanism of transmission of SARS-CoV-2 is still not completely understood. Human-to-human transmission via droplets is the main route of transmission within a susceptible population. Chinese health authorities reported an R0 of 1.4–2.5 on January 23, 2020, to the WHO International Health Regulations (2005) Emergency Committee. Transmission by asym-
ptomatic carriers cannot be ruled out. Actually, it was re-
ported that an asymptomatic family member who traveled from the epidemic center of Wuhan was most likely responsible for a familial cluster of COVID-19 pneumonia once back home. Her reverse transcription polymerase chain reaction (RT-PCR) result was positive for SARS-CoV-2, but her chest CT images did not show significant alterations [15].

Another route of possible viral transmission is the oral-fecal route. The scientific literature showed that SARS-CoV and MERS-CoV are zoonotic in origin; they cause severe respiratory syndrome and are often fatal [12]. Since the beginning of the epidemic in late December 2019, SARS-CoV-2 has now spread to all continents, and as of March 18, 2020, the WHO communicated 179,111 confirmed cases and 7,426 deaths globally (Situation Report-57). In this review, we try to summarize the most recent knowledge about some epidemiological parameters including clinical symptoms, trans-
missibility of the virus, and the incubation period. Fur-
thermore, the molecular diagnostics, protein modeling of the spike glycoprotein, and phylogenesis of the virus will be discussed.

**SARS-CoV-2 Epidemiology**

**Clinical Symptoms**

Patients infected with SARS-CoV-2 can present a wide range of symptoms ranging from mild to severe. Fever, cough, and shortness of breath are the most common symptoms reported in 83, 82, and 31% of patients [13]. In those patients who develop pneumonia, multiple mot-
ting and ground-glass opacity are described on chest X-ray [1, 13]. Patients that develop acute respiratory distress syndrome may worsen rapidly and die of multiple organ failure [13].

It has also been reported that about 2–10% of the pa-
tients with COVID-19 had gastrointestinal symptoms such as vomiting, diarrhea, and abdominal pain [13, 14]. Diarrhea and nausea preceded the development of fever and respiratory symptoms in 10% of patients [13].
Molecular Diagnosis of SARS-CoV-2

Confirmation of cases with suspected SARS-CoV-2 infection is performed by detection of unique viral sequences with nucleic acid amplification tests such as reverse real-time PCR (rRT-PCR). As soon as on January 7, 2020, the Chinese health authorities had declared that a novel coronavirus was responsible for this outbreak of pneumonia in Wuhan, a European network of academic and public laboratories designed an rRT-PCR protocol based on the comparison and alignment of previously available SARS-CoV and bat-related coronavirus genome sequences as well as five sequences derived from the novel coronavirus SARS-CoV-2 made available by the Chinese authorities [23]. Three assays were developed. The first-line assay targets the E gene encoding for the envelope protein, which is common to the Sarbecovirus subgenus, while the second specific assay targets the RdRp gene encoding for RNA-dependent RNA polymerase. This assay contains two probes: one probe, which reacts with the SARS-CoV and SARS-CoV-2 RdRp gene, and a second probe (RdRP_SARSr-P2) which is specific to SARS-CoV-2. Finally, the third additional confirmatory assay targets the nucleocapsid (N) gene. This last assay was not further validated because it is slightly less sensitive [23]. This protocol was adopted in more than 30 European laboratories [24]. Recently, a novel rRT-PCR assay targeting a different region of the RdRp/Hel gene of SARS-CoV-2 has been developed that showed a higher sensitivity and specificity than the RdRp-P2 assay [25].

Currently, several amplification protocols are available on the market and validated for in vitro diagnostic use (CE marked): GeneFinder™ COVID-19 Plus RealAmp Kit (OSANG Healthcare Co., Ltd, South Korea); genesig® Real-Time PCR Coronavirus (COVID-19) (genesig, UK); Allplex™ 2019-nCoV Assay (Seegene, South Korea), etc. Highly sensitive and specific diagnostic assays are key to the identification of cases, contact tracing, identification of the animal source, and implementation of control measures [26–28].

When performing nucleic acid amplification test assays, it is useful to remind ourselves that several factors can be responsible for a negative result in an infected individual, such as the poor quality of a specimen, the time of specimen collection (specimen collected too early or too late during infection), inappropriate handling or shipment of the specimen, and technical reasons.

SARS-CoV-2 Modeling

Coronavirus entry into the host cell is mediated by the transmembrane spike (S) glycoprotein that forms heterotrimers that protrude from the viral surface [29]. The S protein is composed of the two subunits S1 and S2 responsible for binding to the host cell receptor and fusion of the viral and cellular membranes, respectively. Different coronaviruses use different domains within the S1 subunit to enter the cell. These domains are named S^A and S^B. SARS-CoV and SARS-related coronaviruses interact with the angiotensin-converting enzyme 2 (ACE2) via domain S^B to enter target cells [30–34]. It has recently been shown that SARS-CoV-2 binds the ACE2 receptor via the S^B domain similarly to SARS-CoV, and that murine polyclonal antibodies inhibited SARS-CoV-2 entry into the cell mediated by S. These data suggest that cross-neutralizing antibodies targeting conserved S epitopes elicited by vaccination could be used against SARS-CoV-2, SARS-CoV, and SARS-related coronaviruses [35].

Previous studies have shown the presence of positive selective pressure on the Nucleocapsid, Spike glycoprotein, and ORF1a/b regions, while until now no evidence of a positive selective pressure has been found on the Envelope, Membrane, and other ORF proteins.

In the Nucleocapsid region, significant (p < 0.05) pervasive episodic selection was found in 2 sites. In amino acid position 380 of the Wuhan coronavirus sequence there is a Gln residue instead of an Asn, while in amino acid position 410 there is a Thr residue instead of an Ala. Significant (p < 0.05) pervasive negative selection in 6 sites (14%) has been evidenced and confirmed by FUBAR (Fast Unconstrained Bayesian Approximation) analysis [36].

In the Spike glycoprotein region, significant (p < 0.05) pervasive episodic selection was found in 2 different sites (536th and 644th nucleotide position using the reference sequence). In the 536th amino acid position of the Wuhan coronavirus sequence there is an Asn residue instead of an Asp residue, while in amino acid position 644 there is a Thr residue instead of an Ala residue. Significant (p < 0.05) pervasive negative selection in 1,065 sites (87%) has been evidenced and confirmed by FUBAR analysis, suggesting that the S region could be highly conserved [36].

Regarding the sites under positive selective pressure found on the Spike glycoprotein, the results have shown that amino acid position 536 in COVID-19 has an Asn residue, while the Bat SARS-like coronavirus has a Gln

DOI: 10.1159/000507423
residue; the SARS virus, instead, has an Asp residue. In amino acid position 644 of the COVID-19 sequence there is a Thr residue, while the Bat SARS-like virus has a Ser residue; instead, the SARS virus has an Ala residue. Another study highlighted that several key residues responsible for binding of the SARS-CoV receptor-binding domain to the ACE2 receptor were variable in the COVID-19 receptor-binding domain (including Asn439, Asn501, Gln493, Gly485, and Phe486; COVID-19 numbering), and a number of deletion events in amino acid positions 455–457, 463–464, and 485–497 occurred in the bat-derived strains [37].

Also in the ORF1ab region, potential sites under positive selective pressure have been found (p < 0.05). Particularly, in the amino acid position 501, COVID-19 has a Gln residue, the Bat SARS-like coronavirus has a Thr residue, and the SARS virus has an Ala residue. In position 723 of the COVID-19 sequence there is a Ser residue, and the SARS virus has an Ala residue. In amino acid position 1,010, COVID-19 has a Pro residue, the Bat SARS-like coronavirus has a His residue, and the SARS virus has an Ile residue.

As for the residue in position 723 (543 in the nsp3 protein), the COVID-19 sequence displays a Ser, replacing for Gly in the Bat SARS-like and SARS coronaviruses. In this case, it may be argued that this substitution could increase local stiffness of the polypeptide chain both for a steric effect (in contrast to Ser, Gly has no side chain) and for the ability of the Ser side chain to form H-bonds. Moreover, Ser can act as a nucleophile in determined structural environments, such as those of enzymes’ active sites, and can be a phosphorylation site. However, within the I-TASSER model, this position is predicted to have low solvent accessibility.

Regarding the amino acid position 1,010 (corresponding to position 192 of the nsp3 protein), the homologous region of the Bat SARS-like coronavirus and SARS virus has a polar and an apolar amino acid, respectively, while COVID-19 has a Pro residue. In this case, it may be speculated that due to the steric bulge and stiffness of Pro, the molecular structure of COVID-19 may undergo a local conformational perturbation compared to the proteins of the other two viruses. In nsp3, the mutation falls near the polyprotein domain similar to a phosphatase present also in the SARS coronavirus (PDB code 2ACF) playing a key role in the replication process of the virus in infected cells [38]. According to the I-TASSER model, the position is partially accessible to the solvent.

The sites under positive selective pressure in this protein may suggest a possible interpretation of some clinical features of this virus compared to SARS and Bat SARS-like coronavirus. This analysis should find which are probably the most common sites undergoing an amino acid change, providing insight into some important proteins of COVID-19 that are involved in the mechanism of viral entry and viral replication. These data should contribute to improving our understanding of how this virus acts in its pathogenicity. Furthermore, to identify a potential molecular target is fundamental to follow the molecular evolution of the virus, which can suggest some interesting sites for a potential therapy or vaccine.

Fig. 1. HHpred mapping of the homologous structures of all proteins coded by the 2019-nCoV genome. Numbers below the open rectangles refer to all open reading frames (ORFs) detected. Red strips represent the PDB (Protein Data Bank) homologous structures. PDB codes are reported within the corresponding red stripes. Details are given in Table 1.
The structural similarity of the region in which the positive selective pressure occurs, and the stabilizing mutation falling in the endosome-associated protein-like domain of the nsp2 protein, should be probable reasons why this virus is more contagious than SARS. Instead, the destabilizing mutation located near the phosphatase domain of the nsp3 protein may explain why viral replication is slower than in SARS with a longer incubation period. Anyway, further studies are needed on this aspect [39].

The availability of protein structural information is an essential prerequisite for the interpretation of biological phenomena. In this case, knowledge of the virus’s protein structure would greatly enhance the possibility of understanding the biological meaning of the observed mutations. Now, only the X-ray structure of COVID-19 nsp5 protease (PDB code 6LU7) is available, although it is expected that many other structures will become available soon. In the meantime, homology modeling could provide preliminary structural clues. Homology modeling needs structural templates sharing sufficient sequence similarity to the targets. In Figure 1 and Table 1, a list of potential templates for homology modeling of the proteins coded by the COVID-19 genome is displayed. The structures with the largest coverage and the greatest sequence identity have been incorporated into Figure 1 and Table 1. According to this list, it is evident that most of the viral proteins are at modeling distance from PDB structures. This information should be exploited as soon as possible.

### Table 1. PDB codes of all the homologous models showing similarity with the relative SARS-CoV-2 region, reporting the protein function, the sequence position within the viral ORFs, and the percentage of identity

<table>
<thead>
<tr>
<th>PDB code</th>
<th>Protein</th>
<th>Sequence position within virus ORF</th>
<th>% identity</th>
</tr>
</thead>
<tbody>
<tr>
<td>2GDT</td>
<td>nsp1 from SARS coronavirus</td>
<td>13–127</td>
<td>86%</td>
</tr>
<tr>
<td>3LD1</td>
<td>nsp2a from IBV</td>
<td>105–564</td>
<td>11%</td>
</tr>
<tr>
<td>2IDY</td>
<td>nsp3 from SARS coronavirus</td>
<td>818–929</td>
<td>76%</td>
</tr>
<tr>
<td>2ACF</td>
<td>nsp3 from SARS coronavirus</td>
<td>1,025–1,197</td>
<td>72%</td>
</tr>
<tr>
<td>2W2G</td>
<td>human SARS coronavirus unique domain</td>
<td>1,231–1,494</td>
<td>75%</td>
</tr>
<tr>
<td>2KV</td>
<td>human SARS coronavirus unique domain</td>
<td>1,368–1,561</td>
<td>75%</td>
</tr>
<tr>
<td>2KAF</td>
<td>nsp3 from human coronavirus, C-domain</td>
<td>1,496–1,561</td>
<td>70%</td>
</tr>
<tr>
<td>4YPT</td>
<td>nsp3 from murine hepatitis virus</td>
<td>1,449–1,872</td>
<td>27%</td>
</tr>
<tr>
<td>4MOW</td>
<td>structure papain-like protease from human SARS coronavirus</td>
<td>1,563–1,881</td>
<td>82%</td>
</tr>
<tr>
<td>2K87</td>
<td>putative RNA-binding protein from human SARS coronavirus</td>
<td>1,906–2,021</td>
<td>81%</td>
</tr>
<tr>
<td>3GZF</td>
<td>nsp4 from feline coronavirus, C-domain</td>
<td>3,164–3,260</td>
<td>36%</td>
</tr>
<tr>
<td>5CSO</td>
<td>protease complex from human SARS-3CL protease complex</td>
<td>3,264–3,569</td>
<td>96%</td>
</tr>
<tr>
<td>5MK5</td>
<td>DHBN domain from human BLM helicase</td>
<td>3,636–4,042</td>
<td>11%</td>
</tr>
<tr>
<td>3UB0</td>
<td>nsp7 and 8 complex from feline coronavirus</td>
<td>3,857–3,942</td>
<td>41%</td>
</tr>
<tr>
<td>2AHM</td>
<td>nsp9 from human SARS coronavirus</td>
<td>3,942–4,140</td>
<td>97%</td>
</tr>
<tr>
<td>3EE7</td>
<td>nsp9 from human SARS coronavirus</td>
<td>4,141–4,253</td>
<td>96%</td>
</tr>
<tr>
<td>5NM</td>
<td>nsp16/nsp10 complex from human betacoronavirus</td>
<td>4,254–4,394</td>
<td>59%</td>
</tr>
<tr>
<td>2XQP chain B</td>
<td>nsp16/nsp10 complex from human SARS coronavirus</td>
<td>4,264–4,385</td>
<td>98%</td>
</tr>
<tr>
<td>6NUR</td>
<td>nsp12 from human SARS coronavirus</td>
<td>4,394–5,329</td>
<td>96%</td>
</tr>
<tr>
<td>6YIT</td>
<td>nsp13 from human SARS coronavirus</td>
<td>5,325–5,926</td>
<td>100%</td>
</tr>
<tr>
<td>5CBT</td>
<td>nsp14/nsp10 complex from human SARS coronavirus</td>
<td>5,927–6,453</td>
<td>95%</td>
</tr>
<tr>
<td>5YVD</td>
<td>nsp15 from MERS coronavirus</td>
<td>6,454–6,799</td>
<td>51%</td>
</tr>
<tr>
<td>2XQP chain A</td>
<td>nsp16/nsp10 complex from human SARS coronavirus</td>
<td>6,800–7,089</td>
<td>92%</td>
</tr>
<tr>
<td>6ACC</td>
<td>spike glycoprotein from human SARS coronavirus</td>
<td>7,098–8,312</td>
<td>76%</td>
</tr>
<tr>
<td>5XZ9</td>
<td>protein E from human SARS coronavirus</td>
<td>8,648–8,710</td>
<td>87%</td>
</tr>
<tr>
<td>5ISD</td>
<td>domain 245 from salmonella global</td>
<td>8,724–8,817</td>
<td>14%</td>
</tr>
<tr>
<td>4HXT</td>
<td>receptor TatC in DDM from <em>Aquifex aeolicus</em></td>
<td>8,943–9,022</td>
<td>19%</td>
</tr>
<tr>
<td>1XAK</td>
<td>ORF7 from human SARS coronavirus</td>
<td>9,017–9,098</td>
<td>88%</td>
</tr>
<tr>
<td>4UD1</td>
<td>nucleocapsid from MERS coronavirus, domain N-term</td>
<td>9,253–9,420</td>
<td>50%</td>
</tr>
<tr>
<td>6ICZ</td>
<td>human P complex</td>
<td>9,422–9,454</td>
<td>36%</td>
</tr>
<tr>
<td>2CJR</td>
<td>nucleocapsid protein SARS</td>
<td>9,492–9,609</td>
<td>96%</td>
</tr>
</tbody>
</table>

PDB, Protein Data Bank; ORF, open reading frame.
SARS-CoV-2 Phylogeny

Phylogenetic analysis of the SARS-CoV-2 genomes showed that the novel coronavirus responsible for the pneumonia outbreak in Wuhan, China, belongs to the Betacoronavirus genus, subgenus Sarbecovirus [37]. Within the Betacoronavirus genus, 2019-nCoV (SARS-CoV-2) is distant from SARS-CoV (about 79% identity) and MERS-CoV (about 50% identity) responsible for the 2002–2003 [4] and 2012 [7] epidemics, respectively, but closely related (88% identity) to the two bat-derived (SARS)-like coronaviruses bat-SL-CoVZC45 and bat-SL-CoVZXC21 [37].

The origin of the virus is still unclear; however, genomic analysis suggests that SARS-CoV-2 is most closely related to viruses previously identified in bats (Fig. 2). It is plausible that there were other intermediate animal transmissions before its introduction into humans. However, there is no evidence of snakes as an intermediary [36].

Using 74 publicly shared novel coronavirus (nCoV) genomes, we examined genetic diversity to infer the date of the common ancestor and the rate of spread. The high similarity of the genomes suggests they share a recent common ancestor. Otherwise, we would expect a greater number of differences between the samples. The jump from bats to humans most likely occurred in late November or early December 2019 (November 25, 2019; 95% HPD: September 28, 2019; December 21, 2019) [40].

Previous research on related coronaviruses suggests that these viruses accumulate between 1 and 3 changes in their genome per month (rates of $3 \times 10^{-4}$ to $1 \times 10^{-3}$ per site per year). Molecular clock calibration estimated the evolutionary rate of the SARS-CoV-2 whole genome sequences at $6.58 \times 10^{-3}$ substitutions per site per year (95% HPD: $5.2 \times 10^{-3}$ to $8.1 \times 10^{-3}$).

Fig. 2. Phylogenetic tree of all the 2019-nCoV sequences available on February 9, 2020 ($n = 74$, cartooned outbreak clade), plus $n = 1$ Bat CoV-like sequence, $n = 5$ Human SARS, $n = 4$ Bat SARS-like sequences, and $n = 2$ from MERS. Numbers along the branches represent a bootstrap value $> 0.90$. The scale bar represents expected substitutions per nucleotide site.
Fig. 3. Maximum clade credibility phylogeny estimated using all the complete and near-complete 2019-nCov genome sequences available up to February 9, 2020 (n = 74). Clade posterior probabilities are shown at well-supported nodes. The colors represent different locations.
The outbreak first started in Wuhan, China, but cases have been identified in many East and South-East Asian countries, the USA, Australia, the Middle East, and Europe. Vietnam, Japan, and Germany have reported transmission within the country, albeit always with a known link to Wuhan, China (Fig. 3).

Conclusions

This study is a picture of the current research on molecular evolution, epidemiology, and diagnostics in response to the outbreak of COVID-19. Many studies have been published within different scientific disciplines with the intent to control and prevent this pandemic. Phylogenetic analysis and homology modeling add new knowledge together with epidemiological and diagnostic methods. Studies exploring the genome and the structure of the viral proteins are essential in order to define prevention and control measures to minimize the impact of the outbreak. All this knowledge will pave the way for the development of a vaccine and antiviral therapy.

Disclosure Statement

The authors have no conflicts of interest to declare.

Funding Sources

There was no funding for this review.

Author Contributions


References

DO: 10.1159/000507423


