Infection, Genetics and Evolution 85 (2020) 104556

Contents lists available at ScienceDirect

Infection, Genetics and Evolution

journal homepage: www.elsevier.com/locate/meegid

Research paper

# Quasi-species nature and differential gene expression of severe acute respiratory syndrome coronavirus 2 and phylogenetic analysis of a novel Iranian strain

Abozar Ghorbani[a], Samira Samarfard[b,*], Amin Ramezani[c], Keramatollah Izadpanah[a], Alireza Afsharifar[a], Mohammad Hadi Eskandari[d], Thomas P. Karbanowicz[b], Jonathan R. Peters[b]

[a] Plant Virology Research Centre, College of Agriculture, Shiraz University, Shiraz, Iran
[b] Queensland Biosciences Precinct, The University of Queensland, St Lucia 4072, Queensland, Australia
[c] Shiraz Institute for Cancer Research, School of Medicine, Shiraz University of Medical Sciences, Shiraz, Iran
[d] Department of Food Science and Technology, College of Agriculture, Shiraz University, Shiraz, Iran

ARTICLE INFO

*Keywords:*
SARS-CoV-2
Inter-host
Single nucleotide polymorphism
Phylogenetic
Open reading frame
Nucleocapsid

ABSTRACT

A novel coronavirus related to severe acute respiratory syndrome virus, (SARS-CoV-2) is the causal agent of the COVID-19 pandemic. Despite the genetic mutations across the SARS-CoV-2 genome being recently investigated, its transcriptomic genetic polymorphisms at inter-host level and the viral gene expression level based on each Open Reading Frame (ORF) remains unclear. Using available High Throughput Sequencing (HTS) data and based on SARS-CoV-2 infected human transcriptomic data, this study presents a high-resolution map of SARS-CoV-2 single nucleotide polymorphism (SNP) hotspots in a viral population at inter-host level. Four throat swab samples from COVID-19 infected patients were pooled, with RNA-Seq read retrieved from SRA NCBI to detect 21 SNPs and a replacement across the SARS-CoV-2 genomic population. Twenty-two RNA modification sites on viral transcripts were identified that may cause inter-host genetic diversity of this virus. In addition, the canonical genomic RNAs of N ORF showed higher expression in transcriptomic data and reverse transcriptase quantitative PCR compared to other SARS-CoV-2 ORFs, indicating the importance of this ORF in virus replication or other major functions in virus cycle. Phylogenetic and ancestral sequence analyses based on the entire genome revealed that SARS-CoV-2 is possibly derived from a recombination event between SARS-CoV and Bat SARS-like CoV. Ancestor analysis of the isolates from different locations including Iran suggest shared Chinese ancestry. These results propose the importance of potential inter-host level genetic variations to the evolution of SARS-COV-2, and the formation of viral quasi-species. The RNA modifications discovered in this study may cause amino acid sequence changes in polyprotein, spike protein, product of ORF8 and nucleocapsid (N) protein, suggesting further insights to understanding the functional impacts of mutations in the life cycle and pathogenicity of SARS-CoV-2.

## 1. Introduction

Severe acute respiratory syndrome-related coronavirus (SARS-CoV-2), COVID-19 clinically presents the symptoms with fever and mild respiratory illness, with ~20% of cases requiring hospital intervention, of which 5% require intensive support (Wu and Mc Googan, 2020; Huang et al., 2020; Zhou et al., 2020). As of mid-August 2020, there have been 21, 294,845 reported cases globally with 761,779 deaths (World health organization, 2020). Coronaviruses (COVs) are the largest group of none-segmented, single-stranded, positive-sense ribonucleic acid (+ssRNA) viruses in the order *Nidovirales* (Siddell et al., 2019). They are classified within the *Coronaviridae* family, subfamily *Coronavirinae* and cause enzootic infections in a broad range of vertebrates including mammals and birds.

Phylogenetic analysis of full-length genomes of various coronavirus species has revealed that SARS-CoV-2 fall within the subgenus *Sarbecovirus* of the genus *Betacoronavirus*. Betacoronavirus are further divided into four lineages (A–D). Lineage B includes SARS-CoV-2 and about 200 viral sequences (Letko et al., 2020). The complete ssRNA genome sequence of SARS-CoV-2 contains 29,891 nucleotides, encoding 7986 amino acids that make viral structural proteins (SPs) and non-structural proteins (NSPs). The genome comprises of 12 putative

[*] Corresponding author.
*E-mail address:* s.samarfard@uq.edu.au (S. Samarfard).

https://doi.org/10.1016/j.meegid.2020.104556
Received 12 June 2020; Received in revised form 30 August 2020; Accepted 9 September 2020
Available online 13 September 2020

functional open reading frames (ORFs) that are arranged in order ORF1/ab which is translated to NSPs (Wu et al., 2020a). NSPs of SARS-CoV-2 including papain-like protease (NSP3), cysteine 3C-like proteinase (NSP5), the primary RNA-dependent RNA polymerase (NSP12), and helicase/triphosphatase (NSP13), have enzymatic functions during the viral life cycle (Wu et al., 2020b; Chan et al., 2020). The partially overlapping 5′-terminal ORF1a/b encodes the large replicase polyproteins pp1a and pp1ab, that are likely to be involved in the viral transcription and replication. The −1 ribosomal frameshift upstream of the ORF1a stop codon, allows continued translation of ORF1b to pp1ab, that is proteolysis cleaved into 16 NSPs, which are essential in forming the replicase/transcriptase complex (RTC) (Chan et al., 2020; Sola et al., 2015). The 3′-terminal ORFs of SARS-CoV-2 genome encode SPs, including spike glycoprotein (S, ORF2), envelope (E, ORF4), membrane (M, ORF5) and nucleocapsid (N, ORF9a). These proteins are essential for SARS-CoV-2 assembly and infection, and accessory proteins (3a, 6, 7a, 7b, 8, and 10) that are expressed from nine predicted sub-genomic RNAs (Wu et al., 2020b; Chan et al., 2020).

High genetic diversity of RNA viruses at both inter- and intra-host level lead to multiple circulating quasi-species of low/ high frequency, due to the error-prone nature of their genomic replications (Li et al., 2007). Heterogeneity in virus virulence and in host factors, as well as polymorphic virulent quasi-species, can affect the extent of disease severity, antiviral immune response, viral phenotype, and the sensitivity of molecular and serological diagnostic assays (Vignuzzi et al., 2006). The high mutation rate of RNA viruses drives virulence modulation, viral evolvability, and genome variability, allowing viruses to escape host immunity and to develop drug resistance (Duffy, 2018). The mutations that occur during viral replication can form micro variants that diverge from a master sequence by the emergence of single nucleotide polymorphisms (SNPs) within a population. These SNPs may be the origin of viral quasi-species formation that can lead to resistance-breaking strains in some hosts (Domingo et al., 2012; Seguin et al., 2014). In acute RNA viruses, mutation hotspots mainly occur in immunogenic sites to enable the virus to escape the host antiviral immune system, therefore, from a public-health perspective, understanding the mutation rate of the SARS-CoV-2 at interhost levels as it spreads through the population is imperative. (Rogozin and Pavlov, 2003).

This study uses SNP prediction based on RNA-Seq data and the extent of molecular divergence between SARS-CoV-2 specific reads mapped to RefSeq to understand mutational factors driving the evolution of SARS-CoV-2 and the pattern of spread of SNPs in the viral population via SARS-CoV-2 transcriptomic data analyses. The expression levels (transcripts per Kilobase Million, TPM) of SARS-Cov-2 ORFs in RNA-seq data were also explored to provide new insights into the virus replication strategy. This study also provides a phylogenetic analysis of SARS-CoV-2 including an Iranian isolate that was recently sequenced.

## 2. Materials and methods

### 2.1. Data collection and pre-processing

The SARS-CoV-2 transcriptome was previously delineated by Illumina MiniSeq runs on iSeq 100 sequencer using total RNA extracted from hCov-19 infected patients' throat swab (Fang et al., 2020). Samples were collected by Fang et al. (2020) from SARS-CoV-2 infected patients at the Central Hospital of Wuhan. Raw RNA sequencing (RNA-Seq) data for the four throat swab samples (Fang et al., 2020) were downloaded from the National Centre for Biotechnology Information, Sequence Read Archive (NCBI SRA) under BioSample SRA: PRJNA616446 (Table 1). Four throat swab samples were collected from four individual patients, with analysis performed on each technical replicate (Fang et al., 2020). Data analysis on fastq files were conducted with CLC Genomics Workbench (version 12, QIAGEN, Venlo, The Netherlands). The reads were adaptor trimmed (Illumina MiniSeq) and quality trimmed (using the default parameters: bases below 15 nt were trimmed, ambiguous nucleotides maximal 2).

**Table 1**
Raw data statistics of SARS-CoV-2 human infected libraries.

| Accession number | Total reads | Reads mapped to virus | Percentage of mapped-reads |
|---|---|---|---|
| SRR11454606 | 11, 336,944 | 3616 | 0.03 |
| SRR11454609 | 17,121,629 | 66,420 | 0.39 |
| SRR11454610 | 14,337,950 | 126,390 | 0.88 |
| SRR11454611 | 1,405,599 | 3383 | 0.24 |

### 2.2. SARS-CoV-2 gene expression using RNA-Seq data

Clean reads were mapped to the SARS-CoV-2 reference genome (GenBank: NC_045512.2), with the ORFs were annotated and TPM was calculated for each ORF using CLC Genomics Workbench (version 12, QIAGEN, Venlo, The Netherlands). Means and standard deviation for four replications were calculated using Microsoft Excel 2013.

### 2.3. Validation of SARS-CoV-2 gene expression by reverse transcriptase quantitative PCR (RT-qPCR)

For validation of differential expression of SARS-CoV-2 genes, total RNA was extracted from 70 COVID-19 infected patients using Exgene™ Viral DNA/RNA (GeneAll, Korea) following the manufacturer's instructions. UltraPlex™ 1-Step ToughMix® (Quantabio, USA) and specific primers and probes for N and RdRp genes (Table 2), were used for expression analysis. Data were expressed relative to the expression of the Human RNase P as an internal control gene. The threshold cycle (Ct) number was calculated from log scale amplification curves by ABI QuantStudio (Applied Biosystems, USA) software. RT-qPCR conditions were 50 °C for 10 min, 95 °C for 1 min for initial denaturation, followed by 40 cycles of 95 °C for 10 s, and 60 °C for 30 s. The RT-qPCR experiment was performed in triplicate. Amplification efficiencies were calculated and included in data normalization. Data normalization was performed using pfaffl formula (Balotf et al., 2012).

### 2.4. ORFs and SNP prediction

Virus-mapped reads were used for variant discovery using CLC genomic workbench 12. Low-frequency variant detection tools from CLC genomic workbench 12 were used to obtain SNPs from the viral population. SNP discovery Quality filter Neighborhood was set to radius 5, minimum central quality 20 and minimum Neighborhood quality 15. Minimum frequency was selected on 5% for the whole of the virus population in four samples and each sample individually. SNPs were validated, with their position in each ORF annotated using Geneious Prime 2019 (Biomatters, New Zealand). The SNPs that cause changes to amino acids were determined. The correlation of number of SNP in each gene and the gene length was done using Minitab version 17. The distribution of SNP and their frequency between samples have been assessed and compared between the samples. A Pearson correlation of number of SNP in each gene and the gene length was calculated using excel 2013.

### 2.5. Phylogeny, ancestral and consensus sequence reconstruction

All SARS-CoV-2 whole genome sequences obtained from human hosts with geographical annotations were obtained from NCBI and Global initiative on sharing all influenza data (GISAID). SARS-CoV-2 whole-genome sequences were aligned with the ClustalW method and a phylogenetic tree was constructed using MEGA7 following the neighbor-joining method, maximum composite likelihood-parameter distance matrix, bootstrap values of 1000 replicates and with a 70% threshold score (Kumar et al., 2016). The ancestral sequence of the

**Table 2**
Primers and probes that were used for RT-qPCR.

| Primers and Probes | Target | Sequence | Amplicon length (bp) |
|---|---|---|---|
| RdRP_F | RNA-dependent RNA polymerases | GTCTCTATAGAAATAGAGATGTTGACACA | 134 |
| RdRP_ R | | ACCTTGAGATGCATAAGTGCTATTGA | |
| RdRP_ P | | FAM –AATGATGATACTCTCTGACGATGCT-BHQ | |
| N gene_F | Nucleoprotein ORF | GACCCCAAAATCAGCGAAAT | 72 |
| N gene_R | | TCTGGTTACTGCCAGTTGAATCTG | |
| N gene_P | | FAM-ACCCCGCATTACGTTTGGTGGACC-BHQ | |
| RNase P_F | Human Ribonuclease P (Internal control) | AGATTTGGACCTGCGAGCG | 65 |
| RNase P_R | | GAGCGGCTGTCTCCACAAGT | |
| RNase P_P | | ROX –TTCTGACCTGAAGGCTCTGCGCG-BHQ | |

virus was constructed by using the maximum likelihood (ML) algorithm based on the generated phylogenetic tree in MEGA 7.

## 3. Results and discussion

### 3.1. Draft genome of SARS-CoV-2 and SNP profile

Using RNA-Seq data, this study was able to determine the SNPs on an unprecedented scale for the SARS-CoV-2 population. Identification of these SNPs is critical to understanding SARS-CoV-2 variation capacity. The abundance of reads from total RNA-Seq that mapped to the SARS-CoV-2 reference genome provided about 98–100% coverage across the viral genome, with robust sequencing depth (3383–126,390 reads) of the whole viral genome (Table 1). The average fold coverage of viral reads was sufficient to detect low-abundance microbial species genome from metagenomic datasets, implying the reliability of the assembled draft viral genome for SNP determination and further analyses (Albertsen et al., 2013). A total of 199,809 of 32,865,178 reads, from four biological replicates were mapped to the SARS-CoV-2 genome (Table 1). Single nucleotide polymorphisms were determined by annotating ORFs for the virus population in four samples. The SNP profiling revealed that the ORFs encoding the RdRp and S proteins, ORF8 and nucleoprotein (N) of SARS-CoV-2 have undergone mutations within this population at inter-host level. At least 22 sites displayed substantial differences with frequency ranging from 5.06% to 99.8% across the mapped reads when applied with a threshold of 5% for SNP detection, suggesting potential RNA modifications (Table 3 and Fig. 1). Twenty-one low- and higher- frequency SNPs were identified, with variable density across the viral genome and a replacement mutation at a low frequency of 6.4% (Table 3). The Pearson correlation of number of SNP in each gene and the gene length calculated $R^2 = 0.998$ with $P < 0.01$, which showed frequency varied according to the gene length and more frequent in the longest gene. The positions of SNPs in the coding regions of SARS-CoV-2 annotated ORFs were in RdRp ORF (14 SNPs), S ORF (4 SNPs), ORF8 (2 SNPs), N ORF (1 SNP) (Table.2 and Fig. 1). The SNP C → T with variant frequency ranging from 5% to 99.8% was the most abundant SNP type that was positioned only in polyprotein and S protein regions of the genome (Table 3). The only polymorphisms that cause changes in the deduced encoded amino acid sequence were positioned at 5122, 9512, 10,843, 11,206 and 11,876 of RdRp ORF and the whole of SNPs in S, ORF8 and N ORFs (Table 3 and Fig. 1).

The comparative analysis of frequency and distribution of SNPs identified between RNA-Seq reads can be influenced by the sequencing depth However, it was not the case for our study and the outcome of our analysis showed that SRR11454606 sample (with a lower sequencing coverage than the other samples) exhibited more genome-wide SNPs (Supplementary Fig. 1 and Supplementary Table 1) which may be related to host or the virus plasticity. Recent studies have shown that SARS-CoV-2 harbor diversity both between strains or individuals (interhost variation) as well as within a single individual (intra-host variation) (Karamitros et al., 2020). The difference in frequency and distribution of the identified SNVs between throat swab samples in this

**Table 3**
Single-nucleotide polymorphisms (SNPs) among a population of human SARS-COV-2[a] genome RNA-Seq reads.

| Gene | Reference Position | Type of Variation | Reference | Allele | Frequency (%) | Amino acid Change[c] |
|---|---|---|---|---|---|---|
| Poly_protein | 885 | SNP | C | T | 13.17 | – |
| Poly_protein | 2910 | SNP | A | G | 5.05 | – |
| Poly_protein | 5122 | SNP | G | A | 5.33 | + |
| Poly_protein | 6645 | SNP | T | C | 5.06 | – |
| Poly_protein | 7002 | SNP | C | T | 6.95 | – |
| Poly_protein | 8094 | SNP | T | C | 8.86 | – |
| Poly_protein | 8517 | SNP | C | T | 99.80 | – |
| Poly_protein | 9512 | SNP | C | T | 5.74 | + |
| Poly_protein | 10,843 | SNP | G | A | 5.06 | + |
| Poly_protein | 11,206 | SNP | C | T | 5.23 | + |
| Poly_protein | 11,876 | SNP | C | A | 25.25 | + |
| Poly_protein | 21,199 | SNP | G | C | 6.12 | – |
| Poly_protein | 21,209 | SNP | T | G | 5.76 | – |
| Poly_protein | 21,220 | Replacement | AG | T | 6.45 | – |
| Poly_protein | 21,258 | SNP | T | A | 5.47 | – |
| Spike_protein | 24 | SNP | G | C | 6.25 | + |
| Spike_protein | 2254 | SNP | C | T | 38.54 | + |
| Spike_protein | 2464 | SNP | C | T | 7.96 | + |
| Spike_protein | 3017 | SNP | C | T | 13.77 | + |
| ORF8[b] | 184 | SNP | G | C | 6.58 | + |
| ORF8 | 251 | SNP | T | C | 98.92 | + |
| N_protein | 610 | SNP | G | A | 6.43 | + |

[a] Severe acute respiratory syndrome coronavirus 2.
[b] Open Reading Frame.
[c] "+"Changed amino acid, "– "doesn't change amino acid.

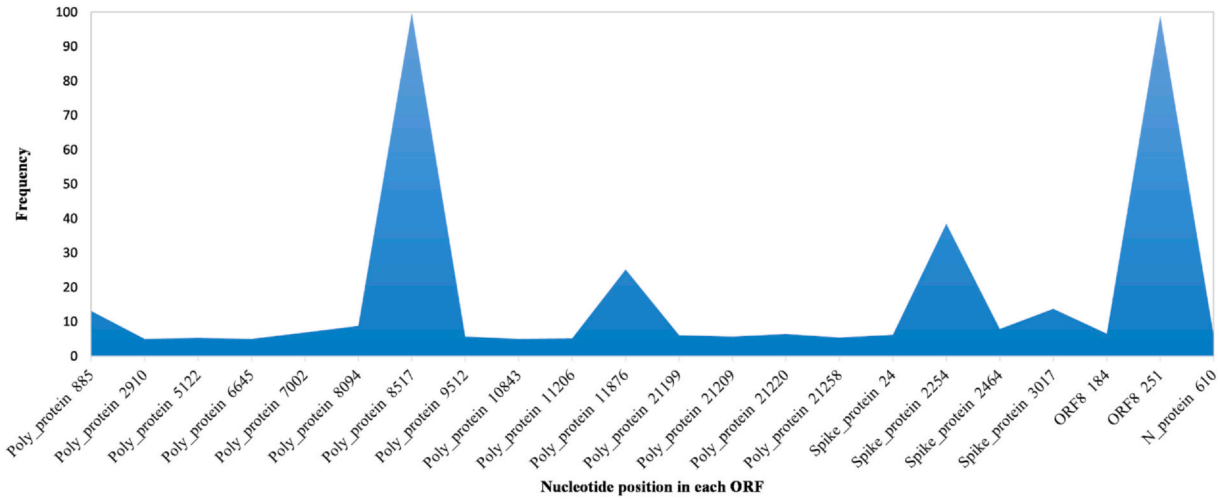

**Fig. 1.** Frequency of the single-nucleotide polymorphism (SNP) positions on the ORFs of SARS-CoV-2 genome in RNA-Seq reads in Human.

study (same strain origin) may also refer to the antigenic variability of SARS-CoV-2 and subsequent immune evasion.

The ORF1/ab encodes polyproteins are likely to be involved in the viral transcription and replication (Chan et al., 2020), but it is not known if the identified amino acid change in ORF1/ab affects these viral processes in the variant of SARS-CoV-2 isolated from throat swab samples. This study observed four SNPs in the spike glycoprotein (S protein), located in the peptide signal and S2 domain, indicating that some sites of S protein might be subjected to positive selection (Zhan et al., 2020). The S glycoprotein of COVs plays a crucial role in the binding and attachment of the virion to the cell membrane, through the host ACE2 receptor (Xiao et al., 2003). Therefore, SARS-CoV-2 S glycoprotein is an important determinant of tissue and cell tropism as well as host range (Millet and Whittaker, 2015). Although the real functional impact of the low-frequency SNPs (C > T) at S protein sequence level remains unclear, the mutation 24 positioned in signal peptide domain and other SNPs located in S2 domain could modify the viral tropism, suggesting new hosts or increasing SARS-CoV-2 pathogenesis (Shang et al., 2020; Millet and Whittaker, 2015). Among these hotspots, one low-frequency SNP in position 610 is located within the N protein which is probably associated with an overall increased mutation rate. The N protein of SARS-CoV-2 is vital for packaging the positive-strand viral RNA genome into helical ribonucleocapsid (RNP) over its interactions with the viral genome and membrane protein M (Chan et al., 2020). The N protein provides a potential vaccine antigen, as it is important in viral immune response. Determining the high-frequency SNPs encoding N genes of different SARS-CoV-2 strains will be key in developing a potential vaccine (Zhao et al., 2005).

Throughout viral replication, hundreds of viral progenies are produced that vary at least at one position and the subsequent rounds of replication generate a more complex mutant distribution that includes variants lying farther away from each other in the viral sequence frame (Lauring and Andino, 2010). This group of mutants forms a “cloud” of variants called quasispecies. RNA viruses have a quasispecies nature with a high mutation rate within infected hosts (Lauring and Andino, 2010). A viral quasispecies includes large numbers of genome variants forming the population structure of viruses that are genetically linked over mutation, interact with each other at a functional level, and jointly contribute to form the main characteristics of a viral population (Lauring and Andino, 2010). The high mutation rate in RNA viruses leads to a high level of intra-host variants (Ni et al., 2016; Domingo et al., 2012). The SNPs and substitution observed in quasispecies of SARS-CoV-2 reflect the inter-patient capacity of the polymorphic quasispecies which may increase rapidly during the outbreak and cause viral immunological escape, resistance to anti-viral drugs and affect the sensitivity of the molecular diagnostics assays. In this study, those SNPs that were detected at low frequency implies viral variations of low impact on the functionality of the genome. However, the frequencies of SNPs in a viral population can be largely affected by the virus population size and epidemic characteristics (Noh et al., 2017; Karamitros et al., 2016). Based on the previous studies SARS-CoV shared 99.8% sequence homology with SARS-like CoV, with a total of 202 single-nucleotide (nt) variations (SNVs) identified across the genome (Song et al., 2005). The SARS-CoV-2 mutations at inter-host level may attribute to the viral survival and immune evasion in infected cells because the human immune system is found to be less responsive to RNAs with SNPs (Karikó et al., 2005). It is yet to be examined if the SNPs detected in ORFs and aa sequence modifications detected in the current study are unique to SARS-CoV-2 or conserved in other taxonomically related coronaviruses.

### 3.2. Expression of SARS-CoV-2 proteins in throat swab cells

At least one Gig of clean data from each sample obtained from transcriptome sequencing, was used as a query to analyze the viral gene expression level after quality control, mapping reads to the SARS-CoV-2 Refseq and viral ORFs annotations. Differentially expressed genes (DEGs) were characterized for each sample (adjusted-value $p < 0.01$) (Fig. 2). Based on the number of specific transcripts (TPM) identified in RNA-Seq, SARS-CoV-2 N which is positioned toward the 3′ terminus of the genome was the most highly expressed gene than the 5′ terminal genes encoding polyproteins and S protein (Fig. 3). Quantitative comparison of TPM reads showed that the N RNA is the most abundantly expressed transcript, followed by ORFs 10, 8, M, 7a, 1a/b ant other SARS-Cov-2 ORFs (Fig. 3). One-way ANOVA $P < 0.01$ revealed that there is no significant difference in viral transcript levels between genes encoding polyprotein, ubiquitin ligases (ORF10), M, and accessory proteins (ORF7a, ORF7b and ORF8), whereas their expression levels were higher than other ORFs (Fig. 2). There are no results on the SARS-CoV gene expression level in patient tissue, and it is unclear which accessory genes of SARS-CoV-2 are highly expressed. In this study, ORF10 encoding a putative ubiquitin ligase is the most highly expressed accessory gene of SARS-CoV-2 when compared with the other viral auxiliary genes. The viral RNA replication has been evolved toward an equilibrium at which a heterogeneous population of viral RNAs (quasispecies) is reproduced with high efficiency (Holland et al., 1992; Domingo et al., 12). The viral RdRP is up regulated during the evolution of RNA viruses, since RdRP is the central enzyme during this process and has the optimal combination of RNA synthesis efficiency and nucleotide incorporation fidelity (Holland et al., 1982; Snijder et al.,

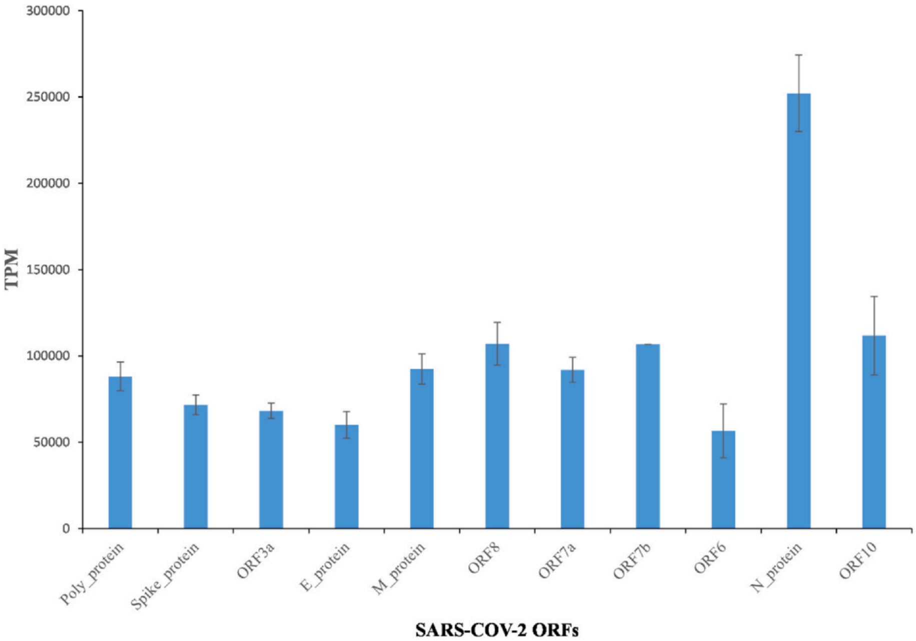

**Fig. 2.** Expression levels (Reads per kilobase of transcript per million mapped reads, RPKM) of SARS-COV-2 ORFs in RNA-Seq data in infected patients. The mean and standard deviation of four biological replicates is shown. One-way ANOVA ($p < 0.01$) was used for statistical analysis.

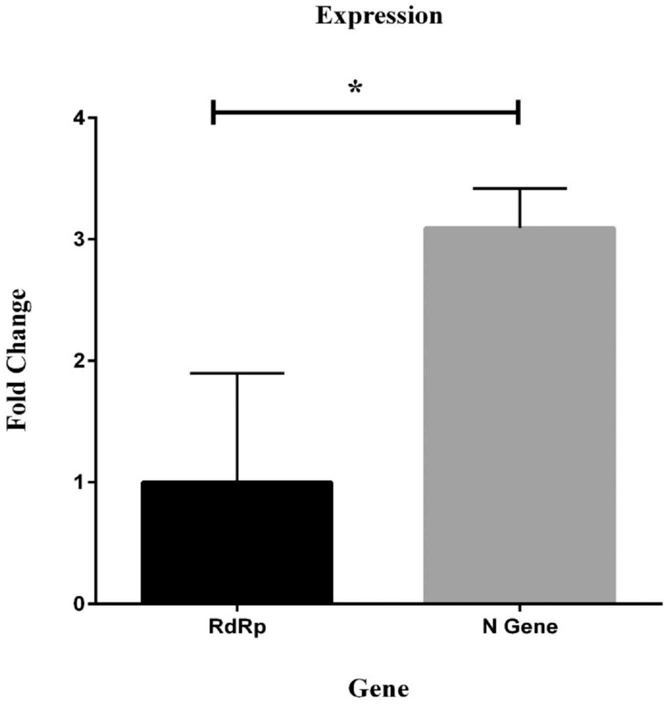

**Fig. 3.** Expression level of RdRP and N ORFs in 70 patients that were infected by SARS-CoV-2 using RT-qPCR. Data were compared with *t*-test analysis method.

2016). Earlier reports have indicated that upon viral entry, the N-protein of the SARS-CoV is primarily distributed in the cytoplasm, after which they expressed heterologously or in infected cells (Surjit et al., 2005; You et al., 2005; Rowland et al., 2005). The co-expression of the M and N viral proteins was reported as a minimal requirement for the formation of SARS-COV virus-like particles in transfected human 293 renal epithelial cells (Huang et al., 2004). Association of M and N proteins stabilizes the nucleocapsid (N protein-RNA complex), and the internal core of virions to promote the completion of viral assembly (Chan et al., 2020; Fehr and Perlman, 2015; Fehr et al., 2016; Narayanan et al., 2000). The higher expression level of M gene than those of E and S genes implies the association and binding of M and N for the completion of viral assembly in infected cells.

### 3.3. Confirmation of DEGs by RT-qPCR

To validate the differential expression profiles of viral genes obtained by RNA-Seq analysis, RT-qPCR was performed on selected DEGs of SARS-COV-2. The RT-qPCR results on the comparative expression level of N and RdRP genes were consistent with those of the RNA-Seq analysis. The RT-qPCR findings demonstrated the same relative regulation of DEGs of the virus as the RNA-Seq data (Fig. 3). Consistent with virus replication during the infection time course, N displayed a higher mRNA level in throat swab cells, indicating the boosted viral replication upon the viral infection. After infection, new genomic RNAs are formed and structural genes are up regulated, then assembly of particles occurs. Assembly and release of virions are the last stages of the virus life cycle (Chan et al., 2020). It has been reported that coronaviruses use an RdRP processivity factor to expedite replication of their RNA genome. The function of SARS-COV RNA polymerase was revealed to be linked with the capability of an up-regulated nsp7/nsp8/nsp12 complex that associates with the activity of nsp14-exonuclease for removing terminal mismatches from an RNA duplex (Bouvet et al., 2012). Since the N gene of COVs is highly immunogenic, un-glycosylated, and is highly expressed in infected cells, it may be an ideal target for developing diagnostic tools that can detect the SARS-CoV-2, as only one SNP was observed in this part of the viral genome based on the present data analysis (Shi et al., 2003; Guan et al., 2004).

### 3.4. Phylogenetic and ancestry studies of SARS-COV-2, Iranian isolate

The whole-genome-based phylogenetic trees for selected SARS-CoV-2 strains reported from different locations are deduced using the NJ method. The consensus tree was derived by bootstrapping values of 1000 replicates and with a 50% threshold score based on the NJ algorithm (Kumar et al., 2016). The robustness of the tree topology was estimated by branch support and the placement of root of the SARS-CoV-2 NJ tree was considered by introducing different out-groups

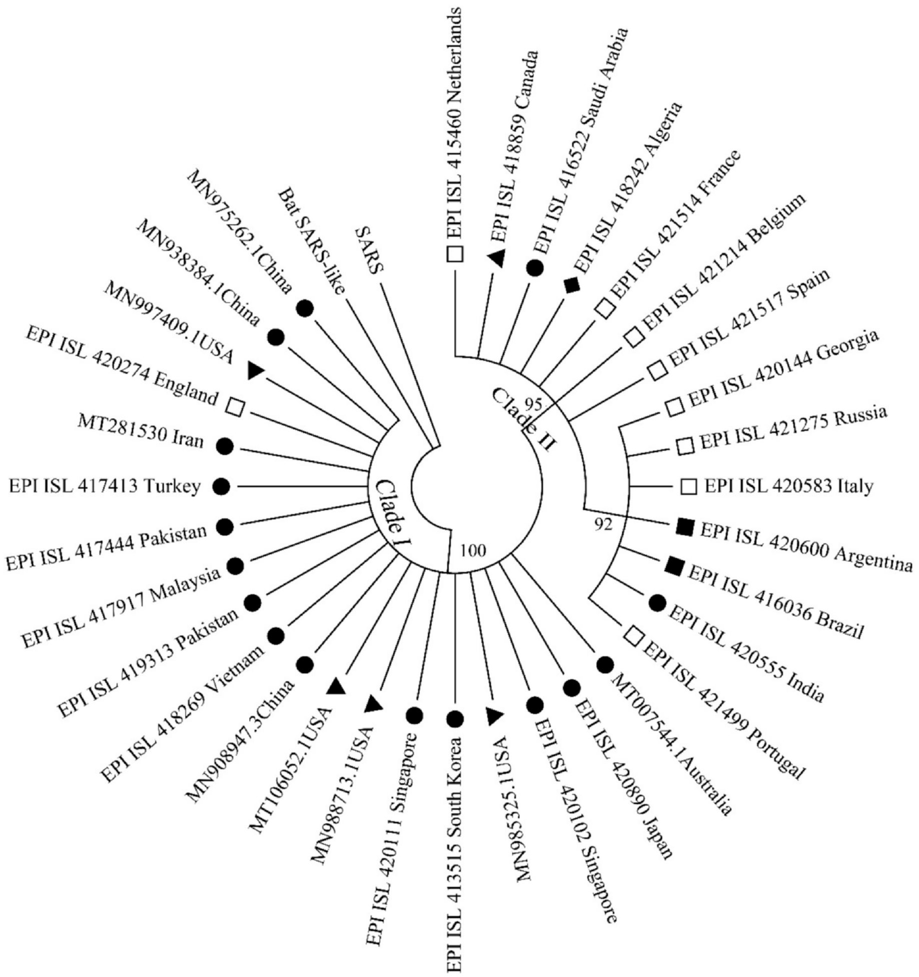

**Fig. 4.** Phylogenetic analysis of full-length genomes SARS-CoV-2 Iranian isolate and isolates from different locations. □: European isolates. ■: South American isolates. ●: Asian Isolates. ◆: African Isolates. ▲: North American Isolates. SARS and bat-SARS-like are out-groups.

including SARS-COV and Bat SARS-like virus. Since the branching order of the NJ tree shows location of virus linkage, in principle the NJ tree provides the details that COVID-19 spread from one place to another. The phylogenetic tree shows that the branches of the NJ tree are mainly located in the same geographical continent and includes two distinct phylogenetic clades of SARS-CoV-2 isolates (Fig. 4), corresponding to subgroups I and II. The clade I mostly include the Asian SARS-CoV-2 sequences. Whilst the phylogenetic clade II is more diverse than clade I in terms of geographic origins of viral strains and includes two clusters particularly with the root distributed in Europe (Fig. 4). The Iranian genome sequence (MT281530) appeared, in contrast, to be in a different cluster of the clade I including a Turkish genome sequence (EPIISL417413) (Fig. 4). Based on the position and the evolutionary relationship of strains presented in the SARS-CoV-2 NJ tree, as well as the presence of Chinese isolates in both clades, it is deduced that the SARS-CoV-2 may have begun to spread to several regions of the world before its outbreak in Wuhan (Gao et al., 2020). To trace back the potential time of the evolution involving ancestral lineages of SARS-CoV-2, this hypothesis was tested further by joint maximum likelihood reconstruction of ancestral sequences based on the entire genome for all SARS-CoV-2 strains presented in the NJ tree and compared with the whole genome phylogeny (Fig. 5). The findings presented did not support the hypothesis, indicating that the COVID-19 may have begun to spread to several regions in the world before its outbreak in Wuhan. Based on the reconstruction of ancestral sequences SARS-CoV-2 may have been derived from a recombination event between two different COVs including SARS-CoV and Bat SARS-like CoV. Furthermore, the most recent common ancestor in the recombinant region of the clade leading to the Chinese SARS-CoV-2 isolates (MN975262 and MN938384) is the ancestor of other isolates from different locations.

## 4. Conclusion

As more SARS-CoV-2 isolate sequences become available, stronger lineage variation may occur over the pandemic time. However, evolutionary factors require further time for further recombination within the viral population. A caveat of this remains the limited number of samples, and the limited number of SARS-CoV-2 genome sequences applied for SNPs prediction based on transcriptome and phylogenetic studies, respectively. This study delineated the mutation positions of SARS-CoV-2 based on unambiguous mapping SRA reads to RefSeq, and investigated the origin of viral quasispecies formation through the analysis of SNPs in the SARS-CoV-2 population structure. It was determined that highly expressed ORFs are a prerequisite for understanding the functional relationship of viral structural and non-structural proteins, replication mechanism, and host-viral interactions involved in pathogenicity.

As in other COVs, a high mutation rate, including SNP, in SARS-CoV-2 may allow rapid viral evolution forming variants with altered biological characteristics including new host specificity and drug sensitivity. The SNPs detected in this study may provide a solid basis for understanding variant generation at inter-host level. To understand further about intra-host level variation and the evolutionary selection pressure on SARS-CoV-2 further investigations are required in detail

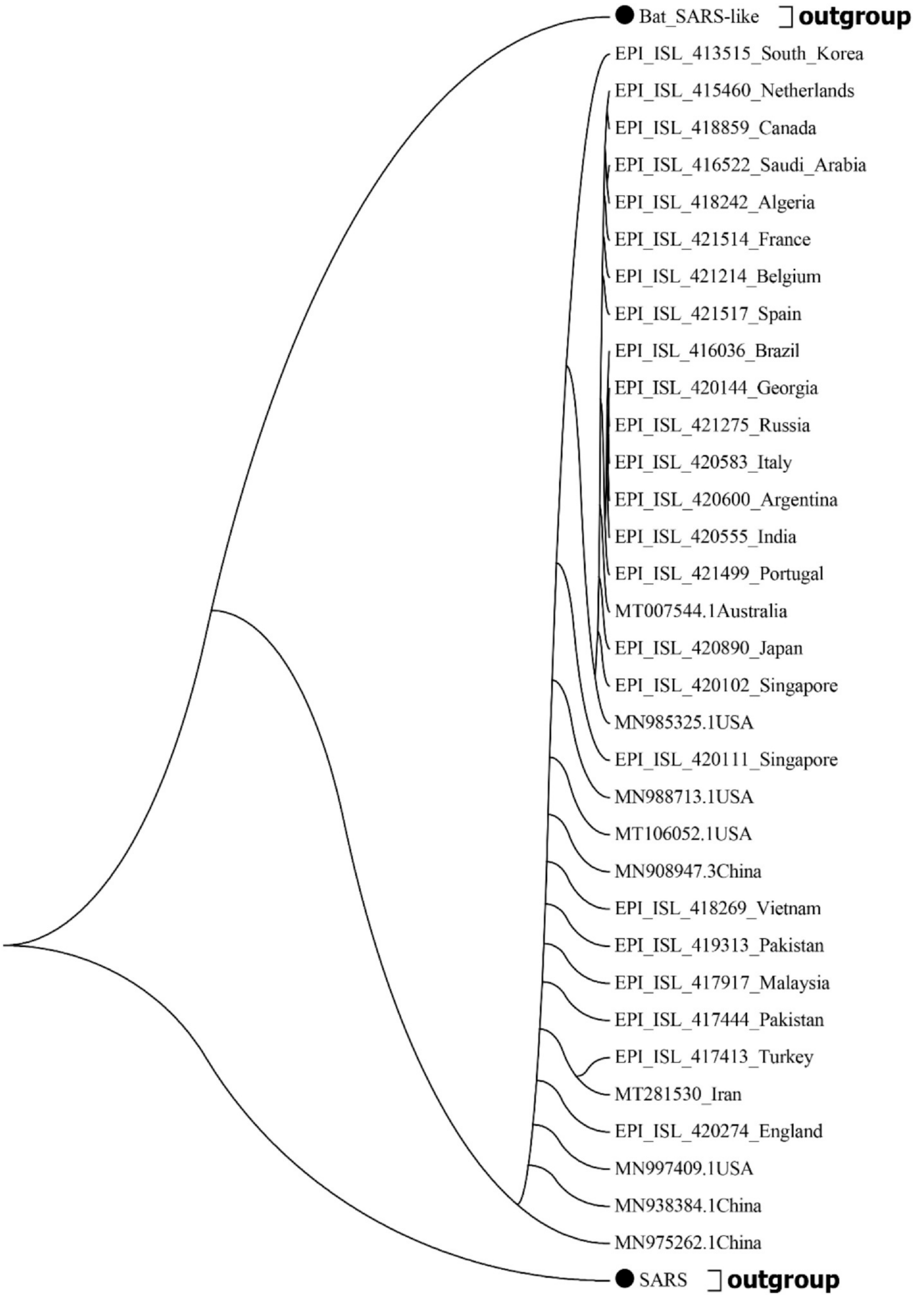

**Fig. 5.** Ancestor analysis of full-length genomes SARS-CoV-2 Iranian isolate and isolates from different location. SARS and bat-SARS-like are out-groups.

based on the sequence alignment of many strains from different locations. However, new molecular features of SARS-CoV-2 particularly co-expression of SARS-CoV-2 proteins will need to be studied further in different cell types that have an intact interferon system. Moreover, it is still unclear if the SNPs detected in annotated ORFs in this current study are unique to SARS-CoV-2 or conserved in other taxonomically related coronaviruses. A comparative study must be conducted on the distribution and functional significance of SNPs in the SARS-CoV-2 populations to gain further understanding of SARS-CoV-2 quasi-species impacts on viral pathogenicity and replication in infected cells. To conclude, this data provides a platform with new directions to investigate the mechanisms that play a key role in the pathogenicity of SARS-CoV.

Supplementary data to this article can be found online at https://doi.org/10.1016/j.meegid.2020.104556.

## Declaration of Competing Interest

All authors report no conflict of interest related to the submitted work.

## Acknowledgements

The corresponding author evinces her deep gratitude to the Shiraz University of Medical Sciences for providing samples and facilities for the completion of this work.

## References

Albertsen, M., Hugenholtz, P., Skarshewski, A., Nielsen, K.L., Tyson, G.W., Nielsen, P.H., 2013. Genome sequences of rare, uncultured bacteria obtained by differential coverage binning of multiple metagenomes. Nat. Biotechnol. 31, 533–538.

Balotf, S., Niazi, A., Kavoosi, G., Ramezani, A., 2012. Differential expression of nitrate reductase in response to potassium and sodium nitrate: realtime PCR analysis. Australian Aust. J. Crop Sci. 6, 130–134.

Bouvet, M., Imbert, I., Subissi, L., Gluais, L., Canard, B., Decroly, E., 2012. RNA 3′-end mismatch excision by the severe acute respiratory syndrome coronavirus non-structural protein nsp10/nsp14 exoribonuclease complex. Proc. Natl. Acad. Sci. 109, 9372–9377.

Chan, J.F.W., Kok, K.H., Zhu, Z., Chu, H., To, K.K.W, Yuan, S., Yuen, K.Y., 2020. Genomic characterization of the 2019 novel human-pathogenic coronavirus isolated from a patient with atypical pneumonia after visiting Wuhan. Emerg. Microbes Infect. 9, 221–236.

Domingo, E., Sheldon, J., Perales, C., 2012. Viral quasispecies evolution. Microbiol. Mol. Biol. Rev. 76, 159–216.

Duffy, S., 2018. Why are RNA virus mutation rates so damn high? PLoS Biol. 16, e3000003.

Fang, B., Liu, L., Yu, X., Li, X., Ye, G., Xu, J., Zhang, L., Zhan, F., Liu, G., Pan, T., Shu, Y., 2020. Genome-wide data inferring the evolution and population demography of the novel pneumonia coronavirus (SARS-CoV-2). bioRxiv. https://doi.org/10.1101/2020.03.04.976662.

Fehr, A.R., Perlman, S., 2015. Coronaviruses: an overview of their replication and pathogenesis. In: Coronaviruses. Humana Press, New York, NY, pp. 1–23.

Fehr, A.R., Channappanavar, R., Jankevicius, G., Fett, C., Zhao, J., Athmer, J., Meyerholz, D.K., Ahel, I., Perlman, S., 2016. The conserved coronavirus macrodomain promotes virulence and suppresses the innate immune response during severe acute respiratory syndrome coronavirus infection. MBio 7 (6) (e01721–16).

Gao, Y., Li, T., Luo, L., 2020. Phylogenetic study of 2019-nCoV by using alignment-free method. arXiv (preprint arXiv:2003.01324).

Guan, M., Chen, H.Y., Foo, S.Y., Tan, Y.J., Goh, P.Y., Wee, S.H., 2004. Recombinant protein-based enzyme-linked immunosorbent assay and immunochromatographic tests for detection of immunoglobulin G antibodies to severe acute respiratory syndrome (SARS) coronavirus in SARS patients. Clin. Diagn. Lab. Immunol. 11, 287–291.

Holland, J., Spindler, K., Horodyski, F., Grabau, E., Nichol, S., VandePol, S., 1982. Rapid evolution of RNA genomes. Science 215, 1577–1585.

Holland, J.J., De La Torre, J.C., Steinhauer, D.A., 1992. RNA virus populations as quasispecies. In: Holland, J.J. (Ed.), Genetic Diversity of RNA Viruses. Current Topics in Microbiology and Immunology. vol 176 Springer, Berlin, Heidelberg.

Huang, Y., Yang, Z.Y., Kong, W.P., Nabel, G.J., 2004. Generation of synthetic severe acute respiratory syndrome coronavirus pseudoparticles:implications for assembly and vaccine production. J. Virol. 78, 12557–12565.

Huang, C., Wang, Y., Li, X., Ren, L., Zhao, J., Hu, Y., Zhang, L., Fan, G., Xu, J., Gu, X., Cheng, Z., 2020. Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. Lancet 395, 497–506.

Karamitros, T., Paraskevis, D., Hatzakis, A., Psichogiou, M., Elefsiniotis, I., Hurst, T., Geretti, A.M., Beloukas, A., Frater, J., Klenerman, P., Katzourakis, A., 2016. A contaminant-free assessment of endogenous retroviral RNA in human plasma. Sci. Rep. 6, 33598.

Karamitros, T., Papadopoulou, G., Bousali, M., Mexias, A., Tsiodras, S., Mentis, A., 2020. SARS-CoV-2 exhibits intra-host genomic plasticity and low-frequency polymorphic quasispecies. J. Clin. Virol. 131, 104585. https://doi.org/10.1016/j.jcv.2020.104585.

Karikó, K., Buckstein, M., Ni, H., Weissman, D., 2005. Suppression of RNA recognition by Toll-like receptors: the impact of nucleoside modification and the evolutionary origin of RNA. Immunity 23, 165–175.

Kumar, S., Stecher, G., Tamura, K., 2016. MEGA7: molecular evolutionary genetics analysis version 7.0 for bigger datasets. Mol. Biol. Evol. 33, 1870–1874.

Lauring, A.S., Andino, R., 2010. Quasispecies theory and the behavior of RNA viruses. PLoS Pathog. 6, e1001005.

Letko, M., Marzi, A., Munster, V., 2020. Functional assessment of cell entry and receptor usage for SARS-CoV-2 and other lineage B betacoronaviruses. Nat. Microbiol. 5, 562–569.

Li, B., Gladden, A.D., Altfeld, M., Kaldor, J.M., Cooper, D.A., Kelleher, A.D., Allen, T.M., 2007. Rapid reversion of sequence polymorphisms dominates early human immunodeficiency virus type 1 evolution. J. Virol. 81, 193–201.

Millet, J.K., Whittaker, G.R., 2015. Host cell proteases: critical determinants of coronavirus tropism and pathogenesis. Virus Res. 202, 120–134.

Narayanan, K., Maeda, A., Maeda, J., Makino, S., 2000. Characterization of the coronavirus M protein and nucleocapsid interaction in infected cells. J. Virol. 74, 8127–8134.

Ni, M., Chen, C., Qian, J., Xiao, H.X., Shi, W.F., Luo, Y., Wang, H.Y., Li, Z., Wu, J., Xu, P.S., Chen, S.H., 2016. Intra-host dynamics of Ebola virus during 2014. Nat. Microbiol. 1, 1–9.

Noh, J.Y., Yoon, S.W., Kim, D.J., Lee, M.S., Kim, J.H., Na, W., Song, D., Jeong, D.G., Kim, H.K., 2017. Simultaneous detection of severe acute respiratory syndrome, Middle East respiratory syndrome, and related bat coronaviruses by real-time reverse transcription PCR. Arch. Virol. 162, 1617–1623.

Rogozin, I.B., Pavlov, Y.I., 2003. Theoretical analysis of mutation hotspots and their DNA sequence context specificity. Mutat. Res-Rev. Mutat. 544, 65–85.

Rowland, R.R., Chauhan, V., Fang, Y., Pekosz, A., Kerrigan, M., Burton, M.D., 2005. Intracellular localization of the severe acute respiratory syndrome coronavirus nucleocapsidprotein: absence of nucleolar accumulation during infection and after expression as a recombinant protein in Vero cells. J. Virol. 79, 11507–11512.

Seguin, J., Rajeswaran, R., Malpica-Lopez, N., Martin, R.R., Kasschau, K., Dolja, V.V., Otten, P., Farinelli, L., Pooggin, M.M., 2014. De novo reconstruction of consensus master genomes of plant RNA and DNA viruses from siRNAs. PLoS One 9, e88513.

Shang, J., Wan, Y., Liu, C., Yount, B., Gully, K., Yang, Y., Auerbach, A., Peng, G., Baric, R., Li, F., 2020. Structure of mouse coronavirus spike protein complexed with receptor reveals mechanism for viral entry. PLoS Pathog. 16, e1008392.

Shi, Y., Yi, Y., Li, P., Kuang, T., Li, L., Dong, M., Ma, Q., Cao, C., 2003. Diagnosis of severe acute respiratory syndrome (SARS) by detection of SARS coronavirus nucleocapsid antibodies in an antigen-capturing enzyme-linked immunosorbent assay. J. Clin. Microbiol. 41, 5781–5782.

Siddell, S.G., Walker, P.J., Lefkowitz, E.J., Mushegian, A.R., Adams, M.J., Dutilh, B.E., Gorbalenya, A.E., Harrach, B., Harrison, R.L., Junglen, S., Knowles, N.J., 2019. Additional changes to taxonomy ratified in a special vote by the international committee on taxonomy of viruses (October 2018). Arch. Virol. 164, 943–946.

Snijder, E.J., Decroly, E., Ziebuhr, J., 2016. The nonstructural proteins directing coronavirus RNA synthesis and processing. In: Advances in Virus Research. vol. 96. Academic Press, pp. 59–126.

Sola, I., Almazan, F., Zuniga, S., Enjuanes, L., 2015. Continuous and discontinuous RNA synthesis in coronaviruses. Annu. Rev. Virol. 2, 265–288.

Song, H.D., Tu, C.C., Zhang, G.W., Wang, S.Y., Zheng, K., Lei, L.C., Chen, Q.X., Gao, Y.W., Zhou, H.Q., Xiang, H., Zheng, H.J., 2005. Cross-host evolution of severe acute respiratory syndrome coronavirus in palm civet and human. Proc. Natl. Acad. Sci. 102, 2430–2435.

Surjit, M., Kumar, R., Mishra, R.N., Reddy, M.K., Chow, V.T., Lal, S.K., 2005. The severe acute respiratory syndrome coronavirus nucleocapsid protein is phosphorylated and localizes in the cytoplasm by 14-3-3-mediated translocation. J. Virol. 79, 11476–11486.

Vignuzzi, M., Stone, J.K., Arnold, J.J., Cameron, C.E., Andino, R., 2006. Quasispecies diversity determines pathogenesis through cooperative interactions in a viral population. Nature 439, 344–348.

World health organization, 2020. Situation Report – 130. Coronavirus Disease 2019 (COVID-19).

Wu, Z., Mc Googan, J.M., 2020. Characteristics of and important lessons from the coronavirus disease 2019 (COVID-19) outbreak in China: summary of a report of 72 314 cases from the Chinese Center for Disease Control and Prevention. JAMA. 323, 1239–1242.

Wu, C., Liu, Y., Yang, Y., Zhang, P., Zhong, W., Wang, Y., Wang, Q., Xu, Y., Li, M., Li, X., Zheng, M., 2020a. Analysis of therapeutic targets for SARS-CoV-2 and discovery of potential drugs by computational methods. Acta Pharm. Sin. B 10, 766–788.

Wu, F., Zhao, S., Yu, B., Chen, Y.M., Wang, W., Song, Z.G., Hu, Y., Tao, Z.W., Tian, J.H., Pei, Y.Y., Yuan, M.L., 2020b. A new coronavirus associated with human respiratory disease in China. Nature 579, 265–269.

Xiao, X., Chakraborti, S., Dimitrov, A.S., Gramatikoff, K., Dimitrov, D.S., 2003. The SARS-CoV S glycoprotein: expression and functional characterization. Biochem. Biophys. Res. Commun. 312, 1159–1164.

You, J., Dove, B.K., Enjuanes, L., DeDiego, M.L., Alvarez, E., Howell, G., Heinen, P., Zambon, M., Hiscox, J.A., 2005. Subcellular localization of the severe acute respiratory syndrome coronavirus nucleocapsid protein. J. Gen. Virol. 86, 3303–3310.

Zhan, X.Y., Zhang, Y., Zhou, X., Huang, K., Qian, Y., Leng, Y., Yan, L., Huang, B., He, Y., 2020. Molecular Evolution of SARS-CoV-2 Structural Genes: Evidence of Positive Selection in Spike Glycoprotein. (BioRxiv).

Zhao, P., Cao, J., Zhao, L.J., Qin, Z.L., Ke, J.S., Pan, W., Ren, H., Yu, J.G., Qi, Z.T., 2005. Immune responses against SARS-coronavirus nucleocapsid protein induced by DNA vaccine. Virology 331, 128–135.

Zhou, P., Yang, X.L., Wang, X.G., Hu, B., Zhang, L., Zhang, W., Si, H.R., Zhu, Y., Li, B., Huang, C.L., Chen, H.D., 2020. Discovery of a Novel Coronavirus Associated with the Recent Pneumonia Outbreak in Humans and its Potential Bat Origin. (BioRxiv).