



Article

# Evolutionary Analysis of Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2) Reveals Genomic Divergence with Implications for Universal Vaccine Efficacy

Nanda Kumar Yellapu <sup>1</sup>, Shachi Patel <sup>1</sup>, Bo Zhang <sup>1</sup>, Richard Meier <sup>1</sup>, Lisa Neums <sup>1</sup>, Dong Pei <sup>1</sup>, Qing Xia <sup>1</sup>, Duncan Rotich <sup>1</sup>, Rosalyn C. Zimmermann <sup>2</sup>, Emily Nissen <sup>1</sup>, Shelby Bell-Glenn <sup>1</sup>, Whitney Shae <sup>1</sup>, Jinxiang Hu <sup>1</sup>, Prabhakar Chalise <sup>1</sup>, Lynn Chollet-Hinton <sup>1</sup>, Devin C. Koestler <sup>1,\*</sup> and Jeffery A. Thompson <sup>1,\*</sup>

<sup>1</sup> Department of Biostatistics & Data Science, University of Kansas Medical Center, 3901 Rainbow Boulevard, Kansas City, KS 66160, USA; nyellapu@kumc.edu (N.K.Y.); spatel14@kumc.edu (S.P.); b021z055@kumc.edu (B.Z.); rmeier2@kumc.edu (R.M.); lneums@kumc.edu (L.N.); dpei@kumc.edu (D.P.); q508x072@kumc.edu (Q.X.); duncancheru@gmail.com (D.R.); e617n596@kumc.edu (E.N.); sbell5@kumc.edu (S.B.-G.); wwwhite7@kumc.edu (W.S.); jhu2@kumc.edu (J.H.); pchalise@kumc.edu (P.C.); lhinton@kumc.edu (L.C.-H.)

<sup>2</sup> Department of Cancer Biology, University of Kansas Medical Center, 3901 Rainbow Boulevard, Kansas City, KS 66160, USA; rzimmermann@kumc.edu

\* Correspondence: dkoestler@kumc.edu (D.C.K.); jthompson21@kumc.edu (J.A.T.)

Received: 23 September 2020; Accepted: 2 October 2020; Published: 8 October 2020



**Abstract:** Coronavirus disease (COVID-19), caused by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), is one of the pressing contemporary public health challenges. Investigations into the genomic structure of SARS-CoV-2 may inform ongoing vaccine development efforts and/or provide insights into vaccine efficacy to fight against COVID-19. Evolutionary analysis of 540 genomes spanning 20 different countries/territories was conducted and revealed an increase in the genomic divergence across successive generations. The ancestor of the phylogeny was found to be the isolate from the 2019/2020 Wuhan outbreak. Its transmission was outlined across 20 countries/territories as per genomic similarity. Our results demonstrate faster evolving variations in the genomic structure of SARS-CoV-2 when compared to the isolates from early stages of the pandemic. Genomic alterations were predominantly located and mapped onto the reported vaccine candidates of structural genes, which are the main targets for vaccine candidates. S protein showed 34, N protein 25, E protein 2, and M protein 3 amino acid variations in 246 genomes among 540. Among identified mutations, 23 in S protein, 1 in E, 2 from M, and 7 from N protein were mapped with the reported vaccine candidates explaining the possible implications on universal vaccines. Hence, potential target regions for vaccines would be ideally chosen from the structural regions of the genome that lack high variation. The increasing variations in the genome of SARS-CoV-2 together with our observations in structural genes have important implications for the efficacy of a successful universal vaccine against SARS-CoV-2.

**Keywords:** COVID-19; SARS-CoV-2; phylogenetic analysis; genomic divergence; vaccine development

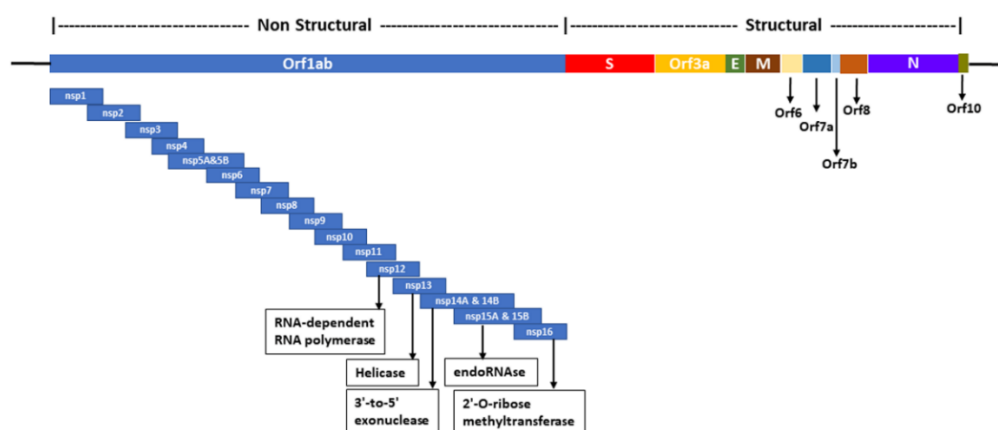
## 1. Introduction

Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) belongs to the viral family *Coronaviridae*, which includes a large number of viruses found in mammals and birds. The first human

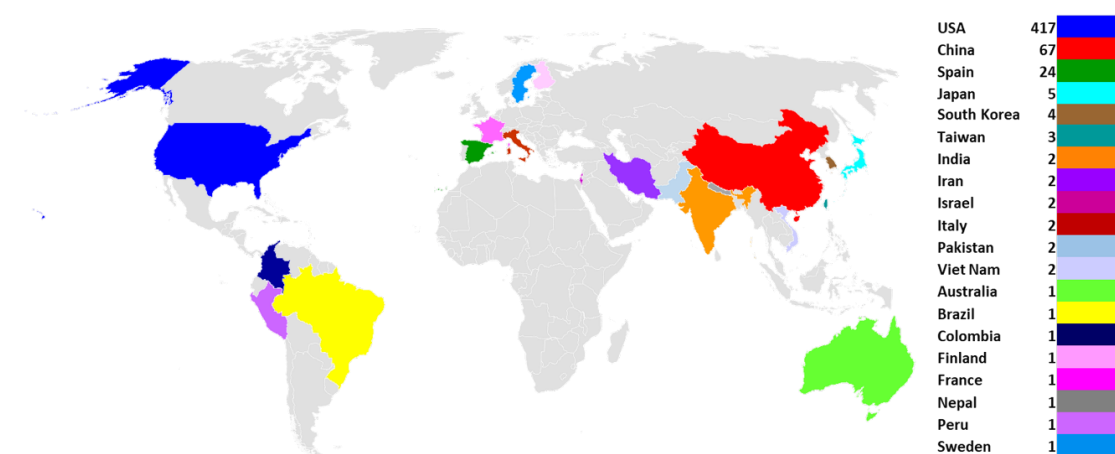
coronavirus was identified in the 1960s after an outbreak caused a large number of respiratory infections [1,2]. The number of scientific contributions related to SARS-CoV-2 has increased exponentially after its emergence in southern China in December 2019. Epidemiological investigations and genome sequencing experiments revealed SARS-CoV-2 as the etiological agent of the COVID-19 pandemic [3]. Prior to the emergence of SARS-CoV-2, six other human coronaviruses have been reported:  $\alpha$ -coronaviruses 229E and NL63,  $\beta$ -coronaviruses OC43 and HKU1, Middle East respiratory syndrome coronavirus (MERS-CoV), and SARS-associated coronavirus (SARS-CoV). Among those six, SARS-CoV and MERS-CoV have been shown to be associated with severe respiratory disease, high rates of acute lung injury, and correspondingly high mortality rates [4–6]. SARS-CoV-2 is divergent from all known human coronaviruses, including the 2012 MERS-CoV, which is believed to have originated from Saudi Arabia [7] and is associated with a rapid rate of transmission from human to human. Though both fall under the Betacoronavirus lineage-c, analysis of genome sequence, size, and organization reveals that MERS-CoV significantly differs from SARS-CoV [8,9].

The first infection of SARS-CoV-2 was reported from the city of Wuhan in Hubei province of China [10,11]. This led to the COVID-19 outbreak, which began in late December 2019 and continues to this day [12]. As of August 2020, more than 19 million confirmed cases of SARS-CoV-2 and 730,000 deaths have been reported. The global emergence and exponential increase of SARS-CoV-2 cases demands rapid scientific contributions to control and mitigate its impact.

The full genomic sequence of SARS-CoV-2 and its structural organization were first reported by the Yongzhen Zhang team in China [13]. The genome is arranged as 5'-untranslated region (UTR)-replicase complex (orf1ab)-structural proteins (Spike(S)-Envelope(E)-Membrane(M)-Nucleocapsid(N)) - 3'-UTR (Figure 1). The sequence was deposited in the National Center for Biotechnology Information (NCBI) GenBank in January 2020 and the annotation was subsequently provided (NC\_045512) [14]. A total of 540 complete genome sequences from 20 different countries/territories with annotated information from NCBI were analyzed in the current study to provide plausible insights into the genomic divergence of SARS-CoV-2 (Figure 2). The pressing public health issues associated with this illness motivated us to conduct an evolutionary analysis of SARS-CoV-2 across different countries/territories, starting from the reference genome reported from China. The evolutionary analysis described herein helps to shed light on the genomic changes across both time and space (e.g., different countries/territories), as well as expanding our understanding of how genomic variations of SARS-CoV-2 may relate to the development/efficacy of vaccines and therapeutics.



**Figure 1.** Genomic structure of SARS-CoV-2. The entire genome is classified as nonstructural and structural regions that produce multiple open reading frames (ORFs).



**Figure 2.** Distribution of SARS-CoV-2 genomes. A total of 540 genomes from 20 different countries are represented with 20 different color patterns.

To date, there have been several studies that have conducted phylogenetic analyses of SARS-CoV-2 based on a varying number of genomic sequences [15–23]. Phylogenetic analysis of 160 SARS-CoV-2 genomes revealed three central variants distinguished by random amino acid variations [24]. A four-genome phylogeny from Chile revealed two different viral variants coming from Europe and Asia [25]. The characterization and phylogenetic analysis of the first three genomes from Italy revealed a single amino acid variation in the surface glycoprotein [26]. A report from Europe on the phylogenetic analysis of two SARS-CoV-2 genomes revealed the introduction of novel variants describing the initial stages of viral evolution [27]. The Nextstrain database is continuously updating the information on the phylogenetic analysis and genomic divergence of SARS-CoV-2 with concomitant updates on the evolutionary changes [28]. Phylogenetic frameworks are essential tools to identify virus lineages that contribute to their active spread. Rambaut et al. proposed a rational and dynamic virus nomenclature for naming the expanding phylogenetic diversity of SARS-CoV-2 [29]. Their method was made tractable by constraining the number and depth of hierarchical lineage labels and focusing on active virus lineages that are spreading to wider locations. This nomenclature will assist in tracking and understanding the patterns and determinants of the global spread of SARS-CoV-2. In addition to the phylogenetics and divergence, in our current study we aimed to derive the impact of the observed evolutionary changes on the protein functionality and therapeutic interventions. The evolutionary analysis of 540 genomes from 20 different countries/territories described here allowed us to detect a greater number of variations, and correspondingly, the potential impact of such variations on the protein sequences and their implications on vaccine development.

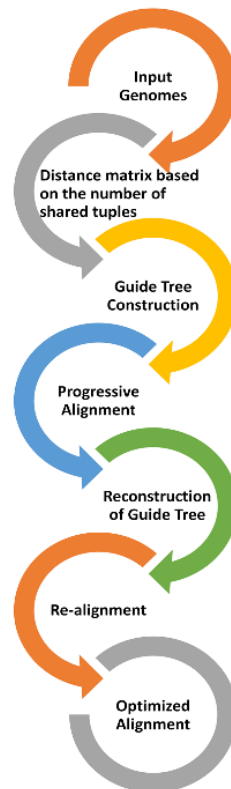
## 2. Materials and Methods

### 2.1. Genome Sequence Data

Aiming to derive the genomic divergence of SARS-CoV-2 across different countries/territories, 540 complete genome sequences with annotated information were retrieved from NCBI—Virus datahub (<https://www.ncbi.nlm.nih.gov/labs/virus/vssi/#>). The NCBI refseqs are supplied through three distinct pipelines (computed annotation, Entrez genomes, and LocusLink supported pipelines) curated on an ongoing basis [30]. The NCBI viral genome resource curates the submitted viral genome sequences with forced recalibration of the data to better provide extant sequence representations with enhanced reference information. This, in turn, increases the emphasis on leveraging the genome sequence data [31]. Such validated SARS-CoV-2 genome sequences were retrieved in FASTA format as available on 12 April 2020, and the sequence information is provided as supporting information (Files S1 & S2).

## 2.2. Phylogenetic Network Analysis

A full-genomic sequence alignment was performed for the 540 sequences using MAFFT v7.4.2. server [32]. The alignment was carried out by the FFT-NS-2 progressive method that performs rapid multiple sequence alignment based on a fast Fourier transform (FFT) [33]. This method uses a default scoring matrix derived from Kimura's two-parameter model [34] (Figure 3). The optimized multiple sequence alignment was further analyzed in MEGA-X V.10.1.7 [35].



**Figure 3.** Multiple sequence alignment process by MAFFT v7.4.2. Multiple sequence alignment was carried out in a multistage process to generate the final accurate alignment. Subsequent phylogenetic analysis was conducted using MEGA-X.

We used the annotated complete genome of SARS-CoV-2 isolate Wuhan-Hu-1 as the reference sequence (NC\_045512) [13] for the alignment of other complete genome sequences. The flanking regions of the alignment on both ends were truncated to approximately 100 base pairs (bp). Phylogenetic analysis was carried out in the MEGA-X environment using the maximum likelihood estimate of the phylogenetic reconstruction. The phylogeny was tested using a bootstrap-based approach with 100 replications. Evolutionary descendants were inferred using the Tamura–Nei model [36]. The initial tree was generated for the heuristic search by applying the Neighbor-Joining method to a matrix of pairwise distances estimated using the Maximum Composite Likelihood (MCL) approach [37]. The lengths of the branches were measured based on the number of substitutions per site. Codon positions included were 1st, 2nd, 3rd, and Noncoding with the involvement of all 540 genome sequences. This resulted in a total of 30,291 positions in the final dataset.

A nucleotide substitution matrix was generated to determine the probability of specific nucleotide substitutions. Substitution patterns and rates were estimated under the Tamura–Nei model [36]. Transition/transversion bias was estimated using the Maximum Likelihood method, where the substitution pattern and rates were estimated under the Kimura 2-parameter model [38]. Further, to clearly estimate the variation between each pair of genomes among the 540 considered in our analysis,

a pairwise distance matrix was generated using the MCL method [39]. The number of base substitutions per site between each sequence was derived as a pairwise distance matrix. Pairwise distances among 540 genomes were visualized by generating a heatmap using `gplot:heatmap2` function [40] in the R statistical programming language. The clustering dendrogram was constructed using hierarchical clustering (`stat:hclust`) [41].

### 2.3. Visualization of Genomic Variations

The complete variations among the 540 genomes were further visualized in Jalview workbench to identify genomic variations such as SNPs and indels [42]. The alignment was applied with a color index based on the percent identity so that identical regions were masked and the nucleotide variations were exposed out of the alignment. This also facilitated visualization of the overall alignment of a large number of genomes in a single window.

### 2.4. Reflection of Genomic Variations in Structural Proteins

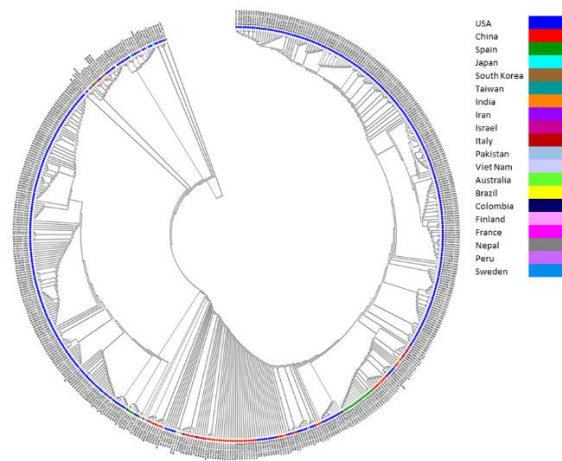
Variations arising in the genome subsequently cause changes in the amino acids in the resulting proteins. We focused on amino acid variations in the structural proteins, specifically, surface glycoprotein (S), envelope protein (E), membrane glycoprotein (M), and nucleocapsid (N) proteins, as these proteins are essential targets for developing universal and/or multi-epitope vaccine candidates, as well as potential drug targets [43–47]. The aligned genomic sequences were translated into protein sequences using MEGA-X software and the amino acid variations were identified.

## 3. Results

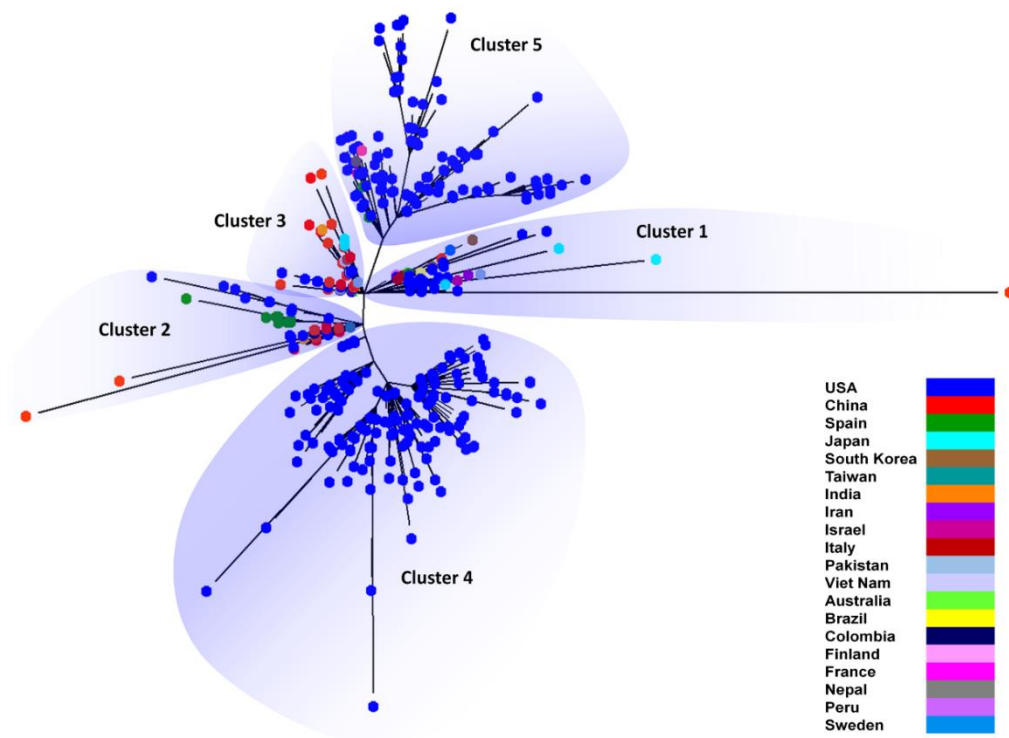
### 3.1. Phylogenetic Analysis of SARS-CoV-2 Genome Sequences

Initially, a rooted circular cladogram was constructed for 540 SARS-CoV-2 genomes to explain the roots and spread of COVID-19 across the 20 different countries/territories (Figure 4). The root node started with the isolate from China (LR757997) obtained from the Wuhan outbreak. The spread of the nodes started successively, with the isolates from other countries/territories in the following order: USA, Japan, Israel, Pakistan, Iran, Brazil, South Korea, Italy, Spain, Sweden, Australia, Finland, France, Peru, Taiwan, Vietnam, India, Nepal, and Colombia. The order of branching was defined based on the pairwise similarity across the 540 genomes. The nodes at the end of the circular cladogram represent the isolates with an increasing amount of variation in the genomic structure when compared to the reference genome.

In order to understand the ancestry and descendants among the 540 isolates, an unrooted phylogenetic tree was constructed (Figure 5). We observed five distinct clusters, among which clusters 4 and 5 were composed nearly entirely of USA isolates and were observed to exhibit variable branch lengths, implying a varying degree of genomic divergence among USA isolates. Interestingly, the most distant nodes were formed by China isolates creating the base rooting nodes, that is, cluster 1 of the tree, and to represent their unique genetic isolation from other nodes of the tree. Several other USA isolates were observed to form distant nodes in clusters 2 and 3, which cover all other countries/territories representing their genetic identity. These nodes would represent the early members or the ancestors in the unrooted tree. The two USA clusters, clusters 4 and 5, were observed to evolve through cluster 2 and cluster 3, respectively, indicating the spread of USA strains from multiple sources. The root bases that indicate the origin of infection were formed by China isolates in clusters 1 (LR757997) and 2 (MT226610, MT123292).



**Figure 4.** Circular cladogram of 540 genomes of SARS-CoV-2. The cladogram explains the base and spread of SARS-CoV-2 across 20 different countries/territories. The root node was found with an isolate from the Wuhan outbreak, China, and the spread was observed to other countries/territories in the order of USA, Japan, Israel, Pakistan, Iran, Brazil, South Korea, Italy, Spain, Sweden, Australia, Finland, France, Peru, Taiwan territory, Vietnam, India, Nepal, and Colombia.

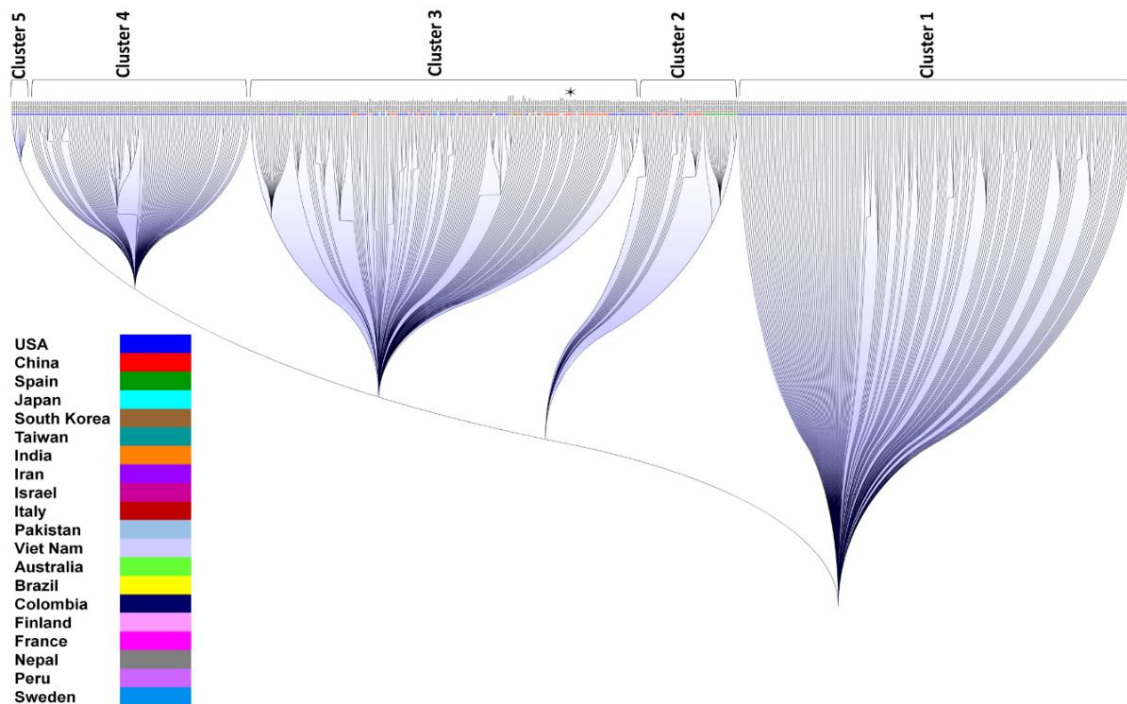


**Figure 5.** Unrooted phylogram showing the clusters of SARS-CoV-2 isolates across 20 different countries/territories. Cluster 1 represents the root cluster formed with the isolate from the Wuhan outbreak, China.

These phylograms were further subjected to bootstrapping in order to find out the most reliable phylogeny in terms of nucleotide variations. The group of observed genomic variations tended to form clusters. A rooted bootstrap phylogenetic tree was constructed to infer the phylogenetic divergence among the 540 isolates (Figure 6). The bootstrap phylogeny showed five distinct clusters with clear branching points. Clusters 1, 4, and 5 include only USA isolates which are distantly rooted together. This would imply more divergence from one USA cluster to another USA cluster. Cluster 2 contains



all Spain isolates and few isolates from China and the USA. This cluster also includes one isolate from India, Columbia, and Taiwan. Cluster 3 is predominantly composed of isolates from China and the USA. The isolates from the other 18 countries/territories were also included in this cluster, along with the reference genome from China. This represents the cluster with internal divergence among the isolates.

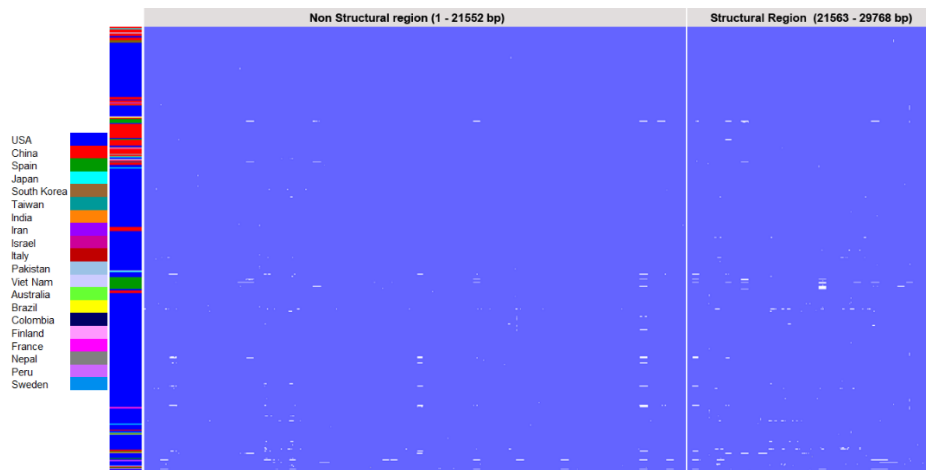


**Figure 6.** Bootstrap phylogram showing the clusters of SARS-CoV-2 isolates among 20 different countries/territories. A total of 540 genomes of SARS-CoV-2 formed as five distinct clusters. The reference genome from the China outbreak was observed to form in cluster 3 represented with a star. The internal branching depicts the genomic divergence within each cluster.

We observed many internal branches with variable branch lengths in the USA clusters of both the unrooted and rooted trees, indicating rapid genetic evolution in the SARS-CoV-2 genomes of USA strains.

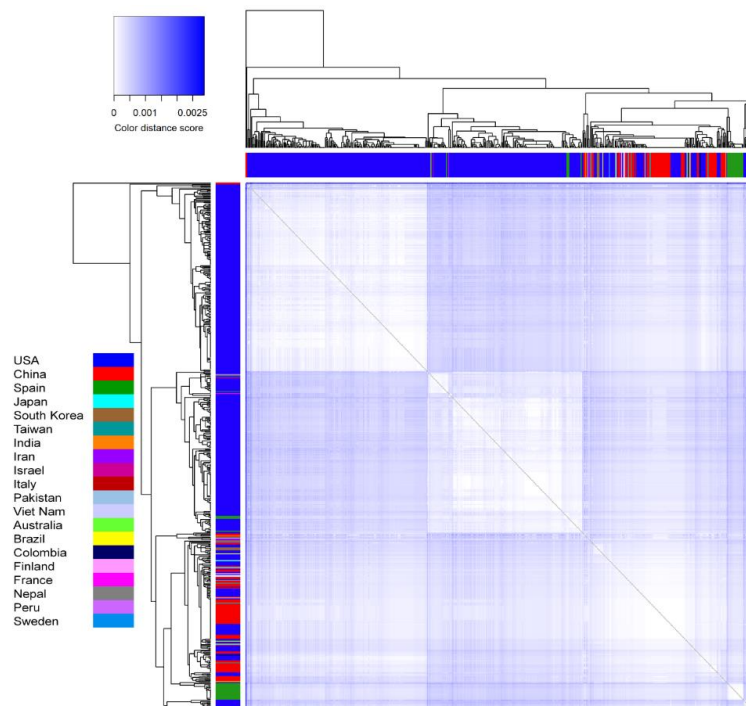
### 3.2. Multiple Sequence Alignment of SARS-CoV-2 Genome Sequences

To understand the genomic variations, we performed a multiple sequence alignment for the 540 complete genome sequences. We observed a large number of nucleotide variations throughout the genomic structure, including point mutations and deletions (Figure 7 & File S3). In order to analyze these variations, a nucleotide substitution matrix was generated using the Tamura–Nei method [39]. We found that transitions were far more likely than transversions in these data (approximately 67.58% of substitutions vs. 32.42%, respectively). Nevertheless, given that about a third of the substitutions are transversions, this represents a high tendency towards changes that are more likely to affect the resulting protein [48]. However, this makes the overall transition-to-transversion ratio approximately 2.08. Although the timeframe for accumulating these substitutions is limited, it is interesting that this ratio appears higher than in the SARS coronavirus outbreak in 2003, which had a ratio of approximately 1.1 [49]. However, it is lower than some influenza A viruses [50]. Furthermore, we note that C > T and G > A transitions are more likely than T > C or A > G, which is concordant with the overall nucleotide frequency in these sequences (A: 29.89%, T: 32.12%, C: 18.36%, G: 19.62%). The substitution and Transition/transversion matrices are provided in the supporting information (Files S5 & S6).



**Figure 7.** Multiple sequence alignment of 540 SARS-CoV-2 complete genomes. The graphical description of alignment of 540 genomes in a single window from Jalview workbench. The identical regions are masked with blue color and the variations highlighted in white color. The small white bars indicate the deletions in the genomes whereas dots indicate the single nucleotide variations. The 540 genomes are represented countrywise as color bars on the left side of the alignment.

Variations between each pair of genomes were calculated by generating a pairwise distance matrix (File S4). The number of base substitutions per site was calculated for each pair of the 540 genome sequences. All ambiguous positions were removed from each sequence pair (pairwise deletion option). A total of 30,291 positions were identified and the values are represented as a heat map (Figure 8) and the corresponding R-code is provided in the supporting information as File S7. The observations from the distance matrix were consistent with the phylogeny where a large number of genomic variations were observed in the USA strains, representing genomic divergence.



**Figure 8.** Heat map representing the pairwise distances among 540 genomes. The distance values among each pair of genomes are represented as color gradient. The sequences are represented as color side bars based on the country.



### 3.3. Genomic Variations as Amino Acid Variations in the Structural Proteins

To understand the impact of genomic variations observed in the 540 genomes, we translated surface glycoprotein (S), envelope protein (E), membrane glycoprotein (M), and nucleocapsid (N) structural genes into protein sequences and assessed amino acid variations. Among the 540 genomes, 246 showed amino acid variations in at least one of these four structural genes. The S protein, a major research focus for antigenic determinants, was observed to show a high rate of amino acid variations. Among 540 genomes, 202 showed 34 types of amino acid changes in the S proteins such as L5F, A27V, Y28N, T29I, H49Y, S50L, L54F, N74K, E96D, D111N, F157L, G181V, S221W, T240I, S248R, A344S, A348T, R408I, G476S, V483A, H519Q, A520S, A570V, D614G, H655Y, Q675H, F797C, A930V, D936Y, S940F, A1078V, D1168H, N1178D, and D1259H. The D614G variation was observed in 160 genomes, indicating its conserved nature. The next highest number of amino acid variations was observed in the N-structural protein, where 65 genomes showed 25 types of variations such as D3Y, N4D, P6T, P13L, P14L, S23T, A35T, P46S, D128Y, R185C, S194L, S197L, S202N, R203K, G204R, T205I, A207G, G215S, S232T, G238C, T271I, Q289H, S327L, D343V, and P344S. The maximum frequency of occurrence was found with the R203K variation, observed in 21 genomes, G204R in 19 genomes, and S197L in 15 genomes, indicating the probable conserved nature of these variations in the N-gene. Only three genomes showed two types of amino acid variations such as L37R and P71L in the E protein. Six genomes showed A2S, V70I, and T175M amino acid variations in the M protein (Supplementary File S8).

### 3.4. Mapping of Mutations to the Reported Vaccine Candidates

In order to find out the impact of identified mutations and their implications on vaccine development, we have investigated the reported vaccine candidates from previous studies [51–56]. The identified mutations in the structural proteins were mapped with these vaccine candidates and it was found that 23 mutations in S protein, 1 in E, 2 from M, and 7 from N protein were observed to map with the reported B-cell and T-cell epitopes (Supplementary File S9).

## 4. Discussion

In the present study, we observed genomic divergence in SARS-CoV-2 across a relatively short timeframe, based on 540 publicly available and validated genomes collected across 20 different countries/territories. The genomic variations observed in structural genes could serve as useful information for the vaccine development community. The results of this study may also help to address the consequences of genomic diversity in SARS-CoV-2 and its effects on immunogenic response in different patients. The approach we followed to derive the genomic divergence among SARS-CoV-2 genomes establishes several lines of inquiry for the variable immunogenic responses that would be an obstacle for developing a successful vaccine.

The rooted circular cladogram explained the root and spread of SARS-CoV-2 starting from China to 20 different countries/territories. This forms the basis to understand the genesis and dispersion of SARS-CoV-2 infection among 20 different countries/territories, which was defined based on the genomic similarities among 540 genomes. The evolution of ancestors and descendants was further evaluated by the unrooted tree that showed five distinct clusters. The USA clusters were observed to be derived from the clusters of all other countries/territories providing a possible evidence of origin from multiple sources. The root bases formed by China isolates are further demonstrating the root of infection, which was already observed in the circular cladogram. The bootstrapping tree showed similar clustering to that of the unrooted phylogram, indicating the reliability of the clustering process. The varying degrees of branch lengths within the clusters of the bootstrap phylogram indicate high rates of genomic divergence.

Genomic variants, such as indels and substitutions, observed in the multiple sequence alignment of 540 genomes indicate ongoing genetic evolution when compared to the reference genome. The genomic

divergence is still continuing as evident from Nextstrain (<https://nextstrain.org/sars-cov-2/>). Such a high frequency of variations in the genome arising in a short period of time could impact the efficacy of therapeutics and vaccines against COVID-19. There exists a greater need to understand how genomic changes brought about by indels in the genome could impact the antigenicity through protein sequence and structural changes. Hence, these variations may be considered during development of drugs and vaccines against SARS-CoV-2. Identification of critical nucleotide changes reflecting as amino acid changes in the proteins would alter the conformation of the proteins that leads to changes in pathogenicity and antigenicity. Such genetic variations would contribute to diverse antigenic properties resulting in variable immunogenic responses in the patients [57,58]. This would cause an involuntary impact on the rational design of a successful vaccine.

In order to ascertain these variations, we have further evaluated the structure of the SARS-CoV-2 genome and provide interpretations based on our analysis and observations. SARS-CoV-2 contains positive single-stranded RNA genomes with at least six open reading frames. The entire genome is grouped into nonstructural and structural regions (Figure 1). The nonstructural region represents a long ORF1ab that encodes replicase polyproteins required for the replication and transcription of the viral genome. It contains the nonstructural genes nsp1 to nsp16 that encode proteins such as papain-like proteinase (PL); 3-chymotrypsin-like proteinase (3CLPro); RNA-dependent RNA polymerase (RdRp); helicase, 3'-to-5' exonuclease; endoRNase; and 2'-O-ribose methyltransferase, which are required for the biochemical and molecular functions of the virus within the host. Hence, the nonstructural genes serve as attractive targets for antiviral drugs [59–61]. Further, nsp7 and nsp13 are associated with T-cell immune response [62]. Most of the T-cell epitopes are encoded by structural genes [63]. The structural region encodes four major structural proteins such as S, E, M, and N proteins. These four structural proteins represent ideal targets for the development of universal vaccines as they represent the majority of potential B- and T-cell epitopes [43–47]. These proteins form the protective layer of the virus and are exposed to the host environment at its primary stages of attack. Hence, the induction of immunogenicity in the host primarily depends on these proteins, making them the ideal vaccine candidates. Naturally, variations in the genome can cause changes in amino acids. We have investigated these amino acid changes in the S, E, M, and N proteins (Table S1). The identified mutations were further mapped with the previous vaccine reports where potential vaccine candidates were reported in the structural proteins of SARS-CoV-2 [51–56]. A total of 23 mutations in S protein, 1 in E, 2 from M, and 7 from N protein were mapped with the reported vaccine candidates (Table S2). The variations, observed in the proteins, disturb the antigenic determinants and could be responsible for the wide variety of immunogenic responses in each patient. Mutation rate drives viral evolution and genome variability, thereby enabling viruses to escape host immunity and to develop drug resistance. Maria et al. analyzed 220 genomic sequences and stated that the virus is evolving [64]. European, North American, and Asian strains might coexist, each of them characterized by a different mutation pattern. In addition to the mutations in structural proteins, mutations in the RdRp are significant as the virus mutagenic capability depends on the fidelity of RdRp [64]. Over 100 mutations of S protein were studied for their impact on the infectivity and antigenicity by Qianqian et al., and they found that D614G mutation is more infectious [65]. This mutation is also associated with higher viral loads. D614G was also reported to be consistently increasing at regional levels, indicating its fitness advantage [66]. Surprisingly, the same mutation was observed in 160 genomes among 540 in our current study, and also mapped with the vaccine candidates. Another mutation, V483A from S protein, reported to be markedly resistant to monoclonal antibodies (mABs), was observed among six genomes and mapped with the vaccine candidates. These large numbers of amino acid variations in the structural genes suggest that the genomic variations could present challenges in terms of vaccine and treatment development. Such an increasing genomic divergence poses numerous challenges to the research community to fight against COVID-19. Most of the current research on COVID-19 vaccination is based on the identification and characterization of the virulent proteins such as structural proteins of SARS-CoV-2 that interfere with innate and adaptive immune response and are also involved in the interactions with macrophages,

T-lymphocytes, and dendritic and epithelial cells. Such immunogenic interactions modulate the host response against viruses to combat pathogenesis [43,67]. The increasing number of mutations causing increase in the genomic divergence would continue to be a challenge in the treatment and vaccine design strategies [58]. The conserved regions of the proteins with no frequency of mutations contribute to stable immunogenicity. Zhang et al. reported that immunodominant (ID) sites of S protein were found to be evolutionarily highly conserved, contributing to potential immunogenicity [68]. Nevertheless, the amino acid variations in SARS-CoV-2 might change the immunogenicity of ID sites, suggesting the careful consideration of epitopes for vaccine design [69].

The findings from the current study highlight the potential impact of genomic variations on protein changes that may stymie the vaccine development process. Information about the possible sites of nucleotide changes and conserved regions of the structural proteins may help other researchers consider specific regions in the proteins that would be avoided as targets for a universal vaccine against SARS-CoV-2 [58,70]. This study also illuminates the important changes for the mechanistic understanding of pathogenicity of SARS-CoV-2 and supports continuing surveillance of mutations to aid with development of a universal vaccine and immunological interventions.

## 5. Conclusions

Currently, most of the world is in the grip of the COVID-19 pandemic, and a vaccine or targeted treatment is urgently needed. Our current study represents an analysis of 540 SARS-CoV-2 complete genomes, collected across 20 different countries/territories. Ongoing genomic divergence was observed among the genomes. In addition, large numbers of nucleotide variations were observed throughout the genome. We analyzed the impact of genomic variations on the structural region of the genome, which is the main target for the development of vaccine candidates. This study suggests these variations be considered in the development of a universal vaccine for COVID-19. We conclude that the continued genomic divergence across successive generations arising due to larger number of nucleotide variations could hinder the development of a universal vaccine. The vaccine research communities could adopt this information to avoid the regions with variations to achieve a successful vaccine against SARS-CoV-2.

**Supplementary Materials:** The following are available online at <http://www.mdpi.com/2076-393X/8/4/591/s1>. File S1. The list of 540 complete genome sequences used in the current study. File S2. The 540 full genome sequences in FASTA format. File S3. The multiple sequence alignment of 540 genomes. File S4. The distance matrix showing the pairwise distances of 540 genomes. File S5. Substitution matrix showing the substitutional probabilities of 540 genomes. File S6. Transition–Transversion matrix showing the transition/transversion bias of 540 genomes. File S7. The R-code generated for the construction of heat map to represent the pairwise distance among 540 genomes. File S8. Genomic variations of SARS-CoV-2 reflected as amino acid changes in the structural proteins. File S9. Mutations mapped onto the reported vaccine candidates.

**Author Contributions:** N.K.Y. designed the project, provided expertise to the bioinformatics concepts, and wrote the paper. S.P. worked on the protein translation of structural genes of 540 genomes and identified the amino acid variations. B.Z. performed the hierarchical clustering analysis and generated the heatmap to represent the distance matrix of 540 genomes. R.M., L.N., D.P., Q.X., D.R., R.C.Z., E.N., S.B.-G., W.S., J.H., P.C., L.C.-H. assisted with the interpretation of the study findings and edited the manuscript. D.C.K. supervised the analysis, participated in the interpretation of the study findings, and edited the manuscript. J.A.T. supervised the analysis, participated in the interpretation of the study findings, and edited the manuscript. All authors have read and agreed to the published version of the manuscript.

**Funding:** Research reported in this publication was supported by the Kansas IDeA Network of Biomedical Research Excellence Bioinformatics Core, supported in part by the National Institute of General Medical Science (url: <https://www.nigms.nih.gov/>) award P20GM103428 (supports DCK, RM, and DP), and the Kansas Institute for Precision Medicine Quantitative Omics Core, supported by National Cancer Institute (url: <https://www.cancer.gov/>) grant P20GM130423 (supports DCK, PC, DP, NKY). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

**Conflicts of Interest:** Authors declare no conflicts of interests.

## References

1. Fehr, A.R.; Perlman, S. Coronaviruses: An Overview of Their Replication and Pathogenesis. *Methods Mol. Biol.* **2015**, *1282*, 1–23. [[CrossRef](#)] [[PubMed](#)]
2. Kahn, J.S.; McIntosh, K. History and Recent Advances in Coronavirus Discovery. *Pediatr. Infect. Dis. J.* **2005**, *24*, S223–S227. [[CrossRef](#)] [[PubMed](#)]
3. Lu, R.; Zhao, X.; Li, J.; Niu, P.; Yang, B.; Wu, H.; Wang, W.; Song, H.; Huang, B.; Zhu, N.; et al. Genomic characterisation and epidemiology of 2019 novel coronavirus: Implications for virus origins and receptor binding. *Lancet* **2020**, *395*, 565–574. [[CrossRef](#)]
4. Chan, J.F.-W.; Lau, S.K.P.; To, K.K.-W.; Cheng, V.C.C.; Woo, P.C.Y.; Yuen, K.-Y. Middle East Respiratory Syndrome Coronavirus: Another Zoonotic Betacoronavirus Causing SARS-Like Disease. *Clin. Microbiol. Rev.* **2015**, *28*, 465–522. [[CrossRef](#)]
5. Cheng, V.C.C.; Lau, S.K.P.; Woo, P.C.Y.; Yuen, K.-Y. Severe Acute Respiratory Syndrome Coronavirus as an Agent of Emerging and Reemerging Infection. *Clin. Microbiol. Rev.* **2007**, *20*, 660–694. [[CrossRef](#)] [[PubMed](#)]
6. Su, S.; Wong, G.; Shi, W.; Liu, J.; Lai, A.C.; Zhou, J.; Liu, W.; Bi, Y.; Gao, G.F. Epidemiology, Genetic Recombination, and Pathogenesis of Coronaviruses. *Trends Microbiol.* **2016**, *24*, 490–502. [[CrossRef](#)]
7. Al-Hajjar, S.; Memish, Z.A.; McIntosh, K. Middle East Respiratory Syndrome Coronavirus (MERS-CoV): A Perpetual Challenge. *Ann. Saudi Med.* **2013**, *33*, 427–436. [[CrossRef](#)]
8. Penttinen, P.M.; Kaasik-Aaslav, K.; Friaux, A.; Donachie, A.; Sudre, B.; Amato-Gauci, A.J.; Memish, Z.A.; Coulombier, D. Taking stock of the first 133 MERS coronavirus cases globally—Is the epidemic changing? *Eurosurveillance* **2013**, *18*, 20596. [[CrossRef](#)]
9. Zaki, A.M.; Van Boheemen, S.; Bestebroer, T.; Osterhaus, A.; Fouchier, R.A.M. Isolation of a Novel Coronavirus from a Man with Pneumonia in Saudi Arabia. *New Engl. J. Med.* **2012**, *367*, 1814–1820. [[CrossRef](#)]
10. WHO. *Situation Report -5 25 January 2020*; World Health Organization: Geneva, Switzerland, 2019; p. 251.
11. Hui, D.S.; Azhar, E.E.; Madani, T.A.; Ntoumi, F.; Kock, R.; Dar, O.; Ippolito, G.; McHugh, T.D.; Memish, Z.A.; Drosten, C.; et al. The continuing 2019-nCoV epidemic threat of novel coronaviruses to global health—The latest 2019 novel coronavirus outbreak in Wuhan, China. *Int. J. Infect. Dis.* **2020**, *91*, 264–266. [[CrossRef](#)]
12. Wang, C.; Horby, P.W.; Hayden, F.G.; Gao, G.F. A novel coronavirus outbreak of global health concern. *Lancet* **2020**, *395*, 470–473. [[CrossRef](#)]
13. 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.; et al. A new coronavirus associated with human respiratory disease in China. *Nat. Cell Biol.* **2020**, *579*, 265–269. [[CrossRef](#)] [[PubMed](#)]
14. Benson, D.A.; Boguski, M.; Lipman, D.J.; Ostell, J. GenBank. *Nucleic Acids Res.* **1994**, *22*, 3441–3444. [[CrossRef](#)] [[PubMed](#)]
15. Wang, J.-T.; Lin, Y.-Y.; Chang, S.-Y.; Yeh, S.-H.; Hu, B.-H.; Chen, P.-J.; Chang, S.-C. The role of phylogenetic analysis in clarifying the infection source of a COVID-19 patient. *J. Infect.* **2020**, *81*, 147–178. [[CrossRef](#)] [[PubMed](#)]
16. Bartolini, B.; Rueca, M.; Gruber, C.E.M.; Messina, F.; Carletti, F.; Giombini, E.; Lalle, E.; Bordi, L.; Matusali, G.; Colavita, F.; et al. SARS-CoV-2 Phylogenetic Analysis, Lazio Region, Italy, February–March 2020. *Emerg. Infect. Dis.* **2020**, *26*, 26. [[CrossRef](#)]
17. Lopes, L.R.; de Mattos Cardillo, G.; Paiva, P.B. Molecular evolution and phylogenetic analysis of SARS-CoV-2 and hosts ACE2 protein suggest Malayan pangolin as intermediary host. *Braz. J. Microbiol.* **2020**, *1*, 1–7. [[CrossRef](#)]
18. Nemudryi, A.; Nemudraia, A.; Surya, K.; Wiegand, T.; Buyukyoruk, M.; Wilkinson, R.; Wiedenheft, B. Temporal detection and phylogenetic assessment of SARS-CoV-2 in municipal wastewater. *Cell Rep. Med.* **2020**, 100098. [[CrossRef](#)]
19. Hassan, S.S.; Choudhury, P.P.; Roy, B. Molecular phylogeny and missense mutations at envelope proteins across coronaviruses. *Genome* **2020**, *112*, 4993–5004. [[CrossRef](#)]
20. Pillay, S.; Giandhari, J.; Tegally, H.; Wilkinson, E.; Chimukangara, B.; Lessells, R.J.; Moosa, M.-Y.; Mattison, S.; Gazy, I.; Fish, M.; et al. Whole Genome Sequencing of SARS-CoV-2: Adapting Illumina Protocols for Quick and Accurate Outbreak Investigation during a Pandemic. *Genes* **2020**, *11*, 949. [[CrossRef](#)]



21. Guruprasad, L. Evolutionary relationships and sequence-structure determinants in human SARS coronavirus-2 spike proteins for host receptor recognition. *Proteins Struct. Funct. Bioinform.* **2020**. [[CrossRef](#)]
22. Sheikh, J.A.; Singh, J.; Singh, H.; Jamal, S.; Khubaib, M.; Kohli, S.; Dobrindt, U.; Rahman, S.A.; Ehtesham, N.Z.; Hasnain, S.E. Emerging genetic diversity among clinical isolates of SARS-CoV-2: Lessons for today. *Infect. Genet. Evol.* **2020**, *84*, 104330. [[CrossRef](#)] [[PubMed](#)]
23. Wen, F.; Yu, H.; Guo, J.; Li, Y.; Luo, K.; Huang, S. Identification of the hyper-variable genomic hotspot for the novel coronavirus SARS-CoV-2. *J. Infect.* **2020**, *80*, 671–693. [[CrossRef](#)] [[PubMed](#)]
24. Forster, P.; Forster, L.; Renfrew, C.; Forster, M. Phylogenetic network analysis of SARS-CoV-2 genomes. *Proc. Natl. Acad. Sci. USA* **2020**, *117*, 9241–9243. [[CrossRef](#)] [[PubMed](#)]
25. Castillo, A.E.; Parra, B.; Tapia, P.; Acevedo, A.; Lagos, J.; Andrade, W.; Arata, L.; Leal, G.; Barra, G.; Tambley, C.; et al. Phylogenetic analysis of the first four SARS-CoV-2 cases in Chile. *J. Med. Virol.* **2020**. [[CrossRef](#)]
26. Zehender, G.; Lai, A.; Bergna, A.; Meroni, L.; Riva, A.; Balotta, C.; Tarkowski, M.; Gabrieli, A.; Bernacchia, D.; Rusconi, S.; et al. Genomic characterization and phylogenetic analysis of SARS-COV-2 in Italy. *J. Med. Virol.* **2020**. [[CrossRef](#)]
27. Stefanelli, P.; Faggioni, G.; Presti, A.L.; Fiore, S.; Marchi, A.; Benedetti, E.; Fabiani, C.; Anselmo, A.; Ciammaruconi, A.; Fortunato, A.; et al. Whole genome and phylogenetic analysis of two SARS-CoV-2 strains isolated in Italy in January and February 2020: Additional clues on multiple introductions and further circulation in Europe. *Eurosurveillance* **2020**, *25*, 2000305. [[CrossRef](#)]
28. Hadfield, J.; Megill, C.; Bell, S.M.; Huddleston, J.; Potter, B.; Callender, C.; Sagulenko, P.; Bedford, T.; Neher, R.A. Nextstrain: Real-time tracking of pathogen evolution. *Bioinformatics* **2018**, *34*, 4121–4123. [[CrossRef](#)]
29. Rambaut, A.; Holmes, E.C.; O’Toole, Á.; Hill, V.; McCrone, J.T.; Ruis, C.; Du Plessis, L.; Pybus, O.G. A dynamic nomenclature proposal for SARS-CoV-2 lineages to assist genomic epidemiology. *Nat. Microbiol.* **2020**, 1–5. [[CrossRef](#)]
30. Pruitt, K.D.; Tatusova, T.; Brown, G.R.; Maglott, D.R. NCBI Reference Sequences (RefSeq): Current status, new features and genome annotation policy. *Nucleic Acids Res.* **2011**, *40*, D130–D135. [[CrossRef](#)]
31. Brister, J.R.; Ako-Adjei, D.; Bao, Y.; Blinkova, O. NCBI viral genomes resource. *Nucleic Acids Res.* **2014**, *43*, D571–D577. [[CrossRef](#)]
32. Katoh, K.; Standley, D.M. MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. *Mol. Biol. Evol.* **2013**, *30*, 772–780. [[CrossRef](#)]
33. Katoh, K. MAFFT: A novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res.* **2002**, *30*, 3059–3066. [[CrossRef](#)]
34. Kimura, M. *Kimura’s Two-Parameter Model of Models of DNA Evolution. Inferring Phylogenies*; Sinauer Associates, Inc.: Sunderland, MA, USA, 1980.
35. Kumar, S.; Stecher, G.; Li, M.; Knyaz, C.; Tamura, K. MEGA X: Molecular Evolutionary Genetics Analysis across Computing Platforms. *Mol. Biol. Evol.* **2018**, *35*, 1547–1549. [[CrossRef](#)] [[PubMed](#)]
36. Tamura, K.; Nei, M. Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. *Mol. Biol. Evol.* **1993**, *10*, 512–526. [[CrossRef](#)]
37. Lele, S.; Taper, M.L. A composite likelihood approach to (co)variance components estimation. *J. Stat. Plan. Inference* **2002**, *103*, 117–135. [[CrossRef](#)]
38. Kimura, M. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *J. Mol. Evol.* **1980**, *16*, 111–120. [[CrossRef](#)] [[PubMed](#)]
39. Tamura, K.; Nei, M.; Kumar, S. Prospects for inferring very large phylogenies by using the neighbor-joining method. *Proc. Natl. Acad. Sci. USA* **2004**, *101*, 11030–11035. [[CrossRef](#)]
40. Warnes, G.R.; Bolker, B.; Bonebakker, L.; Gentleman, R.; Liaw, W.H.A.; Lumley, T.; Maechler, M.; Magnusson, A.; Moeller, S.; Schwartz, M. *gplots: Various R programming tools for plotting data*. 2015. Available online: <https://rdrr.io/cran/gplots/> (accessed on 4 October 2020).
41. Murtagh, F.; Legendre, P. Ward’s Hierarchical Agglomerative Clustering Method: Which Algorithms Implement Ward’s Criterion? *J. Classif.* **2014**, *31*, 274–295. [[CrossRef](#)]
42. Waterhouse, A.M.; Procter, J.B.; Martin, D.M.A.; Clamp, M.; Barton, G.J. Jalview Version 2—a multiple sequence alignment editor and analysis workbench. *Bioinformatics* **2009**, *25*, 1189–1191. [[CrossRef](#)]



43. Ahmed, S.F.; Quadeer, A.A.; McKay, M.R. Preliminary Identification of Potential Vaccine Targets for the COVID-19 Coronavirus (SARS-CoV-2) Based on SARS-CoV Immunological Studies. *Viruses* **2020**, *12*, 254. [[CrossRef](#)] [[PubMed](#)]
44. Prajapat, M.; Sarma, P.; Shekhar, N.; Avti, P.; Sinha, S.; Kaur, H.; Kumar, S.; Bhattacharyya, A.; Kumar, H.; Bansal, S.; et al. Drug targets for corona virus: A systematic review. *Indian J. Pharmacol.* **2020**, *52*, 56–65. [[CrossRef](#)] [[PubMed](#)]
45. Chen, W.-H.; Strych, U.; Hotez, P.J.; Bottazzi, M.E. The SARS-CoV-2 Vaccine Pipeline: An Overview. *Curr. Trop. Med. Rep.* **2020**, *7*, 61–64. [[CrossRef](#)]
46. Lee, C.H.-J.; Koohy, H. In silico identification of vaccine targets for 2019-nCoV. *F1000Research* **2020**, *9*, 145. [[CrossRef](#)]
47. Ou, X.; Liu, Y.; Lei, X.; Li, P.; Mi, D.; Ren, L.; Guo, L.; Guo, R.; Chen, T.; Hu, J.; 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. [[CrossRef](#)] [[PubMed](#)]
48. Guo, C.; McDowell, I.C.; Nodzenski, M.; Scholtens, D.; Allen, A.S.; Lowe, W.; Reddy, T.E. Transversions have larger regulatory effects than transitions. *BMC Genom.* **2017**, *18*, 394. [[CrossRef](#)]
49. Zhao, Z.; Li, H.; Wu, X.; Zhong, Y.; Zhang, K.Q.; Zhang, Y.; Boerwinkle, E.; Fu, Y. Moderate mutation rate in the SARS coronavirus genome and its implications. *BMC Evol. Biol.* **2004**, *4*, 21. [[CrossRef](#)]
50. Pauly, M.D.; Procario, M.C.; Lauring, A.S. A novel twelve class fluctuation test reveals higher than expected mutation rates for influenza A viruses. *eLife* **2017**, *6*, 26437. [[CrossRef](#)]
51. Bhattacharya, M.; Sharma, A.R.; Patra, P.; Ghosh, P.; Sharma, G.; Patra, B.C.; Lee, S.-S.; Chakraborty, C. Development of epitope-based peptide vaccine against novel coronavirus 2019 (SARS-COV-2): Immunoinformatics approach. *J. Med. Virol.* **2020**, *92*, 618–631. [[CrossRef](#)]
52. Enayatkhani, M.; Hasaniazad, M.; Faezi, S.; Gouklani, H.; Davoodian, P.; Ahmadi, N.; Einakian, M.A.; Karmostaji, A.; Ahmadi, K. Reverse vaccinology approach to design a novel multi-epitope vaccine candidate against COVID-19: An in silico study. *J. Biomol. Struct. Dyn.* **2020**, 1–16. [[CrossRef](#)]
53. Grifoni, A.; Sidney, J.; Zhang, Y.; Scheuermann, R.H.; Peters, B.; Sette, A. A Sequence Homology and Bioinformatic Approach Can Predict Candidate Targets for Immune Responses to SARS-CoV-2. *Cell Host Microbe* **2020**, *27*, 671–680.e2. [[CrossRef](#)] [[PubMed](#)]
54. Kalita, P.; Padhi, A.K.; Zhang, K.Y.; Tripathi, T. Design of a peptide-based subunit vaccine against novel coronavirus SARS-CoV-2. *Microb. Pathog.* **2020**, *145*, 104236. [[CrossRef](#)] [[PubMed](#)]
55. Poran, A.; Harjanto, D.; Malloy, M.; Arieta, C.M.; Rothenberg, D.A.; Lenkala, D.; Van Buuren, M.M.; Addona, T.A.; Rooney, M.S.; Srinivasan, L.; et al. Sequence-based prediction of SARS-CoV-2 vaccine targets using a mass spectrometry-based bioinformatics predictor identifies immunogenic T cell epitopes. *Genome Med.* **2020**, *12*, 1–15. [[CrossRef](#)] [[PubMed](#)]
56. Rahman, M.S.; Hoque, M.N.; Islam, M.R.; Akter, S.; Alam, A.R.U.; Siddique, M.A.; Saha, O.; Rahaman, M.; Sultana, M.; Hossain, M.A. Epitope-based chimeric peptide vaccine design against S, M and E proteins of SARS-CoV-2 etiologic agent of global pandemic COVID-19: An in silico approach. *PeerJ* **2020**, *8*, e9572. [[CrossRef](#)]
57. Yee, P.T.I.; Poh, C.L. Impact of genetic changes, pathogenicity and antigenicity on Enterovirus- A71 vaccine development. *Virology* **2017**, *506*, 121–129. [[CrossRef](#)]
58. Peeters, M.; Toure-Kane, C.; Nkengasong, J.N. Genetic diversity of HIV in Africa: Impact on diagnosis, treatment, vaccine development and trials. *AIDS* **2003**, *17*, 2547–2560. [[CrossRef](#)]
59. Yin, C. Genotyping coronavirus SARS-CoV-2: Methods and implications. *Genome* **2020**, *112*, 3588–3596. [[CrossRef](#)]
60. Kim, Y.; Lovell, S.; Tiew, K.-C.; Mandadapu, S.R.; Alliston, K.R.; Battaile, K.P.; Groutas, W.C.; Chang, K.-O. Broad-Spectrum Antivirals against 3C or 3C-Like Proteases of Picornaviruses, Noroviruses, and Coronaviruses. *J. Virol.* **2012**, *86*, 11754–11762. [[CrossRef](#)]
61. Gao, K.; Nguyen, D.D.; Wang, R.; Wei, G.-W. Machine intelligence design of 2019-nCoV drugs. *bioRxiv* **2020**. [[CrossRef](#)]
62. Le Bert, N.; Tan, A.T.; Kunasegaran, K.; Tham, C.Y.L.; Hafezi, M.; Chia, A.; Chng, M.H.Y.; Lin, M.; Tan, N.; Linster, M.; et al. SARS-CoV-2-specific T cell immunity in cases of COVID-19 and SARS, and uninfected controls. *Nat. Cell Biol.* **2020**, *584*, 1–10. [[CrossRef](#)]

63. Liu, W.J.; Zhao, M.; Liu, K.; Xu, K.; Wong, G.; Tan, W.; Gao, G.F. T-cell immunity of SARS-CoV: Implications for vaccine development against MERS-CoV. *Antivir. Res.* **2017**, *137*, 82–92. [[CrossRef](#)] [[PubMed](#)]
64. Pachetti, M.; Marini, B.; Benedetti, F.; Giudici, F.; Mauro, E.; Storici, P.; Masciovecchio, C.; Angeletti, S.; Ciccozzi, M.; Gallo, R.C.; et al. Emerging SARS-CoV-2 mutation hot spots include a novel RNA-dependent-RNA polymerase variant. *J. Transl. Med.* **2020**, *18*, 179. [[CrossRef](#)] [[PubMed](#)]
65. Li, Q.; Wu, J.; Nie, J.; Zhang, L.; Hao, H.; Liu, S.; Zhao, C.; Zhang, Q.; Liu, H.; Nie, L.; et al. The Impact of Mutations in SARS-CoV-2 Spike on Viral Infectivity and Antigenicity. *Cell* **2020**, *182*, 1284. [[CrossRef](#)] [[PubMed](#)]
66. Korber, B.; Fischer, W.; Gnanakaran, S.; Yoon, H.; Theiler, J.; Abfalterer, W.; Hengartner, N.; Giorgi, E.E.; Bhattacharya, T.; Foley, B.; et al. Tracking Changes in SARS-CoV-2 Spike: Evidence that D614G Increases Infectivity of the COVID-19 Virus. *Cell* **2020**, *182*, 812–827. [[CrossRef](#)] [[PubMed](#)]
67. Enjuanes Sánchez, L.; Zúñiga Lucas, S.; Castaño-Rodríguez, C.; Gutierrez-Alvarez, J.; Cantón, J.; Solá Gurpegui, I. Chapter eight-Molecular basis of Coronavirus virulence and vaccine development. *Sci. Direct* **2016**, *96*, 245–286.
68. Zhang, B.-Z.; Hu, Y.-F.; Chen, L.-L.; Yau, T.; Tong, Y.-G.; Hu, J.-C.; Cai, J.-P.; Chan, K.-H.; Dou, Y.; Deng, J.; et al. Mining of epitopes on spike protein of SARS-CoV-2 from COVID-19 patients. *Cell Res.* **2020**, *30*, 702–704. [[CrossRef](#)]
69. Yuan, M.; Wu, N.C.; Zhu, X.; Lee, C.-C.D.; So, R.T.Y.; Lv, H.; Mok, C.K.P.; Wilson, I.A. A highly conserved cryptic epitope in the receptor binding domains of SARS-CoV-2 and SARS-CoV. *Science* **2020**, *368*, 630–633. [[CrossRef](#)]
70. Goudsmit, J.; Back, N.K.T.; Nara, P.L. Genomic diversity and antigenic variation of HIV-1: Links between pathogenesis, epidemiology and vaccine development. *FASEB J.* **1991**, *5*, 2427–2436. [[CrossRef](#)]



© 2020 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0/>).