Phylogenetic analysis of 48 early SARS-CoV-2 genomes

Maggie Z.X. Xiao, BSc[1]; Dylan Whitney, BASc[1]

[1] Faculty of Medicine, University of Alberta, Edmonton, Alberta

Abstract

Background: First isolated in December 2019, SARS-CoV-2 is the agent responsible for the ongoing breakout of COVID-19.

Method: We curated an assembly of the first 48 full-length SARS-CoV-2 genomes isolated and sequenced across the world and performed a phylogenetic network analysis to monitor the emergence of genomic divergence in the global SARS-CoV-2 population.

Results: We identified regions of the genome that have accumulated mutations producing non-synonymous changes at the protein level, suggesting ongoing adaptation of SARS-CoV-2 to its novel human host. We identified a strong L84S mutational signal in ORF8 (present in 29.16% of genomes) together with 12 variable sites in the region encoding non-structural protein Nsp3 that represent the strongest putative regions under selection in our dataset. We did not detect mutations in the coronavirus spike protein, which is reassuring for the vaccines currently available or are ongoing large-scale clinical trials.

Conclusion: Our analysis provides a snapshot in time of a rapidly evolving pandemic based on available data. Our results are in line with previous findings that point to a common ancestor isolated in Wuhan that is likely to have circulated and spread worldwide.

Introduction

Severe Acute Respiratory Syndrome Coronavirus-2 (SARS-CoV-2) is a novel coronavirus confirmed as the causative agent of coronavirus disease 2019 (COVID-19) that began in Wuhan, China in late 2019 and spread worldwide.[1,2] As of 05 January 2021, there have been more than 86 million confirmed cases of SARS-CoV-2 infection with close to 2 million deaths attributed to the virus in 218 countries and territories.[3] SARS-CoV-2 is thought to have evolved from SARS-like bat coronaviruses,[2] before being introduced to human populations during a spillover event and transmitted from person-to-person.[4] SARS-CoV-2 is a single-stranded positive-sense RNA virus and belongs to the β-coronavirus family consisting of many other human-associated pathogens including SARS-CoV and the Middle East respiratory syndrome (MERS)-CoV that both underwent host jumps into humans from bat reservoirs.[5] The ~30-kb non-segmented genome of SARS-CoV-2 encodes 14 major open reading frames (ORFs), which are further processed into 4 structural proteins (glycoprotein spike [S], membrane [M], envelope [E], and nucleocapsid [N]), 16 non-structural proteins (nsp1–nsp16), and at least 8 accessory proteins (ORF3a, ORF6, ORF7a, ORF7b, ORF8, ORF9b, ORF9c, and ORF10).[6]

The arrival of the COVID-19 vaccine has brought with it new hope, but there have been discussions of whether mRNA and adenovirus-based vaccines are able to withstand the force of viral evolution and protect us from new variants of SARS-CoV-2.[7] RNA viruses have higher mutation rates compared to DNA viruses and antigens on the surface of SARS-CoV-2 are particularly sensitive to mutation and strain drift.[8] The three most advanced vaccines (Pfizer/BioNTech, Oxford/AstraZeneca, and Moderna) focus on the surface spike protein, which mediates SARS-CoV-2 entry into cells.[9] There is concern that if mutations occurring within the immunodominant epitopes take hold and the spike sequence alters excessively, then new vaccines will be required.[10]

In this work, we analyze the genomic diversity in the first 48 full-length SARS-CoV-2 specimens sequenced across the globe to capture early global demography and report on changes to the sequence of SARS-CoV-2 over the earlier course of this pandemic. The analysis of genetic sequence data from viruses is increasingly recognized as an important tool in infectious disease epidemiology and allows us to characterize regions of the genome with high rates of mutation.[11] Monitoring the build-up of mutations may indicate ongoing adaptations of the virus to its host and has the potential to inform on targets for drugs and vaccines.[12] We focus in particular on highly variable genomic spots and mutations of high prevalence, as they are likely candidates of regions under selection in our assemblies.

Results

The first whole-genome sequence (WH-01, MN908947) was deposited on GenBank in early December 2019. While there are now thousands of SARS-CoV-2 whole-genome sequences available at the time of writing (01 August 2020), only 90 SARS-CoV-2 whole and partial genomes from 14 countries had been sequenced and made publicly available on GenBank and the National Genomics Data Center (NGDC) at the time of analysis (19 February 2020) thanks to worldwide efforts of contributing laboratories. At this earlier point

Corresponding Author:
Maggie Z.X. Xiao
zixuan2@ualberta.ca

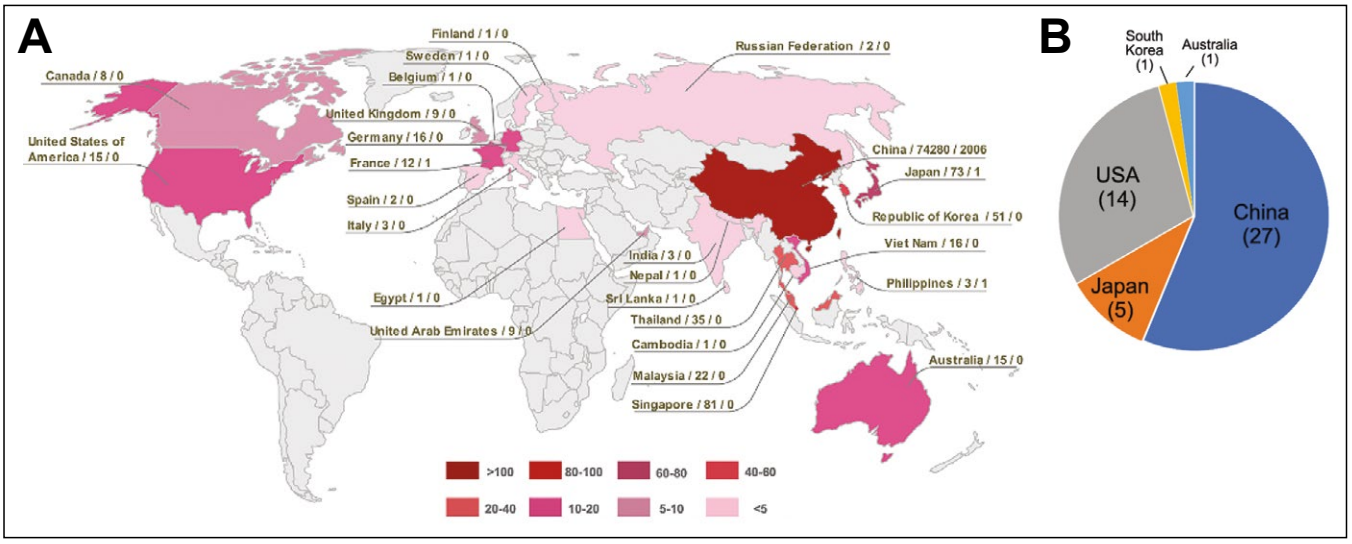

Figure 1. A. Geographical tracking and mapping of COVID-19 (country/number of confirmed cases/numbers of death cases). B. Pie-chart summarizing the 48 available SARS-CoV-2 complete genomes isolated in different geographical regions of the world from December 2019 to February 2020.

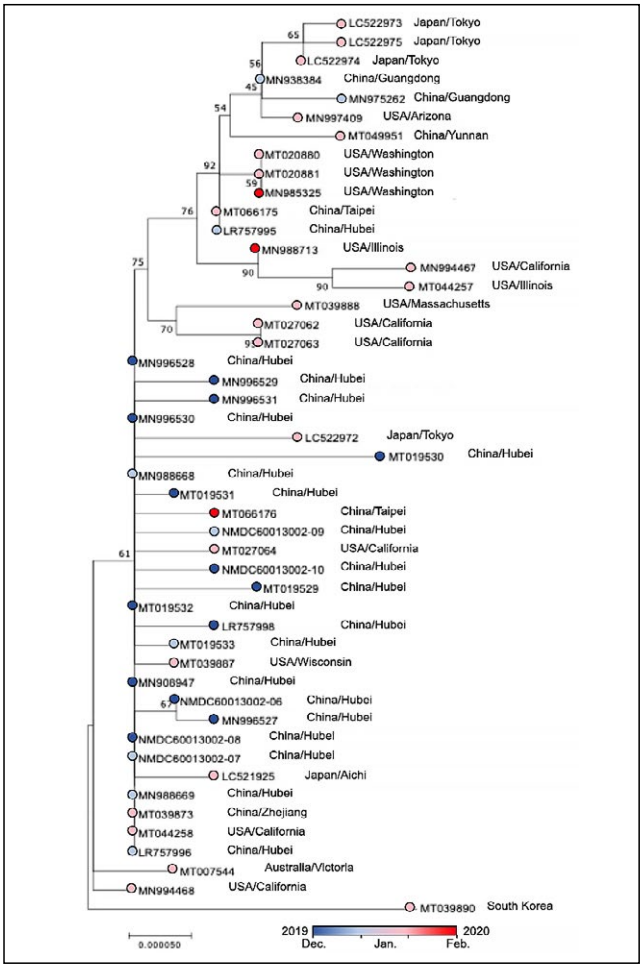

Figure 2. A phylogenetic tree of 48 SARS-CoV-2 genomes constructed based on the Maximum Composite Likelihood model in MEGA. Each strain is labeled with their GenBank/NGDC number followed by the region of collection. The collection date corresponds with the colors noted on the legend bar.

in time, there were 75,285 confirmed cases of COVID-19 and 2,009 related deaths, of which 74,279 cases were isolated in China and the remaining 1,006 cases were reported from 25 other countries and territories (Figure 1A). After removing low-quality sequences and selecting only full-length sequences (>29, 000 bp), 48 SARS-CoV-2 whole-genome sequences were obtained from GenBank and the NGDC for phylogenetic network analysis (Figure 1B).[13]

The genomic diversity of these early sequences is represented as a maximum likelihood phylogeny in linear layout (Figure 2). 47 of 48 virus genomes (98% of all sequences in our dataset) are disposed amongst one clade, suggesting the ancestral genome appears to have originated in Wuhan, China, and is likely to have spread worldwide. The South Korean isolate SNU01 separated as a distinct phylogenetic lineage and shares 99.88%-99.97% homology with the other strains.

We then analyzed and annotated all SARS-CoV-2 sequences compared with the reference Wuhan genome (WH-01). The distribution and frequency of unique nucleotide changes compared to the reference WH-01 strain are shown (Figure 3A). We identified 73 variable nucleotide sites in the SARS-CoV-2 genome alignment, including 31 synonymous mutations that did not alter the sequence of the encoded protein and 42 missense mutations that resulted in 42 variable protein sites (Figure 3B and 4A). The viral genome encodes at least 28 proteins and amongst the samples used in this dataset, 14 viral proteins remained highly conserved compared to the reference WH-01 strain, while the other 14 viral proteins accumulated variable residues compared to the reference (Figure 4A). Among these viral proteins, nonstructural protein 3 (Nsp3) accumulated the greatest number of variable sites (12 sites) and the L84S substitution in ORF8 occurs most frequently (29.17%).

Next, we used Motif Scan (MyHits, Switzerland) to characterize the functional domain, motif, and molecular patterns of each SARS-CoV-2 viral protein as a means to determine the putative structure and function of the coronavirus proteins and to investigate the possible roles they may play in the virus replication cycle. The predicted functional domains and motifs are summarized in Figure 4B-C and include ubiquitin-like domains (Ubl), ADP-ribose

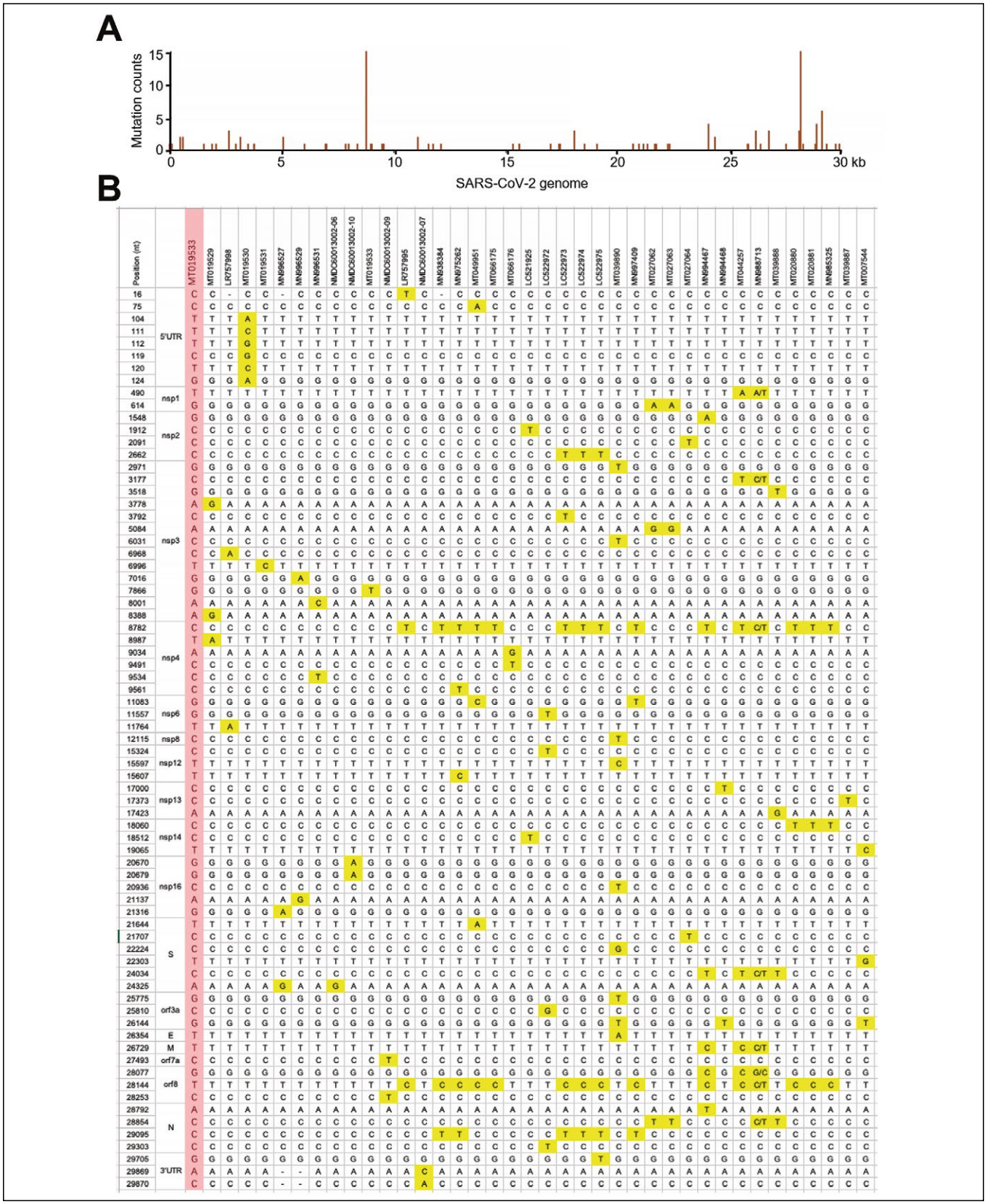

Figure 3. A. The distribution of variable nucleotide sites identified from pairwise whole-genome alignments of 48 available SARS-CoV-2 complete genomes. B. ClustalW multiple sequence alignment of 47 individual strains against the reference WH-01 sequence. Clusters of nucleotide substitutions are highlighted in yellow.

phosphatase domain (ADRP), papain-like protease domain (PLpro), SUMOylation site, and phosphorylation sites by TBK1 and Akt. In contrast to the SARS-CoV identified in 2003, some of the functional regions and motifs (i.e. the predicted SUMOylation site in Nsp3) are unique to SARS-CoV-2, suggesting post-translational regulation may contribute to viral mechanisms of pathogenesis and immune evasion. Interestingly, several bursts of amino acid substitutions have also occurred in other functional regions. The mutations T255I in Nps13 and S194L in the nucleocapsid protein are predicted to disrupt the predicted phosphorylation sites of TBK1 and Akt, respectively, and are likely to have functional implications due to disruption of post-translational modifications.

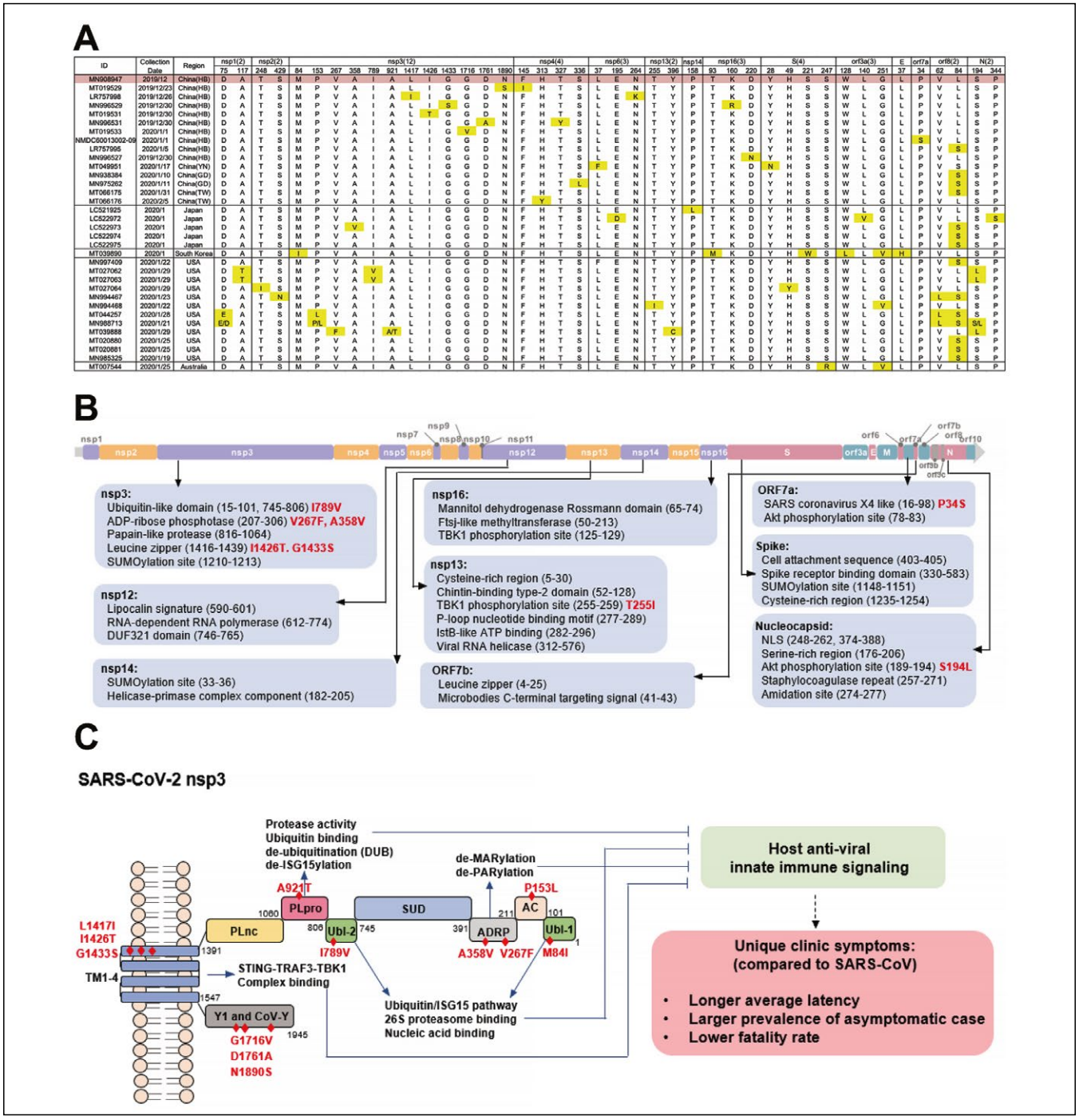

Figure 4. A. Visual depictions of the multiple sequence alignment for 14 SARS-CoV-2 protein by ClustalW, with mutations mapped onto the WH-01 protein sequence. The canonical motifs in SARS-CoV-2 were identified by MotifScan. B. Schematic presentation of the SARS-CoV-2 proteome organization (Wuhan/WH-01/2019, MN908947). C. Schematic representation of the structure and functions of the large multi-domain Nsp3. Twelve single amino acid substitutions across different strains are labeled in red.

Discussion

To analyze the genomic diversity that has emerged in the global population of SARS-CoV-2, we performed a phylogenetic network analysis of the first 48 SARS-CoV-2 genomes isolated and sequenced from across the world. We identify regions of the SARS-CoV-2 genome that have remained largely invariant compared to the reference WH-01 sequence, and others that have already accumulated variable sites. We focus in particular on mutations that are frequent, as they are likely candidates for SARS-CoV-2 divergence and can have important consequences for human hosts.[14] We identified a strong L84S mutational signal in ORF8 (present in 29.16% of genomes) together with 12 variable sites in the ORF encoding Nsp3 that represent the strongest putative regions under selection in our dataset. It is inevitable that viruses will accumulate mutations following replication and passage in human populations. Based on their effect on fitness, mutations can be advantageous (allowing the virus to adapt more to the human host), deleterious (decrease viral fitness), or neutral (effect too small to affect selection).[15,16] While our analysis reveals that the virus has undergone several mutations, it is not immediately clear whether a more transmissible early form of SARS-CoV-2 has emerged. Therefore, the construction of virus-host interactomes will be instrumental to determine mutant virus fitness relative to the reference WH-01 strain.[17] Of note, we did not identify mutations to the SARS-CoV-2 spike protein, which is the known mediator of host-cell entry and is commonly used in inactivated vaccines to elicit protective neutralizing antibodies.[10] Lack of mutational changes to the coronavirus spike protein is reassuring for vaccine candidates in development.

Amongst the samples used in this dataset, Nsp3 accumulated the most variable sites (12 sites, Figure 4A). Similar to the original SARS virus, Nsp3 is an essential component of the replication complex and is reported to block the host immune response.[18] We used Motif Scan to scan the WH-01 Nsp3 sequence for matches to all known motifs in the protein sequence database. Based on pattern searching results, SARS-CoV-2 Nsp3 is likely to contain the eight domains that exist in all known coronaviruses, including: ubiquitin-like domain (Ubl)-1, Glu-rich acidic domain (AC), ADP-ribose-1 phosphatase domain (ADRP), SARS unique domain (SUD), Ubl-2, papain-like protease (PLpro), nucleic acid binding domain (PLnc), transmembrane domain (TM)1-4, and Y domains.[18] Most interestingly, substitutions in the N-terminal region of Nsp3 (M84I, P153L, V267F, A358V, I789V, and A921T) are found only in the sequences reported by China before 1 January 2020, and substitutions in Nsp3 C-terminal region (L1417I, I1426T, G1433S, G1716V, D1761A, and N1890S) are only in the sequences reported by countries outside of China (USA, Japan, and South Korea) after 1 January 2020. This mutational pattern might be shaped by an unknown transmission route or might be driven by geographical environmental pressures.[19] Since both the ubiquitination and ISG15ylation domains of the original SARS Nsp3 have been reported to play critical roles in the inhibition of type I interferon signaling and generation of inflammatory cytokines,18 substitutions in Ubl-1 (M84I), ADRP (V267F and A358V), Ubl-2 (I789V), PLpro (A921T), and TM1-4 (L1417I, I1426T, and G1433S) might modulate how Nsp3 counteracts host innate immunity. Substitutions may additionally contribute to some of the unique clinical features and epidemiological characteristics of SARS-CoV-2 in comparison with the original SARS virus, including its high transmissibility, longer latency period, and lower-case fatality rates.[20] Further characterization of the immune-modulatory mechanism of Nsp3 and other SARS-CoV-2 viral proteins may address these knowledge gaps.

Conclusion

Monitoring how the virus is adapting to its human host is of crucial importance for guiding the ongoing development of vaccines and therapeutics. Our analyses here presented an early snapshot in time and some of the genetic diversity of present SARS-CoV-2 strains in circulation was likely un-sampled. Nonetheless, as additional sequencing assemblies become available to the research community, we present here a plausible pipeline for phylogenetic

Table 1. Contributing sequences from GenBank and NGDC

GenBank/NGDC ID	Region	GenBank/NGDC ID	Region	GenBank/NGDC ID	Region
LC522973	Japan/Tokyo	MT027062	USA/California	LR757998	China/Hubei
LC522975	Japan/Tokyo	MT027063	USA/California	MT019533	China/Hubei
LC522974	Japan/Tokyo	MN996528	China/Hubei	MT039887	USA/Wisconsin
MN938384	China/Guangdong	MN996529	China/Hubei	MN908947	China/Hubei
MN975262	China/Guangdong	MN996531	China/Hubei	NMDC60013002-06	China/Hubei
MN997409	USA/Arizona	MN996530	China/Hubei	MN996527	China/Hubei
MT049951	China/Yunnan	LC522972	Japan/Tokyo	NMDC60013002-08	China/Hubei
MT030880	USA/Washington	MT019530	China/Hubei	NMDC60013002-07	China/Hubei
MT020881	USA/Washington	MN988668	China/Hubei	LC521925	Japan/Aichi
MN985325	USA/Washington	MT019531	China/Hubei	MN988669	China/Hubei
MT066175	China/Taipei	MT066176	China/Taipei	MT039873	China/Zhejiang
LR757995	China/Hubei	NMDC60013002-09	China/Hubei	MT044258	USA/C
MN988713	USA/Illinois	MT027064	USA/California	LR757996	China/Hubei
MN994467	USA/California	NMDC60013002-10	China/Hubei	MT007544	Australia/Victoria
MT044257	USA/Illinois	MT019529	China/Hubei	MN994468	USA/California
MT039888	USA/Massachusetts	MT019532	China/Hubei	MT039890	South Korea

network analysis to study the evolutionary trajectory of SARS-Cov-2. We believe it is important to continue to monitor for potential signatures of divergent strains and distinct geographic variants with biological differences in virulence and transmission rate to guide vaccine development.

Methods

Data acquisition

After removing low-quality sequences (i.e. low coverage reads) and using only full-length sequences (>29, 000 bp), 48 SARS-CoV-2 whole-genome sequences were downloaded from GenBank and the Chinese National Genomics Data Center (NGDC) Genome Warehouse for downstream analysis on February 19 2020. A full table of the GenBank/NGDC and region of collection is provided in Table 1.

Phylogenetic analysis

We used the open-source software Molecular Evolutionary Genetics Analysis (MEGA) to construct a phylogenetic tree from the aligned sequences with bootstrapping values set to 1,000. MEGA uses the maximum likelihood algorithm, which is a standard statistical technique based on the maximum-parsimony method. For this study, we used the first whole-genome assembly (WH-01, MN908947) deposited on NGDC Genome warehouse in December 2019 as the reference sequence. ClustalW multiple sequence alignment was used to align all 47 genome sequences over the WH-01 reference. Motif Scan (MyHits, SIB, Switzerland, https://myhits.sib.swiss/cgi-bin/motif_scan) was used as a sequence-based protein domain search tool to scan the WH-01 Nsp3 sequence for matches to all known motifs in the protein sequence database. Motif Scan is a fast, web-based analysis tool that allows users to search for patterns conserved in protein sequences.

Data availability

The nucleotide sequences of the SARS-CoV-2 genomes used in this analysis are publicly available from the GenBank database and the NGDC.

Acknowledgement

We gratefully acknowledge the laboratories contributing the sequences from GenBank and NGDC on which this research is based. A table of the contributing sequences is in Table 1.

Author contribution

All authors fulfilled all conditions required for authorship and have approved this submission. Z.X. conceived of the research, designed the study, and wrote the manuscript. D.W. helped make substantial contributions to the interpretation of the data and revised the manuscript critically.

References

1. Wu F, Zhao S, Yu B, et al. A new coronavirus associated with human respiratory disease in China. Nature. 2020;3;579(7798):265-269
2. Zhou P, Yang X, Wang X, et al. A pneumonia outbreak associated with a new coronavirus of probable bat origin. Nature (London). 2020 Feb 3;579(7798):270-73
3. World Health Organization. Coronavirus disease (COVID-19) pandemic – World Health Organization. Available at: https://www.who.int/emergencies/diseases/novel-coronavirus-2019. Accessed Nov 4, 2020
4. Chan JF, Yuan S, Kok K, et al. A familial cluster of pneumonia associated with the 2019 novel coronavirus indicating person-to-person transmission: a study of a family cluster. The Lancet (British edition). 2020 Feb;395(10223):514-23
5. Boni MF, Lemey P, Jiang X, et al. Evolutionary origins of the SARS-CoV-2 sarbecovirus lineage responsible for the COVID-19 pandemic. Nature microbiology. 2020 Jul 28;5(11):1408-17
6. Astuti I. Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2): An overview of viral structure and host response. Diabetes Metab Syndr. 2020;14(4):407-12
7. Ma C, Su S, Wang J, et al. From SARS-CoV to SARS-CoV-2: safety and broad-spectrum are important for coronavirus vaccine development. Microbes Infect. 2020;22(6):245-53
8. Duffy S. Why are RNA virus mutation rates so damn high? PLoS Biol. 2018;16(8)
9. Rawat K, Kumari P, Saha L. COVID-19 vaccine: A recent update in pipeline vaccines, their design and development strategies. European Journal of Pharmacology. 2021 Feb;892:173751
10. Gupta T, Gupta SK. Potential adjuvants for the development of a SARS-CoV-2 vaccine based on experimental results from similar coronaviruses. Int Immunopharmacol. 2020;86:106717
11. Nishant KT, Singh ND, Alani E. Genomic mutation rates: What high-throughput methods can tell us. Bioessays. 2009;31(9):912-920
12. Pachetti M, Marini B, Benedetti F, et al. Emerging SARS-CoV-2 mutation hot spots include a novel RNA-dependent-RNA polymerase variant. J Transl Med. 2020;22:18
13. Kumar S, Stecher G, Li M, et al. MEGA X: Molecular Evolutionary Genetics Analysis across Computing Platforms. Mol Biol Evol. 2018;35(6):1547-49
14. Andersen KG, Rambaut A, Lipkin WI, Holmes EC, Garry RF. The proximal origin of SARS-CoV-2. Nature medicine. 2020 Mar 17;26(4):450-52
15. Kaushal N, Gupta Y, Goyal M, et al. Mutational Frequencies of SARS-CoV-2 Genome during the Beginning Months of the Outbreak in USA. Pathogens. 2020;9(7)
16. Loewe L, Hill WG. The population genetics of mutations: good, bad and indifferent. Philosophical transactions. Biological sciences. 2010 Apr 27;365(1544):1153-67
17. Gordon DE, Hiatt J, Bouhaddou M, et al. Comparative host-coronavirus protein interaction networks reveal pan-viral disease mechanisms. Science. 2020 Oct 15
18. Nsp3 of coronaviruses: Structures and functions of a large multi-domain protein. Antiviral Research. 2018 Jan 1;149:58-74
19. Karlsson EK, Kwiatkowski DP, Sabeti PC. Natural selection and infectious disease in human populations. Genetics. 2014 Apr 29;15(6):379-93
20. Guan W, Ni Z, Hu Y, et al. Clinical Characteristics of Coronavirus Disease 2019 in China. N Engl J Med. 2020 Mar 30;382(18):1708-20