 doi:10.20944/preprints202004.0137.v1

Genome Wide Analysis of Severe Acute Respiratory Syndrome Coronavirus-2 Implicates World-Wide Circulatory Virus Strains Heterogeneity

M. Rafiul Islam[1*], M. Nazmul Hoque[1*,2], M. Shaminur Rahman[1*], J. Akter Puspo[1], Masuda Akther[1], Salma Akter[1,3], A. S. M. Rubayet-Ul-Alam[1,4], Munawar Sultana[1], Keith A. Crandall[5], M. Anwar Hossain[1,6**]

[1]Department of Microbiology, University of Dhaka, Dhaka 1000, Bangladesh

[2]Department of Gynecology, Obstetrics and Reproductive Health, Bangabandhu Sheikh Mujibur Rahman Agricultural University, Gazipur-1706, Bangladesh

[3]Department of Microbiology, Jahangirnagar University, Savar, Dhaka-1342, Bangladesh

[4]Department of Microbiology, Jashore University of Science and Technology, Jashore 7408, Bangladesh

[5]Computational Biology Institute, Milken Institute School of Public Health, The George Washington University, USA

[6]Present address: Vice-Chancellor, Jashore University of Science and Technology, Jashore 7408, Bangladesh

*Equal contribution

**Corresponding to:

M. Anwar Hossain, PhD
Professor
Department of Microbiology
University of Dhaka, Dhaka, Bangladesh
E-mail: hossaina@du.ac.bd

 doi:10.20944/preprints202004.0137.v1

Abstract

Severe acute respiratory syndrome coronavirus-2 (SARS-CoV-2), a novel evolutionarily divergent RNA virus etiological agent of COVID-19, is responsible for present devastating pandemic respiratory illness. To explore the genomic signatures, we comprehensively analyzed 2,492 complete and/or near-complete genome sequences of SARS-CoV-2 strains reported from across the globe to the GISAID database up to 30 March 2020. Genome-wide annotations revealed 1,407 nucleotide-level mutations at different positions throughout the entire genome of SARS-CoV-2. Moreover, nucleotide deletion analysis found nine deletions throughout the genome, including in polyprotein (n=6), ORF10 (n=1) and 3´-UTR (n=2). Evidence from the systematic gene-level mutational and protein profile analyses revealed a large number of amino acid (aa) substitutions (n=722), making the viral proteins heterogeneous. Notably, residues of receptor-binding domain (RBD) having crucial interactions with angiotensin-converting enzyme 2 (ACE2), and cross-reacting neutralizing antibody were found to be conserved among the analyzed SARS-CoV-2 strains, except for replacement of Lysine with Arginine at 378 position of the cryptic epitope of a Shanghai isolate, hCoV-19/Shanghai/SH0007/2020 (EPI_ISL_416320). Our method of genome annotation is a promising tool for monitoring and tracking the epidemic, the associated genetic variants, and their implications for the development of effective control and prophylaxis strategy.

Key words: SARS-CoV-2, Genomes, Nucleotide, Amino-acid, Mutations, Replacements

 doi:10.20944/preprints202004.0137.v1

Introduction

Severe acute respiratory syndrome (SARS) is an emerging pneumonia-like respiratory disease of human, which was reported to be re-emerged in Wuhan city of China in December 2019[1]. The identified causative agent is found to be a highly contagious novel beta-coronavirus 2 (SARS-CoV-2). Similar to other known SARS-CoV and SARS-related coronaviruses (SARSr-CoVs)[2,3], the viral RNA genome of the novel SARS-CoV-2 encodes several smaller open reading frames (ORFs) such as ORF1ab, ORF3a, ORF6, ORF7a, ORF7b, ORF8 and ORF10 located in the 3′ region of the genome. These ORFs are predicted to encode for the replicase polyprotein, the spike (S) glycoprotein, envelope (E), membrane (M), nucleocapsid (N) proteins, accessory proteins, and other non-structural proteins (nsp)[3-5].

However, the ongoing rapid transmission and global spread of SARS-CoV-2 have raised critical questions about the evolution and adaptation of the viral population driven by mutations, deletions and/or recombination as it spreads across the world encountering diverse host immune systems and various counter-measures[6]. Initial phylogenomic analysis of three super-clades (S, V, and G) isolated from the outbreaks of distinct geographic locations (China, USA and Europe) within SARS-CoV-2 showed little evidence of local/regional adaptation, suggesting instead that viral evolution is mainly driven by genetic drift and founder events[7]. Nevertheless, several reports predict possible adaptation at the nucleotide, amino acid (aa), and structural heterogeneity in the viral proteins, especially the spike (S) protein[8,9]. Interestingly, Shen et al. reported even intra-host viral evolution among the patients after infection, which might be related to its virulence, transmissibility, and/or evolution do to immune response[10]. However, the previous reports have the limitations of considering a very few representative complete genomes covering only a few countries, targeting clade/group based consensus sequence, comparison to the Wuhan Refseq genome, and focusing on the structural proteins. Therefore, our study has targeted the genome-wide mutational spectra for inference on evolution of the viral population.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 9 April 2020 doi:10.20944/preprints202004.0137.v1

To decipher the genetic variations, we retrieved two thousand four hundred and ninety-two (n = 2,492) complete or near-complete genomes of SARS-CoV-2 available at the global initiative on sharing all influenza data (GISAID) (https://www.gisaid.org/) up to 30 March 2020. These SARS-CoV-2 sequences belonged to the infected patients from 58 countries of seven continents (Supplementary Fig. 1, Supplementary Data 1). We aligned the SARS-CoV-2 genome sequences using MAFFT online server[11], and the complete genome sequence SARS-CoV-2 Wuhan-Hu-1 strain (Accession NC_045512, Version NC_045512.2) was used as a reference genome. Multiple sequence alignments were finally opened with MEGA 7[12] to remove all ambiguous and low-quality sequences. Amino-acid heterogeneity analysis was performed with Fingerprint, a web-based protein profile analysis tool[13]. Finally, the aligned sequences were visualized using Unipro-UGENE 1.26.1 to visualize the deletions with respect to the reference genome[14].

Results and Discussions

Nucleotide sequence alignment revealed a total of 1,407 mutations (synonymous vs nonsynonymous ratio = 2.8:1) across the entire set of genomes of the SARS-CoV-2 strains compared to the NCBI reference strain, Wuhan-Hu-1 (Accession NC_045512). Of the identified nucleotide substitutions, 655 were found in polyprotein regions of SARS-CoV-2 while the structural proteins had 337 nucleotide mutations comprising 173, 30, 25 and 109 mutations in spike (S) glycoprotein, membrane (M), envelop (E) and nucleocapsid (N) proteins, respectively. In addition, the ORF3a, ORF6, ORF7ab, ORF8, and ORF10 had a total of 64, 8, 39, 19 and 16 nucleotide-level variations, respectively whereas 105, 158 and 6 nucleotide mutations were sequentially detected in 5´-UTR, 3´-UTR, and spacer regions. (Supplementary Data 2).

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 9 April 2020 doi:10.20944/preprints202004.0137.v1

Moreover, in the ORFs regions total 1148 nucleotide mutations were observed and among them 416 were missense mutations. In the ORF1ab polyprotein, 74, 69, 57, 35 and 29 aa substitutions have been identified in the nsp3, nsp4, nsp2, nsp12 and nsp5, respectively (Supplementary Data 2). In case of spike protein, 12 aa substitutions have been found in the receptor-binding domain (RBD) at 331 to 524 residues of S1 subunit (V341I, A348T, N354D, D364Y, V367F, K378R, Q409E, A435S, G476S, V483A, and Y508H in Wales, USA, Shenzhen, Shenzhen, Hong Kong/France, Shanghai, Guangdong, Finland, USA, USA, and France, respectively), among which three substitutions occurred in the positions between 424 to 494 comprising the receptor-binding motif (RBM). Sarkar et al. identified a unique mutation in the S glycoprotein (A930V)[15] in the Indian SARS-CoV-2, which was absent in other related SARS-CoV-2 strains from different geographical regions. Additionally, we identified twelve and five aa replacements in the Heptad Repeat-1 and 2 (HR1, HR2) comprising 892-1013 and 1145-1195 positions in the S2 subunit, respectively. We observed five aa replacements in the membrane (M) protein in the topological domains (1-18 and 71-78), and the envelope (E) protein and nucleocapsid (N) protein had 10 and 75 aa replacements, respectively.

In a recent study, Yuan et al. demonstrated the molecular insights of a cross-reacting neutralizing antibody CR3022 to the highly conserved cryptic epitope[16] (86% conserved between SARS-CoV and SARS-CoV-2) on the receptor-binding domain (RBD) of the SARS-CoV-2 spike protein, but distal to the receptor-binding site linked to the angiotensin-converting enzyme 2 (ACE2) of host cell surface during infecting the cells. Significantly, our results identified one amino acid substitution (K378R) among the 28 residues of that epitope in hCoV-19/Shanghai/SH0007/2020 (EPI_ISL_416320) strain, which was isolated on 28 January, 2020 from China. The aa residues (D54 and E56) on the paratope of antibody interact with the Lysine

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 9 April 2020 doi:10.20944/preprints202004.0137.v1

situated on the epitope, however the replaced Arginine found in hCoV-19/Shanghai/SH0007/2020 may have a superior binding to the neutralizing antibody because of stronger electrostatic interactions, such as salt-bridges and hydrogen bonds compared to lysine[17]. At this point, we must stress that the changing of this amino acid is not directly linked to immunological pressure after applying the convalescent plasma therapy on the patients of China started from 9 February 2020, rather can be an evolutionary purifying selection possibly increasing stronger interaction with antibody[18]. In addition, it should be mentioned that the residues associated with ACE2 binding to the spike protein (439N, 449Y, 453Y, 455L, 456F, 475A, 486F, 487N, 489Y, 493Q, 495Y, 498Q, 499T, 500N, 501G, 504Y) have not shown any variation in the mutational analyses[16].

Besides the site-specific mutations, nine deletions of ranged nucleotides were found in polyprotein (n=6), ORF10 (n=1) and 3´-UTR (n=2) in 15 strains reported from Japan, USA, England, Canada, Netherlands, Wuhan and Australia (Fig. 1). These nucleotide deletions can influence potentially the tertiary structures and functions of the polyprotein, S, M and E proteins which may play important role in virus-host interactions for infections, pathogenesis as well as immune-modulations[5,19-21]. Notably, Liu et al. identified two common deletions with a very low frequency at 23585–23599 (aa- QTQTN) positioned at the upstream of the polybasic cleavage site of S1-S2, and 23596–23617 (aa- NSPRRAR) including the polybasic cleavage site in the clinical samples and cell-isolated virus strain[22]. According to that study, no such deletion was reported elsewhere, and our results did not find that in the genomic data deposited in the public database either.

Remarkably, ORF10 undergoes a deletion (35 nucleotides) including its start codon, and instead, the start codon of adjacent spacer region can probably be used for protein coding (Fig. 1g). Among these deletions, three were reported previously in three strains (Japan/AI/I-

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 9 April 2020 doi:10.20944/preprints202004.0137.v1

004/2020|EPI_ISL_407084, USA/WI1/2020|EPI_ISL_408670, Australia/VIC01/2020|EPI_ISL_406844)[5] and an additional 382-nucleotide deletion was also reported spanning from downstream of ORF7b to ORF8 region (not shown in figure) in three SARS-CoV-2 strains (Singapore/14/2020|EPI_ISL_414380, Singapore/12/2020|EPI_ISL_414378, Singapore/13/2020|EPI_ISL_414379) reported earlier from Singapore with an impact of enhanced transcription of the subsequent N protein[23] indicating ongoing evolution of the virus. Notably, the specific functions of these accessory proteins, ORF8 and ORF10 remained mysterious. Nevertheless, recent *in-silico* analyses predicted the binding of the ORF8 to the porphyrin, and antagonistic attacking of ORF10 protein to the heme that dissociates the iron from porphyrin. These effects are linked to the disease manifestations and clinical outcomes of patients, hence the deletions may have possible roles in virulence and pathogenicity[22].

Another noteworthy finding of our present systematic analysis is the identification of 722 aa replacements at different positions of the entire SARS-CoV-2 genome. We found 404 aa replacements in polyprotein regions of SARS-CoV-2 and 215 in the structural proteins comprising 114, 15, 10 and 76 replacements in spike (S) glycoprotein, membrane (M), envelop (E) and nucleocapsid (N) proteins, respectively. Besides structural proteins, 48, 5, 25, 15 and 10 aa replacements were identified in ORF3a, ORF6, ORF7ab, ORF8 and ORF10, respectively (Supplementary Data 2). The protein sequence heterogeneity profile obtained from Fingerprint analysis revealed significant heterogeneity in the aa residues within immunologically important structural proteins, S, E and M of different strains of SARS-CoV-2 (Fig. 2). Also, the high morbidity and mortality of immune-compromised patients infected by COVID-19 can be correlated with the disastrous consequences of these mutations[5].

In conclusion, our study reveals a number of unreported mutations in the genomes of SARS-CoV-2, which cover both mismatch and deletion mutations both in translated and

 doi:10.20944/preprints202004.0137.v1

untranslated regions. Further investigations should focus on *in-silico* structural validations and subsequent phenotypic consequences of the deletions and/or mismatches in transmission dynamics of the current epidemics and the immediate implications of these genomic markers to develop potential prophylaxis and mitigation for tackling the pandemic COVID-19 crisis. Moreover, the identification of the conformational changes in mutated protein structures and untranslated cis-acting elements is of significance for studying the virulence, pathogenicity and transmissibility of SARS-CoV-2. This mutational diversity should be investigated by further studies, including their metabolic functional pathway analysis.

Figures

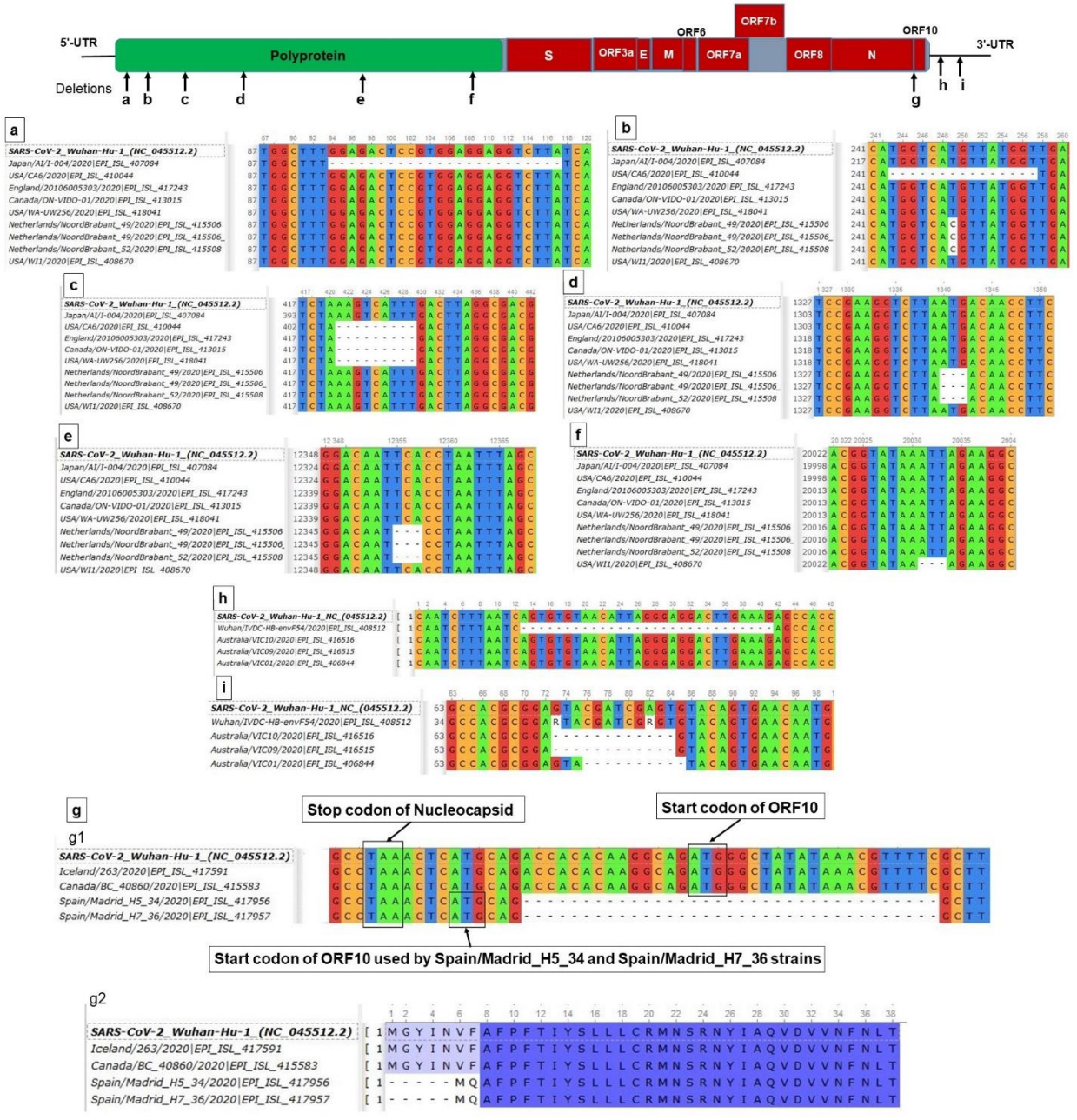


Fig. 1 Genomic deletion analysis of SARS-CoV-2. Genomic deletion analysis of SARS-CoV-2 strains identified (a) twenty-four nucleotide deletions in nsp1 of polyprotein from Japanese strain, (b) fifteen-nucleotide deletions in nsp1 of polyprotein of a USA strain, (c) nine-nucleotide deletion in nsp1 of polyprotein of strains from USA, England and Canada, three-nucleotide deletions in (d) nsp2 and nsp8 (e) of polyprotein of Netherlands strains, (f) three-nucleotide deletion in nsp15 of

 doi:10.20944/preprints202004.0137.v1

polyprotein of USA strain, (g; g1) number of nucleotides deletion including start codon position of ORF10 of Spain strains, and the start codon in spacer position, has been used for ORF10 coding, (g; g2) as a result five aa residues deletion in those strains starting from position 1 to 5. Deletion of (h) twenty-nine nucleotides reported from Wuhan, and (i) nucleotides deletion in 3'-UTR of strains belonged to Australia. The position of nucleotide represents the starting position from each ORFs.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 9 April 2020 doi:10.20944/preprints202004.0137.v1

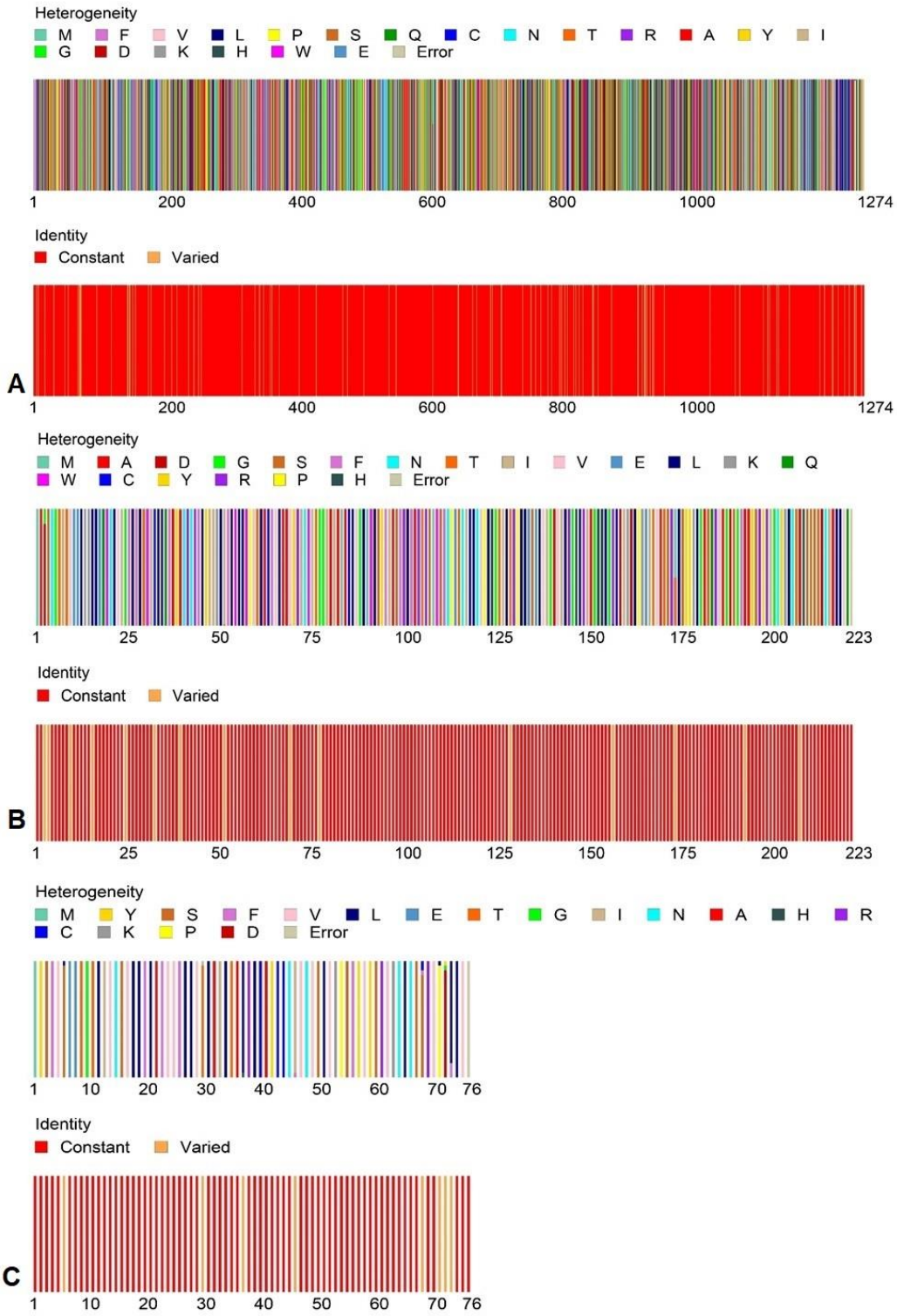
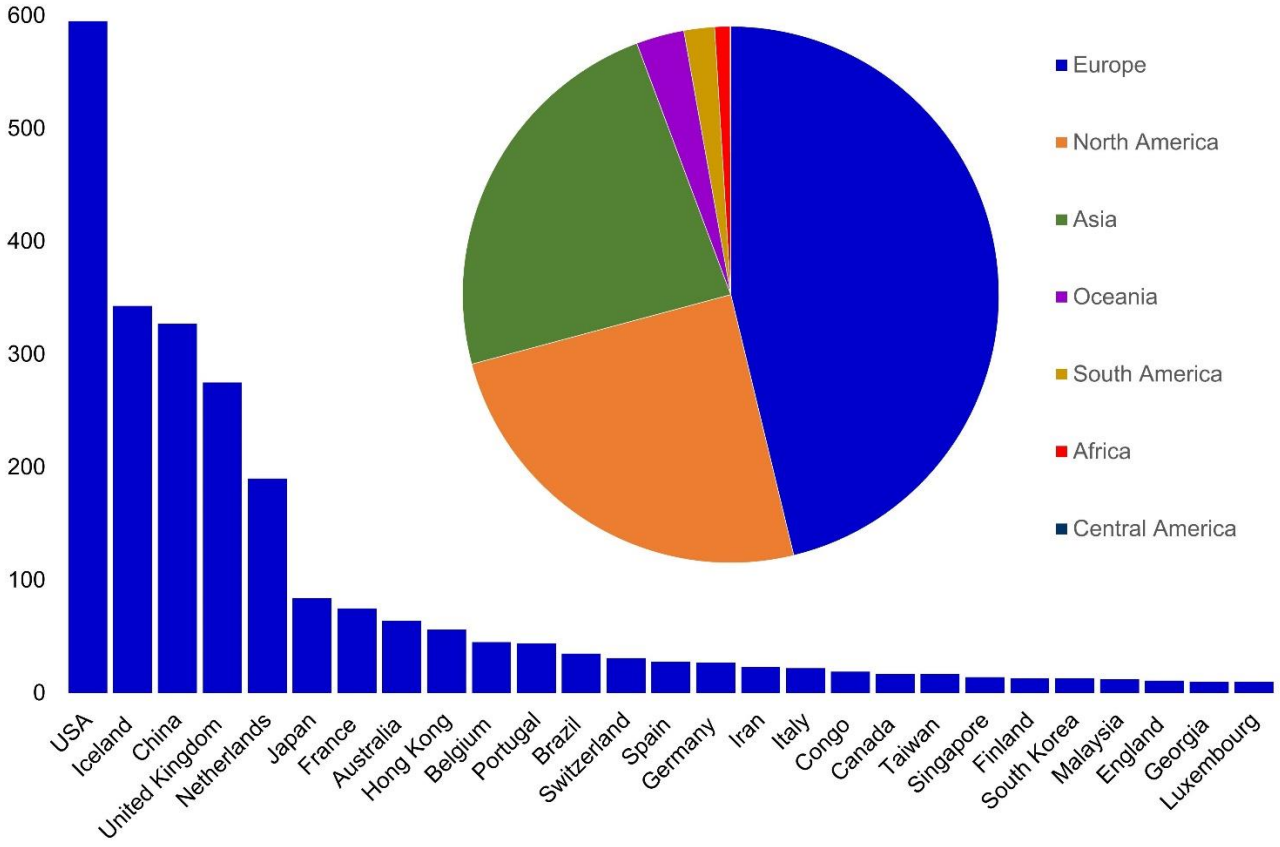


Fig. 2 Amino-acid (aa) residues heterogeneity in (A) S, (B) E and (C) M proteins of SARS-CoV-2. The Fingerprint protein analysis showed that aa residues in S, M and E proteins varied due to change and/or substitutions in their positions.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 9 April 2020 doi:10.20944/preprints202004.0137.v1

Supplementary Figure



Supplementary Fig. 1 Geographic distribution of SARS-CoV-2 strains. We retrieve 2492 SARS-CoV-2 sequences from the GISAID belonged to the infected patients from 58 countries across seven continents. Countries having > 20 sequences are depicted in bar plot, and rest are available in Supplementary Data 1.

 doi:10.20944/preprints202004.0137.v1

Conflict of interest statement

The authors declare no competing interests.

Author contributions

MRI, MNH, MSR, JA, MA, ASMRUA, and SA conducted the overall study and also drafted the manuscript. MNH finally compiled the manuscript. MS, KAC and MAH contributed intellectually to the interpretation and presentation of the results.

Supplementary Material

Supplementary information supporting the findings of the study are available in this article as Supplementary Data files, or from the corresponding author on request.

References

1. Zhu, N., et al. "A novel coronavirus from patients with pneumonia in China, 2019." *New England J. Med.* (2020).
2. Cotten, M. et al. Transmission and evolution of the Middle East respiratory syndrome coronavirus in Saudi Arabia: a descriptive genomic study. *The Lancet* 3**82(9909)**, 1993-2002 (2013).
3. Walls, A. C. et al. Structure, function, and antigenicity of the SARS-CoV-2 spike glycoprotein. *Cell* (2020).
4. 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*. **12(3)**, 254 (2020).
5. Phan, T. (2020). Genetic diversity and evolution of SARS-CoV-2. *Infect. Genet. Evol*. **81**, 104260 (2020).
6. Rahman, M. S. et al. 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. *bioRxiv*. (2020), doi: https://doi.org/10.1101/2020.03.30.015164.
7. Chiara, M., Horner, D. S. & Pesole, G. Comparative genomics suggests limited variability and similar evolutionary patterns between major clades of SARS-Cov-2. *bioRxiv* (2020), https://doi.org/10.1101/2020.03.30.016790.
8. Sardar, R., Satish, D., Birla, S. & Gupta, D. (2020). Comparative analyses of SAR-CoV2 genomes from different geographical locations and other coronavirus family genomes reveals unique features potentially consequential to host-virus interaction and pathogenesis. *bioRxiv*. https://doi.org/10.1101/2020.03.21.001586.
9. Armijos-Jaramillo, V., Yeager, J., Muslin, C. & Perez-Castillo, Y. (2020). SARS-CoV-2, an evolutionary perspective of interaction with human ACE2 reveals undiscovered amino

acids necessary for complex stability. *bioRxiv*. (2020), https://doi.org/10.1101/2020.03.21.001933.

10. Shen, Z. et al. Genomic diversity of SARS-CoV-2 in Coronavirus Disease 2019 patients. *Clinical Infect. Dis.* (2020).

11. Katoh, K., Misawa, K., Kuma, K. & Miyata, T. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res.* **30(14)**, 3059–3066 (2002), https://doi.org/10.1093/nar/gkf436.
12. Kumar, S., Stecher, G. & Tamura, K. MEGA7: molecular evolutionary genetics analysis version 7.0 for bigger datasets. *Mol. Biol. Evol.* **33(7)**, 1870-1874 (2016).
13. Goyal, A. et al. Identification of an Ideal-like Fingerprint for a Protein Fold using Overlapped Conserved Residues based Approach. *Sci. Rep.* **4**, 5643 (2015).
14. Okonechnikov, K., Golosova, O., Fursov, M. & Ugene Team. Unipro UGENE: a unified bioinformatics toolkit. *Bioinformatics* **28(8)**, 1166-1167 (2012).
15. Sardar, R., Satish, D., Birla, S. & Gupta, D. (2020). Comparative analyses of SAR-CoV2 genomes from different geographical locations and other coronavirus family genomes reveals unique features potentially consequential to host-virus interaction and pathogenesis. *bioRxiv*. (2020), doi: https://doi.org/10.1101/2020.03.21.001586.
16. Yuan, M. et al. (2020). A highly conserved cryptic epitope in the receptor-binding domains of SARS-CoV-2 and SARS-CoV. *Science* **eabb7269**, doi: 10.1126/science. abb7269.
17. Sokalingam, S., Raghunathan, G., Soundrarajan, N. & Lee, S. G. A Study on the Effect of Surface Lysine to Arginine Mutagenesis on Protein Stability and Structure Using Green Fluorescent Protein. *PLoS ONE* **7(7)**, e40410 (2012).
18. Casadevall, A. & Pirofski, L. A. The convalescent sera option for containing COVID-19. *The Journal Clin. Invest*. **130(4)**, (2020).

19. Li, W. et al. Bats are natural reservoirs of SARS-like coronaviruses. *Science* **310**, 676–679 (2005).

20. Xu, Y. et al. Characteristics of pediatric SARS-CoV-2 infection and potential evidence for persistent fecal viral shedding. *Nat. Med.* 1-4 (2020).

21. Zhou, P. et al. A pneumonia outbreak associated with a new coronavirus of probable bat origin. *Nat*. 1-4 (2020).

22. Liu, S.et al. Interaction between heptad repeat 1 and 2 regions in spike protein of SARS-associated coronavirus: implications for virus fusogenic mechanism and identification of fusion inhibitors. *The Lancet* **363(9413)**, 938-947 (2004).

23. Su, Y. et al. Discovery of a 382-nt deletion during the early evolution of SARS-CoV-2. *bioRxiv*. (2020)