A gapless unambiguous RNA metagenome-assembled genome sequence of a unique SARS-CoV-2 variant encoding spike S813I and ORF1a A859V substitutions

May Sherif Soliman
Clinical Pathology Department, Faculty of Medicine, Cairo University, Cairo, Egypt
May AbdelFattah
Clinical Pathology Department, Faculty of Medicine, Cairo University, Cairo, Egypt
Soad M. N. Aman
Department of Microbiology and Immunology, Faculty of Pharmacy, Cairo University, Cairo, Egypt
Lamyaa M. Ibrahim
Department of Microbiology and Immunology, Faculty of Pharmacy, Cairo University, Cairo, Egypt
Ramy Karam Aziz (raziz1@gmail.com)
Department of Microbiology and Immunology, Faculty of Pharmacy, Cairo University, Cairo, Egypt

Short Report

Keywords: COVID-19, pandemic, genomic epidemiology, high-throughput sequencing, single-nucleotide polymorphism, spike protein

Posted Date: November 13th, 2020

DOI: https://doi.org/10.21203/rs.3.rs-98061/v2

Version of Record: A version of this preprint was published on November 30th, 2020. See the published version at https://doi.org/10.1089/omi.2020.0194.

Abstract

The novel severe acute respiratory syndrome corona virus 2 (SARS-CoV-2) is causing an unprecedented pandemic, threatening global health, daily life, and economy. Genomic surveillance continues to be a critical effort towards tracking the virus and containing its spread, and more genomes from diverse geographical areas and different time points are needed to provide an appropriate representation of the virus evolution. We here report the successful assembly of one single gapless, unambiguous contiguous sequence representing the complete viral genome from a nasopharyngeal swab of an infected healthcare worker in Cairo, Egypt. The sequence has all typical features of SARS-CoV-2 genomes, with no protein-disrupting mutations; however, three mutations are worth highlighting and future tracking: a synonymous mutation causing a rare spike S-813-I variation) and two less frequent ones leading to an A41V variation in NSP3, encoded by ORF1a (ORF1a A895V), and a Q677H variation in the spike protein. Both affected proteins, S and NSP3, are relevant to vaccine and drug development. While the genome, named CU_S3, belongs to the prevalent global genotype, marked by the D614G spike variation, the combined variations in the spike proteins and ORF1a have not been observed in any of the 197,000 genomes reported to date. Future studies will assess the biological, pathogenic, and epidemiologic implications of this set of genetic variations.

Background And Motivation

Since the first case of COVID-19 in December 2019, the novel severe acute respiratory syndrome corona virus 2 (SARS-CoV-2) has been circulating among humans (Li et al., 2020) causing a historically unmatched pandemic that is threatening global health, lifestyle, and economy.

In context, two of the most unforgettable pandemics (the medieval bubonic plague—'the *Black Death*'—and the 1918 influenza) occurred at times neither bacteria nor viruses were properly characterized, and thus remained with no defined etiological agent (neither identified morphologically/microscopically nor—of course—genetically). Conversely, the early 21st century SARS epidemic of 2002-2003 was promptly identified and well contained, and the success of its containment was, at least in part, due to genomics. The SARS virus genome was sequenced and tracked properly; cases were traced and isolated in a proportionate way; and the severe disease, which had a high case fatality rate, vanished within two years (de Wit et al., 2016).

On the other hand, the new SARS-like virus, SARS-CoV-2 is intensely challenging all containment measures for several epidemiological reasons (e.g., longer incubation period, transmission from presymptomatic or mildly symptomatic individuals, and—sadly—politics).

As of November 11 2020, over 197,000 genome sequences have been made available (with an hourly increase), in an extraordinary speed, through different databases, including the GISAID website (GISAID), which allows swift, immediate, and open sharing of data. To the best of our knowledge, no organism or virus has been sequenced at this throughput—not even close. Continuous sequencing allows tracking

mutations and their resulting genome variants, following trends of transmission, and—most importantly—getting prepared for potential biologically significant variants of the virus that may increase the its spread, virulence, or immune escape.

Here we report a complete genomic sequence of SARS-CoV-2, assembled from a metagenomic sequence of a nasopharyngeal swab obtained from a healthcare worker in a hospital in Egypt. The reported sequence encodes a unique combination of amino acid variations that affect the sequence of the replicase and spike proteins.

Materials And Methods

Ethics statement.All protocols were approved by the Cairo University anti-COVID-19 Task Force on 15 April 2020. Sampling, informed consent forms, and all sequencing protocols have been approved by the Ethical committee and Institutional Review Board of the Faculty of Medicine, Cairo University (IRB Approval: 07072018).

Sample information.The sequenced swab was obtained on June 19th, 2020 from a 40-year-old woman, a healthcare worker, who had typical symptoms of mild COVID-19 pneumonia and a 38.5 ˚C fever. After confirmation of SARS-CoV-2 positivity by real-time PCR, the patient's consent was obtained for full sequencing of the extracted RNA.

Library preparation and Sequencing. Total RNA was directly extracted from the swab with Qiagen QIAamp Viral RNA Mini Kit (Qiagen, Valencia, CA, USA), depleted for human, bacterial, and mitochondrial ribosomal RNA by Ribo-Zero plus (Illumina, La Jolla, CA, USA). Illumina MiSeq was used for sequencing a shotgun metagenomic library prepared by Illumina's TruSeq RNA Library Prep Gold Kit (v2) protocol.

Assembly and mapping. The sequences were mapped to SARS-CoV-2 genomes by Blast+ local alignment (Camacho et al., 2009). Mapped hits were filtered and assembled in the PATRIC platform (Davis et al., 2020) (accessed on 9/9/2020) by the built-in SPAdes *de novo* assembly algorithm (Bankevich et al., 2012).

Bioinformatics analysis. All genomes used in multiple sequence alignments or phylogenetic analysis were obtained from GISAID or NCBI Virus portals (Brister et al., 2015). Only genomes labeled as 'complete' were used. The data set was stringently filtered for ambiguous sequences (no more than 3 Ns) and for redundancy (all duplicate/100% identical genomes were filtered out). The stand-alone MUSCLE program (Edgar, 2004a, b) was used for multiple sequence alignment of the full nucleotide sequence of the filtered SARS-CoV-2 genomes. FastTree (Price et al., 2009, 2010) was used to compute phylogenetic distances, with a general-time reversible (GTR) substitution model, and FigTree (v.1.4.4) was used for the tree visualization (Rambaut, 2018). Other phylogenetic and phylogeographic comparisons were implemented from built-in tools in the GISAID (GISAID), NextStrain (Hadfield et al., 2018), and CoV-Glue (Singer et al., 2020) sites. CoVsurver, an application available at the GISAID website, was used to map spike amino acid variations to the 3D structure of a spike protein trimer.

Data availability. The CU_S3 sequence was deposited in both GISAID and NCBI Virus databases and was given the accession IDs **EPI_ISL_529032** and **MT990450**, respectively.

Results

The MiSeq high-throughput sequencing of the rRNA-depleted shotgun RNA library, obtained from the patient's nasopharyngeal swab extract, generated over 5 million high-quality RNA sequence reads. Out of 15,000 reads with positive BlastN hits (>90% identity) to SARS-CoV-2, the complete viral genome of SARS-CoV-2 was successfully assembled into a single gapless contiguous sequence with no ambiguous bases. The coverage of any given genomic position ranged from 15-50x.

Brief genome description. We named the assembled genome CU002b-S3 (or CU_S3 for short).Its sequence has 29,792 bases with typical features of SARS-CoV-2 genomes (Dabravolski and Kavalionak, 2020; Li et al., 2020) and includes all its open reading frames (ORFs), ORF1a through ORF10, without any disruptive mutations. The genome has been classified as part of *clade GR* (according to GISAID classification); of the pangolin *lineage B.1.1.1* (Rambaut et al., 2020); or of the *major clade 20B* (according to the new NextStrain classification (Hadfield et al., 2018))—formerly known as *clade A2a*.

Unique features of CU002b-S3. The genome of CU002b-S3 (CU_S3) has a mutation, previously unseen in Egypt and Africa, that caused the substitution of a highly conserved serine residue (at position 813 of the spike protein) by isoleucine (Fig. 1A)—a variation only reported in 24 out of 157,853 genomes (http://cov-glue.cvr.gla.ac.uk/#/project/replacement/S:S:813:I (Singer et al., 2020)), i.e., ~ 0.015% frequency.

Other than this unique variation, the spike protein in the CU_S3 genome carries three other variations from the 'reference' Wuhan isolate, the most popular of which is the D614G variation that is becoming prevalent worldwide (~ 88% frequency), and two more substitutions at positions 12 and 677 of the spike protein (S12F and Q677H, Fig. 1B and C) with frequencies of ~ 0.11% and ~ 0.19%, respectively . Another rather rare variation is an A-to-V substitution in residue 859 of ORF1a (amino acid 41 in NSP3), the one that encodes the replicase, and that has been seen in only 45/ 157,853 (~0.03%) global isolates (http://cov-glue.cvr.gla.ac.uk/#/project/replacement/NSP3:A:41:V), including one from Malaysia and one from Italy (Fig. 1D and 2A), and only one from Africa/Egypt (isolate CUNCI-HGC4I031).

To date, none of the 197,265 publicly available SARS-CoV-2 genome sequences has the unique combination of mutations observed in CU_S3. Specifically, the combination of S: S813I and ORF1a A858V variations encoded by this genome has no precedent in any public genomic sequence.

Phylogeny and phylogeography. Although genomic sequences reported to date remain highly closely related, we used multiple sequence alignment followed by phylogenetic analysis, at the nucleotide level, to analyze CU_S3 and trace its microevolution, in context of local and global genomes.

To visualize the phylogeny CU_S3, we compared it to selected genomes from all around the globe representing different major clades (Fig. 2A) and to all non-redundant genomic sequences from Egypt

that were available until October 1st (Fig. 2B). As of that date, the most closely related local isolate was one from the same city and hospital (GUNCI-HGC4I031) and the most closely related global isolate was one from Malaysia (Fig. 2).

Discussion

We here report the successful metagenome-based assembly of a single gapless contiguous sequence representing the complete viral genome of a SARS-CoV-2 isolate from an Egyptian patient. The genome, named CU002b-S3 (CU_S3 for short), was analyzed for mutations that resulted in amino acid variations, and was phylogenetically compared to representative local and global genomes.

In spite of the millions of cases worldwide and the tens of thousands of sequenced genomes, SARS-CoV-2 genome is relatively stable with an estimated mutation rate of ~24 base per genome per year (GISAID); thus phylogenetic analyses do not lead to strongly distinct clusters, except for the major clades (Rambaut et al., 2020) that have been defined so far in several studies (reviewed in Dabravolski and Kavalionak (2020) and monitored by NextStrain (Hadfield et al., 2018)).

What we find intriguing and worth sharing, while we continue to monitor isolates from Egyptian patients, is that the specific isolate reported here (CU_S3) encodes a rare S813I variation, not seen in Africa. Although the biological significance of this variation is yet to be explored, and although serine and isoleucine are not dramatically different, they still are physicochemically distinct, as isoleucine is larger in size and less polar than serine.

Of interest, the rare cases in which this S813I variation has been reported (24 cases) are mostly scattered and polyphyletic (belonging to different clades, with no evidence of phylogenetic contiguity). Such observation suggests that this variation results from spontaneous mutations that may occur independently and repeatedly (akin to convergent evolution, but with no particular selective pressure favoring the event—yet).

Additionally, while all 24 genomes with reported nonsynonymous mutations leading to the spike S813I variation also have the D614G variation, CU_S3 is the only genome sequence, in all public databases, to have both mutations causing S: S813I and ORF1a: A859V (NSP3: A41V) variations.

Numerous reports focused on the spike D614G variation (Díez-Fuertes et al., 2020; Korber et al., 2020; Plante et al., 2020; Zhou et al., 2020). Although it is not the most frequent variation, it is quite prevalent, and its occurrence in the spike protein made it an attractive target for speculations about an effect on viral transmissibility, infection efficiency, virulence, immunogenicity, immune evasion, and even disease prognosis (Grubaugh et al., 2020; Hou et al., 2020; Korber et al., 2020; Zhou et al., 2020). Yet, the only solid experimental evidence supports a slightly higher transmissibility but no significant structural effects leading to virulence or immune evasion (Díez-Fuertes et al., 2020; Grubaugh et al., 2020; Korber et al., 2020; Plante et al., 2020). Quite interestingly, D614G, Q677H, and S813I, three of the four variations in the CU_S3 isolate, occur all in the middle of the spike protein, away from its receptor-binding domain (which

is logically the most impactful on infectivity, drug susceptibility, and neutralizing antibody binding, Fig. 3). However, suggestions that these variations may affect the protein flexibility or stability are valid and remain to be structurally confirmed by X-ray crystallography or cryogenic electron microscopy. As for the S12F variation, it doesn't show in the available protein structure (Fig. 3) as it is in the N-terminal region.

In conclusion, using a direct, amplification-free RNA metagenome/metatranscriptome sequencing approach, we fully assembled and identified a novel variant of SARS-CoV-2, with a novel non-synonymous mutation (causing a rare spike S-813-I variation) and two less frequent ones (NSP3:A41V) and (S:Q677H). Both affected proteins have significance in vaccine and drug development, respectively. To the best of our knowledge, no other sequenced genomes combine mutations that lead to S: S813I and ORF1a: A859V (NSP3: A41V) variations.

Future studies are needed, in parallel with relentless genomic surveillance programs, to assess the biological, pathogenic, and epidemiologic implications of this set of genetic variations. Whether such variant is going to expand and be seen further depends on an increase of genomic sampling as well as potential fitness advantages conferred by the observed amino acid substitutions.

Declarations

Acknowledgments:

The authors are grateful to Cairo University (CU) President, Prof. Mohamed Osman Elkhosht for establishing the COVID-19 CU fund, and for launching and supporting the anti-COVID-19 task force. We thank the CU anti-COVID-19 task force, led by Prof. Omneya Khalil (Dean, Faculty of Pharmacy Cairo University) and Prof. Neveen Soliman (Vice-Dean, Faculty of Medicine, Cairo University).

Funding:

The work was fully funded by Cairo University, as part of the COVID-19 fund, established since April 2020. The funder has no interference in the scientific content of this letter.

Conflict of interest statement:

None of the authors have any personal or financial conflicts of interests regarding this work.

References

Bankevich A, Nurk S, Antipov D, et al. (2012). SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. J Comput Biol 19, 455-477.

Brister JR, Ako-Adjei D, Bao Y, Blinkova O (2015). NCBI viral genomes resource. Nucleic Acids Res 43, D571-577.

Camacho C, Coulouris G, Avagyan V, et al. (2009). BLAST+: architecture and applications. BMC Bioinformatics 10, 421.

Dabravolski SA, Kavalionak YK (2020). SARS-CoV-2: Structural diversity, phylogeny, and potential animal host identification of spike glycoprotein. J Med Virol.

Davis JJ, Wattam AR, Aziz RK, et al. (2020). The PATRIC Bioinformatics Resource Center: expanding data and analysis capabilities. Nucleic Acids Res 48, D606-d612.

de Wit E, van Doremalen N, Falzarano D, Munster VJ (2016). SARS and MERS: recent insights into emerging coronaviruses. Nat Rev Microbiol 14, 523-534.

Díez-Fuertes F, Iglesias-Caballero M, García Pérez J, et al. (2020). A founder effect led early SARS-COV-2 transmission in Spain. J Virol.

Edgar RC (2004a). MUSCLE: a multiple sequence alignment method with reduced time and space complexity. BMC Bioinformatics 5, 113.

Edgar RC (2004b). MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res 32, 1792-1797.

GISAID (https://www.gisaid.org/), Last accessed 15 October 2020.

Grubaugh ND, Hanage WP, Rasmussen AL (2020). Making sense of mutation: What D614G means for the COVID-19 pandemic remains unclear. Cell 182, 794-795.

Hadfield J, Megill C, Bell SM, et al. (2018). Nextstrain: real-time tracking of pathogen evolution. Bioinformatics 34, 4121-4123.

Hou YJ, Chiba S, Halfmann P, et al. (2020). SARS-CoV-2 D614G variant exhibits enhanced replication *ex vivo* and earlier transmission *in vivo*. bioRxiv.

Korber B, Fischer WM, Gnanakaran S, et al. (2020). Tracking changes in SARS-CoV-2 spike: Evidence that D614G increases infectivity of the COVID-19 virus. Cell 182, 812-827.e819.

Li H, Liu SM, Yu XH, Tang SL, Tang CK (2020). Coronavirus disease 2019 (COVID-19): current status and future perspectives. Int J Antimicrob Agents 55, 105951.

Plante JA, Liu Y, Liu J, et al. (2020). Spike mutation D614G alters SARS-CoV-2 fitness. Nature.

Price MN, Dehal PS, Arkin AP (2009). FastTree: computing large minimum evolution trees with profiles instead of a distance matrix. Mol Biol Evol 26, 1641-1650.

Price MN, Dehal PS, Arkin AP (2010). FastTree 2–approximately maximum-likelihood trees for large alignments. PLoS One 5, e9490.

Rambaut A (2018). FigTree v.1.4.4. http://tree.bio.ed.ac.uk/software/figtree/, Last accessed 11 Nov. 2020.

Rambaut A, Holmes EC, O'Toole Á, et al. (2020). A dynamic nomenclature proposal for SARS-CoV-2 lineages to assist genomic epidemiology. Nat Microbiol.

Singer J, Gifford R, Cotten M, Robertson D (2020). CoV-GLUE: A Web Application for Tracking SARS-CoV-2 Genomic Variation. Preprints 2020.

Zhou B, Thao TTN, Hoffmann D, et al. (2020). SARS-CoV-2 spike D614G variant confers enhanced replication and transmissibility. bioRxiv.

Figures

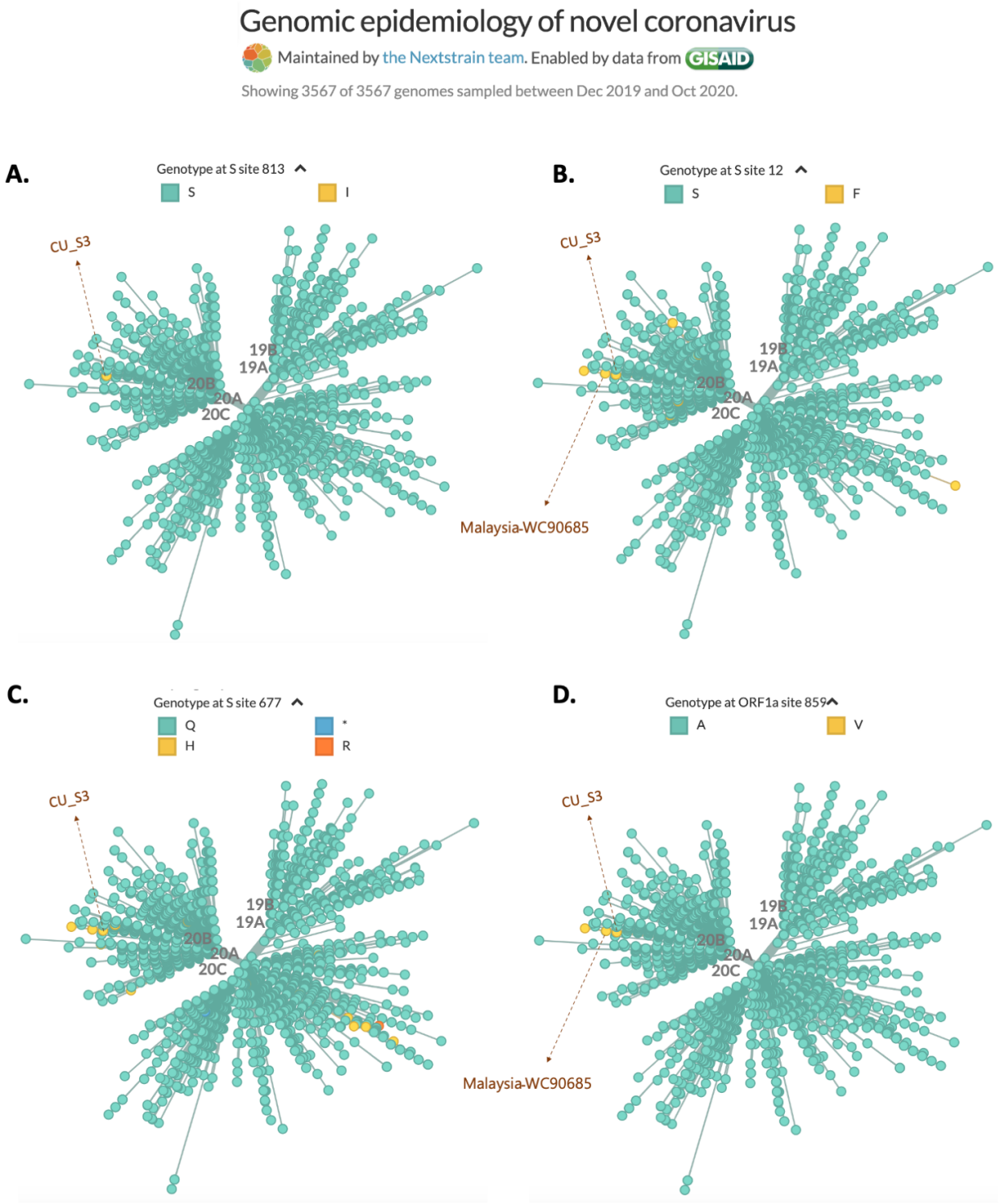

Figure 1

Phylogenetic analysis of 3,567 representative genomes of the global SARS-CoV-2 surveillance, generated by the NextStrain algorithm (Hadfield et al., 2018) using the GISAID dataset (GISAID). The trees indicate the uniqueness of the genome of CU002b-S3 isolate in being the only one among all representative genomes with the rare S-813-I amino acid substitution (A). Three other relatively rare variations are

shown: B. S: S12F, C. S: Q677H; D. ORF1a A859V (NSP3: A41V). The tree screenshots were captured on October 15, 2020.

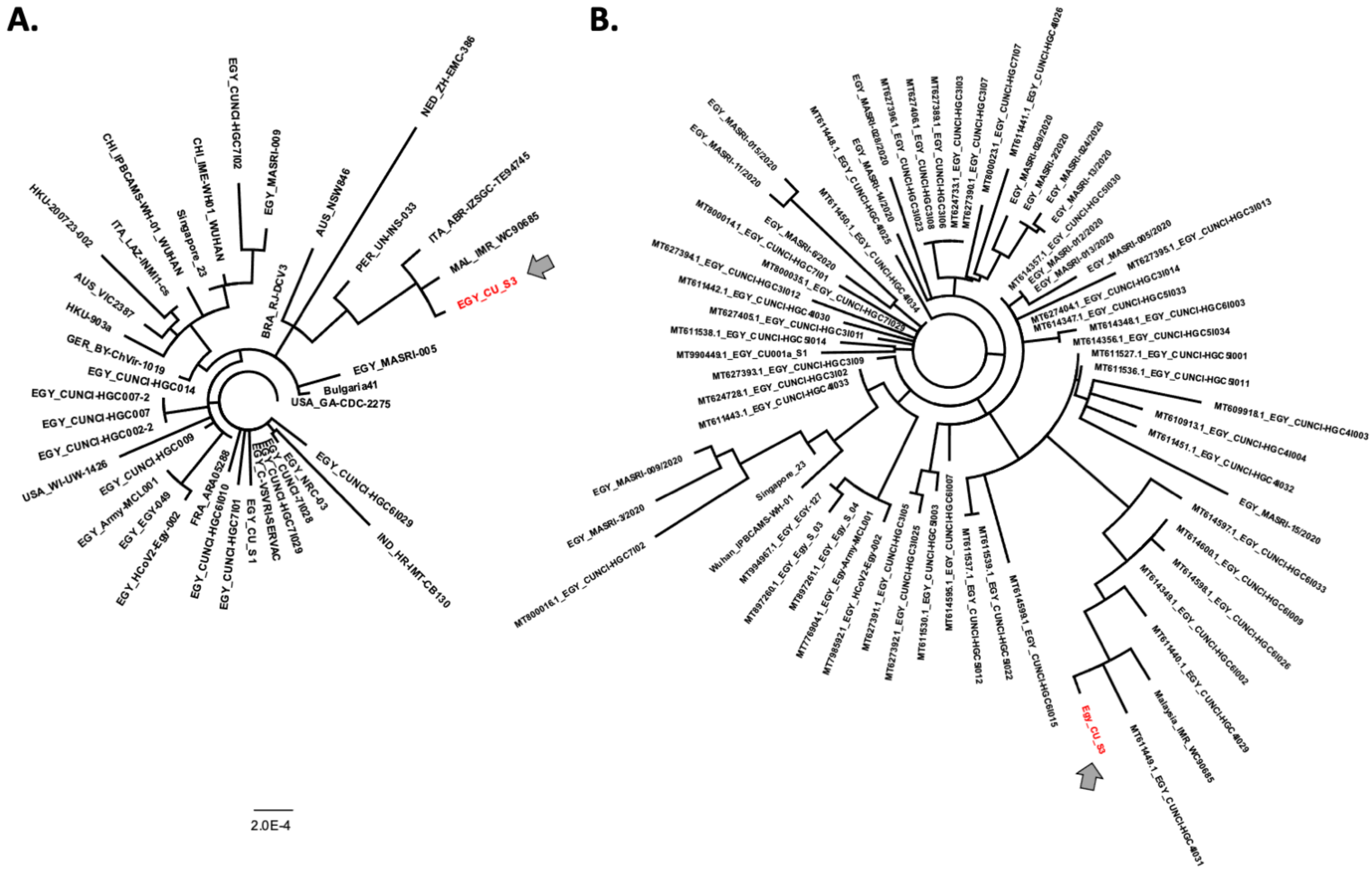

Figure 2

Maximum likelihood phylogenetic analysis of CU_S3 genome (ID on tree: EGY_CU_S3 or full ID: hCoV-19/Egypt/C-CU002b-S3/2020) in context of (A) selected representative members of different clades distributed around the globe and (B) 60 non-redundant genome sequences from Egypt as well as three global reference genomes. CU_S3 is marked by an arrow in both trees.

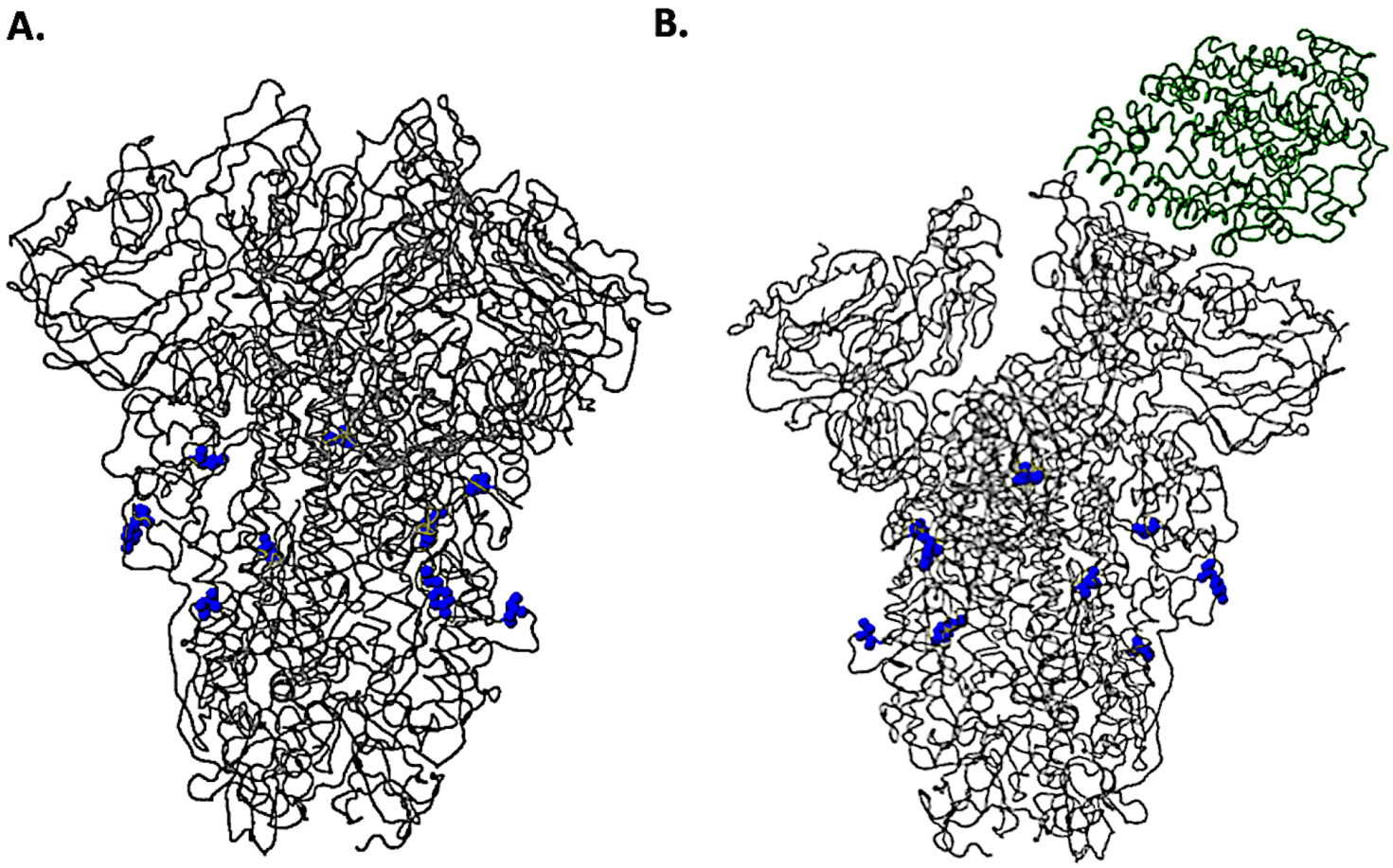

Figure 3

CoVsurver analysis results of substitutions in the spike protein of CU_S3. Three of the four reported spike variations in this work (D614G, Q677H, and S813I) are mapped to the 3-D structure of: A. the spike glycoprotein trimer (PDB: 6acc, EM 3.6 Angstrom) with RBD in down conformation and B. the spike glycoprotein trimer (PDB: 6acj, EM 4.2 Angstrom) in complex with host cell receptor ACE2 (green ribbon).