 doi:10.20944/preprints202003.0320.v1

Molecular analyses of over hundred sixty clinical isolates of SARS-CoV-2: Insights on likely origin, evolution and spread, and possible intervention

Salma Jamal, PhD[1#], Jasdeep Singh, PhD[1#], Javaid Ahmad Sheikh, PhD[2#], Hina Singh, PhD[1#], Mohd. Khubaib, PhD[1#], Sunil Kohli, MD[3], Ulrich Dobrindt, PhD[4], Syed Asad Rahman, PhD[5*], Nasreen Zafar Ehtesham, PhD[6], Seyed Ehtesham Hasnain, PhD[1*]

[1] JH-Institute of Molecular Medicine, Jamia Hamdard, New Delhi, India

[2] Department of Biotechnology, Jamia Hamdard, New Delhi, India

[3] Hamdard Institute of Medical Sciences & Research, Jamia Hamdard, New Delhi, India

[4] Institute of Hygiene, Mendelstrasse 7, 48149 Münster, Germany

[5] BioInception Pvt. Ltd, Swift House Ground Floor, 18 Hoffmanns Way, Chelmsford, Essex CM1 1GU, United Kingdom

[6] ICMR-National Institute of Pathology, Safdarjung Hospital Campus, New Delhi, India

# These authors contributed equally.

*Corresponding Author

Email:

SEH: seyedhasnain@gmail.com, vc@jamiahamdard.ac.in

SAR: asad.rahman@bioinceptionlabs.com

Keywords

Envelope protein, Furin cleavage site, Hydrophobic aggregation and rapid transmission, Immediate ancestor, Pangolins coronaviruses, SARS-CoV-2, Spike protein

 doi:10.20944/preprints202003.0320.v1

Abstract

We are witnessing the severe third outbreak mediated by coronaviruses affecting global public health with unprecedented economic consequences. A better understanding of its phylogenetics, exploration of sequence features and mutational changes could unveil its genealogy to gain insights into the mechanism of transmission and development of possible interventions. Our comparative genomic analyses of >160 isolates of SARS-CoV-2 reveal phylogenetic kinship with other coronaviruses and emergence of evolutionary divergence in clinical isolates. t-SNE-based clustering revealed different clades but no continent specific clusters. Amino acid substitutions at RBD of spike protein provide possible reasons for rapid transmission. Few proteins specific to SARS-CoV-2 were identified which could have implications as therapeutic targets and diagnostic biomarkers. Virtual screening identified repurposed drugs, known nutraceuticals, for specific interventions. These phylogenetic observations reveal the ancestry and computational studies reveal the emergency measures to interject this emerging pathogen that pose threat to whole of mankind.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 23 March 2020 doi:10.20944/preprints202003.0320.v1

Introduction

We are currently going through a pandemic caused by a novel coronavirus causing severe respiratory illness and pneumonia-like disease, renamed as COVID-19 [1,2]. The etiological agent, SARS-CoV-2 (Severe Acute Respiratory Syndrome-CoronaVirus-2), was presumably so termed due to phylogenetic similarity to SARS-CoV, the causative agent of SARS that caused an epidemic in Guangdong, China in 2003 [3,4]. More than 1.9 lakh infections and 7804 fatalities (as of March 18, 2020) have been reported in >160 countries, within a short span since initial reports [5]. The human to human transmission is rapidly disseminating due to higher population density in some regions of China, the worst affected area [6]. China has a great deal of transport links to major cities around the world that led to various other outbreak epicentres, further complicating the transmission scenario [7]. This third coronavirus outbreak in the last two decades is more transmissible than SARS-CoV [8] and MERS-CoV [9] (Middle East Respiratory Syndrome Coronavirus epidemic in 2012, Saudi Arabia) however, fatal only in patients with underlying illness [10,11]. These characteristics make it a perfect pathogen for a looming pandemic that could have serious public health and economic implications.

Coronaviruses infections have a wide range of symptoms from a mild cold to severe ARDS (Acute Respiratory Distress Syndrome). The uncontrolled virus replication eventually leads to cytokine storm and severe inflammatory disease of lungs that then spread across to other organs. Though the mechanism of this selective severity is unknown but genetic and environmental factors do play a role with data now suggesting gender disparity in fatalities [12].

The latest reports of disease outbreak with significant fatalities in many other countries, Italy, South Korea, Iran, France and USA being the latest, is likely to have global health consequences [13]. This burgeoning crisis globally suggests that the focus now need to be on reducing the outbreaks as containment may not necessarily be a useful measure. The real test will be the introduction of systems to control and reduce the speed of transmission in countries with poor health care infrastructure [14]. Measures like travel bans, suspending public transport and banning public gathering, though effective in China would unlikely affect other countries [15] given the latest evidence of low risk of transmission of this virus from China to Africa and South America [16].

The in-depth molecular phylogenetic analyses is expected to reconstruct the evolutionary history of the pathogen. Along with epidemiological data, these analyses could unveil the event(s) of zoonotic transmission and the evolutionary changes thereafter during human to

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 23 March 2020 doi:10.20944/preprints202003.0320.v1

human transmission. This information could be exploited to prevent future inter-species spill over of viruses as well as unveil plethora of targets to develop interventions and diagnostics.

Methodology

Phylogenetic Analysis

Genomes of clinical isolates for available coronavirus sequences (n=220) were retrieved from GISAID (https://www.gisaid.org) database as on 6th Mar 2020. These were manually curated to remove redundancy (100% identical sequences were removed using mmseq2 software [17]). Non-redundant (n=167) were retained for analysis. Genomes of other coronaviruses (Bat, SARS-CoV and MERS-CoV) were obtained from NCBI viral genome database (https://www.ncbi.nlm.nih.gov/). Whole-genome multiple sequence alignments (MSA) of all coronavirus genomes were performed by using MUSCLE [18] and MAFFT [19] software as part of the in-house pipeline at BioInception and EMBL-EBI. Phylogenetic analysis of all clinical isolates and other coronavirus genomes was carried out using FigTree (http://tree.bio.ed.ac.uk/software/figtree/) and iTOL (https://itol.embl.de) [20] software based on the neighbour-joining method. To find conserved and variable regions in SARS-CoV-2 CDS, protein similarity (cut-off = 0.3) based clustering was done among all samples using in-house protocols. Coding ORF's were predicted by Glimmer [21] and mmseq2 [17] was used to perform homology search. These similarity and identity scores were used to generate a similarity matrix for clustering. The unique proteins/signatures were identified on the basis of binary matrix projected using cluster analysis employing R (https://cran.r-project.org/) software (hclust library, ward.D2 clustering method). The protein sequences of the Spike protein of SARS-COV-2 and RaTG13 were retrieved from the NCBI database and their 3D structures were generated using SWISS MODEL [22]. The crystal structure of host Angiotensin converting enzyme 2 (ACE2) receptor was downloaded from Protein Data Bank (PDB). Tertiary structural alignment of models was done using PyMol. Mutations among the clinical isolates and w.r.t. RaTG13 were mapped using sequence alignment of spike proteins using SeaView [23].

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 23 March 2020 doi:10.20944/preprints202003.0320.v1

Modelling Spike-ACE2 interaction and drug repurposing

The data mediated flexible docking of Spike proteins (SARS-COV-2 and RaTG13) and host ACE2 receptor was executed using HADDOCK [24] web server. The docked protein-protein complexes (SARS-CoV-2+ACE2 and RaTG13+ACE2) were subjected to molecular dynamics (MD) simulations so as to explore the physical movement of their atoms with respect to time. The complexes were placed in a cubic box filled with simple point charge water model and Na^+ and Cl^- ions were included to neutralize the total charge of the system. Further energy minimization was performed on the solvated system followed by NVT and NPT equilibration and 10ns MD simulations were run using GROMACS [25].

In addition to this, using a library of FDA approved drugs, virtual screening was carried out against receptor binding domain (RBD) of spike protein of SARS-CoV-2 to identify probable drugs that can be repurposed against COVID-19. The crystal structure of SARS-CoV-2 RBD complexed with host ACE2 receptor was obtained from PDB (PDB ID: 6VW1). FDA approved library of drugs was retrieved from DrugBank [26] database. Prior to the screening, the protein and ligands were pre-processed followed by grid generation centred on the RBD-ACE2 interacting residues and molecular docking was performed using Schrodinger suite.

Transmission mechanism

To gain insights into possible transmission mechanisms of SARS-CoV-2 compared to SARS-CoV and MERS-CoV putatively mediated through spike surface protein, trimer models were simulated in water (0.1 M NaCl) for ~20 ns. Protein-protein docking of spike proteins of SARS-CoV, MERS-CoV and SARS-CoV-2 was done using ClusPro. Comparative changes in surface hydrophobicity and surface-water contacts of spike proteins were used as primary parameters to gain insights into transmission mechanisms. Data were plotted using Origin V8.5.

Amino acid conservation analyses

In-house pipeline was used for mapping conserved amino acid residues in envelope proteins and spike proteins of all bat coronaviruses and COVID-19 clinical isolates by using multiple sequence alignment files. Two conservation score schemes, KABAT[27] and VALDAR[28] were used (https://www.compbio.dundee.ac.uk/aacon/), the former calculates conservation based on antibody response whereas the latter extracts information from the residue environment.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 23 March 2020 doi:10.20944/preprints202003.0320.v1

These were used to construct the relative evolutionary conservation for each site in the envelope protein and spike protein, ranging from 0 to 1, where 1 is total conservation and 0 is no conservation.

Results

Phylogenetic Analysis of SARS-CoV-2 Clinical Isolates

Phylogenetic analysis by various groups initially revealed that this novel clade has the highest similarity (82.3%) to already published SARS-CoV [29]. Our phylogenetic analyses involving 167 clinical isolates obtained globally, corroborated by others, depict that SARS-CoV-2 has much higher similarity (96.1%) with the bat coronavirus genome RaTG13. The sequence of bat RaTG13 virus isolated by the Wuhan Institute of Virology back in 2013 from a faecal swab of *Rhinolophus affinis* (reported in [2]), appeared in the public domain [2] after the COVID-19 epidemic. SARS-CoV-2 remarkably is evolutionarily distant from MERS and other known bat coronaviruses hence unlikely to have originated from them. Genomes of all these clinical isolates are highly similar (>99 %) and fall in the same clade, with the exception of the France and Finland, isolate which was distinct from both RaTG13 and other SARS-CoV-2 clinical isolates, an observation which is difficult to explain (possibly a sequencing error) in the absence of any epidemiological information (Figure S1). t-SNE based clustering [30] again shows that RATG13 and SARS-CoV-2 are very close to each other when compared to other coronavirus genomes (Figure S2) as observed in dendrogram based clustering. Further phylogenetic analyses of protein sequences (functional similarity) also are in line with the above observations (Figure 1A). The t-SNE based analysis revealed interesting transmission dynamics as there was no specific cluster in any of the continents which could be correlated to demographic features (Figure 1B). Continent based clades showed diverse phenotype possibly related to rapid transmission possibly via intercontinental travel in this globalised era. Moreover, molecular phylogenetics also revealed two separate clades in clinical isolates that denotes the emergence of evolutionary diversity. It is clear from the strain diversity between COVID-19 genomes that the virus is evolving as we see two distinct clades (Figure 1C). This would mean that there are specific mutations in the ORF regions which might impact the function of proteins and virulence of these strains - something to consider for interventions when it comes to diagnostic biomarkers, vaccines and drugs.

Comparative genomics of all the clinical isolates of SARS-CoV-2, reference coronavirus genomes and RaTG13 genome revealed conservation of ORF1ab, Spike (S), ORF3a, ORF5, ORF7a and ORF8 in all coronaviruses. At the same time, ORF4, ORF6, and ORF7b were found to be specific to SARS-CoV-2 clinical isolates, RaTG13, SARS-CoV and very few other bat and pangolin coronaviruses (Figure 2, figure S3 and table S1). In addition to clinical isolates in each cluster, ORF 4 was present in SARS-CoV and bat coronaviruses; ORF 6 was present SARS-CoV and pangolin coronaviruses, and ORF 7b was present in RaTG13 and pangolin coronaviruses.

Role of amino acid substitutions in spike protein in enhancing SARS-CoV-2 infectivity

MSA of genomes reveals an active Furin cleavage site 'PRRAR' in SARS-CoV-2 isolates that were absent in RaTG13 genome despite a high degree of identity among amino acids on both the sides of this insertion (Figure S4). Considering the evolutionary proximity to RaTG13 and a high degree of sequence similarity in clinical isolates of SARS-CoV-2, exploring this gain of function could reveal the mechanism of inter-species transmission crossover. Taking clues from SARS-CoV and MERS-CoV, it can be safely assumed that inheritance of this gene segment will enhance the infectivity, as this spike glycoprotein would be possibly cleaved during virus entry/exit and thereby infect neighbouring cells [31]. Apart from this Furin cleavage site, we observed a mutational cluster at the receptor binding site of spike protein in SARS-CoV-2 (Figure 3). MD simulations revealed that the substituted amino acids enhanced the interaction with ACE2 receptor thus impacting pathogenicity, as also confirmed now by surface plasmon resonance (SPR) data [32] (Figure S5, table S2 and S3). Insights from MD simulations corroborate that SARS-CoV-2 spike proteins have higher propensity to form a hydrophobic collapse than MERS-CoV and SARS-CoV [33] (Figure S6).

Repurposed Drugs as an Emergency Measure

Virtual screening using FDA approved drug library targeting the RBD domain of SARS-CoV-2 spike protein led to the identification of NADH (DB00157) and Rutin (DB01698) that can be likely inhibitors against SARS-CoV-2 (Table S4). NADH formed hydrogen bond interactions with Asp30 and Ala387 of ACE2 receptor and Lys403, Arg408, Gln409, Gly496 and Tyr505 of SARS-CoV-2 RBD. The hydrophobic network forming residues include Asn33, His34, Glu37, Lys353, Pro389 of ACE2 and Asp405 and Tyr495 of RBD.

Rutin formed hydrogen bonds with Asn33, His34 and Gln388 of ACE2 and Lys403, Asp405, Gln409 of RBD of SARS-CoV-2. Asp30, Glu37, Ala387, Pro389, Phe390, Arg393 of ACE2

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 23 March 2020 doi:10.20944/preprints202003.0320.v1

and Asp406, Arg408, Val417, Ile41, Tyr453 spike RBD formed hydrophobic interactions. The binding of drugs with the residues which have been identified as key residues for interaction with ACE2, point towards the strong inhibitory potential of these FDA approved drugs (Figure 4).

Amino Acid Conservation Profiling

In light of the functional importance of envelope and spike protein, we performed an amino acid conservation analysis of all the bat coronaviruses and SARS-CoV-2 using KABAT [27] (based on antibody response) and VALDAR [28] (based on the properties of neighbouring amino acid residues) scoring schemes. Amino acid conservation analyses of envelope protein, specific to SARS-CoV-2, reveal that in comparison to other Bat coronavirus genomes, envelope sequences are highly conserved in SARS-CoV-2 (Figure S7 A). We see two specific mutations in the envelope proteins of the clinical isolates in BetaCoV_Canada (S6L) and BetaCoV_SouthKorea (L37H) (Figure S7 B).

A variable N-terminal domain and a comparatively conserved C-terminal domain is in line with our earlier observation that there are significant amino acid substitutions in N-terminal of spike protein and more specifically in the region 350-650 which encompasses RBD (Figure S7 C). Comparison of a representative clinical isolate (Wuhan) with other bat coronaviruses reveals thorough low-grade loss of conservation in amino acids, more so in the N-terminal region with a sharp spike around 680 residue pointing to Furin cleavage insertion/deletion (Figure 5A). Some recent reports have suggested pangolins to be intermediate hosts, however, our analysis reveals that there is a sharp distinction at several critical amino acids (Furin cleavage site; 681-684) refuting the possibility of pangolin coronaviruses as immediate ancestors of SARS-CoV-2 (Figure 5B). A highly conserved SYLTPGDSSS stretch present only in pangolins, SARS-CoV-2 and RaTG13 and absent in all bat coronaviruses was observed. This stretch forms a conformation cluster at solvent exposed disordered region in the N-terminal domain of spike proteins, previously shown to confer extended host range for murine coronaviruses [34] (Figure S8). Interestingly, on comparing all clinical isolates with RaTG13, the loss of conservation is consistent at specific positions, notably in the RBD (Figure 5C). The sequence similarity among clinical isolates is evident from the very few substitutions observed in the comparative profile of all clinical isolates (Figure 5D, figure S7 D).

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 23 March 2020 doi:10.20944/preprints202003.0320.v1

Discussions

MSA based phylogenetic analyses of clinical isolates (n=167) from COVID-19 patients unravel a typical betacoronavirus like genomic organisation. Phylogenetic divergence from SARS-CoV and MERS-CoV suggests that it is unlikely that SARS-CoV-2 has emerged from any of them. Though, bat coronavirus RaTG13 emerges as potential immediate ancestor owing to its genetic similarity, it seems unlikely due to significant amino acid substitutions that exceeds the divergence expected from an immediate ancestor during the course of natural evolution. Moreover, it is well known that amino acid conservation is an indicator of a sequence being preserved by natural selection. The uniform high-grade loss of conservation at specific amino acids depicts that RaTG13 is unlikely an immediate predecessor of SARS-CoV-2. These amino acid conservation profiles reveal interesting distinctions between evolutionarily conserved strains that need to be further explored. Further, the extraordinary high sequence similarity in ALL (n=167) the clinical isolates is also suggestive of its origin from a single incident. This observation is corroborated by the epidemiological demographics of COVID-19 patients [35,36] and transmission chain thereof by familial clusters [6] or people that have been termed as super spreaders. Moreover, the continued human to human transmission now depicts two distinct types circulating in the population [37]. The ancestral 'S' type seems to be less virulent as compared to alternate 'L' type. The selection pressure is expected to limit the spread of 'L' type, even though it is currently the prevalent form present in COVID-19 patients. The emerging genomic data from clinical isolates along with symptomatic data is expected to reveal the evolutionary track of SARS-CoV-2 and the possible implications of these SNP's in disease pathogenesis.

Amino acid level clustering analyses revealed ORFs, 4, 6 and 7b of SARS-CoV-2 have diverged from other bat coronaviruses as compared to other ORFs that are conserved throughout beta coronavirus family. This unanticipated observation points to likely functional importance of these ORFs in the pathogenesis of SARS-CoV-2. Interestingly indeed, the ORF4 turned out to be very unique in SARS-CoV-2 clinical isolates, with all isolates clustering together along with SARS-CoV and RaTG13. This specificity of SARS-CoV-2 ORF4 could have profound implications in the pathogenesis of COVID-19 as ORF4 corresponds to envelope protein of the coronaviruses and is an indispensable structural protein. This protein, though scarce in virions, is abundantly expressed in ER and Golgi compartments of infected cells where it participates in virus assembly and intracellular

trafficking [38,39]. It consists of a single α-helical transmembrane domain that forms pentameric ion channels important for virus-host interaction [40]. Removal of this protein is deleterious to coronaviruses as SARS-CoV and MERS-CoV deficient in this protein are attenuated [41,42] and being tested as potential vaccines[43]. This protein has an established role in inflammasome activation, which translates into the observed cytokine storm in diseased individuals and thus the causal factor for lung pathology and oedema [44]. Interestingly, the envelope protein of SARS-CoV has a PDZ-binding motif (PBM). PDZ domains being a common structural domain of signal transduction complexes, PBM may allow them to bind over 400 host proteins and thus modulate a range of cellular processes. The envelope protein of infectious bronchitis coronavirus has been reported to play a role in the regulation of ER stress through ion channel activity, modulate release of viruses, induce apoptosis and play role in pathogenesis [45]. It also neutralizes pH of Golgi which in turn modulates secretory pathway of host and helps in evading premature cleavage of Spike protein and thus helps in the efficient release of infectious viruses [46 47]. These important pathological functions of envelope protein make it an important drug target and the gain or loss by the altered sequence in SARS-CoV-2 would be instrumental in exploring the viral fitness of this pathogen. ORF6 homologs are involved in virulence by increasing viral replication [48] and also impact immunity by inhibiting STAT1 induced activation of antiviral response [49]. ORF7b is a highly hydrophobic 43 amino acid protein which is homologous to an accessory but structural component of SARS-CoV virion [50]. ORF7b possesses a transmembrane helical domain (between 9-29AA), and its homologue has Golgi complex retention signal within this domain [51]. Though silencing of SARS-ORF 7a and 7b cumulatively has been correlated with reduced *in-vitro* progeny virus yield [52], *in-vivo* deletion studies have not shown significant variation in replication, host tropism, mortality or morbidity [50]. Feline-CoV ORF7b is maintained in natural strains but is readily lost upon cell culture adaptation of virus [53]. Effectively, the specific role of ORF7b in viral life-cycle and pathogenesis is undefined and warrants further investigation. Collectively, while these data majorly point to the critical roles of the ORFs 4, 6, and 7b in SARS-CoV-2 pathogenicity, further experimental evidence using COVID-19 patient samples will delineate the possible evolutionary significance of these specific ORFs. The data so far are also suggestive of immune system dysregulation in the shape of aberrant cytokine storm in COVID-19 patients that causes pathology and likely enhances the mortality [54]. Whether these clinical features can be attributed to these ORFs, 4, 6 and 7b remains to be seen with possible therapeutic and diagnostic implications.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 23 March 2020 doi:10.20944/preprints202003.0320.v1

An essential feature of coronaviruses is a trimeric transmembrane protein (spike) that forms a trademark crown-like structure, which is critical for entry of the virus into host [55]. This protein contains two distinct domains, RBD and a fusion domain. The cryo-EM structure (3.5 Å-resolution) precisely depicts the pre-fusion and post-fusion conformations of RBD [32]. These domains need to be separated in order for the fusion domain to function for virus entry which is triggered by binding of RBD to host cell receptor ACE2. This binding leads to destabilisation of pre-fusion trimer and separation of domains followed by stabilisation of post-fusion confirmation [32,56].

Various host proteases cleave these sites in coronaviruses that regulate the infectivity and host tropism as typified by highly pathogenic avian influenza virus [57,58]. There is a variation in the cleavage sites in viruses associated with previous epidemics: MERS-CoV has two Furin cleavage sites and SARS-CoV has a replacement of these with cleavage site for alternative proteases [59]. This evolutionary adaptation is attributed to either selection pressure in intermediate hosts or random recombination events. These amino acid adaptations/substitutions also increase the hydrophobicity of the spike protein which is a major surface protein and thus likely to induce viral aggregation *via* hydrophobic collapse which may promote its aerial transmission and likely confer a fitness advantage [60]. These potential adaptations could be the reason for enhanced R_0 values (a measure of viral transmissibility indicating how contagious an infectious disease is) compared to SARS-CoV and MERS-CoV.

Being an important mediator at host pathogen interface, the structural and functional attributes of spike protein in coronaviruses can be exploited as a key target for developing possible interventions [61]. Apart from RBD of spike protein, ‘Furin cleavage site’, that seems to impart pathogenicity to SARS-CoV-2, can be possibly modulated. The functional significance of Furin cleavage site presence in SARS-CoV-2 can be assessed by reverse genetics approach. The mutant SARS-CoV-2 with knocked of Furin motif should have possibly reduced pathogenicity and altered host tropism. Apart from these molecular approaches, an exigent approach to curb the epidemic is a promising area of drug repurposing FDA approved drugs. Virtual screening revealed, known nutraceuticals, ‘NADH’ and ‘Rutin’ as potential inhibitors of RBD-ACE2 interaction, a fundamental step for cell entry. NADH is an oxidative cofactor involved in energy production in the cells and has been suggested to be used as a medication for chronic fatigue syndrome, Alzheimer's disease and Parkinson's. Rutin is a bioflavonoid with a multitude of pharmacological activities and an impeccable

safety profile. Considering the strong binding of these drug molecules to RBD-ACE2 interface, we propose them as potential therapeutic candidates that could be introduced into clinical testing along with other potential antivirals [62]. Apart from disrupting the interaction between RBD and ACE-2, the Furin site in SARS-CoV-2 could also act as a potential drug target, though with caution as these enzymes are involved in a multitude of functions [63-65].

Despite intense efforts, we are yet to comprehend the coronavirus epidemic and host transmission mechanism fully. The close evolutionary relationship with bat coronavirus RaTG13 and divergence from SARS-CoV and MERS-CoV need a more critical evaluation to better understand the coronavirus genealogy. Our in-depth phylogenetic and sequence diversity analyses of all coronaviruses will help design effective interventions that can be accelerated by involving AI & ML based approaches [66]. The use of PDZ domain as target for blocking SARS-CoV-2 induced cellular pathologies could be an interesting strategy for controlling post infection complications. The information about amino acid substitutions in RBD of spike protein of SARS-CoV-2 could be employed to design specific neutralizing antibodies aided by machine learning algorithms [67]. The COVID-19 epidemic is rapidly spreading which can be possibly explained by the increased hydrophobicity of spike protein of SARS-CoV-2. It is therefore tempting to hypothesize that the rapid transmission of SARS-CoV-2 may be difficult to control. The conservation profile of amino acids reveals a missing link in the ancestry and the identification of the ancestor of SARS-CoV-2 will also throw some light on human-animal-environment disease interface that drives cross-species viral transmission [68]. The apparent absence of recent large scale recombination events that could have otherwise explained the origin of SARS-CoV-2 seems very intriguing [69]. Efforts to identify the precursor will also put an end to 'infodemics' of rumours regarding origin of SARS-CoV-2. It is very important not to jeopardize international endeavours to control this epidemic but continue with concerted undertakings by medical practitioners, health care workers and researchers to contain this global threat [70,71]. As the death toll keeps on increasing, there is an urgent need for interventions and various host-directed therapies could be a rescue measure [54]. At the same time, there is a need to tone down the hype and keep hope alive [44].

Conflict of Interest: The authors declare no conflict of interest whatsoever. BioInception has contributed by making available their data analysis pipeline and platform for this research and has disclosed pre-competitive research data in the interest of Public Health. Anyone can

Author Contributions: SEH, SAR and NZE designed the study. SJ, JS and SAR carried out the computational analyses while JAS, SAR, HS, MK, SK, and UD collated, analysed and compiled the results. All authors reviewed the data and contributed in different ways in the writing of this manuscript. All Authors approved the final manuscript.

Acknowledgements

SEH and NZE thank Department of Biotechnology, Government of India. SEH is a JC Bose National Fellow, Department of Science and Technology, Government of India. NZE thanks Indian Council of Medical Research.

References

1. Wu F, Zhao S, Yu B, et al. A new coronavirus associated with human respiratory disease in China. *Nature* 2020.
2. Zhou P, Yang XL, Wang XG, et al. A pneumonia outbreak associated with a new coronavirus of probable bat origin. *Nature* 2020.
3. Enserink M. Update: 'A bit chaotic.' Christening of new coronavirus and its disease name create confusion2020. https://www.sciencemag.org/news/2020/02/bit-chaotic-christening-new-coronavirus-and-its-disease-name-create-confusion (accessed 25/02/2020).
4. Jiang S, Shi Z, Shu Y, et al. A distinct name is needed for the new coronavirus. *The Lancet* 2020.
5. Organization WH. Coronavirus disease 2019 (COVID-19), 2020.
6. Chan JF, Yuan S, Kok KH, et al. A familial cluster of pneumonia associated with the 2019 novel coronavirus indicating person-to-person transmission: a study of a family cluster. *Lancet* 2020; **395**(10223): 514-23.
7. Wu JT, Leung K, Leung GM. Nowcasting and forecasting the potential domestic and international spread of the 2019-nCoV outbreak originating in Wuhan, China: a modelling study. *Lancet* 2020; **395**(10225): 689-97.
8. Ksiazek TG, Erdman D, Goldsmith CS, et al. A novel coronavirus associated with severe acute respiratory syndrome. *N Engl J Med* 2003; **348**(20): 1953-66.
9. Zaki AM, van Boheemen S, Bestebroer TM, Osterhaus AD, Fouchier RA. Isolation of a novel coronavirus from a man with pneumonia in Saudi Arabia. *N Engl J Med* 2012; **367**(19): 1814-20.
10. Chen N, Zhou M, Dong X, et al. Epidemiological and clinical characteristics of 99 cases of 2019 novel coronavirus pneumonia in Wuhan, China: a descriptive study. *Lancet* 2020; **395**(10223): 507-13.
11. Wang D, Hu B, Hu C, et al. Clinical Characteristics of 138 Hospitalized Patients With 2019 Novel Coronavirus-Infected Pneumonia in Wuhan, China. *JAMA* 2020.
12. Zimmer K. Why Some COVID-19 Cases Are Worse than Others. The Scientist. 2020.
13. Munster VJ, Koopmans M, van Doremalen N, van Riel D, de Wit E. A Novel Coronavirus Emerging in China - Key Questions for Impact Assessment. *N Engl J Med* 2020; **382**(8): 692-4.
14. Gilbert M, Pullano G, Pinotti F, et al. Preparedness and vulnerability of African countries against importations of COVID-19: a modelling study. *Lancet* 2020.

15. Jon Cohen KK. The coronavirus seems unstoppable. What should the world do now? Science. 2020.
16. Haider N, Yavlinsky A, Simons D, et al. Passengers' destinations from China: low risk of Novel Coronavirus (2019-nCoV) transmission into Africa and South America. *Epidemiol Infect* 2020; **148**: e41.
17. Mirdita M, Steinegger M, Soding J. MMseqs2 desktop and local web server app for fast, interactive sequence searches. *Bioinformatics* 2019; **35**(16): 2856-8.
18. Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res* 2004; **32**(5): 1792-7.
19. 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* 2002; **30**(14): 3059-66.
20. Letunic I, Bork P. Interactive Tree Of Life (iTOL) v4: recent updates and new developments. *Nucleic Acids Res* 2019; **47**(W1): W256-W9.
21. Delcher AL, Harmon D, Kasif S, White O, Salzberg SL. Improved microbial gene identification with GLIMMER. *Nucleic Acids Res* 1999; **27**(23): 4636-41.
22. Schwede T, Kopp J, Guex N, Peitsch MC. SWISS-MODEL: An automated protein homology-modeling server. *Nucleic Acids Res* 2003; **31**(13): 3381-5.
23. Gouy M, Guindon S, Gascuel O. SeaView version 4: A multiplatform graphical user interface for sequence alignment and phylogenetic tree building. *Mol Biol Evol* 2010; **27**(2): 221-4.
24. Dominguez C, Boelens R, Bonvin AM. HADDOCK: a protein-protein docking approach based on biochemical or biophysical information. *J Am Chem Soc* 2003; **125**(7): 1731-7.
25. Van Der Spoel D, Lindahl E, Hess B, Groenhof G, Mark AE, Berendsen HJ. GROMACS: fast, flexible, and free. *J Comput Chem* 2005; **26**(16): 1701-18.
26. Wishart DS, Knox C, Guo AC, et al. DrugBank: a knowledgebase for drugs, drug actions and drug targets. *Nucleic Acids Res* 2008; **36**(Database issue): D901-6.
27. Kabat EA, Te Wu T, Bilofsky H, Resources NIoHDoR, Health NIo. Sequences of Immunoglobulin Chains: Tabulation and Analysis of Amino Acid Sequences of Precursors, V-regions, C-regions, J-chain and BP-microglobulins, 1979: Department of Health, Education, and Welfare, Public Health Service, National Institutes of Health; 1979.
28. Valdar WSJ, Thornton JM. Protein-protein interfaces: Analysis of amino acid conservation in homodimers. *Proteins: Structure, Function, and Genetics* 2001; **42**(1): 108-24.
29. Zhu N, Zhang D, Wang W, et al. A Novel Coronavirus from Patients with Pneumonia in China, 2019. *N Engl J Med* 2020; **382**(8): 727-33.
30. Maaten Lvd. Accelerating t-SNE using Tree-Based Algorithms. *Journal of Machine Learning Research* 2014; **15**: 3221-45.
31. Belouzard S, Chu VC, Whittaker GR. Activation of the SARS coronavirus spike protein via sequential proteolytic cleavage at two distinct sites. *Proc Natl Acad Sci U S A* 2009; **106**(14): 5871-6.
32. Wrapp D, Wang N, Corbett KS, et al. Cryo-EM structure of the 2019-nCoV spike in the prefusion conformation. *Science* 2020: eabb2507.
33. Andreu-Moreno I, Sanjuan R. Collective Viral Spread Mediated by Virion Aggregates Promotes the Evolution of Defective Interfering Particles. *mBio* 2020; **11**(1).
34. Schickli JH, Thackray LB, Sawicki SG, Holmes KV. The N-terminal region of the murine coronavirus spike glycoprotein is associated with the extended host range of viruses from persistently infected murine cells. *J Virol* 2004; **78**(17): 9073-83.
35. Li Q, Guan X, Wu P, et al. Early Transmission Dynamics in Wuhan, China, of Novel Coronavirus-Infected Pneumonia. *N Engl J Med* 2020.
36. Guan WJ, Ni ZY, Hu Y, et al. Clinical Characteristics of Coronavirus Disease 2019 in China. *N Engl J Med* 2020.
37. Tang X, Wu C, Li X, et al. On the origin and continuing evolution of SARS-CoV-2. *National Science Review* 2020.

38. Schoeman D, Fielding BC. Coronavirus envelope protein: current knowledge. *Virol J* 2019; **16**(1): 69.
39. Castano-Rodriguez C, Honrubia JM, Gutierrez-Alvarez J, et al. Role of Severe Acute Respiratory Syndrome Coronavirus Viroporins E, 3a, and 8a in Replication and Pathogenesis. *mBio* 2018; **9**(3): e02325-17.
40. Surya W, Li Y, Verdia-Baguena C, Aguilella VM, Torres J. MERS coronavirus envelope protein has a single transmembrane domain that forms pentameric ion channels. *Virus Res* 2015; **201**: 61-6.
41. DeDiego ML, Alvarez E, Almazan F, et al. A severe acute respiratory syndrome coronavirus that lacks the E gene is attenuated in vitro and in vivo. *J Virol* 2007; **81**(4): 1701-13.
42. Anthony SJ, Epstein JH, Murray KA, et al. A strategy to estimate unknown viral diversity in mammals. *mBio* 2013; **4**(5): e00598-13.
43. Fett C, DeDiego ML, Regla-Nava JA, Enjuanes L, Perlman S. Complete protection against severe acute respiratory syndrome coronavirus-mediated lethal respiratory disease in aged mice by immunization with a mouse-adapted virus lacking E protein. *J Virol* 2013; **87**(12): 6551-9.
44. Nieto-Torres JL, DeDiego ML, Verdia-Baguena C, et al. Severe acute respiratory syndrome coronavirus envelope protein ion channel activity promotes virus fitness and pathogenesis. *PLoS Pathog* 2014; **10**(5): e1004077.
45. Li S, Yuan L, Dai G, Chen RA, Liu DX, Fung TS. Regulation of the ER Stress Response by the Ion Channel Activity of the Infectious Bronchitis Coronavirus Envelope Protein Modulates Virion Release, Apoptosis, Viral Fitness, and Pathogenesis. *Front Microbiol* 2019; **10**: 3022.
46. Westerbeck JW, Machamer CE. The Infectious Bronchitis Coronavirus Envelope Protein Alters Golgi pH To Protect the Spike Protein and Promote the Release of Infectious Virus. *J Virol* 2019; **93**(11): e00015-19.
47. Stodola JK, Dubois G, Le Coupanec A, Desforges M, Talbot PJ. The OC43 human coronavirus envelope protein is critical for infectious virus production and propagation in neuronal cells and is a determinant of neurovirulence and CNS pathology. *Virology* 2018; **515**: 134-49.
48. Pewe L, Zhou H, Netland J, et al. A severe acute respiratory syndrome-associated coronavirus-specific protein enhances virulence of an attenuated murine coronavirus. *J Virol* 2005; **79**(17): 11335-42.
49. Cheng W, Chen S, Li R, Chen Y, Wang M, Guo D. Severe acute respiratory syndrome coronavirus protein 6 mediates ubiquitin-dependent proteosomal degradation of N-Myc (and STAT) interactor. *Virol Sin* 2015; **30**(2): 153-61.
50. Schaecher SR, Mackenzie JM, Pekosz A. The ORF7b protein of severe acute respiratory syndrome coronavirus (SARS-CoV) is expressed in virus-infected cells and incorporated into SARS-CoV particles. *J Virol* 2007; **81**(2): 718-31.
51. Schaecher SR, Diamond MS, Pekosz A. The transmembrane domain of the severe acute respiratory syndrome coronavirus ORF7b protein is necessary and sufficient for its retention in the Golgi complex. *J Virol* 2008; **82**(19): 9477-91.
52. Akerstrom S, Mirazimi A, Tan YJ. Inhibition of SARS-CoV replication cycle by small interference RNAs silencing specific SARS proteins, 7a/7b, 3a/3b and S. *Antiviral Res* 2007; **73**(3): 219-27.
53. Herrewegh AA, Vennema H, Horzinek MC, Rottier PJ, de Groot RJ. The molecular genetics of feline coronaviruses: comparative sequence analysis of the ORF7a/7b transcription unit of different biotypes. *Virology* 1995; **212**(2): 622-31.
54. Zumla A, Hui DS, Azhar EI, Memish ZA, Maeurer M. Reducing mortality from 2019-nCoV: host-directed therapies should be an option. *Lancet* 2020; **395**(10224): e35-e6.
55. Walls AC, Park Y-J, Tortorici MA, Wall A, McGuire AT, Veesler D. Structure, function and antigenicity of the SARS-CoV-2 spike glycoprotein. *Cell* 2020.
56. Heald-Sargent T, Gallagher T. Ready, set, fuse! The coronavirus spike protein and acquisition of fusion competence. *Viruses* 2012; **4**(4): 557-80.

57. Bertram S, Glowacka I, Steffen I, Kuhl A, Pohlmann S. Novel insights into proteolytic cleavage of influenza virus hemagglutinin. *Rev Med Virol* 2010; **20**(5): 298-310.
58. Letko M, Marzi A, Munster V. Functional assessment of cell entry and receptor usage for SARS-CoV-2 and other lineage B betacoronaviruses. *Nat Microbiol* 2020.
59. Millet JK, Whittaker GR. Host cell proteases: Critical determinants of coronavirus tropism and pathogenesis. *Virus Res* 2015; **202**: 120-34.
60. Andreu-Moreno I, Sanjuan R. Collective Infection of Cells by Viral Aggregates Promotes Early Viral Proliferation and Reveals a Cellular-Level Allee Effect. *Curr Biol* 2018; **28**(20): 3212-9 e4.
61. Hoffmann M, Kleine-Weber H, Schroeder S, et al. SARS-CoV-2 Cell Entry Depends on ACE2 and TMPRSS2 and Is Blocked by a Clinically Proven Protease Inhibitor. *Cell* 2020.
62. Wang M, Cao R, Zhang L, et al. Remdesivir and chloroquine effectively inhibit the recently emerged novel coronavirus (2019-nCoV) in vitro. *Cell Res* 2020; **30**(3): 269-71.
63. Coutard B, Valle C, de Lamballerie X, Canard B, Seidah NG, Decroly E. The spike glycoprotein of the new coronavirus 2019-nCoV contains a furin-like cleavage site absent in CoV of the same clade. *Antiviral Res* 2020; **176**: 104742.
64. Dahms SO, Jiao GS, Than ME. Structural Studies Revealed Active Site Distortions of Human Furin by a Small Molecule Inhibitor. *ACS Chem Biol* 2017; **12**(5): 1211-6.
65. Becker GL, Lu Y, Hardes K, et al. Highly potent inhibitors of proprotein convertase furin as potential drugs for treatment of infectious diseases. *J Biol Chem* 2012; **287**(26): 21992-2003.
66. Stokes JM, Yang K, Swanson K, et al. A Deep Learning Approach to Antibiotic Discovery. *Cell* 2020; **180**(4): 688-702 e13.
67. Fast E, Chen B. Potential T-cell and B-cell Epitopes of 2019-nCoV. *bioRxiv* 2020: 2020.02.19.955484.
68. Kock RA, Karesh WB, Veas F, et al. 2019-nCoV in context: lessons learned? *Lancet Planet Health* 2020.
69. Paraskevis D, Kostaki EG, Magiorkinis G, Panayiotakopoulos G, Sourvinos G, Tsiodras S. Full-genome evolutionary analysis of the novel corona virus (2019-nCoV) rejects the hypothesis of emergence as a result of a recent recombination event. *Infect Genet Evol* 2020; **79**: 104212.
70. Calisher C, Carroll D, Colwell R, et al. Statement in support of the scientists, public health professionals, and medical professionals of China combatting COVID-19. *Lancet* 2020; **395**(10226): e42-e3.
71. The L. COVID-19: fighting panic with information. *The Lancet* 2020; **395**(10224): 537.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 23 March 2020 doi:10.20944/preprints202003.0320.v1

Figures

Figure 1

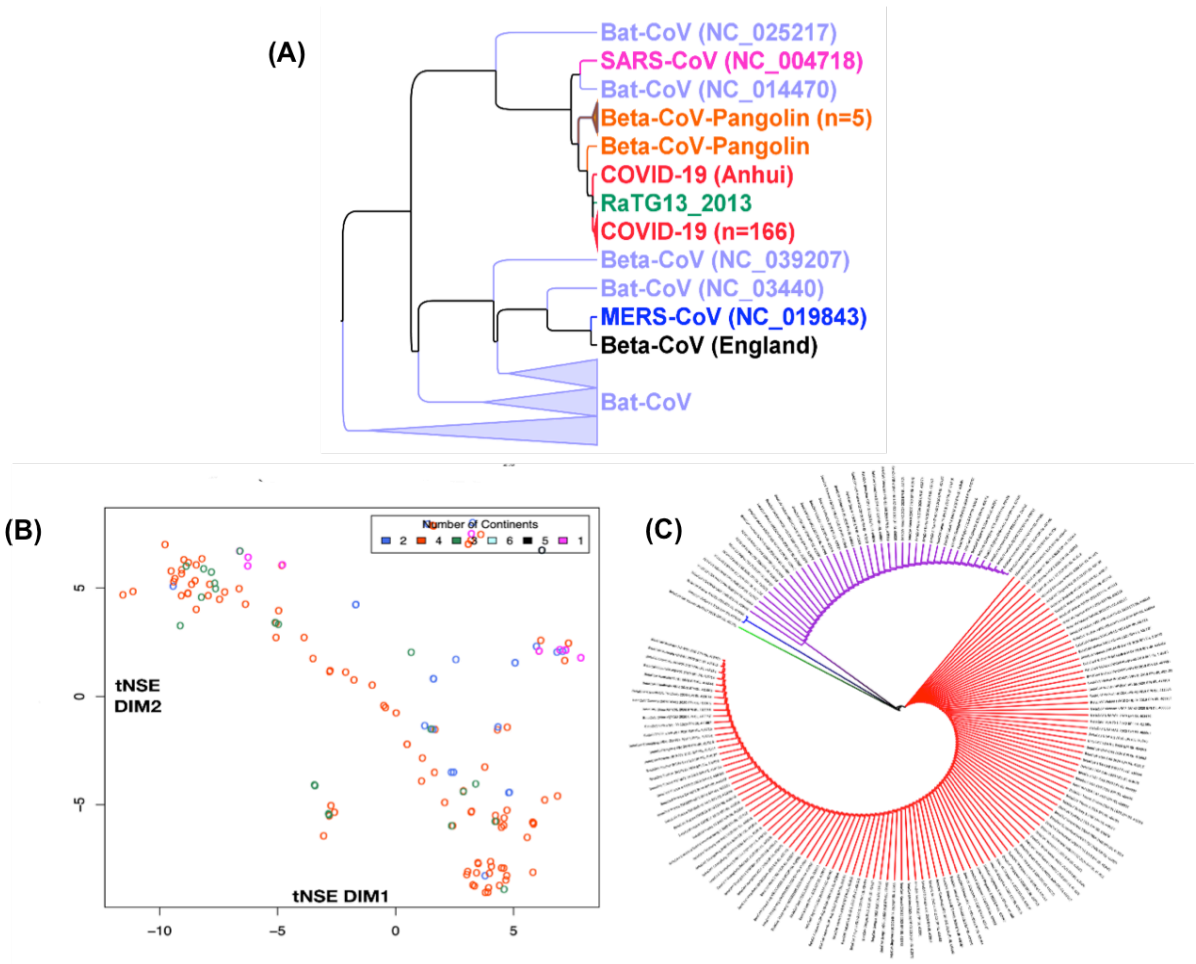

Figure 1. Comparative genomics and t-SNE based clustering of SARS-CoV-2, RaTG13, SARS-CoV, MERS-CoV and other coronaviruses. **(A)** Global similarity between genomes (n=167) based on amino acid (ORF) identity scores highlights strong diversity of COVID-19 strains when compared to other coronavirus strains. It is closer to RaTG13 and less similar to SARS-CoV and MERS-CoV at ORF level (function) **(B)** The t-SNE plot for COVID-19 strains (based on the global ORF similarity between genomes) shows no genomic correlation of the strains across continents - an indication of the effects of human migration in the modern population. It is difficult to correlate virulence rates of strains due to lack of strain-specific data. **(C)** Diversity in clinical isolates indicate that SARS-CoV-2 has divulged into at least two different clades. The purple coloured strains (n=42) are closer to RATG13 than the other red coloured strains (n=117).

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 23 March 2020 doi:10.20944/preprints202003.0320.v1

Figure 2

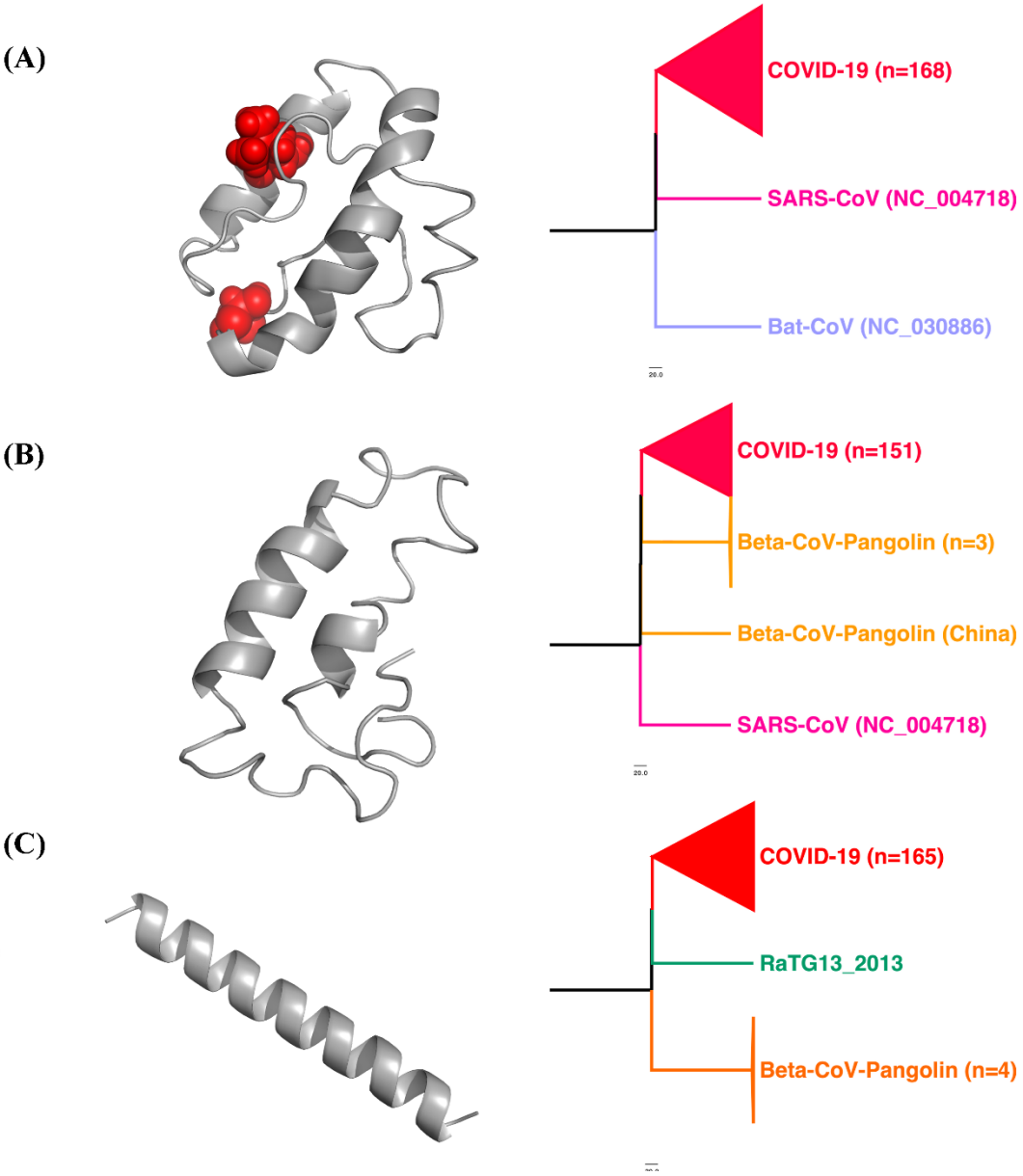

Figure 2. ORF based phylogenetic relationship of COVID-2019 clinical isolates with RaTG13, SARS-CoV, MERS and other coronaviruses. The left panel is the superimposition of proteins, **(A)** ORF4, **(B)** ORF6 and **(C)** ORF7b, of SARS-CoV-2, reference coronavirus genomes and RaTG13. Mutations among the clinical isolates are seen only in ORF 4 (shown in red). The right panel is clustering which shows sequence-based specificity of some of the ORF's **(A)** ORF4, **(B)** ORF6 and **(C)** ORF7b in SARS-CoV-2 clinical isolates.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 23 March 2020 doi:10.20944/preprints202003.0320.v1

Figure 3

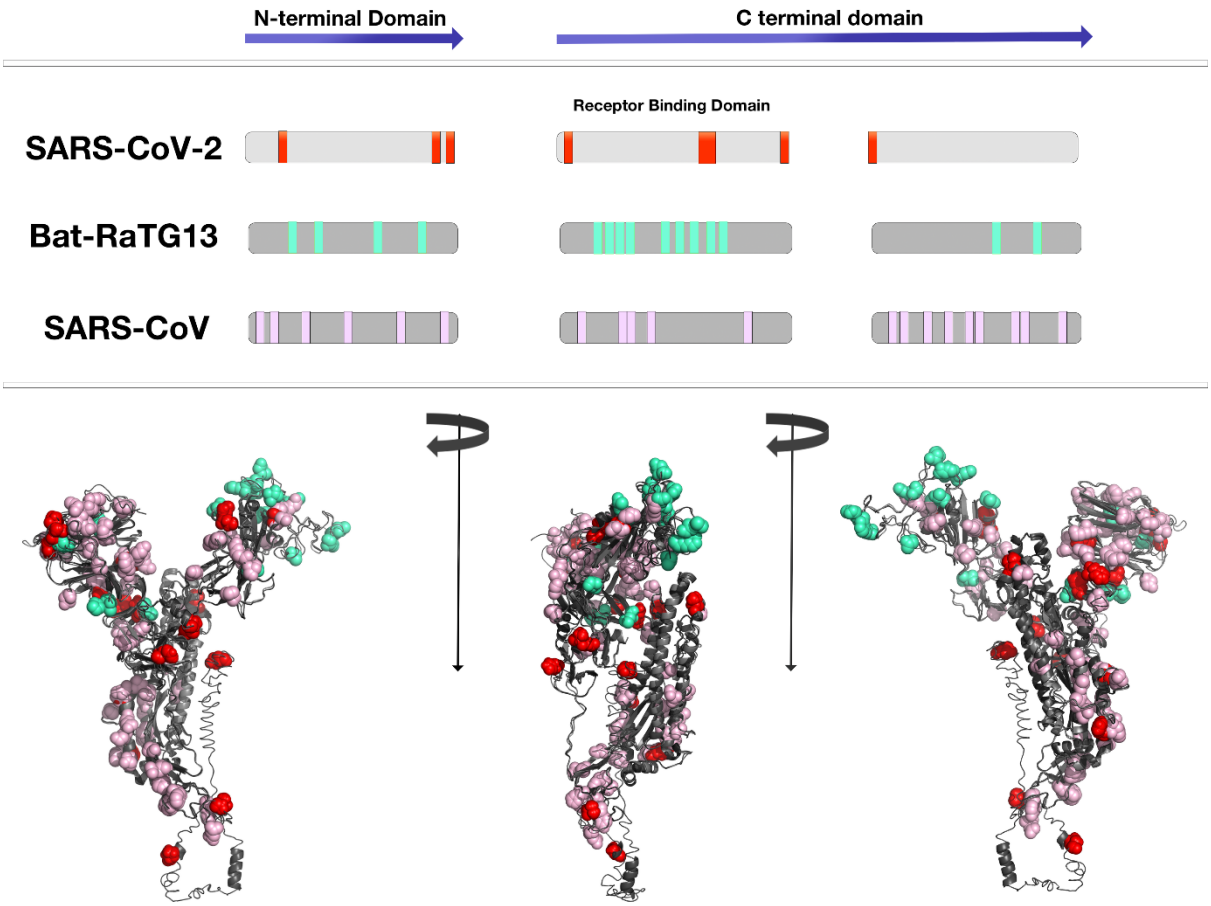

Figure 3. Comparative mutations in spike glycoprotein among RaTG13, SARS-COV and SARS CoV-2 clinical isolates (n=167). Single point amino acid mutations upon a comparison of RaTG13 and SARS-CoV with SARS-CoV-2 are shown in cyan green and light pink spheres, respectively. Mutations among the clinical isolates are shown in red. Comparative structural analysis between and RaTG13 and SARS-CoV-2 highlights crucial mutations restricted to ACE-2 receptor binding domain. However, compared to SARS-CoV mutations in both N-terminal and receptor binding domains were observed. These could lead to the differential binding affinity of the new coronavirus compared to the closely related pathogenic SARS-CoV-2.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 23 March 2020 doi:10.20944/preprints202003.0320.v1

Figure 4

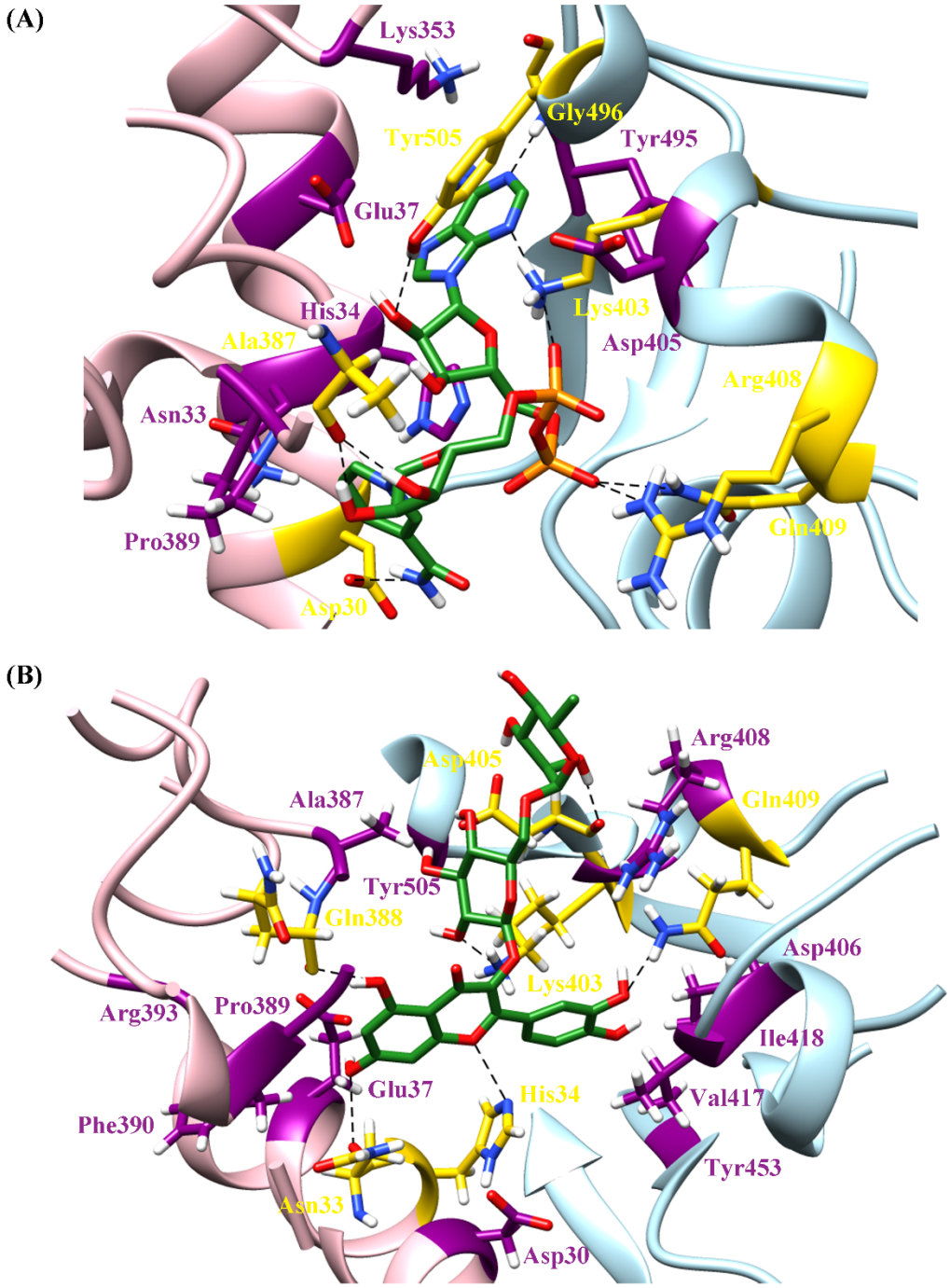

Figure 4. Proteins (spike-ACE2)-drug (A) NADH (B) Rutin interactions. NADH forms eight hydrogen bonds while Rutin forms six hydrogen bonds (shown as black dashed lines) with spike protein of SARS-CoV-2 and host ACE2 receptor suggesting its strong binding affinity and inhibitory potential against COVID-2019. NADH and Rutin are shown in dark green colour, RBD is in cyan and ACE2 is shown in pink colour. Hydrogen bonding residues are shown in yellow colour and hydrophobic network forming residues are shown in dark purple colour.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 23 March 2020 doi:10.20944/preprints202003.0320.v1

Figure 5

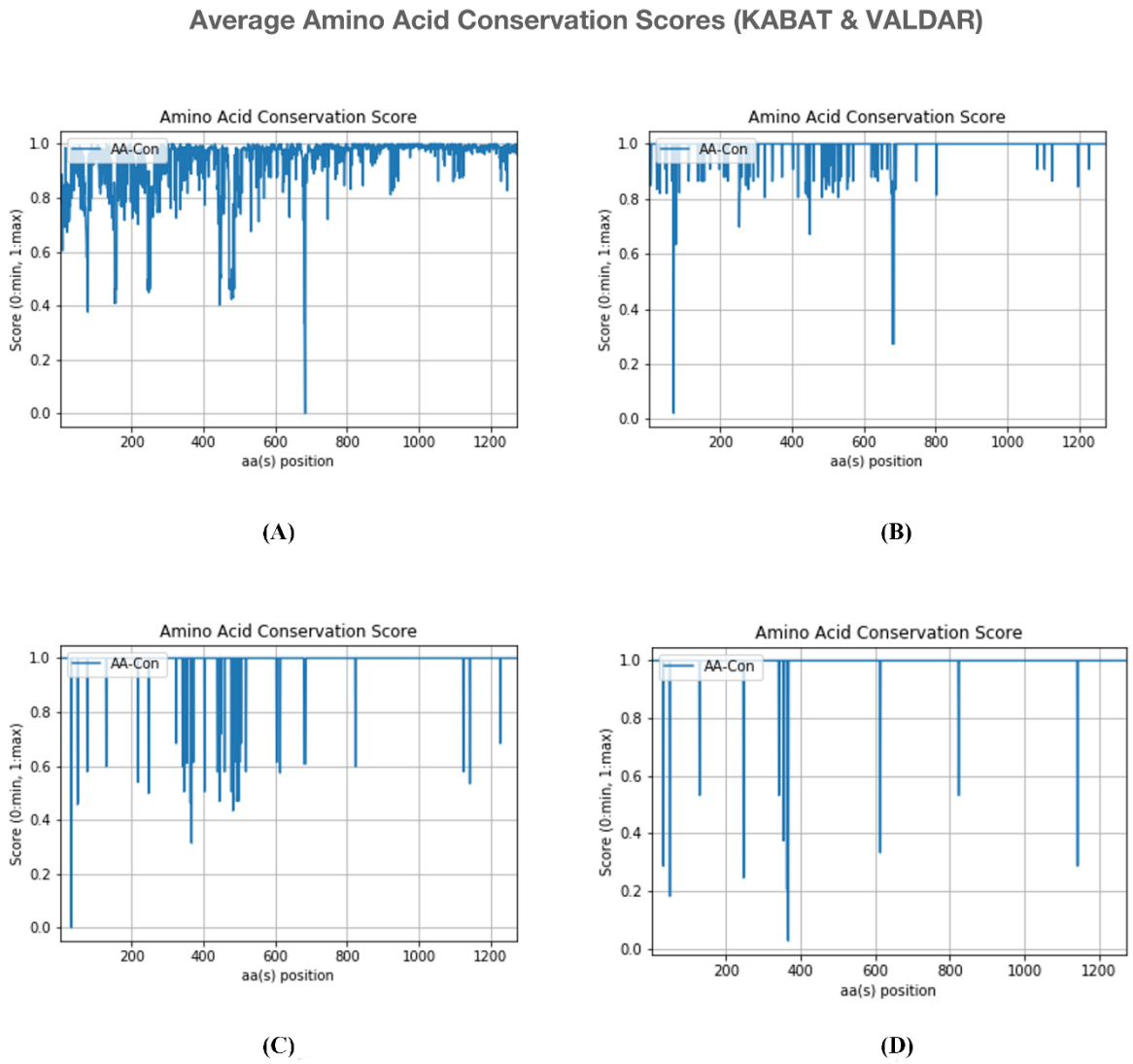

Figure 5. Amino acid conservation analysis of Spike proteins using an average of KABAT and VALDAR score. **(A)** Comparison of bat coronavirus genomes to one of the representative SARS-CoV-2 clinical isolate. **(B)** Comparison of pangolin coronaviruses to one of the representative SARS-CoV-2 clinical isolate. **(C)** Comparison of all clinical isolates with RaTG13. **(D)** Comparative profile of all clinical isolates of SARS-CoV-2. Low level conservation of N-terminal domain comparable to C-terminal domain in spike protein is evident. However, in comparison to other bat coronavirus genomes this sequence is highly conserved in SARS-CoV-2.