PLOS ONE

RESEARCH ARTICLE

Comparative genome analysis of novel coronavirus (SARS-CoV-2) from different geographical locations and the effect of mutations on major target proteins: An *in silico* insight

Mohd Imran Khan[1☯], Zainul A. Khan[2☯], Mohammad Hassan Baig[3☯], Irfan Ahmad[4,5], Abd-ElAziem Farouk[6], Young Goo Song[7]*, Jae-Jun Dong[3]*

1 Department of Biophysics, All India Institute of Medical Sciences, New Delhi, India, **2** Department of Plant Molecular Biology, University of Delhi South Campus, New Delhi, India, **3** Department of Family Medicine, Yonsei University College of Medicine, Seoul, Republic of Korea, **4** Department of Clinical Laboratory Sciences, College of Applied Medical Sciences, King Khalid University, Abha, Saudi Arabia, **5** Research Center for Advanced Materials Science, King Khalid University, Abha, Saudi Arabia, **6** Department of Biotechnology, Faculty of Science, Taif University, Taif, Saudi Arabia, **7** Department of Infectious Diseases, Gangnam Severance Hospital, Yonsei University College of Medicine, Seoul, South Korea

☯ These authors contributed equally to this work.
* IMFELL@yuhs.ac (YGS); S82TONIGHT@yuhs.ac (JJD)

Citation: Khan MI, Khan ZA, Baig MH, Ahmad I, Farouk A-EAziem, Song YG, et al. (2020) Comparative genome analysis of novel coronavirus (SARS-CoV-2) from different geographical locations and the effect of mutations on major target proteins: An *in silico* insight. PLoS ONE 15 (9): e0238344. https://doi.org/10.1371/journal.pone.0238344

Editor: Ghulam Md Ashraf, King Abdulaziz University, SAUDI ARABIA

Received: April 30, 2020

Accepted: August 14, 2020

Published: September 3, 2020

Data Availability Statement: All relevant data are within the manuscript and its Supporting Information files.

Funding: The authors received no specific funding for this work.

Competing interests: The authors of this paper have the journal's policy and have the following competing interests: Abd-ElAziem Farouk is an employee of GmbH. However, GmbH did not

Abstract

A novel severe acute respiratory syndrome-related coronavirus-2 (SARS-CoV-2) causing COVID-19 pandemic in humans, recently emerged and has exported in more than 200 countries as a result of rapid spread. In this study, we have made an attempt to investigate the SARS-CoV-2 genome reported from 13 different countries, identification of mutations in major coronavirus proteins of these different SARS-CoV-2 genomes and compared with SARS-CoV. These thirteen complete genome sequences of SARS-CoV-2 showed high identity (>99%) to each other, while they shared 82% identity with SARS-CoV. Here, we performed a very systematic mutational analysis of SARS-CoV-2 genomes from different geographical locations, which enabled us to identify numerous unique features of this viral genome. This includes several important country-specific unique mutations in the major proteins of SARS-CoV-2 namely, replicase polyprotein, spike glycoprotein, envelope protein and nucleocapsid protein. Indian strain showed mutation in spike glycoprotein at R408I and in replicase polyprotein at I671T, P2144S and A2798V,. While the spike protein of Spain & South Korea carried F797C and S221W mutation, respectively. Likewise, several important country specific mutations were analyzed. The effect of mutations of these major proteins were also investigated using various *in silico* approaches. Main protease (Mpro), the therapeutic target protein of SARS with maximum reported inhibitors, was thoroughly investigated and the effect of mutation on the binding affinity and structural dynamics of Mpro was studied. It was found that the R60C mutation in Mpro affects the protein dynamics, thereby, affecting the binding of inhibitor within its active site. The implications of mutation on structural characteristics were determined. The information provided in this manuscript holds

provide financial support for this study. This does not alter our adherence to PLOS ONE policies on sharing data and materials. There are no patents, products in development or marketed products associated with this research to declare.

Abbreviations: COVID-19, Coronavirus disease 19; MD, Molecular Dynamics simulations; MERS-CoV, Middle-East respiratory syndrome coronavirus; Mpro, Main protease; ORFs, Open reading frames; RdRp, RNA-dependent RNA polymerase; SARS-CoV, Severe acute respiratory syndrome coronavirus; SARS-CoV-2, Severe acute respiratory syndrome-related coronavirus-2; WT, Wildtype.

great potential in further scientific research towards the design of potential vaccine candidates/small molecular inhibitor against COVID19.

Introduction

In the last two decades, three coronaviruses *viz*. severe acute respiratory syndrome coronavirus (SARS-CoV) [1], Middle-East respiratory syndrome coronavirus (MERS-CoV) [2] and SARS-CoV-2 have crossed the species barrier to cause deadly pneumonia in humans. In 2002, SARS-CoV emerged in the Guangdong province of China and spread to five continents, infecting 8,098 people with 774 deaths. In 2012, MERS-CoV emerged in the Arabian Peninsula, transmitted to 27 countries, infecting a total of ~2,494 individuals and claiming 858 lives. The current outbreak of coronavirus disease 19 (COVID-19) caused by SARS-CoV-2, was first reported in December 2019 in Wuhan, Hubei province of China [3, 4] and spread across 200 countries, infecting over 2.5 million people and killed more than 1.5 lakh as of April, 23, 2020. SARS-CoV-2 was declared a pandemic by the World Health Organization on March 12, 2020.

SARS-CoV-2 belongs to the family *Coronaviridae* of genus *Betacoronavirus*, having positive sense strand RNA genome of 26–32 kb size. SARS-CoV-2 genome has six major open reading frames (ORFs) *viz*. replication enzyme coding region (ORF 1a and 1b), E gene (envelope protein), M gene (membrane protein), S gene (spike protein), and N gene (nucleocapsid protein) that are common to coronaviruses and a number of other accessory genes (ORF 3a, 6, 7a, 7b and 8) (Fig 1) [3]. The structural proteins: envelope protein, nucleocapsid protein, spike protein and membrane protein are essential for producing the structurally complete viral particle [5–8]. Entry of coronavirus into host cells is guided by spike glycoprotein. ORF 1a and 1b encode replication enzyme consisting 16 non-structural proteins (nsp1-16) that are highly conserved among the coronaviruses. Main protease (Mpro, also known as 3CLpro) is one of the important nsp encoded by ORF 1a and 1b, play an essential role in the processing of polyproteins and control the replication of coronavirus [9, 10]. RNA-dependent RNA polymerase (RdRp) also known as nsp12, another important replicase catalyze the replication of RNA using viral genomic RNA template [11].

Reports showed that MERS-CoV originated from bats, but the reservoir host fueling spillover to humans is unequivocally dromedary camels [12, 13]. Both SARS-CoV and SARS-CoV-2 are originated from bats, which serve as reservoir host for these two viruses [3, 14, 15]. Raccoon dogs and palm civets have been identified as intermediate hosts for zoonotic transmission of SARS-CoV between bats and humans [16–18], however, the intermediate host of SARS-CoV-2 remains unknown.

Mutation rate is very high in RNA viruses, up to a million times higher than their host, which enhance their virulence and evolvability (formation of new species) [19]. Coronavirus replication is error prone as compared to other RNA viruses and the estimated mutation rate is $4x10^{-4}$ nucleotide substitutions/site/year [20]. The rate of SARS-CoV-2 mediated disease spread and the mortality varies from country to country. Of several reasons affecting the rate of disease spread and mortality, mutations within the SARS-CoV-2 strains is also considered one of the major factors.

This study was conducted to gather additional information on the SARS-CoV-2 sequences from different geographical locations infected with COVID-19. The genome analysis of the SARS-CoV-2 strains from 13 different countries showed a large number of mutations within the major structural proteins. This is the first time we have comprehensively investigated these

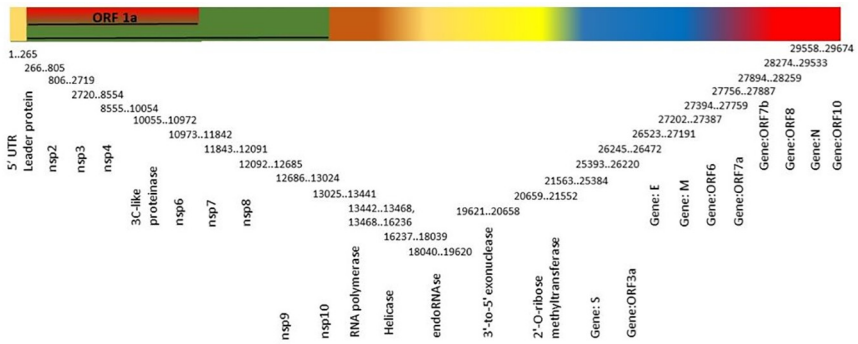

Fig 1. Schematic representation of genome organization of SARS-CoV-2.

https://doi.org/10.1371/journal.pone.0238344.g001

mutations and also discussed their potential roles in the pathogenicity, replication and entry of virus particle. This study provides a deeper insight into the emergence of these mutations within the major structural as well as nsp encoded by the SARS-CoV-2 genome from different countries. Here, molecular dynamics and other *in silico* studies were also performed to investigate the effect of mutations on the dynamics of Mpro. The findings of this study provide a clue for the futuristic development of potential vaccine candidate or therapeutic design against COVID19.

Material and methods

Sequence analysis and stability prediction

The genome sequence of ORF1ab for SARS with reference sequence ID: NC_004718.3 and protein sequence with GenBank ID: AAP41036.1, was retrieved from NCBI database. Similarly, the genome sequence for SARS-CoV-2 with Reference Sequence ID: MT012098.1, MT019529.1, MT039890.1, MT093571.1, MT192772.1, MT126808.1, MT192759.1, MN985325.1, MT007544.1, LC529905.1, MT020781.2, MT072688.1 and MT066156.1 and protein sequence with GenBank ID: QHS34545.1, QHU36823.1, QHZ00378.1, QIC53203.1, QIK50437.1, QIG55993.1, QIK50416.1, QHO60603.1, QHR84448.1, BCB15089.1, QHU79171.2, QIB84672.1 QIA98553.1 for India, China, South-Korea, Sweden, Vietnam, Brazil, Taiwan, USA, Australia, Japan, Finland, Nepal and Italy, respectively, were downloaded from NCBI. Multiple sequence alignment (MSA) and visualization of SARS and SARS-CoV-2 sequences from 13 different countries was performed using Molecular Evolutionary Genetics Analysis (MEGA) version 10.1.8. To delineate and analyze the mutation across different countries, an in-house script written in Perl and Python was used. MUPRO server was used to determine the effect of mutation on various SARS-CoV-2 proteins [21].

Model building

The crystal structure of Mpro protein from SARS-CoV-2 in complex with Boceprevir (pdb id: 7BRP) was taken as a wildtype (WT). The structure of R60C was generated by inserting the Point mutation and modelled using modeler 9v13 [22].

Molecular docking

Boceprevir was retrieved and redocked within the structure of Mpro using CCDC Gold. The RMSD for the crystal and redocked conformation of Boceprevir were compared. Further,

Boceprevir was subjected to dock within the active site of R60C mutant. The poses were visualized on PMV viewer [23].

Molecular dynamics simulations of the protein and their complexes

The structure of Boceprevir in complex with Mpro (WT) and R60C mutant was subjected to energy minimization using Gromacs-5 with the CHARMM27 all atom force field [24–26]. The models were solvated with a SPC/E water model in a cubic periodic box with 1 nm distance from the edge of the complex atoms. The solvated system was neutralized by seven chloride ions. The system was, thereafter, minimized using steepest descent algorithm with convergence criteria of tolerance value 1000 kJ mol^{-1} nm^{-2}. The complete simulation of minimized solvated proteins was performed under periodic boundary condition with time step of 2 fs. Particle mesh Ewald was used for long range electrostatic interactions with an interpolation order of 4 and a Fourier spacing of 0.16. The first phase simulation was conducted under an NVT ensemble for 500 ps by keeping all bonds constrained using the LINCS algorithm for temperature equilibration. The system was heated to 300 K using leap-frog integrator while pressure coupling was set off. A V-rescale thermostat was used to maintain constant temperature for each system, followed by pressure equilibration at 300 K using Parrinello-Rahman pressure coupling algorithm under an isothermal-isobaric ensemble for another 500 ps at 1.0 bar. Isothermal compressibility of the solvent was set to $4.5e^{-5}$ bar^{-1}. Further, simulation of 50000 ps production run was carried out at 300 K and 1 atm pressure for trajectory analysis. The final models obtained at the end of MD were validated and taken for structural analysis.

Results

Sequence analysis and mutation detection

The complete nucleotide sequences of 13 SARS-CoV-2 reported from 13 different countries showed ~82% sequence identity with SARS-CoV. Also, all 13 sequences shared more than 99% sequence identity to each other. Replicase polyprotein (ORF 1ab) of 13 isolates, which are most conserved in all coronaviruses shared maximum identity (87%) with SARS-CoV (NC_0047180), which is less than the threshold value (90%) for demarcation of betacoronavirus species [27, 28]. Phylogenetic analysis revealed that all 13 SARS-CoV-2 identified from different geographical locations clustered together in a single clad as compared to SARS-CoV (Fig 2A and 2B).

Further, we checked the mutation in all major proteins of 13 SARS-CoV-2 sequences and compared with SARS-CoV. ORF 1a and 1b showed 11 changes among all 13 SARS-CoV-2. Indian SARS-CoV-2 sequence showed three changes at 671 (Isoleucine to Threonine), 2144 (Proline to Serine) and 2798 (Alanine to Valine) compared to all other 12 isolates. Here, we also noted two amino acid mutations (in ORF1ab) in each SARS-CoV-2 sequences isolated from China (2708: Asparagine to Serine; 2908: Phenylalanine to Isoleucine), South Korea (902: Methionine to Isoleucine; 6891: Threonine to Methionine) and Sweden (818: Glycine to Serine; 4321: Phenylalanine to Leucine). Brazil and Vietnam isolate showed only one change at 3606 (Leucine to Phenylalanine) and 3323 (Arginine to Cystine), respectively (Table 1). 996 changes have been reported when ORF 1a and 1b amino acid sequences of all 13 SARS-CoV-2 was compared with SARS-CoV (S1 File).

The viral Mpro controls the replication of coronavirus and is a key protein responsible for its life cycle [29–31]. Mpro is an attractive drug discovery target. The analysis of Mpro reveals that there was only one point mutation (R60C) in the Vietnam strain of SARS-CoV-2 (Fig 3A). RdRp, which is another important target for antiviral drugs functions by catalyzing the

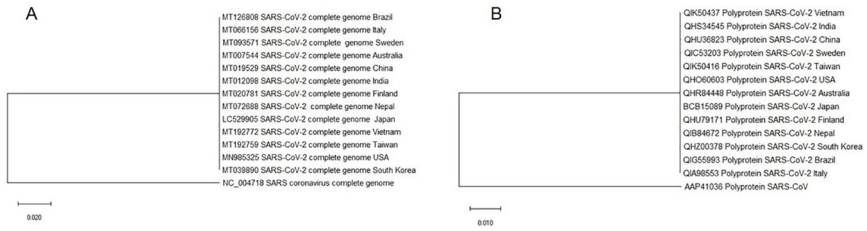

Fig 2. Phylogenetic dendogram showing the relationship of SARS-CoV-2 complete sequence (nucleotide) from different geographical locations (13 no.) with SARS-CoV (A) and amino acid sequence of replicase polyprotein of 13 SARS-CoV-2 with SARS-CoV (B). The evolutionary history was inferred using the Neighbor-Joining method.

https://doi.org/10.1371/journal.pone.0238344.g002

viral RNA synthesis [32]. Only one mutation (A406V) was observed in the RdRp of Indian SARS-CoV-2 isolate (Fig 3B).

Spike proteins are the key surface glycoproteins and are well reported for their prominent role in interaction with host cell receptors [33, 34]. Here, we analyzed the mutations in the spike protein of SARS-CoV-2 from different countries. It was found that this glycoprotein carried five different amino acid mutations at various positions within the investigated SARS-CoV-2 isolates. For instance, India, Finland, Australia, South Korea and Sweden SARS-CoV-2 isolates showed one amino acid change at 408 (Arginine to Isoleucine), 49 (Histidine to Tyrosine), 247 (Serine to Arginine), 221 (Serine to Tryptophan) and 797 (Phenylalanine to Cysteine), respectively (Table 2 and Fig 3C). The value of ΔΔG show that the mutant R408I (0.49732107 kcal/mol) mutation was having stabilization effect on spike protein. It was found that the mutation on the receptor binding domain (RBD) of spike protein increases the stability.

When these 13 SARS-CoV-2 isolates were compared to SARS-CoV sequence, 1338 changes have been reported (S1 File). The analysis of ORF3a showed 3 mutations within different SARS-CoV-2 strains: W128L (South Korea), L140V (Japan), G251V (Australia, South Korea, Brazil, Italy, Sweden) (Table 3).

One amino acid change occurred in each envelope protein of South Korea SARS-CoV-2 isolate at 37 (Leucine to Histidine) and nucleocapsid protein of Japan SARS-CoV-2 isolate at 344 (Proline to Serine) when compared among 13 SARS-CoV-2 isolates (Tables 4 and 5, Fig 3D and 3E), while, 5 and 45 changes has been reported in envelop and nucleocapsid proteins, respectively as compared to SARS-CoV. Deletion of Glycine and Serine occurred at position 70 and 8 in envelop and nucleocapsid proteins, respectively, in all 13 SARS-CoV-2 isolates

Table 1. Amino acid variation in replicase polyprotein of SARS-CoV-2 strains of 13 different countries.

Amino acid	India	China	South Korea	Sweden	Vietnam	Brazil	Taiwan	USA	Australia	Japan	Finland	Nepal	Italy
671	T	I	I	I	I	I	I	I	I	I	I	I	I
818	G	G	G	S	G	G	G	G	G	G	G	G	G
902	M	M	I	M	M	M	M	M	M	M	M	M	M
2144	S	P	P	P	P	P	P	P	P	P	P	P	P
2708	N	S	N	N	N	N	N	N	N	N	N	N	N
2908	F	I	F	F	F	F	F	F	F	F	F	F	F
3323	R	R	R	R	C	R	R	R	R	R	R	R	R
3606	L	L	L	L	L	F	L	L	L	L	L	L	X
4321	F	F	F	L	F	F	F	F	F	F	F	F	F
4798	V	A	A	A	A	A	A	A	A	A	A	A	A
6891	T	T	M	T	T	T	T	T	T	T	T	T	T

https://doi.org/10.1371/journal.pone.0238344.t001

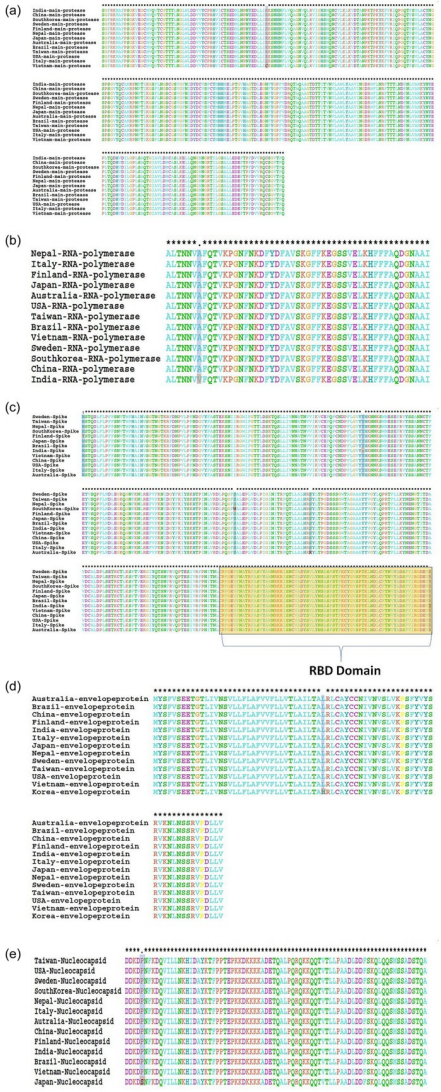

Fig 3. Alignment of SARS-CoV-2 major proteins (A) main protease, (B) RNA-dependent RNA polymerase, (C) spike proteins, (D) envelope proteins and (E) nucleocapsid proteins from different countries.

https://doi.org/10.1371/journal.pone.0238344.g003

when compared to SARS-CoV. MEM glycoprotein did not show any amino acid change among 13 SARS-CoV-2 isolates, while 24 changes occurred when compared to SARS-CoV (S1 File). All the other point mutations occurring within the structural proteins of SARS-CoV-2 isolates from different countries were found to decrease protein stability (Table 6).

Molecular Dynamics (MD) studies

In the present study, we performed the MD simulations for the Boceprevir bound complexes of SARS-CoV-2 Mpro and its R60C mutant to study the effect of mutation on the protein dynamics.

Root-Mean-Square Deviation (RMSD)

The root mean square deviations of the backbone were calculated to analyze the trajectories of Mpro from SARS-CoV-2 and its R60C mutant. In the complex form, a slight fluctuation in the

Table 2. Amino acid variation in spike protein of SARS-CoV-2 strains of 13 different countries.

Amino acid	QHS34546 India	QHU79173 Finland	QHR84449 Australia	QHZ00379 South Korea	QIG55994 Brazil	QIK50438 Vietnam	QIK50417 Taiwan	QIA98554 Italy	QHU36824 China	QHO60594 USA	BCB15090 Japan	QIC53204 Sweden	QIB84673 Nepal
49	H	Y	H	H	H	H	H	H	H	H	H	H	H
145	-	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
221	S	S	S	W	S	S	S	S	S	S	S	S	S
247	S	S	R	S	S	S	S	S	S	S	S	S	S
408	I	R	R	R	R	R	R	R	R	R	R	R	R
797	F	F	F	F	F	F	F	F	F	F	F	C	F

https://doi.org/10.1371/journal.pone.0238344.t002

Table 3. Amino acid variation in ORF 3 encoded protein of SARS-CoV-2 strains of 13 different countries.

Position Amino acid	QHS34546 India	QHU79173 Finland	QHR84449 Australia	QHZ00379 South Korea	QIG55994 Brazil	QIK50438 Vietnam	QIK50417 Taiwan	QIA98554 Italy	QHU36824 China	QHO60594 USA	BCB15090 Japan	QIC53204 Sweden	QIB84673 Nepal
128	W	W	W	L	W	W	W	W	W	W	W	W	W
140	L	L	L	L	L	L	L	L	L	L	V	L	L
251	G	G	V	V	V	G	G	V	G	G	G	V	G

https://doi.org/10.1371/journal.pone.0238344.t003

Table 4. Amino acid variation in envelop protein of SARS-CoV-2 strains of 13 different countries.

Position Amino acid	QHS34546 India	QHU79173 Finland	QHR84449 Australia	QHZ00379 South Korea	QIG55994 Brazil	QIK50438 Vietnam	QIK50417 Taiwan	QIA98554 Italy	QHU36824 China	QHO60594 USA	BCB15090 Japan	QIC53204 Sweden	QIB84673 Nepal
37	L	L	L	H	L	L	L	L	L	L	L	L	L

https://doi.org/10.1371/journal.pone.0238344.t004

Table 5. Amino acid variation in nucleocapsid protein of SARS-CoV-2 strains of 13 different countries.

Position Amino acid	QHS34546 India	QHU79173 Finland	QHR84449 Australia	QHZ00379 South Korea	QIG55994 Brazil	QIK50438 Vietnam	QIK50417 Taiwan	QIA98554 Italy	QHU36824 China	QHO60594 USA	BCB15090 Japan	QIC53204 Sweden	QIB84673 Nepal
344	P	P	P	P	P	P	P	P	P	P	S	P	P

https://doi.org/10.1371/journal.pone.0238344.t005

Table 6. Mutation in SARS-CoV-2 proteins from different geographical locations and their predicted effect on protein stability.

Protein Name	Mutation	Country	Stability effect (MUPRO)
3C-like proteinase (3CLpro)	R60C	Vietnam	DECREASE stability (ΔΔG -1.0163868)
Envelope Protein	L37H	South Korea	DECREASE stability (ΔΔG -2.4215632)
ORF3a	W128L	South Korea	DECREASE stability (ΔΔG -0.39593766)
	L140V	Japan	DECREASE stability (ΔΔG -0.90740107)
	G251V	Australia, South Korea, Brazil, Italy, Sweden	DECREASE stability (ΔΔG -0.45128408)
Spike Protein	H49Y	Finland	DECREASE stability (ΔΔG -0.20900128)
	S221W	Brazil	DECREASE stability (ΔΔG -0.45085799)
	S247R	Australia	DECREASE stability (ΔΔG -1.3464875)
	R408I	India	INCREASE stability (ΔΔG 0.49732107)
	F797C	Sweden	DECREASE stability (ΔΔG -1.501262)
Nucleocapsid	P344S	Japan	DECREASE stability (ΔΔG -1.2252261)
RNA-dependent RNA polymerase	A406V	India	DECREASE stability (ΔΔG -0.76907034)

https://doi.org/10.1371/journal.pone.0238344.t006

backbone RMSD was also noticed (Fig 4A). It was found that the RMSD of mutant was comparatively more stable than its WT. This variation in the average RMSD values suggests that this mutation was affecting the dynamic behavior of Mpro.

Radius of gyration and SASA

Fig 4B illustrates the Rg of Cα atoms plot of the complexed Mpro from SARS-CoV-2 and the Vietnam mutant Mpro (R60C). The R60C mutant shows slightly lower value for Rg as compared to its WT. This suggests that R60C mutation affects the stability of Mpro. Fig 4C shows the change of SASA of native and R60C mutant with time. The greater value of SASA for R60C mutant (in complexed form) was supported by Rg plot [35].

Root Mean Square Fluctuation (RMSF)

RMSF values of native as well as R60C mutant Mpro were calculated to determine the impact of mutation on dynamic behavior of protein at residue level. RMSF plot clearly indicates the fluctuation in residues and showed the existence of higher degree of flexibility in R60C mutant Mpro. It was found that the maximum amino acid fluctuation was in the region 50–76 and 127–222 (Fig 4D). The binding studies also confirm that the residues falling within this region were very much involved in accommodating the inhibitor within the active site of Mpro [29, 31].

Interaction energy and effect on hydrogen bond network

Throughout the MD trajectory, the interaction energy of ligand in complex with the surrounding protein residues of WT and mutant Mpro were calculated. The Lennard–Jones short-

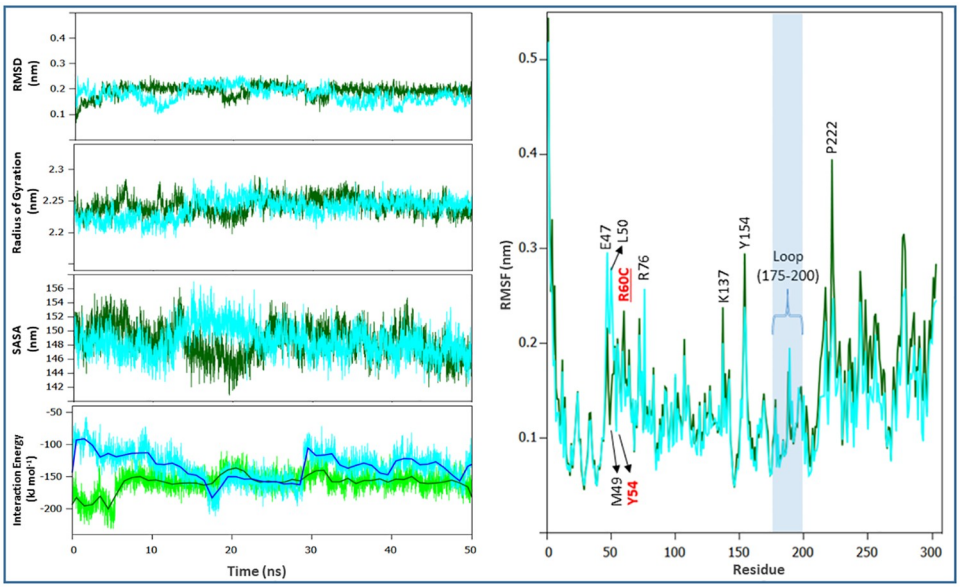

Fig 4. Molecular dynamics of complexed SARS-CoV-2 Mpro and the R60C mutant. Green color indicates the SARS-CoV-2 Mpro while the cyan color indicates the R60C mutant Mpro. (A) Backbone RMSDs of Mpro and its mutated form (B) Rg of Cα atoms (C) Change in Solvent accessible surface area (D) RMSF of the backbone atoms (E) The Lennard–Jones short-range (LJ-SR) and Coulombic short-range (Coul-SR) potential energies.

https://doi.org/10.1371/journal.pone.0238344.g004

range (LJ-SR) and Coulombic short-range (Coul-SR) potential energies were calculated throughout the course of 50 ns of MD simulation (Fig 4E). The average interaction energy for 50 ns are shown in Table 7. It was found that the binding of inhibitor within the active site of Mpro (WT) was stronger as compared to the R60C mutant. The analysis of hydrogen bond network revealed that the R60C mutation also cause disturbance in the interactions with inhibitor as well as other surrounding active site residues of Mpro. A large fluctuation was noticed in the hydrogen bond network of Mpro and its R60C mutant (Fig 5A). It was found that R60C mutation results in the changes in local environment that cascade further to the short helix and loop of the catalytic active site of Mpro.

Discussion

Till date (April 23, 2020), 2.6 million cases of COVID-19 have been reported worldwide. In this study, we analyzed 13 complete sequences of SARS-CoV-2 reported from 13 different countries and compared with SARS-CoV. Phylogenetic analysis showed that SARS-CoV-2 sequences clustered together in a single clad irrespective of their geographic origin, whether from the same continent or neighboring countries. Replicase polyprotein, which are most conserved among coronaviruses, shared 87% amino acid sequence similarity to SARS-CoV, less

Table 7. The Lennard–Jones short-range (LJ-SR) and Coulombic short-range (Coul-SR) potential energies calculated throughout the course of 50 ns of MD simulation.

Complex	Average (kJ/mol)	Total drift (kJ/mol)
LJ-SR:Mpro(WT)- Boceprevir	-158.90	18.54
LJ-SR:Mpro(R60C)- Boceprevir	-134.43	-19.01
Coul-SR: Mpro(WT)- Boceprevir	-65.37	17.45
Coul-SR: Mpro(R60C)- Boceprevir	-59.56	6.49

https://doi.org/10.1371/journal.pone.0238344.t007

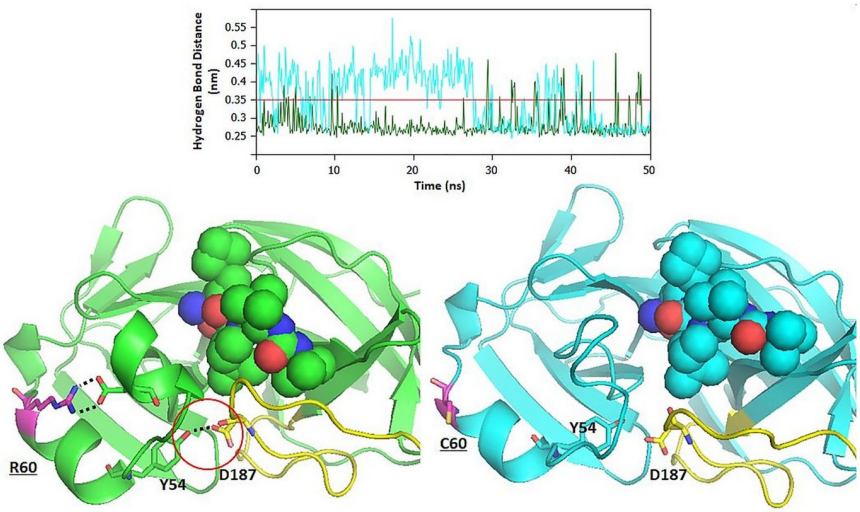

Fig 5. (A) The hydrogen bond network of the Mpro (WT) and R60C mutant. (B) The Structure of WT and R60C mutant Mpro.

https://doi.org/10.1371/journal.pone.0238344.g005

than the threshold value (90%) for demarcation of betacoronavirus species [27]. They belong to new virus species *Severe acute respiratory syndrome-related coronavirus* of genus *Betacoronavirus* [28].

Among all known RNA viruses, coronaviruses consist of the largest genome (26.4 to 31.7 kb) [36, 37]. The large genome size provides more plasticity in accommodating and modifying genes [36–38]. Mutation frequency is very high in RNA viruses, which enhances virulence and responsible for the formation of new species [19]. The high frequency of mutation within the viral genome at different geographical locations may be one of the reasons that SARS-CoV-2 is responsible for change in mortality rate and symptom of the disease [39]. The comparison of amino acid sequences of replicase polyprotein of 13 SARS-CoV-2 showed mutations in India, China, South Korea, Sweden, Vietnam and Brazil strains at different amino acid locations. Earlier report showed the similar result i.e. a single mutation in replicase polyprotein at 3606 (L to F) [40]. We could identify few more single amino acid mutations at different positions in above mentioned SARS-CoV-2 strains (Table 1). The replicase polyprotein codes for nsp2 and nsp3 and it has been suggested in previous research that the mutation in nsp2 and nsp3 play a key role in infectious capability and are responsible for the differentiation mechanism of SARS-CoV-2 [39].

The RBD of spike protein is the region which specifically interact with ACE2 leading to viral entry into the host cell [41–43]. The Indian isolate of SARS-CoV-2 showed mutation within this region where at 408 position, Arginine is replaced by Isoleucine. For several years, the prediction of protein stability via theoretical or experimental approaches has been a profound area of research [44]. Earlier findings suggest that a single point mutation at RBD is responsible for disrupting the antigenic structure, thereby, affecting the binding of RBD to ACE2 [45, 46]. The mutation within this region of spike protein may affect the binding of RBD to its receptor, thus, affecting the viral entry within the host cells. Further, *in silico* studies revealed that this point mutation within the RBD of spike glycoprotein was having stabilization effect on spike protein and found to increase the protein stability (ΔΔG 0.49732107 kcal/mol).

Single amino acid mutation was observed in both Mpro (R60C) of SARS-CoV-2 Vietnam isolate and RdRp (A408V) of SARS-CoV-2 India isolate. The *in silico* findings revealed that the mutations in both strains decrease the stability of protein. The MD simulation studies on

Mpro further confirmed that the point mutation on Mpro affects the stability of proteins as well as the binding of inhibitor. Our *in silico* study found that the catalytic active site of Mpro is surrounded by amino acid residues of a loop (142–145, 175–200), short helix (40–43, 46–50), and beta sheet regions (25–27, 164–167). The R60C mutant lies at helix adjacent to the short helix (H2) that forms the catalytic channel (Fig 5). Substitution of an amino acid with charged side chain to uncharged cysteine residue leads to loss of conserved ionic bond interaction and the effect cascades to other conserved ionic interactions. Loss of conserved ionic interaction was observed between amide nitrogen of arginine and carboxylic oxygen atom of aspartic acid at position 48 of the catalytic channel.

It was found that the short helix H2, that form the catalytic channel, have attained a more flexible loop like conformation in the mutant protein. Conserved hydrogen bonded interactions that stabilizes the catalytic channel L1 loop between Tyr54 OH···Asp187 Oδ1, Tyr54 OH···Asp187 O and Leu50 O···Arg188 NE were lost in the mutant enzyme, thereby, increasing the flexibility of structure forming the binding pocket (Fig 5B). Therefore, the local change of an ordered secondary structure to a more disordered loop like structure have increased the overall flexibility of the secondary structure elements forming the catalytic pocket, thereby, effecting the binding of ligand to residue in the catalytic channel. This is quite evident from the reduced LJ-SD and coulombic-SR interaction energies between the enzyme and ligand in wild and mutant complexes (Fig 4E and Table 7). The role of important active site residues, discussed in this study, has been reported before [47, 48]. The RMSF plot also reveals the key role of these residues in accommodating the inhibitor within the active site of Mpro.

Envelop protein plays an important role in the assembly of viral genome and the formation of ion channels (IC), responsible for virus-host interaction, which is mainly associated with pathogenesis [5, 49]. We detected one amino acid mutation L37H in transmembrane domain (TMD) of envelop protein of SARS-CoV-2 South Korea isolate. The TMD is hydrophobic in nature consisting mainly hydrophobic amino acids, while the mutation in TMD at 37 position changes hydrophobic to hydrophilic amino acid which changes the integrity of TMD. Earlier report showed that mutations within the TMD domain of envelop protein completely disrupted IC activity [50]. This might be one of the reasons for slow spreading/low pathogenicity of SARS-CoV-2 in South Korea.

Nucleocapsid protein of coronavirus is necessary for RNA replication, transcription and genome packaging [51, 52]. A mutation P344S in nucleocapsid protein has been detected in SARS-CoV-2 Japan strain. The P344S mutation on nucleocapsid was found to decrease the protein stability (ΔΔG -1.2252261). This mutation is located in carboxy-terminal RNA-binding domain (CTD) of nucleocapsid protein. Earlier studies showed that CTD is responsible for oligomerization [53].

It was also revealed that among all the genomes studied in this study, the Indian SARS-CoV-2 isolates were carrying maximum mutation. The Indian isolates were carrying the R408I on the spike protein while A406V on the RdRp and several mutations on the replicase polyprotein of SARS-CoV-2. It is expected that these large number of mutations among the SARS-CoV-2 may affect the vaccine/inhibitor development against these isolates.

Conclusion

To conclude, SARS-CoV-2 complete sequences from 13 countries were analyzed and compared with SARS-CoV. We identified country specific mutations in major proteins (replicase polyprotein, spike protein, envelop protein and nucleocapsid protein). Further, molecular dynamics and other *in silico* studies revealed that mutations decrease the stability of protein and also hinders the binding of inhibitor. Mutation R408I in spike protein of Indian strain has

significant influence on RBD domain of spike protein and this point mutation has a stabilization effect on the spike protein. The findings of the present study could help for the design of potential vaccine candidates/small molecular inhibitor against COVID19.

Supporting information

S1 File.
(ZIP)

Acknowledgments

The authors are very grateful to Prof. Indranil Dasgupta, University of Delhi South Campus, New Delhi, India, for giving suggestions to improve the manuscript.

Author Contributions

Conceptualization: Zainul A. Khan, Mohammad Hassan Baig, Young Goo Song, Jae-Jun Dong.

Data curation: Mohd Imran Khan.

Formal analysis: Zainul A. Khan, Mohammad Hassan Baig, Irfan Ahmad.

Investigation: Mohd Imran Khan, Zainul A. Khan.

Methodology: Mohd Imran Khan, Zainul A. Khan.

Supervision: Young Goo Song, Jae-Jun Dong.

Writing – original draft: Zainul A. Khan, Mohammad Hassan Baig, Irfan Ahmad, Abd-ElAziem Farouk.

Writing – review & editing: Mohd Imran Khan, Mohammad Hassan Baig, Abd-ElAziem Farouk, Young Goo Song, Jae-Jun Dong.

References

1. Drosten C, Gunther S, Preiser W, van der Werf S, Brodt HR, Becker S, et al. Identification of a novel coronavirus in patients with severe acute respiratory syndrome. N Engl J Med. 2003; 348(20):1967–76. https://doi.org/10.1056/NEJMoa030747 PMID: 12690091.
2. 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. https://doi.org/10.1056/NEJMoa1211721 PMID: 23075143.
3. Zhou P, Yang XL, Wang XG, Hu B, Zhang L, Zhang W, et al. A pneumonia outbreak associated with a new coronavirus of probable bat origin. Nature. 2020; 579(7798):270–3. https://doi.org/10.1038/s41586-020-2012-7 PMID: 32015507.
4. Zhu N, Zhang D, Wang W, Li X, Yang B, Song J, et al. A Novel Coronavirus from Patients with Pneumonia in China, 2019. N Engl J Med. 2020; 382(8):727–33. https://doi.org/10.1056/NEJMoa2001017 PMID: 31978945.
5. Schoeman D, Fielding BC. Coronavirus envelope protein: current knowledge. Virol J. 2019; 16(1):69. https://doi.org/10.1186/s12985-019-1182-0 PMID: 31133031.
6. Masters PS. The molecular biology of coronaviruses. Adv Virus Res. 2006; 66:193–292. https://doi.org/10.1016/S0065-3527(06)66005-3 PMID: 16877062.
7. Wang C, Zheng X, Gai W, Zhao Y, Wang H, Wang H, et al. MERS-CoV virus-like particles produced in insect cells induce specific humoural and cellular imminity in rhesus macaques. Oncotarget. 2017; 8(8):12686–94. https://doi.org/10.18632/oncotarget.8475 PMID: 27050368.
8. Mortola E, Roy P. Efficient assembly and release of SARS coronavirus-like particles by a heterologous expression system. FEBS Lett. 2004; 576(1–2):174–8. https://doi.org/10.1016/j.febslet.2004.09.009 PMID: 15474033.

9. Lindner HA, Fotouhi-Ardakani N, Lytvyn V, Lachance P, Sulea T, Menard R. The papain-like protease from the severe acute respiratory syndrome coronavirus is a deubiquitinating enzyme. J Virol. 2005; 79 (24):15199–208. https://doi.org/10.1128/JVI.79.24.15199-15208.2005 PMID: 16306591.

10. Shimamoto Y, Hattori Y, Kobayashi K, Teruya K, Sanjoh A, Nakagawa A, et al. Fused-ring structure of decahydroisoquinolin as a novel scaffold for SARS 3CL protease inhibitors. Bioorg Med Chem. 2015; 23(4):876–90. https://doi.org/10.1016/j.bmc.2014.12.028 PMID: 25614110.

11. te Velthuis AJ, Arnold JJ, Cameron CE, van den Worm SH, Snijder EJ. The RNA polymerase activity of SARS-coronavirus nsp12 is primer dependent. Nucleic Acids Res. 2010; 38(1):203–14. https://doi.org/10.1093/nar/gkp904 PMID: 19875418.

12. Haagmans BL, Al Dhahiry SH, Reusken CB, Raj VS, Galiano M, Myers R, et al. Middle East respiratory syndrome coronavirus in dromedary camels: an outbreak investigation. Lancet Infect Dis. 2014; 14 (2):140–5. https://doi.org/10.1016/S1473-3099(13)70690-X PMID: 24355866.

13. Memish ZA, Mishra N, Olival KJ, Fagbo SF, Kapoor V, Epstein JH, et al. Middle East respiratory syndrome coronavirus in bats, Saudi Arabia. Emerg Infect Dis. 2013; 19(11):1819–23. https://doi.org/10.3201/eid1911.131172 PMID: 24206838.

14. Ge XY, Li JL, Yang XL, Chmura AA, Zhu G, Epstein JH, et al. Isolation and characterization of a bat SARS-like coronavirus that uses the ACE2 receptor. Nature. 2013; 503(7477):535–8. https://doi.org/10.1038/nature12711 PMID: 24172901.

15. Hu B, Zeng LP, Yang XL, Ge XY, Zhang W, Li B, et al. Discovery of a rich gene pool of bat SARS-related coronaviruses provides new insights into the origin of SARS coronavirus. PLoS Pathog. 2017; 13(11): e1006698. https://doi.org/10.1371/journal.ppat.1006698 PMID: 29190287.

16. Guan Y, Zheng BJ, He YQ, Liu XL, Zhuang ZX, Cheung CL, et al. Isolation and characterization of viruses related to the SARS coronavirus from animals in southern China. Science. 2003; 302 (5643):276–8. https://doi.org/10.1126/science.1087139 PMID: 12958366.

17. Kan B, Wang M, Jing H, Xu H, Jiang X, Yan M, et al. Molecular evolution analysis and geographic investigation of severe acute respiratory syndrome coronavirus-like virus in palm civets at an animal market and on farms. J Virol. 2005; 79(18):11892–900. https://doi.org/10.1128/JVI.79.18.11892-11900.2005 PMID: 16140765.

18. Wang M, Yan M, Xu H, Liang W, Kan B, Zheng B, et al. SARS-CoV infection in a restaurant from palm civet. Emerg Infect Dis. 2005; 11(12):1860–5. https://doi.org/10.3201/eid1112.041293 PMID: 16485471.

19. Duffy S. Why are RNA virus mutation rates so damn high? PLoS Biol. 2018; 16(8):e3000003. https://doi.org/10.1371/journal.pbio.3000003 PMID: 30102691.

20. Salemi M, Fitch WM, Ciccozzi M, Ruiz-Alvarez MJ, Rezza G, Lewis MJ. Severe acute respiratory syndrome coronavirus sequence characteristics and evolutionary rate estimate from maximum likelihood analysis. J Virol. 2004; 78(3):1602–3. https://doi.org/10.1128/jvi.78.3.1602-1603.2004 PMID: 14722315.

21. Cheng J, Randall A, Baldi P. Prediction of protein stability changes for single-site mutations using support vector machines. Proteins. 2006; 62(4):1125–32. https://doi.org/10.1002/prot.20810 PMID: 16372356.

22. Webb B, Sali A. Comparative Protein Structure Modeling Using MODELLER. Curr Protoc Bioinformatics. 2016; 54:5.6.1–5.6.37. https://doi.org/10.1002/cpbi.3 PMID: 27322406.

23. Sanner MF. Python: a programming language for software integration and development. J Mol Graph Model. 1999; 17(1):57–61. PMID: 10660911.

24. Lindorff-Larsen K, Maragakis P, Piana S, Eastwood MP, Dror RO, Shaw DE. Systematic validation of protein force fields against experimental data. PLoS One. 2012; 7(2):e32131. https://doi.org/10.1371/journal.pone.0032131 PMID: 22384157.

25. Beauchamp KA, Lin YS, Das R, Pande VS. Are Protein Force Fields Getting Better? A Systematic Benchmark on 524 Diverse NMR Measurements. J Chem Theory Comput. 2012; 8(4):1409–14. https://doi.org/10.1021/ct2007814 PMID: 22754404.

26. Guvench O, Mallajosyula SS, Raman EP, Hatcher E, Vanommeslaeghe K, Foster TJ, et al. CHARMM additive all-atom force field for carbohydrate derivatives and its utility in polysaccharide and carbohydrate-protein modeling. J Chem Theory Comput. 2011; 7(10):3162–80. https://doi.org/10.1021/ct200328p PMID: 22125473.

27. Stoye J, Blomberg J, Coffin J, Fan H, Hahn B, Neil J. ICTV 9th Report.(2011). International Committee on Taxonomy of Viruses.

28. Coronaviridae Study Group of the International Committee on Taxonomy of V. The species Severe acute respiratory syndrome-related coronavirus: classifying 2019-nCoV and naming it SARS-CoV-2. Nat Microbiol. 2020; 5(4):536–44. https://doi.org/10.1038/s41564-020-0695-z PMID: 32123347.

29. Anand K, Ziebuhr J, Wadhwani P, Mesters JR, Hilgenfeld R. Coronavirus main proteinase (3CLpro) structure: basis for design of anti-SARS drugs. Science. 2003; 300(5626):1763–7. https://doi.org/10.1126/science.1085658 PMID: 12746549.

30. Chen YW, Yiu CB, Wong KY. Prediction of the SARS-CoV-2 (2019-nCoV) 3C-like protease (3CL (pro)) structure: virtual screening reveals velpatasvir, ledipasvir, and other drug repurposing candidates. F1000Res. 2020; 9:129. https://doi.org/10.12688/f1000research.22457.2 PMID: 32194944.

31. Jin Z, Du X, Xu Y, Deng Y, Liu M, Zhao Y, et al. Structure of M(pro) from COVID-19 virus and discovery of its inhibitors. Nature. 2020. https://doi.org/10.1038/s41586-020-2223-y PMID: 32272481.

32. Gao Y, Yan L, Huang Y, Liu F, Zhao Y, Cao L, et al. Structure of the RNA-dependent RNA polymerase from COVID-19 virus. Science. 2020. https://doi.org/10.1126/science.abb7498 PMID: 32277040.

33. Li F, Li W, Farzan M, Harrison SC. Structure of SARS coronavirus spike receptor-binding domain complexed with receptor. Science. 2005; 309(5742):1864–8. https://doi.org/10.1126/science.1116480 PMID: 16166518.

34. Lu R, Zhao X, Li J, Niu P, Yang B, Wu H, et al. Genomic characterisation and epidemiology of 2019 novel coronavirus: implications for virus origins and receptor binding. Lancet. 2020; 395(10224):565–74. https://doi.org/10.1016/S0140-6736(20)30251-8 PMID: 32007145.

35. Baig MH, Sudhakar DR, Kalaiarasan P, Subbarao N, Wadhawa G, Lohani M, et al. Insight into the effect of inhibitor resistant S130G mutant on physico-chemical properties of SHV type beta-lactamase: a molecular dynamics study. PLoS One. 2014; 9(12):e112456. https://doi.org/10.1371/journal.pone.0112456 PMID: 25479359.

36. Woo PC, Lau SK, Lam CS, Lai KK, Huang Y, Lee P, et al. Comparative analysis of complete genome sequences of three avian coronaviruses reveals a novel group 3c coronavirus. J Virol. 2009; 83(2):908–17. https://doi.org/10.1128/JVI.01977-08 PMID: 18971277.

37. Mihindukulasuriya KA, Wu G, St Leger J, Nordhausen RW, Wang D. Identification of a novel coronavirus from a beluga whale by using a panviral microarray. J Virol. 2008; 82(10):5084–8. https://doi.org/10.1128/JVI.02722-07 PMID: 18353961.

38. Woo PC, Huang Y, Lau SK, Yuen KY. Coronavirus genomics and bioinformatics analysis. Viruses. 2010; 2(8):1804–20. https://doi.org/10.3390/v2081803 PMID: 21994708.

39. Angeletti S, Benvenuto D, Bianchi M, Giovanetti M, Pascarella S, Ciccozzi M. COVID-2019: The role of the nsp2 and nsp3 in its pathogenesis. J Med Virol. 2020. https://doi.org/10.1002/jmv.25719 PMID: 32083328.

40. Wang C, Liu Z, Chen Z, Huang X, Xu M, He T, et al. The establishment of reference sequence for SARS-CoV-2 and variation analysis. J Med Virol. 2020. https://doi.org/10.1002/jmv.25762 PMID: 32167180.

41. Tai W, He L, Zhang X, Pu J, Voronin D, Jiang S, et al. Characterization of the receptor-binding domain (RBD) of 2019 novel coronavirus: implication for development of RBD protein as a viral attachment inhibitor and vaccine. Cell Mol Immunol. 2020. https://doi.org/10.1038/s41423-020-0400-4 PMID: 32203189.

42. Du L, He Y, Zhou Y, Liu S, Zheng BJ, Jiang S. The spike protein of SARS-CoV—a target for vaccine and therapeutic development. Nat Rev Microbiol. 2009; 7(3):226–36. https://doi.org/10.1038/nrmicro2090 PMID: 19198616.

43. Kuhn JH, Li W, Choe H, Farzan M. Angiotensin-converting enzyme 2: a functional receptor for SARS coronavirus. Cell Mol Life Sci. 2004; 61(21):2738–43. https://doi.org/10.1007/s00018-004-4242-5 PMID: 15549175.

44. Wang L, Veenstra DL, Radmer RJ, Kollman PA. Can one predict protein stability? An attempt to do so for residue 133 of T4 lysozyme using a combination of free energy derivatives, PROFEC, and free energy perturbation methods. Proteins. 1998; 32(4):438–58. PMID: 9726415.

45. Prabakaran P, Gan J, Feng Y, Zhu Z, Choudhry V, Xiao X, et al. Structure of severe acute respiratory syndrome coronavirus receptor-binding domain complexed with neutralizing antibody. J Biol Chem. 2006; 281(23):15829–36. https://doi.org/10.1074/jbc.M600697200 PMID: 16597622.

46. Babcock GJ, Esshaki DJ, Thomas WD Jr., Ambrosino DM. Amino acids 270 to 510 of the severe acute respiratory syndrome coronavirus spike protein are required for interaction with receptor. J Virol. 2004; 78(9):4552–60. https://doi.org/10.1128/jvi.78.9.4552-4560.2004 PMID: 15078936.

47. Gurung AB, Ali MA, Lee J, Farah MA, Al-Anazi KM. Unravelling lead antiviral phytochemicals for the inhibition of SARS-CoV-2 M(pro) enzyme through in silico approach. Life Sci. 2020; 255:117831. https://doi.org/10.1016/j.lfs.2020.117831 PMID: 32450166.

48. Joshi T, Joshi T, Sharma P, Mathpal S, Pundir H, Bhatt V, et al. In silico screening of natural compounds against COVID-19 by targeting Mpro and ACE2 using molecular docking. Eur Rev Med Pharmacol Sci. 2020; 24(8):4529–36. https://doi.org/10.26355/eurrev_202004_21036 PMID: 32373991.

49. Ruch TR, Machamer CE. The coronavirus E protein: assembly and beyond. Viruses. 2012; 4(3):363–82. https://doi.org/10.3390/v4030363 PMID: 22590676.

50. Nieto-Torres JL, DeDiego ML, Verdia-Baguena C, Jimenez-Guardeno JM, Regla-Nava JA, Fernandez-Delgado R, et al. Severe acute respiratory syndrome coronavirus envelope protein ion channel activity promotes virus fitness and pathogenesis. PLoS Pathog. 2014; 10(5):e1004077. https://doi.org/10.1371/journal.ppat.1004077 PMID: 24788150.

51. Hsin WC, Chang CH, Chang CY, Peng WH, Chien CL, Chang MF, et al. Nucleocapsid protein-dependent assembly of the RNA packaging signal of Middle East respiratory syndrome coronavirus. J Biomed Sci. 2018; 25(1):47. https://doi.org/10.1186/s12929-018-0449-x PMID: 29793506.

52. Masters PS. Coronavirus genomic RNA packaging. Virology. 2019; 537:198–207. https://doi.org/10.1016/j.virol.2019.08.031 PMID: 31505321.

53. Lo YS, Lin SY, Wang SM, Wang CT, Chiu YL, Huang TH, et al. Oligomerization of the carboxyl terminal domain of the human coronavirus 229E nucleocapsid protein. FEBS Lett. 2013; 587(2):120–7. https://doi.org/10.1016/j.febslet.2012.11.016 PMID: 23178926.