Decoding the structure of RNA-dependent RNA-polymerase (RdRp), understanding the ancestral relationship and dispersion pattern of 2019 Wuhan Coronavirus

Abbas Khan  
Shanghai Jiao Tong University

Mazhar Khan  
University of Science and Technology of China (USTC)

Shoaib Saleem  
Quaid-i-Azam University

Zainib Babar  
Shanghai Jiao Tong University

Arif Ali  
Shanghai Jiao Tong University

Abdul Aziz Khan  
Shanghai Jiao Tong University

Zain Sardar  
University of Swat

Fahad Hamayun  
Shanghai Jiao Tong University

Syed Shujait Ali  
University of Swat

Dong-Qing Wei  
 dqwei@sjtu.edu.cn  
Shanghai Jiao Tong University

Research Article

Keywords: SAR-COV-2, RdRp, homology modeling, ancestral relationship, Biogeography

Posted Date: April 29th, 2020

DOI: https://doi.org/10.21203/rs.3.rs-25334/v1

License: © This work is licensed under a Creative Commons Attribution 4.0 International License. Read Full License
Abstract

Most recently, an outbreak of severe pneumonia caused by the infection of 2019-nCoV, a novel coronavirus first identified in Wuhan, China, imposes serious threats to public health. Upon infecting host cells, coronaviruses assemble a multi-subunit RNA-synthesis complex of viral non-structural proteins (nsp) responsible for the replication and transcription of the viral genome. Therefore, the role and inhibition of nsp12 are indispensable. Since there is no crystallographic structure of RdRp is available, so, here, we present the 3-dimensional structure of the 2019-nCoV nsp12 polymerase using a computational approach. nsp12 of 2019-nCoV possesses an architecture common to all viral polymerases as well as a large N-terminal extension. This structure illuminates the assembly of the coronavirus core RNA-synthesis machinery, provides key insights into nsp12 polymerase catalysis and fidelity, and acts as a template for the design of novel antiviral therapeutics. Besides, the experimental structure could reveal the organization in a more sophisticated way. Furthermore, the ancestral state reconstruction suggests the possible evolution of nCoV in Wuhan China and its dispersal to the USA. The result of our analyses postulates the possible dispersal of nCoV from the USA and Shenzhen back to Wuhan. This disclosing of valuable knowledge regarding the 3D structure of 2019-nCoV nsp12 architecture, ancestral relation, and dispersion pattern could help to design effective therapeutic candidates against the coronaviruses and design robust preventive measurements.

Introduction

The viruses of the family Coronaviridae are now notoriously famous for their diseases causing capabilities in birds, humans & mammals. The corona virion typically composed of RNA enclosed in enveloped protein, having glycoprotein spikes, is capable of infecting a broad range of hosts, including humans. Coronaviruses, as the number of variants and diversity increases in this family, based on similarities are classified into four sub-genera, designated as alpha (α), beta (β), gamma (γ) & delta (δ)[1]. So far, the β coronaviruses (CVs) are known to cause infections in humans including common colds and primarily affecting the respiratory system. Bats are associated with the CVs pandemics in the human population, bats harbor the virus and are believed to be immune to the viral infection itself, promoting the mutations that are crucial for the CVs pathogenicity [2]. The spike-like glycoprotein (S), giving the virus its corona like appearance is vital for their pathogenicity and helps them to attach with the host cell surface receptors and also delimits the hosts’ range for the CVs [3].

The CVs genome, ranging from 27 to 32 kilo-basis, are positive-sense single-stranded RNA (+ssRNA) coding for, ORF1a & ORF1b, the poly-proteins involved in RNA polymerization (RNA-dependent RNA-polymerases) (RdRp) and also for modulation of host responses[4,5]. Fatal diseases causing zoonotic strains in this family are severe acute respiratory syndrome (SARS) and the Middle East respiratory syndrome (MERS) [6]. Additionally, there are four more strains, which are reported to be disease-causing in humans, mainly common colds in individuals with immunodeficiency (229E, HKU1, NL63 & OC43)[4].
The 2019-novel-corona-virus (2019-nCoV) that emerged in Wuhan in 2019 belongs to a bat derived Coronaviridae family, that have gained the transmission capability from animals to humans and from human to human, due to which 2019-nCoV became so lethal and caused global emergency[7]. The 2019-nCoV is an enveloped RNA virus with the distinctive corona like shape protein spikes (usually about nine to twelve nanometers) capable of attachment to host cells. The 2019-nCoV potentially causes “novel corona-virus-infected pneumonia” or NCIP, the disease of lower respiratory tract having common cold-like symptoms with fever chest congestion leading to difficulty in breathing[8].

The 2019-nCoV has an 86.9% similarity with the genome of bat-like SARS-CVs and was classified as a distinctive subclade in the subgenus of sarbecovirus having typical β-CVs genome organization[8]. The 2019-nCoV genome, like other CoVs constitutes a 5¢ untranslated region (UTR) replicase-complex Orf1a & Orf1ab followed by protein-encoding genes for the spike (S), membrane (M), envelope (E), nucleic capsid (N) and a 3¢ UTR [9].

The non-structural proteins (nsp) from 1 to 16 of CoVs has a vital role in their replication, while the functions of certain nsps remain elusive. The structural proteins are indispensable for viral assembly and infection, while S protein for spike has distinctive variations and helps in the attachment to the host cell surface proteins[10,11]. The M protein having transmembrane domains binds to the nucleocapsid and shaping the virion [12,13]. The E protein is indispensable for viral pathogenesis and is responsible for virion assembly and budding [14,15]. The N protein comprising of two domains, having the capability of binding with virion genome and nsp-3 protein triggering replicase-transcriptase complex and viral genome encapsulation [16-18].

To elucidate the 3D structure of 2019-nCoV, here we applied the computational modeling approach to give an insight into the domain architecture of nsp12. The classification of the RdRp domain into fingers, palm, thumb, and decoding the conserved motifs in the N-terminal domain and RdRp domains provides a better insight into the exact mechanism. Furthermore, we also exposed the familial relationships and its dispersion pattern. This disclosing of necessary knowledge regarding the 3D structure of 2019-nCoV nsp12 architecture, ancestral relation and dispersion pattern could help to design effective therapeutic candidates against the coronaviruses and design robust preventive measurements.

**Material And Methods**

The primary amino acid sequence of RNA-dependent RNA-polymerase (RdRp) of the recent 2019-nCOV was retrieved from NCBI (https://www.ncbi.nlm.nih.gov/)[19] using accession number QHD43415.1. The RdRp (nsp12) was extracted from the polyprotein using information from UniProt (https://www.uniprot.org/)[20]. The sequence was submitted to the Robetta server (http://new.robetta.org/)[21] for comparative modeling. Sequences submitted to the server are parsed into putative domains for structure prediction, and structural models are produced using either comparative modeling or de novo prediction methods. If a positive match is found using BLAST, PSI-BLAST, FFAS03, or 3D-jury to a protein with the known structure, it is used as a template for comparative modeling. If no
match is found, predictions of the structure are rendered using the Rosetta de novo fragment insertion process.

For structural validation, Ramachandran plot servers were accessed to quantify the quality of the predicted models. Based on the model's scores the top model was selected for further analysis. Pymol was used for structure visualization and analysis. Domains architectures and conserved motifs were highlighted using the Pymol visualization tool. For the electrostatic potential APBS module in Pymol was utilized. Multiple sequence alignment was performed using clustal Omega. These sequences are comprised of porcine delta coronavirus (AMN91620.1), HuCoVNL63 (AVA26872.1), porcine epidemic diarrhea virus (AKJ21970.1), infectious bronchitis virus (AAS00078.1), murine hepatitis virus (YP_209229.2), HuCoV-HKU1 (ARB07596.1), MERS-CoV (ATG84853.1), and SARS-CoV (AAP33696.1) were used for multiple sequence alignment. Online web server the Consurf (https://consurf.tau.ac.il/) was utilized to calculate the conservancy score of each amino acid.

**Phylogeny-based historical biogeography of Coronavirus:**

Information about the distribution area and sequences (obtained from databases in this study) are essential for the inferences of historical biogeography of Coronavirus. Keeping the recent outbreak in mind, we divided the distribution ranges into 9 areas based on the availability of sequences. The distribution areas are A (Australia), B (Thailand), C (Shenzhen and the adjoining regions), D (Shanghai and the adjoining regions), E (Wuhan), F (Beijing), G (Japan), H (USA) and I (Europe). The Sequences of Coronaviruses (RdRp region) were retrieved from the NCBI database and were aligned with the help of Clustal software. This data matrix was used for generating trees file using Beast software. Trees file was uploaded to tree annotator software for obtaining a Maximum clade credibility tree with posterior probability and branch length information. The annotated tree was visualized on Figtree software. The trees and distribution file was used for constructing ancestral states on phylogeny. The ancestral states were obtained by RASP software using BBM (Bayesian Binary Method), and S-DIVA (statistical dispersal and variance analysis) approaches. The output of RASP was edited on adobe Illustrator and Adobe Photoshop.

**Results**

**Structure Prediction and Validation:**

Robetta comparative modeling approach was used to model the 3D structure of RNA-dependent RNA-polymerase (RdRp) of the recent 2019-nCOV. The amino acid sequence of RdRp was extracted from orf1ab polyprotein submitted to NCBI using the accession number QHD43415.1. Five different models were generated from the amino acid sequence. The model generated by the Robetta ab initio modeling server is given in **Figure 1**. All the models were subjected to structural validation using the Ramachandran plot analysis.
The analysis from all the servers revealed that model 1 is the best model. The initial analysis suggested that the 2019-nCoV RdRp structure (sequence) possess 95.77% sequence identity with SAR-CoV (PDB ID: 6NUR) while 86.27% sequence identity with 6NUS[27]. Model 1 was selected for further analyses. All the validation scores are given in **Table 1**. The superimposed structure of model1 with 6NUR and 6NUS is given in **Figure 2**.

**Table 1:** Showing the structure validation scores by different rampage server.

| S. No | Model   | Ramachandran Plot |
|-------|---------|-------------------|
|       |         | Favored Region    | Allowed Region | Outlier Region |
| 1.    | Model 1 | 98.9%             | 1.1%           | 0.0%           |
| 2.    | Model 2 | 98.4%             | 1.6%           | 0.0%           |
| 3.    | Model 3 | 98.3%             | 1.6%           | 0.1%           |
| 4.    | Model 4 | 93.7%             | 4.9%           | 1.4%           |
| 5.    | Model 5 | 97.2%             | 2.6%           | 0.2%           |

**Domains Architecture of RdRp:**

The nsp12 comprises a polymerase domain from 398–931 amino acid position, which assumes a structure similar to other polymerases that resemble a cupped "right hand." The polymerase domain further comprises three more domains, including fingers, palm, and thumb. The fingers domain consists of amino acids from 398–581 and 628–687; a palm domain consists of amino acids from 582–687 and 688–815, while the thumb domain consists of amino acids from 816-931. The N-terminal extension, which is reported to be present in all CoVs, is also a part of 2019-nCoV. This region consists of amino acids from 1-397. All these domains are coloured differently and presented in **Figure 3**.

**RdRp (2019-nCoV) Possess two highly conserved metal-binding sites**

Previous studies reported that the RdRp enzymes have two metal-binding sites coordinated by four residues each. The four residues are enriched by Histidine repeats. In the case of the previously reported cryo-EM structure, these residues include His295, Cys301, Cys306, and Cys310, while the second is in the fingers domain and is coordinated by Cys487, His642, Cys645, and Cys646. Herein, these residues also coordinate the two metal-binding sites. All these eight amino acids, coordinating the metals, are reported to be highly conserved in all the RdRp. Both of these metal-binding sites are distal to known active sites as well as protein-protein and protein–RNA interactions. Thus, rather than being directly involved in enzymatic activity, these ions are expected to be structural components of the folded protein. The involvement of intrinsic zinc ions in nsp12 is reflective of bound zinc atoms in coronavirus nsp3, nsp10, nsp13, and nsp14 and leads to the common use of zinc ions to fold viral replication functional proteins. **Figure 4** showing the two metal-binding sites given below.

**Electrostatic Potential and conservation analysis of RdRp:**
As given in Figure 5, it can be seen that the outer surface of the predicted model carries a mostly negative electrostatic potential. Nevertheless, a strong positive electrostatic potential was reported as the nucleotide triphosphate (NTP) binding site and the polymerase RNA template site. The RNA exit tunnel is comparatively neutral. A relative neutral electrostatic potential at the nsp7 and nsp8 can be observed. The electrostatic potential of the predicted model is given in Figure 5(A).

Furthermore, sequence conservation analysis using sequences from the coronavirus family reveals that the NTP tunnel, template entry, and primer exit tunnels are the most highly conserved surfaces on nsp12. On the other hand, the polymerase active site is also highly conserved site Figure 5(B). The previous study reported that the nsp12 nidovirus-unique N-terminal extension also has a conserved surface, which may reflect an interface site for the N-terminal disordered domain of nsp12 (1–116).

**The N-terminal extension conserved motifs**

The 2019-nCoV nsp12 is ∼931 amino acids long, which is in distinction to the polymerases of the closely related picornaviruses that are typically closer to 500 amino acids. With the long N-terminal extension the C-terminal region comprises of the polymerase domain. Using the sequence analysis of the RNA polymerase domain from the Nidovirales order revealed three conserved sequence motifs within this region. We have named it $A_N$, $B_N$ and $C_N$ as according to the previous study. In the case of 2019-nCoV, these regions correspond to a different position from other CoVs. Here the $A_N$, $B_N$ and $C_N$ are located at (69-89), (102-130) and (202-222) positions. Previous studies reported that Lys73 residue in the $A_N$ is the proposed active site. The conserved regions of the polymerase harboring this nucleotidyltransferase activity were termed NiRAN9. However, the role and mechanisms of this nucleotidyltransferase activity remain to be elucidated. The identified conserved NiRAN motifs ($A_N$, $B_N$ and $C_N$) and their location coloured differently $A_N$(orange), $B_N$(green) and $C_N$(blue) are given in Figure 6.

Previously it has been also reported that the unique N-terminal extension could be divided into two main separate regions. In the 2019-nCoV nsp12, these regions could correspond to (117-249) known as NiRAN while the other could correspond to (250-398) termed as interface region (as shown in figure 2). A study reported that the NiRAN region specifically the $B_N$ and $C_N$ motifs interacts with the fingers and palm domain of the polymerase region. Therefore, we speculate that this motif ($B_N$) implies a functional role tink the N-terminal and the polymerase domain. The previous study reported that seven amino acids from the NiRAN region are strictly conserved across the Nidovirales order. In the cry-EM structure (PDB ID: 6NUR) having sequence identity 95.78% with the 2019-nCoV reported that four residues, among the seven conserved, include Asp126, Gly214, Asp218 and Phe219 are present to play an important role in the interaction.

**nsp12 RNA-dependent RNA polymerase (RdRp) domain & Catalytic mechanism**

The RdRp region has been reported to have a shape like a right hand with subdomains include fingers, palm and thumb. Herein, we defined these subdomains with their respective residues location. The seven
(A-G) conserved motifs (Figure 7) in the RdRp domain of all the viruses are reported to be involved in the NTP binding, template and catalysis. The superimposition of SARs-CoVs (6NUR and 6NUS) (figure 2) onto the structure of 2019-nCoV RdRp revealed important information regarding the active site and catalysis. Upon accessing the active site comprised of motifs A and C and assisted by motifs B and D, single-stranded RNA template threads its way past to motif G. Incoming NTPs will pass through a tunnel at the back into the active site and communicate with Motif F. Motif E at the base of the thumb interacts with the 3’ nucleotide of the primer strand. The primer-template product of RNA synthesis exits the active site through the RNA exit tunnel. This double-stranded RNA product would interact with the N-terminal region of motif G in the fingers domain via the major groove while a helix (2019-nCoV amino acids 798–846) from the thumb domain interacts with the minor groove.

CoVs own the most giant recognized RNA genomes and entail an RNA synthesis complex with the fidelity to replicate their RNA. In the polymerase active site, the incoming NTPs form a base pair with the template RNA while the 2’ and 3’ hydroxyls form hydrogen bonds with the polymerase. In 2019-nCoV nsp12, the 2’ hydroxyl of the incoming NTP is likely to form hydrogen bonds with the three conserved residues of the CoVs family, Thr662 and Asn673 in motif B. Furthermore, Asp605 in motif A is located to interrogate the 3’ hydroxyl through hydrogen bonding.

The hydrophobic side chain in Motif F facilitates the base pairing. The previous study on SARS-CoV reported Val557 to be of significant importance, which is also located at position (Val557) in the case of 2019-nCoV. GS-5734 a nucleoside analog has been reported to impair the RNA synthesis by targeting the RNA synthesis machinery. Partial resistance to GS-5734 posed by two mutations Phe480Leu and Val557Leu has been reported. The previous study reported that the bulkier leucine side chain at amino acid 557A creates higher stringency for base pairing with the templating nucleotide is likely to be established, enabling nsp12 to exclude this analog from its active site. This residue has also previously been labeled in modulating polymerase fidelity. Amino acid positioned at 480th in motif B (Fingers domain) may influence catalytic-related dynamics. This structure is based on the previous structures, thus based on homology modeling. Therefore, the structure is based on comparative analysis and experimental methods such as x-ray crystallography or Cryo-EM methods could verify. Furthermore, looking into the conservancy of RdRp domains we reconstructed the ancestral relationship and used it for the dispersion pattern of the recent outbreak.

**Ancestral state reconstruction:**

The ancestral reconstructions at node I (node ID 219) suggest that Coronavirus evolved in area E (Wuhan, China) or B (Thailand) based on BBM results with 46% and 39% marginal probability values (MP), respectively with 1.00 posterior probability value. Whereas S-DIVA results postulate the origin of this group in area BE (Thailand+ Wuhan China) with 81% MP values, as shown in Table 2. Similarly, the ancestral states on node II, III, and IV indicates that the ancestors of the species belonging to these nodes were evolved in area E with MP values of 82%, 71% and 63% respectively based on BBM inference, on other hand similar results were obtained by S-DIVA approach with slightly changed MP values for node II,
III and IV (74%, 74% and 61%, respectively as shown in table 2). Both BBM and S-DIVA analyses suggest H as an ancestral state on node V with 91% and 99% MP values respectively.

Table 2: Ancestral state reconstructions by BBM and S-DIVA approach. Where AR=ancestral states, MP=marginal probability value and PP=posterior probability value, respectively.

| Nodes | S-DIVA | BBM |
|-------|--------|-----|
|       | AR     | MP (%) | AR | MP (%) | PP |
| I     | BE/BH  | 81/09  | E/B | 46/39  | 1.00 |
| II    | E/EH   | 74/09  | E/EF| 82/4.0 | 1.00 |
| III   | E/EF   | 74/19  | E/EF| 71/13  | 0.55 |
| IV    | EH/E   | 61/23  | E/EH| 63/21  | 0.85 |
| V     | H      | 99     | H/EH| 90/08  | 0.95 |

Table 3: Dispersal events among various regions obtained from S-DIVA and BBM methods.

| Distribution Range | S-DIVA | BBM |
|--------------------|--------|-----|
| Dispersal from | Dispersal to | Within | Dispersal Range | Dispersal from | Dispersal to | Within |
| A                  | 0.00   | 1.00  | 0.00 | A | 0.00 | 1.00 | 0.00 |
| B                  | 0.00   | 1.00  | 1.00 | B | 0.00 | 2.00 | 1.00 |
| C                  | 10.0   | 11.0  | 08.0 | C | 9.00 | 10.00 | 09.0 |
| D                  | 00.0   | 01.0  | 00.0 | D | 00.0 | 01.0 | 00.0 |
| E                  | 13.0   | 4.00  | 12.0 | E | 13.0 | 6.00 | 11.0 |
| F                  | 2.00   | 6.00  | 4.00 | F | 3.00 | 5.00 | 5.00 |
| G                  | 0.00   | 1.00  | 0.00 | G | 0.00 | 1.00 | 0.00 |
| H                  | 9.00   | 6.00  | 49.0 | H | 9.00 | 5.00 | 50.0 |
| I                  | 0.00   | 3.00  | 0.00 | I | 0.00 | 3.00 | 0.00 |

These ancestral reconstructions suggest three central of diversity and expansions in coronaviruses, Wuhan is the primary center of diversity, whereas Shenzhen and USA are the other two centers (Figure 8 and Table 3). Dispersals events occurred from Wuhan to Thailand, USA, Shenzhen, Shanghai and Beijing. From USA dispersal took place to Area B, C, D E and I (Europe). The dispersal from Shenzhen (C) is unique because for very remote dispersal to A (Australia), Japan (G), USA (H) and Europe (I) occurred via this area as shown in Figure 9B.

As far as the recent new coronaviruses are concerned, they are nested within Bat SARs like coronaviruses and probably evolved in E (Wuhan) region. An early expansion occurred from Wuhan to USA and Shenzhen. From the USA, recent dispersals are identified toward Wuhan and Shenzhen. The most of the long-distance dispersal events in 2019-nCoV took place from Shenzhen to USA, Australia, Finland and Japan.

**Discussion**
Here we described the in-silico characterization of 2019-nCoV proteins RdRp and its interaction with the nsp12, nsp8 and nsp7 and how they interact to initiate the RNA synthesis and polymerization of 2019-nCoV genome. A comparative modeling approach using Robetta was used to predict the protein models from the 2019-nCoV genome, retrieved from the NCBI database. Five models were predicted, and after structure, validation using online servers and Ramachandran plot, the best-predicted model was selected and its sequence identity was compared with experimentally verified similar protein model. Further, the domain architecture RdRp was characterized and melded into fingers, palm and thumb design, and its interaction and binding with the RNA for accurate demonstration of the protein-protein and protein–RNA interactions. The position, binding, structure integrity and association of the zinc metals in the protein structure is also sketched. Computational methods are of great importance in determining the structure and function of proteins, drug binding, exploring the resistance mechanism and biocatalysis[28-30].

Our analysis is in accordance with the previously reported SARS-CoV identical protein, where the nsp12 protein interacts with the CoV RNA, and the required stability is provided with the nsps, including nsp7 and nsp8. Further, the nsp12 is also involved in template recognition. Together with the stability in SARS-CoV the nsp8 also plays its role in polymerization. Biochemical confirmations also demonstrated the de-novo synthesis ability of nsp12, nsp8 and nsp7 complex. Herein, the higher identity between the previously reported SAR-CoV, our model works in the same way due to high amino acid conservancy. The highlighted residues, important domains and conserved motifs will help to identify potent inhibitors and help to control the emerging infections related to Coronaviridae family.

The biogeographic pattern based on BBM and S_DIVA analyses indicates the possible origin of the ancestor of coronaviruses in Wuhan, China or Thailand, however the marginal probability value favour the origin in Wuhan, China. The new coronavirus clade is nested within Bat SARS-like coronaviruses and based on the ancestor reconstruction they are evolved in Wuhan China.

Our analyses suggest the dispersal from Wuhan to USA as well as Shenzhen and Hong Kong. Recently based on our analysis dispersal occurred back to Wuhan China from USA, and it is consistent with the recent outbreak in Wuhan and China.

**Conclusion**

In conclusion, this study decoded the important domains and motifs of RNA-dependent RNA polymerase, which is important for viral replication. Also, it provides a basis for designing novel potent inhibitors targeting the RdRp region. Furthermore, we also exposed the familial relationships and its dispersion pattern. Thus, this study is a significant consideration in future strategies against the outbreaks caused by such viruses.

**Declarations**

**Conflict of Interest**
The authors declare no conflict of interest.

**Funding and Acknowledgment**

Dong-Qing Wei is supported by the grants from the Key Research Area Grant 2016YFA0501703 of the Ministry of Science and Technology of China, the National Natural Science Foundation of China (Contract no. 61832019, 61503244), the Natural Science Foundation of Henan Province (162300410060) and Joint Research Funds for Medical and Engineering and Scientific Research at Shanghai Jiao Tong University (YG2017ZD14). The computations were partially performed at the Center for High-Performance Computing, Shanghai Jiao Tong University.

**Authors Contribution**

AK, ZB, MK and SSA conceptualized the study and did the analysis. AA, AAK, FH, ZS wrote the manuscript. DQW is academic supervisor. He supervised the study.

**References**

1. Spaan W, Cavanagh D, Horzinek M (1988) Coronaviruses: structure and genome expression. Journal of General Virology 69 (12):2939-2952

2. Li W, Shi Z, Yu M, Ren W, Smith C, Epstein JH, Wang H, Crameri G, Hu Z, Zhang H (2005) Bats are natural reservoirs of SARS-like coronaviruses. Science 310 (5748):676-679

3. Masters PS (2006) The molecular biology of coronaviruses. Advances in virus research 66:193-292

4. Su S, Wong G, Shi W, Liu J, Lai AC, Zhou J, Liu W, Bi Y, Gao GF (2016) Epidemiology, genetic recombination, and pathogenesis of coronaviruses. Trends in microbiology 24 (6):490-502

5. Khan A, Saleem S, Idrees M, Ali SS, Junaid M, Kaushik AC, Wei D-Q (2018) Allosteric ligands for the pharmacologically important Flavivirus target (NS5) from ZINC database based on pharmacophoric points, free energy calculations and dynamics correlation. Journal of Molecular Graphics and Modelling 82:37-47

6. Cui J, Li F, Shi Z-L (2019) Origin and evolution of pathogenic coronaviruses. Nature reviews Microbiology 17 (3):181-192

7. Zhou P, Yang X-L, Wang X-G, Hu B, Zhang L, Zhang W, Si H-R, Zhu Y, Li B, Huang C-L (2020) A pneumonia outbreak associated with a new coronavirus of probable bat origin. Nature:1-4

8. Lu R, Zhao X, Li J, Niu P, Yang B, Wu H, Wang W, Song H, Huang B, Zhu N (2020) Genomic characterisation and epidemiology of 2019 novel coronavirus: implications for virus origins and receptor binding. The Lancet

9. Dong N, Yang X, Ye L, Chen K, Chan EW-C, Yang M, Chen S (2020) Genomic and protein structure modelling analysis depicts the origin and infectivity of 2019-nCoV, a new coronavirus which caused a pneumonia outbreak in Wuhan, China. bioRxiv
10. Beniac DR, Andonov A, Grudeski E, Booth TF (2006) Architecture of the SARS coronavirus prefusion spike. Nature structural & molecular biology 13 (8):751-752

11. Delmas B, Laude H (1990) Assembly of coronavirus spike protein into trimers and its role in epitope expression. Journal of Virology 64 (11):5367-5375

12. NaL B, Chan C, Kien F, Siu L, Tse J, Chu K, Kam J, Staropoli I, Crescenzo-Chaigne B, Escriou N (2005) Differential maturation and subcellular localization of severe acute respiratory syndrome coronavirus surface proteins S, M and E. Journal of general virology 86 (5):1423-1434

13. Neuman BW, Kiss G, Kunding AH, Bhella D, Baksh MF, Connelly S, Droese B, Klaus JP, Makino S, Sawicki SG (2011) A structural analysis of M protein in coronavirus assembly and morphology. Journal of structural biology 174 (1):11-22

14. DeDiego ML, Álvarez E, Almazán F, Rejas MT, Lamirande E, Roberts A, Shieh W-J, Zaki SR, Subbarao K, Enjuanes L (2007) A severe acute respiratory syndrome coronavirus that lacks the E gene is attenuated in vitro and in vivo. Journal of virology 81 (4):1701-1713

15. Nieto-Torres JL, DeDiego ML, Verdia-Baguena C, Jimenez-Guardeno JM, Regla-Nava JA, Fernandez-Delgado R, Castano-Rodriguez C, Alcaraz A, Torres J, Aguilella VM (2014) Severe acute respiratory syndrome coronavirus envelope protein ion channel activity promotes virus fitness and pathogenesis. PLoS pathogens 10 (5)

16. Fehr AR, Perlman S (2015) Coronaviruses: an overview of their replication and pathogenesis. In: Coronaviruses. Springer, pp 1-23

17. Chang C-k, Sue S-C, Yu T-h, Hsieh C-M, Tsai C-K, Chiang Y-C, Lee S-j, Hsiao H-h, Wu W-J, Chang W-L (2006) Modular organization of SARS coronavirus nucleocapsid protein. Journal of biomedical science 13 (1):59-72

18. Hurst KR, Koetzner CA, Masters PS (2009) Identification of in vivo-interacting domains of the murine coronavirus nucleocapsid protein. Journal of virology 83 (14):7221-7234

19. Coordinators NR (2017) Database resources of the national center for biotechnology information. Nucleic acids research 45 (Database issue):D12

20. UniProt: the universal protein knowledgebase (2017). Nucleic acids research 45 (D1):D158-D169

21. Kim DE, Chivian D, Baker D (2004) Protein structure prediction and analysis using the Robetta server. Nucleic acids research 32 (suppl_2):W526-W531

22. Ramachandran GN (1963) Stereochemistry of polypeptide chain configurations. J Mol Biol 7:95-99

23. DeLano WL (2002) Pymol: An open-source molecular graphics tool. CCP4 Newsletter on protein crystallography 40 (1):82-92

24. Lerner M, Carlson H (2006) APBS plugin for PyMOL. Ann Arbor: University of Michigan

25. Sievers F, Higgins DG (2014) Clustal omega. Current protocols in bioinformatics 48 (1):3.13. 11-13.13. 16

26. Ashkenazy H, Abadi S, Martz E, Chay O, Mayrose I, Pupko T, Ben-Tal N (2016) ConSurf 2016: an improved methodology to estimate and visualize evolutionary conservation in macromolecules.
27. Kirchdoerfer RN, Ward AB (2019) Structure of the SARS-CoV nsp12 polymerase bound to nsp7 and nsp8 co-factors. Nature communications 10 (1):1-9

28. Khan A, Kaushik AC, Ali SS, Ahmad N, Wei D-Q (2019) Deep-learning-based target screening and similarity search for the predicted inhibitors of the pathways in Parkinson's disease. RSC advances 9 (18):10326-10339

29. Khan A, Muhammad J, Li C-D, Saleem S, Humayun F, Shamas S, Ali SS, Babar Z, Wei D-Q (2019) Dynamics insights into the gain of flexibility by Helix-12 in ESR1 as a mechanism of resistance to drugs in breast cancer cell lines. Frontiers in Molecular Biosciences 6:159

30. Wang Y, Khan A, Chandra Kaushik A, Junaid M, Zhang X, Wei D-Q (2019) The systematic modeling studies and free energy calculations of the phenazine compounds as anti-tuberculosis agents. Journal of Biomolecular Structure and Dynamics 37 (15):4051-4069

**Figures**

*Figure 1*

Structure of RNA-dependent RNA polymerase 2019-nCOV. The Robetta server-generated five models.
Figure 2

The superimposed structure of model 1 (yellow) with 6NUR (green) (left) and 6NUS (cyan) (right). The RMSD of each superimposition was reported to be 0.235Å and 0.258Å.

Figure 3
Shows the domain architecture of RdRp from 2019-nCoV. Gray colour (N-terminal extension) (1-397), green colour (palm domain) (582–687 and 688–815), blue colour (Finger domain) (398–581 and 628–687) while yellow colour represents the thumb domain (816–931).

**Figure 4**

The figure shows the two metal-binding sites on the protein structure. Four residues enriched by the histidine amino acid coordinate both metals. The Zinc ions are given as a sphere in blue colour.
Figure 5

(A) showing the electrostatic potential of each residue. (B) Showing the residue conservancy score calculated by Consurf.
Figure 6

Sequence alignment of 2019-nCoV nsp12 with other related coronaviruses. The above panel is showing the aligned conserved motifs (AN, BN and CN) while the 3D structure has been added to highlight these motifs on the structure. The identified conserved NiRAN motifs (AN, BN and CN) and their location coloured differently AN(orange), BN(green) and CN(blue) are shown.
Figure 7

The above figure shows the conserved motifs in the RdRp domain of nsp12. These motifs play important role in the replication of virus’ RNA.
Figure 8

Ancestral state reconstruction of coronavirus by BBM method. Nodes I to V shown on the phylogenetic tree along with posterior probability values are discussed in the text. Legend colour shows the favoured ancestral state on the node. World map was used to indicate the distribution range of coronavirus, which was divided into nine areas from A to I indicated on a map. Note: The designations employed and the presentation of the material on this map do not imply the expression of any opinion whatsoever on the
part of Research Square concerning the legal status of any country, territory, city or area or of its authorities, or concerning the delimitation of its frontiers or boundaries. This map has been provided by the authors.

Figure 9

Dispersal scenarios based on BBM and S-DIVA methods. (a) Indicates dispersal from E (Wuhan) to other areas. (b) Dispersals from Shenzhen to other areas and (c) dispersals from the USA to Wuhan other
regions. Note: The designations employed and the presentation of the material on this map do not imply
the expression of any opinion whatsoever on the part of Research Square concerning the legal status of
any country, territory, city or area or of its authorities, or concerning the delimitation of its frontiers or
boundaries. This map has been provided by the authors.