Extended Data

**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**

Ning Dong1#, Xuemei Yang1#, Lianwei Ye1#, Kaichao Chen1#, Edward Wai-Chi Chan2, Mengsu Yang3, Sheng Chen 1\*

1Department of Infectious Diseases and Public Health, Jockey Club College of Veterinary Medicine and Life Sciences, City University of Hong Kong, Kowloon, Hong Kong

2State Key Lab of Chirosciences, Department of Applied Biology and Chemical Technology, The Hong Kong Polytechnic University, Hung Hom, Kowloon, Hong Kong;

3Department of Biomedical Science, Jockey Club College of Veterinary Medicine and Life Sciences, City University of Hong Kong, Kowloon, Hong Kong

#contribute equally to the work.

\*Corresponding author: Sheng Chen, City University of Hong Kong, Kowloon, Hong Kong;

Email: [shechen@cityu.edu.hk](mailto:shechen@cityu.edu.hk)

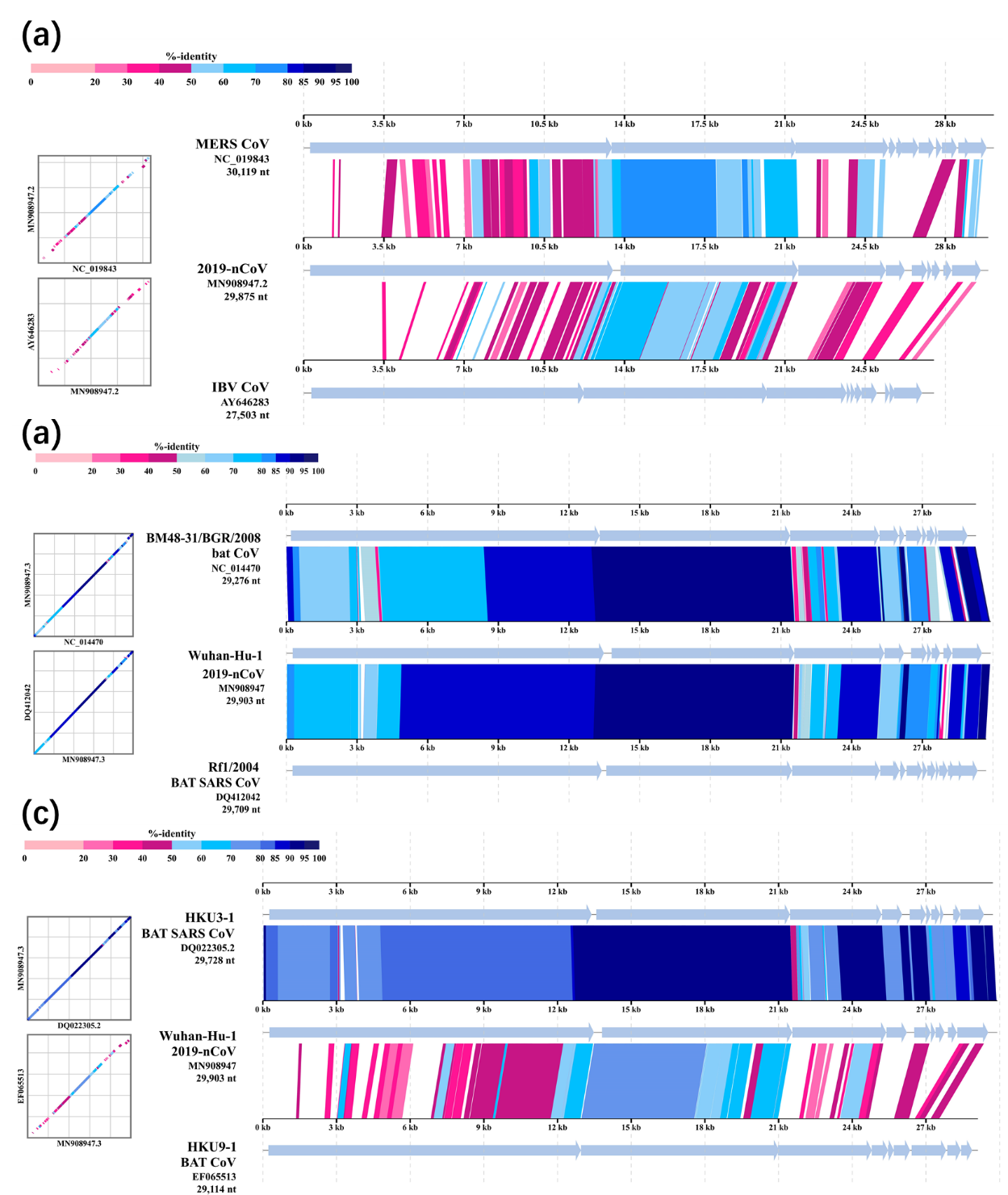
**Keywords:** 2019-nCoV, Genomics, Protein modelling, Origin, Infectivity, Wuhan

**Supplementary Table S1. Assessment of the quality of modeled structures of spike protein and its receptor binding domain (RBD) of different coronavirus using protein structures of the human SARS virus as templates.**

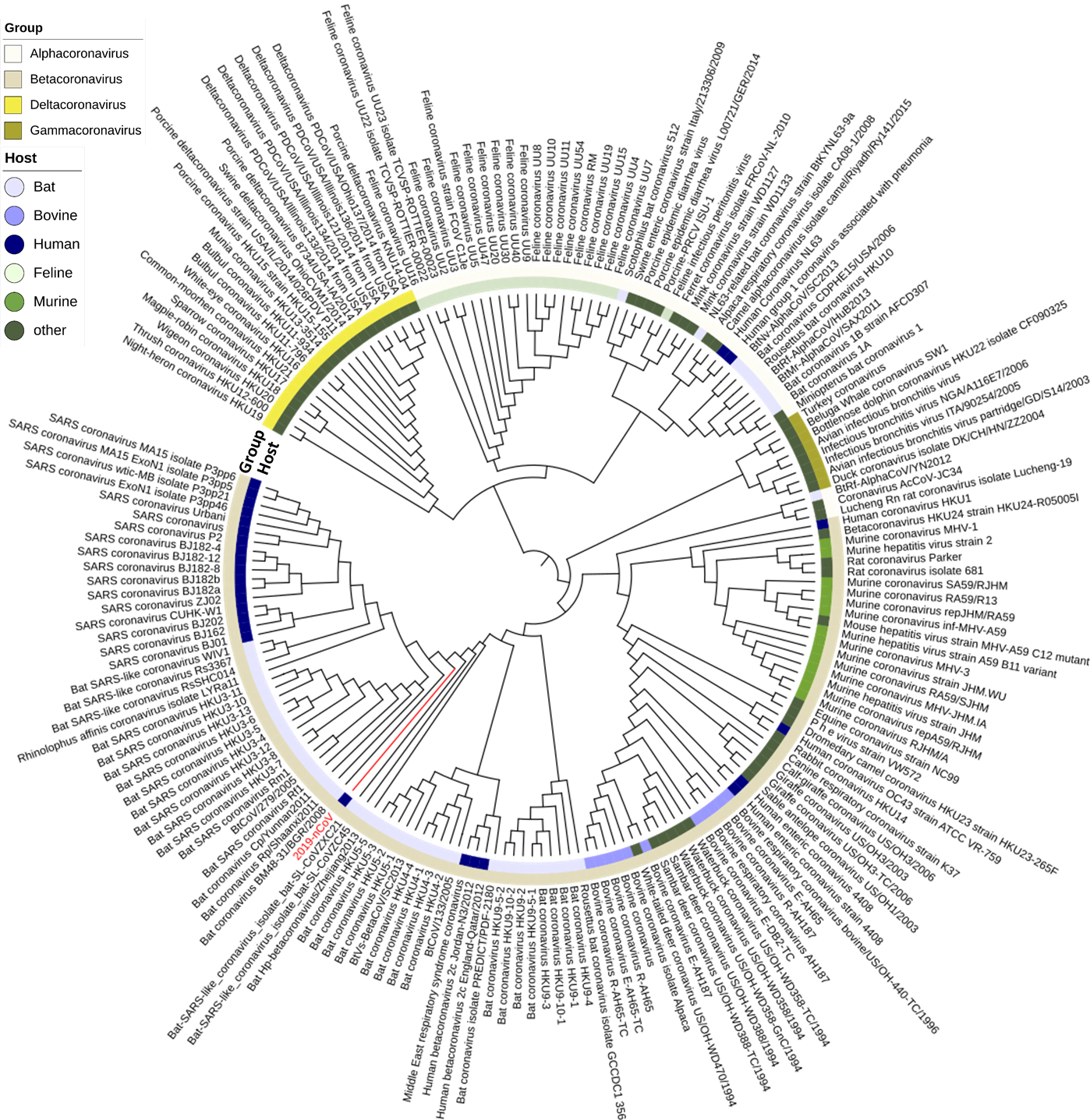
|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **S protein** | **GenBank** | **residues** | **MolProbity Score1** | **Ramachandran Favoured1** | **GMQE2** | **QMEAN2** |
| 2019-nCoV | MN908947 | 1273 | 1.64 | 89.4% | 0.73 | -4.46 |
| BatSARS-likeCoV-WIV1 | KC881007 | 1256 | 1.36 | 90.02% | 0.74 | -3.45 |
| BatSARS-likeCoV | MG772934 | 1245 | 1.64 | 88.96% | 0.73 | -3.91 |
| BatSARSr-CoV | DQ022305 | 1242 | 1.34 | 89.49% | 0.73 | -3.96 |
| BatCoV | JX993987 | 1240 | 1.4 | 90.02% | 0.73 | -3.90 |

1, Structure assessment. MolProbity is all-atom contact analysis based only on properties of the predicted model. Lower numbers indicate better models. Ramachandran favoured indicates energetically favoured regions for backbone dihedral angles against amino acid residues in protein structure. Larger numbers indicate better models.

2, Model evaluation. GMQE (Global Model Quality Estimation) is a quality estimation which combines properties of the target–template alignment and the template search method. The resulting GMQE score is expressed as a number between 0 and 1, larger numbers indicate higher reliability. The QMEAN Z-score provides an estimate of the "degree of nativeness" of the structural features observed in the model on a global scale. QMEAN Z-scores around zero indicate good agreement between the model structure and experimental structures of similar size.



**Supplementary Figure S1. Sequence alignment of 2019-nCoV with Middle East Respiratory Syndrome (MERS) virus: MERS CoV (NC019843) and Avian Infectious Bronchitis (IBV) virus: IBV CoV (AY646283).** The new coronavirus showed very low homology with these two types of coronaviruses.



**Supplementary Figure S2. Phylogenetic analysis of the spike (S) protein of different coronaviruses.** The coronaviruses selected for this analysis are the same as in Figure 1a. The pattern of genetic linkage among sequences of the S protein of various viruses is slightly different from that of the whole genome sequences. In particular, the S protein of 2019-nCoV is at a position further away from the human SARS coronavirus, yet the protein is closer to bat SARS-like coronaviruses than other bat coronaviruses. Different coronavirus groups and the respective hosts are labelled.