

# A database resource for Genome-wide dynamics analysis of Coronaviruses on a historical and global scale

Zhenglin Zhu<sup>1\*</sup>, Kaiwen Meng<sup>2</sup>, Geng Meng<sup>2\*</sup>

1. School of Life Sciences, Chongqing University, Chongqing, China

2. College of Veterinary Medicine, China Agricultural University, Beijing, China

\* Corresponding authors

Zhenglin Zhu, School of Life Sciences, Chongqing University, No.55 Daxuecheng South Road,  
Shapingba, Chongqing, 401331, China.

TEL: (86)23-6512-2686, FAX: (86)23-6512-2689, zhuzl@cqu.edu.cn

Geng Meng, College of Veterinary Medicine, China Agricultural University, Beijing, 100094 China

TEL: (86)10-6273-3466, FAX: (86)10-6273-3466, mg@cau.edu.cn

Keywords: Coronavirus, population genomics, proteomics, database

## Abstract

The recent outbreak of a new zoonotic origin Coronavirus has ring the bell for the potential spread of epidemic Coronavirus crossing the species. With the urgent needs to assist the control of the Coronavirus spread and to provide valuable scientific information, we developed a coronavirus database (CoVdb), an online genomics and proteomics analysis platform. Based on public available coronavirus genomic information, the database annotates the genome of every strain and identifies 780 possible ORFs of all strains available in Genebank. In addition, the comprehensive evaluation of all the published genomes of Coronavirus strains, including population genetics analysis, functional analysis and structural analysis on a historical and global scale were presented in the CoVdb. In the database, the researcher can easily obtain the basic information of a Coronavirus gene with the distribution of the gene among strains, conserved or high mutation regions, possible subcellular location and topology of the gene. Moreover, sliding windows for population genetics analysis results is provided, thereby facilitating genetics and evolutionary analysis at the genomic level. CoVdb can be accessed freely at <http://covdb.popgenetics.net>.

**KEYWORDS:** Coronavirus; population genomics; database

## Introduction

Coronaviridae is a group of positive-sense, single-strand RNA viruses with a likely ancient origin, and human Coronavirus repeatedly emerged during the past hundred years<sup>1</sup>. Coronaviruses are classified into four distinct genera: alpha and beta Coronavirus mainly infect mammals, whereas gamma and delta Coronavirus more circulate in avian hosts<sup>2</sup>. As a potential dangerous zoonotic disease, the previous outbreaks of respiratory syndrome-related Coronavirus (SARS-CoV) and Middle East respiratory syndrome-related Coronavirus (MERS-CoV) have plagued the general public and researchers in the past years<sup>3</sup>. Recently, a new Coronavirus (2019-nCoV), which may originated from wild animals, first identified in Wuhan city, China. Till now it has been resulted in more than fourteen thousand confirmed infections in China<sup>4,5</sup> with the cases number is still increasing. Although we have knowledge and experience in the virology, diagnosis, clinical characteristics, and other aspects related to SARS-CoV and MERS-CoV, there are many unanswered questions about the new emerging 2019-nCoV. The new Coronavirus outbreak in China strongly reminds the continued threat of zoonotic diseases caused by Coronavirus to global health security. Sharing experience and knowledge from across disciplines in the historical and global scale should provide valuable scientific knowledge to fight against the threat of Coronavirus.

The aim of the construction of CoVdb is to provide Coronavirus knowledge, to contribute to global Coronavirus research, especially for the investigation of the emerging 2019-nCoV. For previous works, ViPR<sup>6</sup> and ViralZone<sup>7</sup> are general data resources and is lack of analysis tools in population genomics and evolution. In contrast, CoVdb is specially designed for Coronavirus. It combines, compares, and annotates all the published Coronavirus genomes up to date<sup>8,9,10,11,12,13,14,15,16,17,18,19</sup>. The new developed database provides the convenience for the identification of gene function and identity among the Coronaviridae genomes. CoVdb provides information on subcellular location, functions, proteins topology, as well as population level thorough analysis results. We will be dedicated to keep updating all

the genomic information and optimizing the database.

## Result and discussions

### Data and information

CoVdb extensively collects published Coronavirus genome data (Table S1), including 104 strains are of 780 possible ORFs. In average, there are 5-14 possible protein coding genes in each strain. Although the structure of Coronavirus (Figure 1) is not complex, we still performed a subcellular localization analysis of the Coronavirus genes to predict their roles in the infection process with addition to the previous research. Base on prediction only, 27% of the proteins are predicted to be located in the host nucleus or host cytoplasm (Figure S1). 32% of the Coronavirus genes are predict to be membrane proteins. In gene ontology, Coronavirus genes enrich in association to the membrane (Figure S2). Moreover, based on the population genetics test analysis, we found that 2242 regions are of a significantly low Tajima's  $D^{20}$  and significantly high composite likelihood ratio,  $CLR^{21, 22}$  (Rank Test,  $P$ -value $<0.05$ ), indicating that these genes were recently possibly under positive selection (Table S2). These regions deposited in 400 genes, and most of the regions are located at ORF1 of the virus genome. Among all these 2242 regions, 98 are located in the noncoding region and 30 are located in the coding regions of non-structural proteins. We also found that most high CLR regions (1452, 64.8%) are at the protein ORF1ab (Table S3). These positive selection regions may involve in the adaptive mutations related to the virus replication and infection. The results are informative for future Coronavirus research and epidemiological survey.

We identified 322 Coronavirus gene clusters from the genomes of 104 strains (for details, see Materials and Methods). Using the genomes documented in CoVdb, we generated the phylogenetic tree of Coronavirus (Figure 2). The tree indicates that 2019-nCoV is of close relationship with SARS-CoV and may arrive from bat. Coronavirus strains extracted from human are always surrounded by strains extracted from bats, indicating bats may be the main source leading to the infection of Coronavirus in human world (Figure S3).

## Interface and main functions

The genome browser page follows a style with analysis tracks (Pi, Theta<sup>23</sup>, Tajima's D<sup>20</sup>, and CLR<sup>21, 22</sup>) listed following gene segments (Figure S4). CoVdb are equipped with other general genome browser tools like UCSC genome browser<sup>24</sup>. In addition to basic information, CoVdb show gene information mainly in cell, protein structure and evolutionary signatures.

The search engine in CoVdb is powerful and supports fuzzy search, BLAT and Blast. CoVdb also allows to search by cell location, function, evolutionary test parameters and protein structure parameters (Figure 3). To facilitate personalized gene list analysis, CoVdb provides gene links through inputting a range of genomic locations or a list of genes' accession numbers.

## Conclusion

Dedicated to assist the researcher to combat the pandemic of 2019-nCoV, and to provide a more specialized platform for Coronavirus, we comprehensively gathered data and systematically constructed the Coronavirus Database, CoVdb. With the help of this database, we have successfully developed a novel tool (unpublished) to detect 2019-nCoV and the program is in the process to be put into production. Researchers can conveniently retrieve Coronavirus genomic and gene information from CoVdb. Hopefully, this database will play more important roles in fighting against the infection of Coronavirus.

## Materials and methods

### Gene Information, subcellular localization and topology prediction

Coronavirus sequences and annotations are downloaded from the NCBI genome database. For strains without ORF annotations, we reannotated the genomes through mapping known Coronavirus genes to the genome, requiring identity > 50% and coverage > 80%. Then, we verified the quality of these proteins by known proteins. We kept predicted proteins with both identity and coverage higher than 0.5. According to previous homologous gene identification methods<sup>25, 26</sup>, we also we performed pairwise alignments for all protein sequence, using CD-HIT<sup>27</sup> with the parameters of identity > 90% and coverage > 70%. In this way Coronavirus proteins are clustered into 322 unified gene clusters.

We wrote PERL scripts to automatically BLAST Coronavirus protein sequence against the UniProt protein database<sup>28</sup> and took the hits with the highest scores as the best matches requiring E-value <0.05. Using the accession number of matched UniProt proteins, we retrieved detailed proteomics information from UniProt. In the same way, we also acquired protein 3D structure information from PDB database<sup>29, 30</sup>.

We did the subcellular localization prediction of all Coronavirus genes using an online tool MSLVP<sup>31</sup>. We used TMHMM 2.0<sup>32</sup> to predict the transmembrane helices within protein sequences, and converted output images into PNG format by Magick ([www.imagemagick.org](http://www.imagemagick.org)).

### Evolutionary analysis

We utilized CUDA clustalW<sup>33</sup> to perform a whole genome alignment of all Coronavirus genomes. The results are used to built a phylogenetic tree by FastTree 2.1<sup>34</sup>. We did genomic level alignment by LASTZ<sup>35</sup>, did sequence level alignment by MUSCLE<sup>36, 37</sup> and made phylogenetic trees by FastTree 2.1<sup>34</sup>.

To detect selection signals, we did sliding widow analysis for each genome (window=200 bp, step=50 bp) and did post analysis by VariScan 2.0<sup>38, 39</sup> as well as SweepFinder2<sup>40</sup>. For each gene, we used the median of population genetics test statistics as the corresponding value.

### **The building of CoVdb**

The web interface of CoVdb is on the basis of SWAV<sup>41</sup>. CoVdb also incorporates MSAViewer<sup>42</sup> to display multiple alignments and phylotree.js<sup>43</sup> to show the phylogenetic trees. To fit the requirement to display virus data, we made changes in these two softwares, such as changing parameters to fit virus's dense gene arrangements and adding links within diagrams. The search engine is written by PHP integrated with SQL, BLAT<sup>44</sup> and NCBI BLAST<sup>45</sup>.



## **Data availability**

All CoVdb data are publicly and freely accessible at <http://covdb.popgenetics.net>. Feedback on any aspect of the CoVdb and discussions of Cononavirus gene annotations are welcome by email to [zhuzl@cqu.edu.cn](mailto:zhuzl@cqu.edu.cn) or [mg@cau.edu.cn](mailto:mg@cau.edu.cn).

## **Author contributions**

Z.Z. developed the web interface of the database, collected and compiled the data, K.M. performed the analysis. Z.Z. and G.M. conceived the idea, coordinated the project and wrote the manuscript.

## **Acknowledgments**

This work was supported by grants from the National Key Research and Development Program (2019YFC1604600), the National Natural Science Foundation of China (31200941), the Fundamental Research Funds for the Central Universities (106112016CDJXY290002), the National Natural Science Foundation of HeBei province (19226631D).

## Figure Legends

Figure 1. A diagram displaying the general structure of Coronavirus.

Figure 2. The phylogenetic tree built by Coronavirus genomes in CoVdb. Links with Bootstrap Likelihood Value = 1 are colored by blue and the ones with value > 0.5 is colored by green. The names of 7 major Coronavirus strains are enlarged. They are MERS, SARS, 2019-nCoV, OC43, HKU1, NL63 and 229E. Different hosts from which virus is collected are marked by cycles with different colors.





Figure 3. The main page of CoVdb.

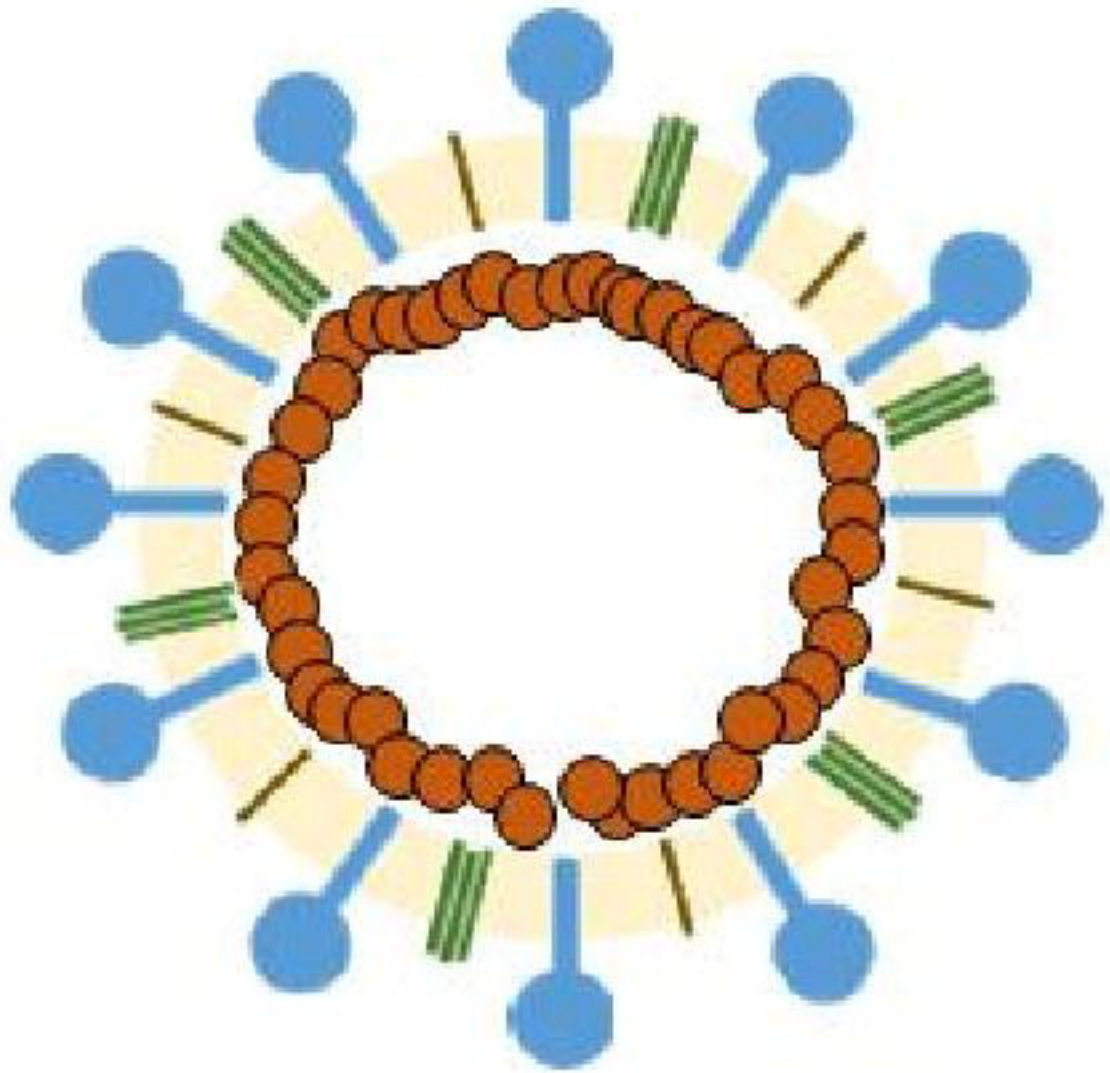
## References

1. Forni D, Cagliani R, Clerici M, Sironi M. Molecular Evolution of Human Coronavirus Genomes. *Trends Microbiol* **25**, 35-48 (2017).
2. Wertheim JO, Chu DK, Peiris JS, Kosakovsky Pond SL, Poon LL. A case for the ancient origin of coronaviruses. *J Virol* **87**, 7039-7045 (2013).
3. de Wit E, van Doremalen N, Falzarano D, Munster VJ. SARS and MERS: recent insights into emerging coronaviruses. *Nat Rev Microbiol* **14**, 523-534 (2016).
4. Lu H, Stratton CW, Tang YW. Outbreak of Pneumonia of Unknown Etiology in Wuhan China: the Mystery and the Miracle. *J Med Virol*, (2020).
5. Hui DS, *et al.* The continuing 2019-nCoV epidemic threat of novel coronaviruses to global health - The latest 2019 novel coronavirus outbreak in Wuhan, China. *Int J Infect Dis* **91**, 264-266 (2020).
6. Pickett BE, *et al.* Virus pathogen database and analysis resource (ViPR): a comprehensive bioinformatics database and analysis resource for the coronavirus research community. *Viruses* **4**, 3209-3226 (2012).
7. Hulo C, *et al.* ViralZone: a knowledge resource to understand virus diversity. *Nucleic Acids Res* **39**, D576-582 (2011).
8. Bourns ME, Brown TD, Foulds IJ, Green PF, Tomley FM, Binns MM. Completion of the sequence of the genome of the coronavirus avian infectious bronchitis virus. *J Gen Virol* **68** ( Pt 1), 57-77 (1987).
9. Coley SE, *et al.* Recombinant mouse hepatitis virus strain A59 from cloned, full-length cDNA replicates to high titers in vitro and is fully pathogenic in vivo. *J Virol* **79**, 3097-3106 (2005).
10. St-Jean JR, Jacomy H, Desforgues M, Vabret A, Freymuth F, Talbot PJ. Human respiratory coronavirus OC43: genetic stability and neuroinvasion. *J Virol* **78**, 8824-8834 (2004).
11. Chouljenko VN, Lin XQ, Storz J, Kousoulas KG, Gorbalenya AE. Comparison of genomic and predicted amino acid sequences of respiratory and enteric bovine coronaviruses isolated from the same animal with fatal shipping pneumonia. *J Gen Virol* **82**, 2927-2933 (2001).
12. van Boheemen S, *et al.* Genomic characterization of a newly discovered coronavirus associated with acute respiratory distress syndrome in humans. *mBio* **3**, (2012).
13. Vlasova AN, Halpin R, Wang S, Ghedin E, Spiro DJ, Saif LJ. Molecular characterization of a new species in the genus Alphacoronavirus associated with mink epizootic catarrhal gastroenteritis. *J Gen Virol* **92**, 1369-1379 (2011).
14. Marra MA, *et al.* The Genome sequence of the SARS-associated coronavirus. *Science* **300**, 1399-1404 (2003).
15. Woo PC, *et al.* Characterization and complete genome sequence of a novel coronavirus, coronavirus HKU1, from patients with pneumonia. *J Virol* **79**, 884-895 (2005).
16. Tang XC, *et al.* Prevalence and genetic diversity of coronaviruses in bats from China. *J Virol* **80**, 7481-7490 (2006).
17. Lau SK, *et al.* Complete genome sequence of bat coronavirus HKU2 from Chinese horseshoe bats revealed a much smaller spike gene with a different evolutionary lineage from the rest of the genome. *Virology* **367**, 428-439 (2007).
18. Chu DK, Peiris JS, Chen H, Guan Y, Poon LL. Genomic characterizations of bat coronaviruses (1A, 1B and HKU8) and evidence for co-infections in *Miniopterus* bats. *J Gen Virol* **89**,

- 1282-1287 (2008).
19. Woo PC, *et al.* Comparative analysis of twelve genomes of three novel group 2c and group 2d coronaviruses reveals unique group and subgroup features. *J Virol* **81**, 1574-1585 (2007).
  20. Tajima F. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* **123**, 585-595 (1989).
  21. Nielsen R, Williamson S, Kim Y, Hubisz MJ, Clark AG, Bustamante C. Genomic scans for selective sweeps using SNP data. *Genome Res* **15**, 1566-1575 (2005).
  22. Zhu L, Bustamante CD. A composite-likelihood approach for detecting directional selection from DNA sequence data. *Genetics* **170**, 1411-1421 (2005).
  23. Watterson GA. On the number of segregating sites in genetical models without recombination. *Theor Popul Biol* **7**, 256-276 (1975).
  24. Karolchik D, Hinrichs, A.S., Kent, W.J. *The UCSC Genome Browser* (2007).
  25. Yue H, *et al.* Genome-Wide Identification and Expression Analysis of the HD-Zip Gene Family in Wheat (*Triticum aestivum* L.). *Genes (Basel)* **9**, (2018).
  26. She R, Chu JS, Wang K, Pei J, Chen N. GenBlastA: enabling BLAST to identify homologous gene sequences. *Genome Res* **19**, 143-149 (2009).
  27. Li W, Godzik A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* **22**, 1658-1659 (2006).
  28. Patient S, Wieser D, Kleen M, Kretschmann E, Jesus Martin M, Apweiler R. UniProtJAPI: a remote API for accessing UniProt data. *Bioinformatics* **24**, 1321-1322 (2008).
  29. Berman HM, *et al.* The Protein Data Bank. *Nucleic Acids Res* **28**, 235-242 (2000).
  30. Burley SK, *et al.* RCSB Protein Data Bank: biological macromolecular structures enabling research and education in fundamental biology, biomedicine, biotechnology and energy. *Nucleic Acids Res* **47**, D464-D474 (2019).
  31. Thakur A, Rajput A, Kumar M. MSLVP: prediction of multiple subcellular localization of viral proteins using a support vector machine. *Mol Biosyst* **12**, 2572-2586 (2016).
  32. Krogh A, Larsson B, von Heijne G, Sonnhammer EL. Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *J Mol Biol* **305**, 567-580 (2001).
  33. Hung CL, Lin YS, Lin CY, Chung YC, Chung YF. CUDA ClustalW: An efficient parallel algorithm for progressive multiple sequence alignment on Multi-GPUs. *Comput Biol Chem* **58**, 62-68 (2015).
  34. Price MN, Dehal PS, Arkin AP. FastTree 2--approximately maximum-likelihood trees for large alignments. *PLoS One* **5**, e9490 (2010).
  35. Harris RS. Improved pairwise alignment of genomic DNA. (ed<sup>^</sup>(eds). Pennsylvania State University (2007).
  36. Edgar RC. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics* **5**, 113 (2004).
  37. Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res* **32**, 1792-1797 (2004).
  38. Hutter S, Vilella AJ, Rozas J. Genome-wide DNA polymorphism analyses using VariScan. *BMC Bioinformatics* **7**, 409 (2006).
  39. Vilella AJ, Blanco-Garcia A, Hutter S, Rozas J. VariScan: Analysis of evolutionary patterns from large-scale DNA sequence polymorphism data. *Bioinformatics* **21**, 2791-2793 (2005).
  40. DeGiorgio M, Huber CD, Hubisz MJ, Hellmann I, Nielsen R. SweepFinder2: increased

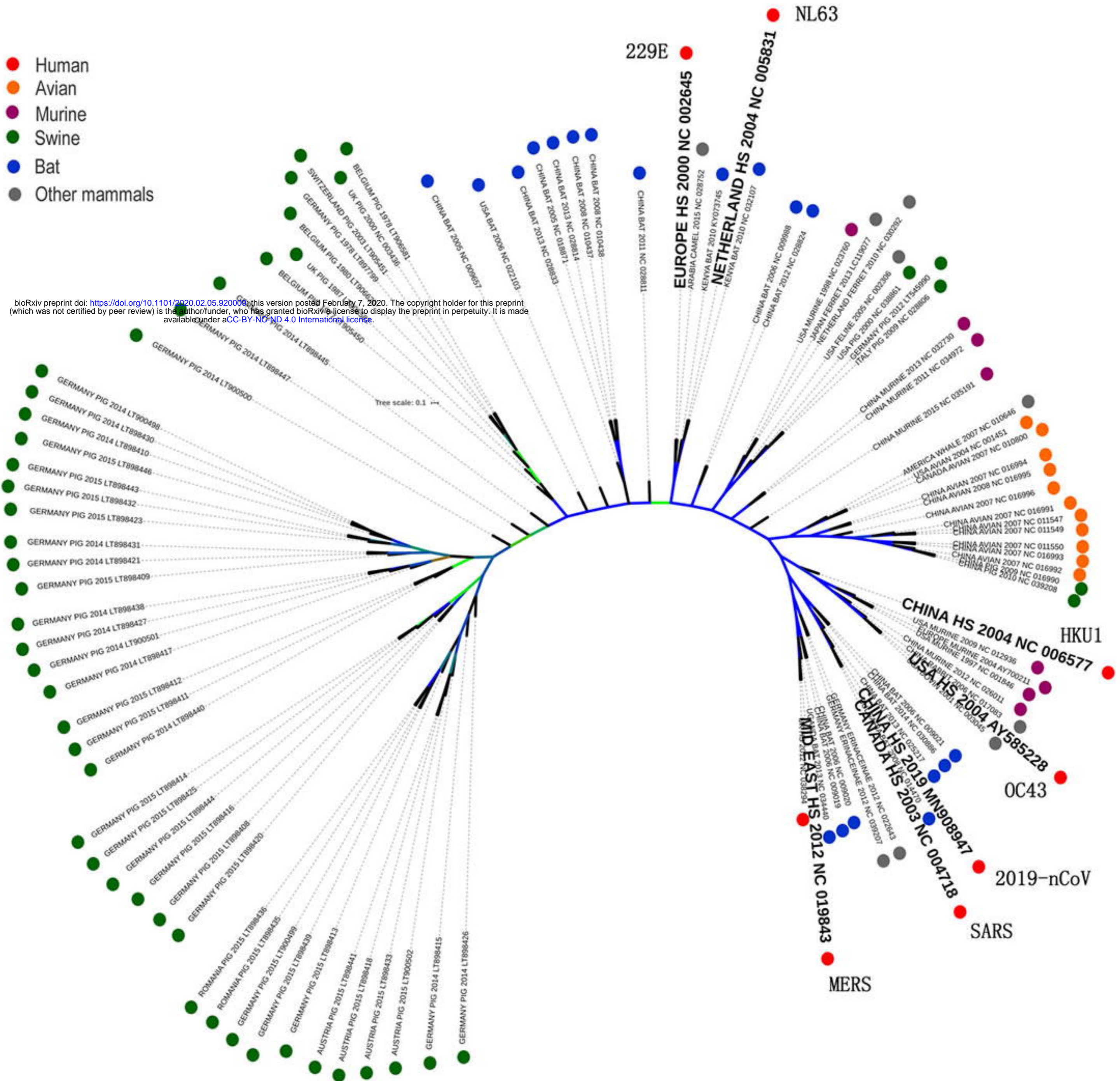
- sensitivity, robustness and flexibility. *Bioinformatics* **32**, 1895-1897 (2016).
41. Zhu Z, Wang Y, Zhou X, Yang L, Meng G, Zhang Z. SWAV: a web-based visualization browser for sliding window analysis. *Sci Rep* **10**, 149 (2020).
  42. Yachdav G, *et al.* MSAViewer: interactive JavaScript visualization of multiple sequence alignments. *Bioinformatics* **32**, 3501-3503 (2016).
  43. Shank SD, Weaver S, Kosakovsky Pond SL. phylotree.js - a JavaScript library for application development and interactive data visualization in phylogenetics. *BMC Bioinformatics* **19**, 276 (2018).
  44. Kent WJ. BLAT--the BLAST-like alignment tool. *Genome Res* **12**, 656-664 (2002).
  45. Johnson M, Zaretskaya I, Raytselis Y, Merezuk Y, McGinnis S, Madden TL. NCBI BLAST: a better web interface. *Nucleic Acids Res* **36**, W5-9 (2008).

-  S protein
-  M protein
-  E protein
-  N protein



- Human
- Avian
- Murine
- Swine
- Bat
- Other mammals

bioRxiv preprint doi: <https://doi.org/10.1101/2020.02.05.920009>; this version posted February 7, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.



bioRxiv preprint doi: <https://doi.org/10.1101/2020.02.05.920009>; this version posted February 7, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

e.g. orflab or AAT84351.1 or hemagglutinin-esterase

SEARCH

BLAT/BLAST

Cell Location

Function

Evolution

Protein Structure

Gene Clusters

