The genomic recombination events may reveal the evolution of coronavirus and the origination of 2019-nCoV

CURRENT STATUS: UNDER REVIEW

nature research

Zhenglin Zhu
School of Life Sciences, Chongqing University
zhuzl@cqu.edu.cn*Corresponding Author*

Kaiwen Meng
College of Veterinary Medicine, China Agricultural University

Geng Meng
College of Veterinary Medicine, China Agricultural University
mg@cau.edu.cn*Corresponding Author*

DOI:
10.21203/rs.3.rs-21488/v1

SUBJECT AREAS
Evolutionary Genetics

KEYWORDS
Coronavirus, genomic recombination, evolution, origination, 2019-nCoV

Abstract

To trace the evolution of coronavirus and reveal the possible origination of the novel pneumonia coronavirus (2019-nCoV), we collected and thoroughly analyzed 2966 publicly available coronavirus genomes, including 182 2019-nCoVs strains. We observed 3 independent recombination events with statistical significance between some isolates from bats and pangolins. In consistence with previous records, we also detected the putative recombination between Bat-CoV-RaTG13 and Pangolin-CoV-2019 covering the receptor bind domain (RBD) of the spike glycoprotein (S protein), which may lead to the origination of 2019-nCoV. Population genetic analyses give estimations indicating that the recombinant region around RBD is possibly undergoing directional evolution. This may result to the adaption of the virus to be infectious in hosts. Not surprisingly, we find that the S protein of coronavirus keeps high diversity among bat isolates, which may provide a genetic pool for the origination of 2019-nCoV.

Introduction

The novel pneumonia coronavirus (2019-nCoV), after the firstly identification in Wuhan, China 1,2, has become pandemic worldwide. Up to now there have been more than 180 thousands confirmed novel coronavirus pneumonia (NCP) cases around the world. For the control and prevention of the disease, efforts have been made to tracing the origins of 2019-nCoV. In the publication of the first genome of 2019-nCoV, bats was considered as the original host of this virus 3. Bat-CoV-RaTG13, a bat coronavirus isolated from *Rhinolophus affinis,* is 96% identical to 2019-nCoV at the whole genome level 4. A pangolin isolate Pangolin-CoV–2019 shares only 91.02% identity in whole genome level to 2019-nCoV, but shows higher sequence identity in the spike glycoprotein (S protein, 97.5 %) coding sequence than Bat-CoV-RaTG13 5. Therefore Pangolin was considered as a potential intermediate host of 2019-nCoV 6–8. It is reported that the receptor binding domain (RBD) of the S protein in 2019-nCoV might be resulted from a recombination event between Bat-CoV-RaTG13 virus and Pangolin-CoV–2019 6,7,9. The RBD-ACE2 binding free energy for 2019-nCoV is significantly lower than that for SARS 10,11, which partially explained the highly infectious activity of the 2019-nCoV. Thus, genomic recombination may be closely related to the origination of nCoV–2019. Statistic analyses of the

genomic recombination between pangolin coronavirus and bat coronavirus should be important for tracing 2019-nCoV's origins. The subsequent evolution of the recombinants is scientifically interesting through in depth analysis in population level for more coronavirus strains.

For the reasons described above, we scanned for available documented coronavirus genomes 12–23 and specially examined possible recombination between 2019-nCoV and coronavirus closely related to 2019-nCoV according to coronavirus' genomic phylogenetic tree 24. For the detection of selection in recombinants, we performed population genetic analyses extensively. Our results depict a possibility for the origination of 2019-nCoV.

Results And Discussion

The multiple alignment of 2966 coronavirus genome sequences has been performed and tried to identify if there is recombination between bat and pangolin coronavirus. In total, we identified 3 independent recombinants. Each of them has evidences from at least six statistic tests (*P-value<0.05)* (Table 1). We have validated the three recombinants by generating their own phylogenetic trees (Figure 1) and pairwise identity plots (Figure S1). Two of three recombination events renovate the structure of two different pangolin coronavirus isolates, Pangolin-CoV–2017 and Pangolin-CoV–2019. A 1260 bp segment in Pangolin-CoV–2017 and a 1182 bp segment in Pangolin-CoV–2019 are possible recombinants from bat isolates (bat-SL-CoVZC45 or bat-SL-CoVZXC21), inferring that the exchanges of genetic materials between coronavirus from bats and coronavirus from pangolins are not rare. One of these two recombinants is within the ORF1 region and the other one is spanning the 3' end of ORF1 and 5' beginning of the S protein (Figure 2A).

Our analysis once again verified that a 228bp long sequence within the S protein (Figure 2A) in 2019-nCoV is of high possibility to be resulted from recombination between Bat-CoV-RaTG13 and Pangolin-CoV–2019 (Table 1, Figure 1D, Figure S1C, S2), although the 2019-nCoV is not isolated and identified from bat or pangolin yet. In whole genome level, Bat-CoV-RaTG13 shows higher identity to 2019-nCoV than Pangolin-CoV–2019. Our analysis suggest the high possibility that 2019-nCoV originated from a bat coronavirus after acquiring a recombinant sequence at the S protein from a pangolin coronavirus (Figure 2B). The S protein recombinant sequence encodes a 76 AA long peptide and locates at the

receptor binding domain (RBD), which may influence the host preference of the virus. This recombination event may play a key role in the origination of 2019-nCoV.

We observed that there is a peak value at the S protein recombinant in Fixation Index (Fst) calculated between human and bat coronavirus. So do that in Fst between human and other hosts, such as camel or cow (Figure 2C, Figure S3). The rise in

differentiation reflected from Fst inferred that the S protein recombination is a usual event and may be important for the coronavirus' adaption to different hosts. We did not observe obvious variation in composite likelihoods (CLR) or Tajima's D within the S protein recombinant among all the 2019-nCoV strains. One explanation for these is that the RBD region is highly conserved for 2019-nCoV. We observed CLR peaks surrounding the S protein recombinant for SARS-CoVs. With more 2019-nCoV samples are sequenced and analyzed, the CLR curve for 2019-nCoV may be changed. There is a sharp decrease of diversity (Pi) in RBD for human, camel or cow isolates. In contrast, for bat coronavirus, Pi values in RBD are high and there is a Pi peak at the S protein recombinant (Figure 2C). Meanwhile, bats are of a higher Tajima's D (–0.27 in median, Wilcoxon rank sum test, *P-value* = 3.428e–09) than that for human coronavirus (–0.73 in median) in the RBD region. These indicated the RBD sequence of the coronavirus isolate from bats kept high diversity, which may provide a genetic pool for the evolution of coronavirus towards 2019-nCoV. It may also explain the fact that most of the human coronavirus originate from bats.

Furthermore, we observed a CLR peak at the ORF1 recombinant for 20190-nCoV (Figure S4). We also observed a CLR peak at the boundary recombinant for human isolates (Figure S5). There is also an Fst peak between human and bat coronavirus at the boundary recombinant (Figure S5). Because of the lack of enough samples, we could not identify selection signals in pangolin coronavirus.

Methods

We collected genomic sequences of coronavirus from NCBI, GISAID (www.gisaid.org) and CoVdb 24. To speed up the progress, we performed whole genome alignments by CUDA ClustalW 25. We did recombination detection by RDP4 26, which used RDP 27, GENECONV 28, Bootscan 29, Maxchi 30, Chimaera 31, SiSscan 32 and 3Seq 33 as test methods for recombinants. We used MEGA 34 to do

local alignments and to build phylogenetic trees. We used online tools in CoVdb 24 to do population genetics analysis. In the platform, Pi 35 and Tajima's D 36 were calculated by VariScan 2.0 37,38. The composite likelihood ratio (CLR) 39,40 was calculated by SweepFineder2 41. There are 173, 216, 38 and 21 samples for 2019-nCoV, SARS-CoV, HKU1-CoV and TGEV-CoV, separately. There are 972, 176, 303, 34 and 90 samples for human, bat, camel, cow and murine coronavirus, separately.

Declarations

Author contributions

Z. Z. 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) and the National Natural Science Foundation of HeBei province (19226631D).

References

1. Lu, H., Stratton, C. & Tang, Y. W. Outbreak of Pneumonia of Unknown Etiology in Wuhan China: the Mystery and the Miracle. *J Med Virol*, doi:10.1002/jmv.25678 (2020).
2. Hui, S. *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, doi:10.1016/j.ijid.2020.01.009 (2020).
3. Lu, R. *et al.* Genomic characterisation and epidemiology of 2019 novel coronavirus: implications for virus origins and receptor binding. *Lancet* **395**, 565-574, doi:10.1016/S0140-6736(20)30251-8 (2020).
4. Zhou, *et al.* A pneumonia outbreak associated with a new coronavirus of probable bat origin. *Nature* **579**, 270-273, doi:10.1038/s41586-020-2012-7 (2020).
5. Zhang, , Wu, Q. & Zhang, Z. Pangolin homology associated with 2019-nCoV. *bioRxiv*

(2020).

6. Xiao, K. *et al.* Isolation and Characterization of 2019-nCoV-like Coronavirus from Malayan Pangolins. *bioRxiv* (2020).
7. Wong, M. C., Javornik Cregeen, S. J., Ajami, N. J. & Petrosino, J. Evidence of recombination in coronaviruses implicating pangolin origins of nCoV-2019. *bioRxiv* (2020).
8. Liu, *et al.* Are pangolins the intermediate host of the 2019 novel coronavirus (2019-nCoV)? *bioRxiv* (2020).
9. Lam, T.-Y. *et al.* Identification of 2019-nCoV related coronaviruses in Malayan pangolins in southern China. *bioRxiv* (2020).
10. He, J., Tao, , Yan, Y., Huang, S.-Y. & Xiao, Y. Molecular mechanism of evolution and human infection with the novel coronavirus (2019-nCoV). *bioRxiv* (2020).
11. Tian, X. *et al.* Potent binding of 2019 novel coronavirus spike protein by a SARS coronavirus-specific human monoclonal antibody. *bioRxiv* (2020).
12. Boursnell, M. E. *et al.* Completion of the sequence of the genome of the coronavirus avian infectious bronchitis virus. *J Gen Virol* **68 (Pt 1)**, 57-77, doi:10.1099/0022-1317-68-1-57 (1987).
13. Coley, E. *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, doi:10.1128/JVI.79.5.3097-3106.2005 (2005).
14. St-Jean, R. *et al.* Human respiratory coronavirus OC43: genetic stability and neuroinvasion. *J Virol* **78**, 8824-8834, doi:10.1128/JVI.78.16.8824-8834.2004 (2004).
15. Chouljenko, N., Lin, X. Q., Storz, J., Kousoulas, K. G. & Gorbalenya, A. E. 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, doi:10.1099/0022-1317-82-12-2927 (2001).

16. van Boheemen, S. *et al.* Genomic characterization of a newly discovered coronavirus associated with acute respiratory distress syndrome in humans. *mBio* **3**, doi:10.1128/mBio.00473-12 (2012).

17. Vlasova, A. N. *et al.* Molecular characterization of a new species in the genus Alphacoronavirus associated with mink epizootic catarrhal gastroenteritis. *J Gen Virol* **92**, 1369-1379, doi:10.1099/vir.0.025353-0 (2011).

18. Marra, A. *et al.* The Genome sequence of the SARS-associated coronavirus. *Science* **300**, 1399-1404, doi:10.1126/science.1085953 (2003).

19. Woo, C. *et al.* Characterization and complete genome sequence of a novel coronavirus, coronavirus HKU1, from patients with pneumonia. *J Virol* **79**, 884-895, doi:10.1128/JVI.79.2.884-895.2005 (2005).

20. Tang, C. *et al.* Prevalence and genetic diversity of coronaviruses in bats from China. *J Virol* **80**, 7481-7490, doi:10.1128/JVI.00697-06 (2006).

21. Lau, S. K. *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, doi:10.1016/j.virol.2007.06.009 (2007).

22. Chu, K., Peiris, J. S., Chen, H., Guan, Y. & Poon, L. L. Genomic characterizations of bat coronaviruses (1A, 1B and HKU8) and evidence for co-infections in Miniopterus bats. *J Gen Virol* **89**, 1282-1287, doi:10.1099/vir.0.83605-0 (2008).

23. Woo, C. *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, doi:10.1128/JVI.02182-06 (2007).

24. Zhu, Z., Meng, K. & Meng, G. A database resource for Genome-wide dynamics

analysis of Coronaviruses on a historical and global scale. *bioRxiv* (2020).

25. Hung, C. L., Lin, S., Lin, C. Y., Chung, Y. C. & Chung, Y. F. CUDA ClustalW: An efficient parallel algorithm for progressive multiple sequence alignment on Multi-GPUs. *Comput Biol Chem* **58**, 62-68, doi:10.1016/j.compbiolchem.2015.05.004 (2015).
26. Martin, P. *et al.* RDP3: a flexible and fast computer program for analyzing recombination. *Bioinformatics* **26**, 2462-2463, doi:10.1093/bioinformatics/btq467 (2010).
27. Martin, & Rybicki, E. RDP: detection of recombination amongst aligned sequences. *Bioinformatics* **16**, 562-563, doi:10.1093/bioinformatics/16.6.562 (2000).
28. Padidam, M., Sawyer, & Fauquet, C. M. Possible emergence of new geminiviruses by frequent recombination. *Virology* **265**, 218-225, doi:10.1006/viro.1999.0056 (1999).
29. Martin, P., Posada, D., Crandall, K. A. & Williamson, C. A modified bootscan algorithm for automated identification of recombinant sequences and recombination breakpoints. *AIDS Res Hum Retroviruses* **21**, 98-102, doi:10.1089/aid.2005.21.98 (2005).
30. Smith, J. M. Analyzing the mosaic structure of genes. *J Mol Evol* **34**, 126-129, doi:10.1007/bf00182389 (1992).
31. Posada, & Crandall, K. A. Evaluation of methods for detecting recombination from DNA sequences: computer simulations. *Proc Natl Acad Sci U S A* **98**, 13757-13762, doi:10.1073/pnas.241370698 (2001).
32. Gibbs, M. J., Armstrong, J. S. & Gibbs, A. J. Sister-scanning: a Monte Carlo procedure for assessing signals in recombinant sequences. *Bioinformatics* **16**, 573-582, doi:10.1093/bioinformatics/16.7.573 (2000).
33. Boni, M. , Posada, D. & Feldman, M. W. An exact nonparametric method for inferring mosaic structure in sequence triplets. *Genetics* **176**, 1035-1047,

doi:10.1534/genetics.106.068874 (2007).

34. Kumar, , Stecher, G., Li, M., Knyaz, C. & Tamura, K. MEGA X: Molecular Evolutionary Genetics Analysis across Computing Platforms. *Mol Biol Evol* **35**, 1547-1549, doi:10.1093/molbev/msy096 (2018).

35. Watterson, G. A. On the number of segregating sites in genetical models without recombination. *Theor Popul Biol* **7**, 256-276, doi:10.1016/0040-5809(75)90020-9 (1975).

36. Tajima, Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* **123**, 585-595 (1989).

37. Hutter, , Vilella, A. J. & Rozas, J. Genome-wide DNA polymorphism analyses using VariScan. *BMC Bioinformatics* **7**, 409, doi:10.1186/1471-2105-7-409 (2006).

38. Vilella, A. J., Blanco-Garcia, A., Hutter, & Rozas, J. VariScan: Analysis of evolutionary patterns from large-scale DNA sequence polymorphism data. *Bioinformatics* **21**, 2791-2793, doi:10.1093/bioinformatics/bti403 (2005).

39. Nielsen, R. *et al.* Genomic scans for selective sweeps using SNP data. *Genome Res* **15**, 1566-1575, doi:10.1101/gr.4252305 (2005).

40. Zhu, L. & Bustamante, C. A composite-likelihood approach for detecting directional selection from DNA sequence data. *Genetics* **170**, 1411-1421, doi:10.1534/genetics.104.035097 (2005).

41. DeGiorgio, M., Huber, D., Hubisz, M. J., Hellmann, I. & Nielsen, R. SweepFinder2: increased sensitivity, robustness and flexibility. *Bioinformatics* **32**, 1895-1897, doi:10.1093/bioinformatics/btw051 (2016).

Table 1

Table 1. Putative Recombinants between bat and pangolin coronavirus. ‘Position’ refers to the start and the end at the reference genome MN908947 (2019-nCoV). ’NS’ means not significant. 410539

to 410543 are Pangolin-CoV-2017 strains. 412860, 410721 and 410544 are Pangolin-CoV-2019 strains. MN996532 is Bat-CoV-RaTG13. 408512, 412900 and 402119 to 402124 are 2019-nCov strains.

Position		Recombinant	Minor	Major	Statistic Tests						
Begin	End				RDP	GENECONV	Bootscan	Maxchi	Chimaera	SiSscan	3Seq
16623	17891	410539	MG772933 MG772934	408512	2.29E-13	1.43E-03	2.59E-11	3.82E-05	2.01E-06	1.26E-11	1.39E-08
		410538		402119							
		410540		402120							
		410541		402121							
		410542		402123							
		410543		402124							
		410721	MG772933	MN996532	6.20E-43	1.75E-12	6.52E-06	2.25E-14	7.05E-09	1.75E-10	1.26E-06
		410544	MG772934	402119							
22870	23099	MN996532	412860	412900	5.80E-14	1.83E-04	1.48E-04	5.02E-03	6.84E-04	NS	1.02E-11

Figures

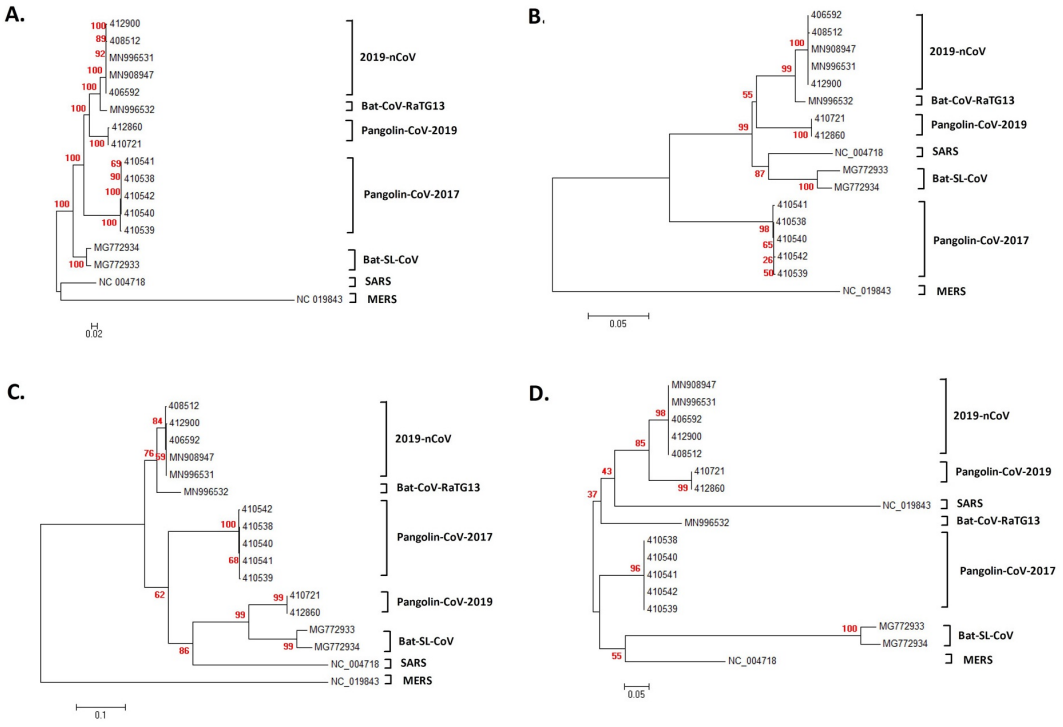

Figure 1

A is phylogenetic tree in the whole genome level. B is a phylogenetic tree built by sequences in the ORF1 recombinant. C is a phylogenetic tree built by sequences in the boundary recombinant. D is a phylogenetic tree built by sequences in the S protein recombinant. The numbers marked in red are the marginal likelihood of the tree.

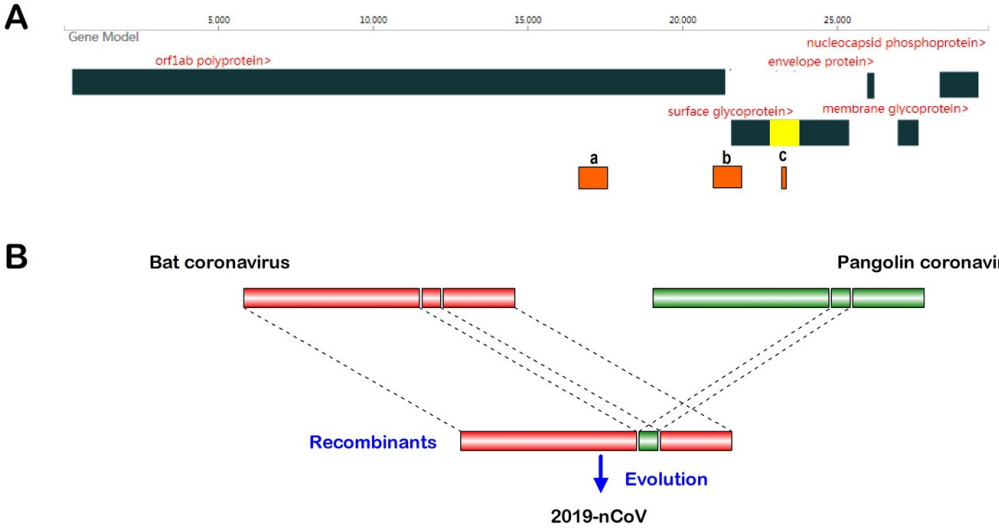
A
5,000
10,000
15,000
20,000
25,000
Gene Model
orf1ab polyprotein>
nucleocapsid phosphoprotein>
envelope protein>
surface glycoprotein>
membrane glycoprotein>
a
b
c
B
Bat coronavirus
Pangolin coronavirus
Recombinants
Evolution
2019-nCoV
C
ORF3a protein>
surface glycoprotein>
1, Bat
2, Camel
3, Cow
4, Human
Pi
0.40
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
21,500
22,000
22,500
23,000
23,500
24,000
24,500
25,000
25,500
Pd
1, Bat
2, Camel
3, Cow
0.60
0.55
0.50
0.45
0.40
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
21,500
22,000
22,500
23,000
23,500
24,000
24,500
25,000
25,500

Figure 2

A is the positions of three recombinants (colored in orange) in the genome of 2019-nCoV, with major proteins marked. ‘a’, ’b’ and ‘c’ refer to ORF1, boundary and S recombinants separately. Yellow represent the RBD in S protein. B is a diagram depicting a possible origination of 2019-nCoV. C is a snapshot of sliding window analysis on Pi (for bat, camel, cow and human coronavirus) and Fst (between human and bat, camel or cow coronavirus). The S protein recombinant an orange box and its region is marked in light blue.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

Supplementary information.pdf