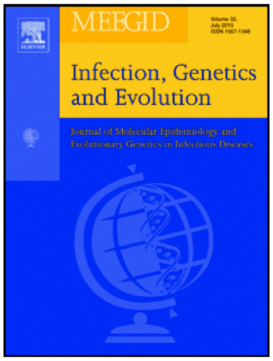

MEEGID
Infection, Genetics and Evolution
Journal of Molecular Epidemiology and Evolutionary Genetics in Infectious Diseases

Predicting the receptor-binding domain usage of the coronavirus based on kmer frequency on spike protein

Zhaozhong Zhu, Zheng Zhang, Wenjun Chen, Zena Cai, Xingyi Ge, Haizhen Zhu, Taijiao Jiang, Wenjie Tan, Yousong Peng

PII: S1567-1348(18)30128-X
DOI: doi:10.1016/j.meegid.2018.03.028
Reference: MEEGID 3463

To appear in: *Infection, Genetics and Evolution*

Received date: 8 March 2018
Revised date: 26 March 2018
Accepted date: 28 March 2018

Predicting the receptor-binding domain usage of the coronavirus based on kmer frequency on spike protein

Zhaozhong Zhu[1, #], Zheng Zhang[1, #], Wenjun Chen[1], Zena Cai[1], Xingyi Ge[1], Haizhen Zhu[1, 2], Taijiao Jiang[3, 4], Wenjie Tan[5, *], Yousong Peng[1, *]

[1] College of Biology, Hunan University, Changsha, China

[2] State Key Laboratory of Chemo/Biosensing and Chemometrics, Hunan University, Changsha, China

[3] Center of System Medicine, Institute of Basic Medical Sciences, Chinese Academy of Medical Sciences & Peking Union Medical College, Beijing, China

[4] Suzhou Institute of Systems Medicine, Suzhou, China

[5] National Institute for Viral Disease Control and Prevention, China CDC, Beijing 100052, China

# These authors contributed equally to this work

* Correspondence: tanwj28@163.com (WT), pys2013@hnu.edu.cn (YP)

To the editor,

The coronavirus is an enveloped, positive-sense, single-stranded RNA virus. It could be classified into four major genera: *Alphacoronavirus*, *Betacoronavirus*, *Gammacoronavirus* and *Deltacoronavirus*, based on serological and genetic studies (Li, 2016). The *Alphacoronavirus* and *Betacoronavirus* mainly infect mammals, whereas the *Gammacoronavirus* and *Deltacoronavirus* mainly infect avians (Tang et al., 2015). The coronavirus poses a serious threat to human health and global security because several coronaviruses could cross-species to infect humans, such as the Severe Acute Respiratory Syndrome Coronavirus (SARS-CoV) and Middle East Respiratory Syndrome Coronavirus (MERS-CoV) (Lu et al., 2015; Peck et al., 2015; Smith, 2006). The SARS-CoV was reported to cause 774 human deaths in 37 countries from 2002 to 2003 (Smith, 2006), while the MERS-CoV is still persistently infecting humans in many countries and have already caused more than 700 deaths around the world (World Health Organization, 2017). How to prevent and control the coronavirus has become a global concern.

The genome of the coronavirus generally encodes more than ten proteins (Peck et al., 2015; Yang et al., 2013). Among them, the spike surface envelope glycoprotein is responsible for binding to host receptors and determines the tissue tropism and host range of the virus to a large extent (Li, 2015, 2016; Lu et al., 2015). The spike protein contains an ectodomain, a transmembrane anchor and a short intracellular tail. Among them, the ecotodomain could be cleaved into a receptor-binding S1 subunit and a membrane-fusion S2 subunit during molecular maturation. The S1 subunit binds to a host receptor for entry into the host cell (Li, 2015, 2016; Qian et al., 2015). Depending on the coronavirus species, the spike protein could bind to either protein receptors or glycans (Li, 2016). Multiple receptors were reported for the coronavirus. This is largely attributed to the double receptor-binding domains (RBD) on the S1 subunit: one RBD is located in the N-terminal (denoted as NTD), while the other is located in the C-terminal (denoted as CTD) (Li, 2016). One coronavirus species generally uses one RBD. Some coronaviruses used NTD, for example, the mouse hepatitis virus (MHV) (Peng et al., 2011), while the others used CTD, such as SARS-CoV (Lu et al., 2015) and MERS-CoV (Lu et al., 2015). Previous studies suggest that the usage of two RBDs could facilitate expansion of host range of the virus (Li, 2015, 2016). However, the mechanism under the RBD usage is still obscure. Besides, RBD usage of most coronavirus species is still unknown. Here, we attempted to develop a computational method for determining RBD usage of the coronavirus based on the protein sequence of S1.

We firstly manually compiled twelve coronavirus species with RBD usage reported from the literature (Table S1). Four coronavirus species used NTD, including the bovine coronavirus (BCoV), MHV, IBV and the human coronavirus OC43 (HCoV-OC43), while the other eight coronavirus species used CTD, including the human coronavirus 229E (HCoV-229E), feline coronavirus (FCoV), bat coronavirus HKU4 (BatCoV-HKU4), human coronavirus HKU1 (HCoV-HKU1), human coronavirus NL63 (HCoV-NL63), MERS-CoV, SARS-CoV and transmissible gastroenteritis virus (TGEV). The protein sequences of the spike protein S1 subunit of these viruses were collected from the NCBI protein database. For convenience, only 800 amino acids in the N-terminal of each spike protein sequence, which covered the S1 subunit of all coronavirus species, were kept for further analysis (Supplementary Methods).

Then, the frequency of kmers (one or two amino acids) was used individually to predict whether a coronavirus used NTD or CTD for binding to the receptor (see Supplementary Methods and Table S2). Most of them achieved a predictive accuracy ranging from 0.6 to 0.8. Surprisingly, we found a pair of amino acids, i.e., “FS”, could discriminate the RBD usage of these 12 coronavirus species with an average predictive accuracy of 97% (Fig. 1A). More specifically, it achieved an accuracy of 100% for BCoV, MHV, HCoV-OC43, BatCoV-HKU4, HCoV-HKU1, HCoV-NL63 and TGEV, and an accuracy of 0.94, 0.87, 0.99, 0.99 and 0.92 for IBV, HCoV-229E, FCoV, MERS-CoV and SARS-CoV, respectively. Analyzing the number of “FS” in the protein sequence of S1 subunit of these viruses, we found that the viruses using NTD generally had less than 3 “FS”s in S1 expect for IBV, while the viruses using CTD generally had 6 or more “FS”s in S1 (Fig. 1A).

Further analysis of the ratio between the observed and expected number of “FS” in S1 protein of these viruses showed that the “FS” was under-represented in the viruses using NTD (Fig. S1), i.e., the observed number of “FS” in S1 was lower than that of the expected; while for the viruses using CTD, the “FS” was generally over-represented in S1. We next analyzed the location of “FS”s on the 3D structure of S1 protein of the coronavirus. Figure 1B&C show the 3D structures for S1 protein of MHV and HCoV-NL63 respectively. For most coronavirus species, the “FS”s (colored in blue) were generally scattered around the S1 protein (Fig. 1B&C and Fig. S2). Few of them were located in or near the receptor-binding interface (colored in red), suggesting that “FS” may not contribute directly to the virus-receptor interaction. One exception is the SARS-CoV, for which there was one “FS” in the interface (Fig. S2A). More efforts are needed to clarify how does the “FS” influence the RBD usage of the coronavirus.

Finally, except for 12 coronavirus species mentioned above, we inferred the RBD usage of all other coronavirus species which had S1 protein sequence available in the NCBI protein database (Table S3), based on the number of “FS” in S1 protein. A total of 31 coronavirus species covering all four major genera were used in prediction. For the virus in *Alphacoronavirus*, except for the Mink coronavirus 1, all the other coronavirus species were predicted to use CTD; while for other genera, most coronavirus species were predicted to use NTD.

Overall, this work provides a simple and effective method for inferring the RBD usage of the coronavirus based on the protein sequence of the spike protein. It may not only help understand the mechanisms behind the RBD usage of the coronavirus, but also help for identification of host receptors for the virus.

ACKNOWLEDGEMENTS

This work was supported by the National Key Plan for Scientific Research and Development of China (2016YFD0500300 and 2016YFC1200200), the National Natural Science Foundation of China (31500126 and 31671371) and the Chinese Academy of Medical Sciences (2016-I2M-1-005).

The authors have declared that no competing interests exist.

REFERENCES

Li, F., 2015. Receptor Recognition Mechanisms of Coronaviruses: a Decade of Structural Studies. Journal of virology 89, 1954-1964.

Li, F., 2016. Structure, Function, and Evolution of Coronavirus Spike Proteins. Annu Rev Virol 3, 237-261.

Lu, G.W., Wang, Q.H., Gao, G.F., 2015. Bat-to-human: spike features determining 'host jump' of coronaviruses SARS-CoV, MERS-CoV, and beyond. Trends in microbiology 23, 468-478.

Peck, K.M., Burch, C.L., Heise, M.T., Baric, R.S., 2015. Coronavirus Host Range Expansion and Middle East Respiratory Syndrome Coronavirus Emergence: Biochemical Mechanisms and Evolutionary Perspectives. Annual Review of Virology, Vol 2 2, 95-117.

Peng, G.Q., Sun, D.W., Rajashankar, K.R., Qian, Z.H., Holmes, K.V., Li, F., 2011. Crystal structure of mouse coronavirus receptor-binding domain complexed with its murine receptor. Proceedings of the National Academy of Sciences of the United States of America 108, 10696-10701.

Qian, Z.H., Ou, X.Y., Goes, L.G.B., Osborne, C., Castano, A., Holmes, K.V., Dominguez, S.R., 2015. Identification of the Receptor-Binding Domain of the Spike Glycoprotein of Human Betacoronavirus HKU1. Journal of virology 89, 8816-8827.

Smith, R.D., 2006. Responding to global infectious disease outbreaks: Lessons from SARS on the role of risk perception, communication and management. Soc Sci Med 63, 3113-3123.

Tang, Q., Song, Y.L., Shi, M.J., Cheng, Y.Y., Zhang, W.T., Xia, X.Q., 2015. Inferring the hosts of coronavirus using dual statistical models based on nucleotide composition. Scientific reports 5.

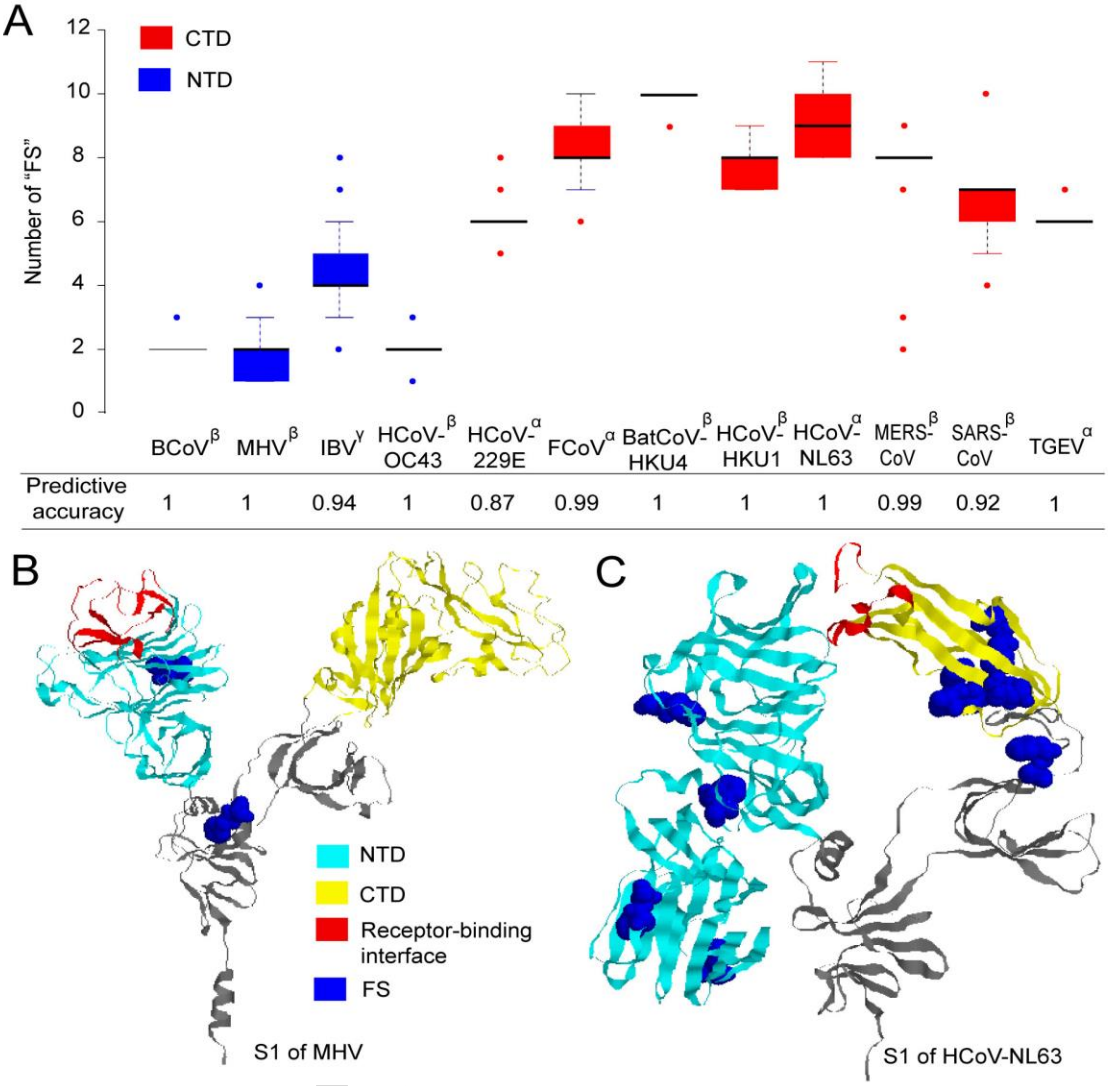
World Health Organization, 2017. Middle East respiratory syndrome coronavirus (MERS-CoV). Available at http://www.who.int/emergencies/mers-cov/en/

Yang, Y., Zhang, L., Geng, H.Y., Deng, Y., Huang, B.Y., Guo, Y., Zhao, Z.D., Tan, W.J., 2013. The structural and accessory proteins M, ORF 4a, ORF 4b, and ORF 5 of Middle East respiratory syndrome coronavirus (MERS-CoV) are potent interferon antagonists. Protein Cell 4, 951-961.

Figure Legends

Figure 1 Predicting the RBD usage of the coronavirus based on the number of "FS" in the protein sequence of the spike protein S1 subunit. (A) The distribution of the number of "FS" in S1 and the predictive accuracy based on the number of "FS" in 12 coronavirus species. The coronavirus species using NTD and CTD were colored in blue and red, respectively. The genus each virus belongs to was labeled in the top right of the virus name. (B) and (C) refer to the 3D structure of S1 subunit for MHV and HCoV-NL63, respectively. The receptor-binding interface was inferred manually from the spike-receptor complex (PDB id: 3r4d for MHV and 3kbh for HCoV-NL63). NTD and CTD were colored in cyan and yellow respectively. The "FS"s were colored in blue.

Figure 1



	BCoV$^{\beta}$	MHV$^{\beta}$	IBV$^{\gamma}$	HCoV-OC43$^{\beta}$	HCoV-229E$^{\alpha}$	FCoV$^{\alpha}$	BatCoV-HKU4$^{\beta}$	HCoV-HKU1$^{\beta}$	HCoV-NL63$^{\alpha}$	MERS-CoV$^{\beta}$	SARS-CoV$^{\beta}$	TGEV$^{\alpha}$
Predictive accuracy	1	1	0.94	1	0.87	0.99	1	1	1	0.99	0.92	1