Profiling the immune vulnerability landscape of the 2019 Novel Coronavirus

James Zhu[1,*], Jiwoong Kim[1,*], Xue Xiao[1,*], Yunguan Wang[1,*], Danni Luo[1], Ran Chen[1], Lin Xu[1], He Zhang[1], Guanghua Xiao[1,2], Xiaowei Zhan[1,3], Tao Wang[1,3,+], Yang Xie[1,2+]

1 Quantitative Biomedical Research Center, Department of Population and Data Sciences, University of Texas Southwefstern Medical Center, Dallas, TX, USA, 75390.
2. Department of Bioinformatics, University of Texas Southwefstern Medical Center, Dallas, TX, USA, 75390.
3. Center for the Genetics of Host Defense, University of Texas Southwefstern Medical Center, Dallas, TX, USA, 75390.

* Co-first authors
+ Corresponding authors: (1) Tao Wang, Ph.D., Quantitative Biomedical Research Center, Department of Population and Data Sciences, UT Southwestern Medical Center, Dallas, TX, 75390, USA; Phone: 214-648-4082; E-mail: Tao.Wang@UTSouthwestern.edu (2) Yang Xie, Ph.D., Quantitative Biomedical Research Center, Department of Population and Data Sciences, UT Southwestern Medical Center, Dallas, TX, 75390, USA; Phone: 214-648-5178; E-mail: Yang.Xie@UTSouthwestern.edu

Funding: This study was supported by Cancer Prevention Research Institute of Texas [CPRIT RP190208/TW]

ABSTRACT

The outbreak of the 2019 Novel Coronavirus (2019-nCoV) has rapidly spread from Wuhan, China to multiple countries, causing staggering number of infections and deaths. A systematic profiling of the immune vulnerability landscape of 2019-nCoV is lacking, which can bring critical insights into the immune clearance mechanism, peptide vaccine development, and antiviral antibody development. In this study, we predicted the potential of all the 2019-nCoV viral proteins to induce class I and II MHC presentation and form linear antibody epitopes. We showed that the enrichment for T cell and B cell epitopes is not uniform on the viral genome, with several focused regions that generate abundant epitopes and may be more targetable. We showed that genetic variations in 2019-nCoV, though fewer for the moment, already follow the pattern of mutations in related coronaviruses, and could alter the immune vulnerability landscape of this virus, which should be considered in the development of therapies. We create an online database to broadly share our research outcome. Overall, we present an immunological resource for 2019-nCoV that could significantly promote both therapeutic development and mechanistic research.

INTRODUCTION

In December 2019, an outbreak of a novel coronavirus (2019-nCoV) was reported in Wuhan, China (*1*). 2019-nCoV rapidly spread to other regions of China, and multiple other countries, with a contagious speed that is much higher than the Severe Acute Respiratory Syndrome (SARS) coronavirus and the Middle East Respiratory Syndrome (MERS) coronavirus (*2*). Despite the lower mortality rate of 2019-nCoV compared with SARS and MERS, the scale of the 2019-nCoV contagion has already caused more casualties than either of them, as of this writing. Early research into 2019-nCoV has mostly described its epidemiological features (*1*, *3*), reported the possible curative effect of remdesivir (*4*), and characterized its basic genomics features (*5*, *6*).

Scant works have reported on the immunological features of 2019-nCoV, which could have significant bearing on the mechanistic studies of viral life cycle. Such analyses could also inform anti-viral immuno-therapeutic development, which can be either T cell-based or B cell-based. Antibodies can neutralize viral infectivity in a number of ways, such as interference with binding to receptors, block uptake into cells, *etc*. For SARS-CoV, the human ACE-2 protein is the functional receptor, and anti-ACE2 antibody can block viral replication (*7*). On the other hand, previous studies have indicated a crucial role of both $CD8^+$ and $CD4^+$ T cells in SARS-CoV clearance (*8*, *9*), while Janice Oh *et al* also observed that development of SARS-CoV specific neutralizing antibodies requires $CD4^+$ T helper cells (*8*). In fact, there are examples of vaccines for influenza that contain both antibody and T cell inducing components (*10*, *11*).

In this work, we performed a bioinformatics profiling of the class I and class II MHC binding potentials of the 2019-nCoV proteins, and also a profiling of the potentials of the linear epitopes of the viral proteins to induce antibodies. We correlated this immune vulnerability map of the

2019-nCoV proteins with their possible mutational hotspots. We made the analyses publicly available as a resource to the research community, in the form of the 2019-nCoV Immune Viewer: https://qbrc.swmed.edu/projects/2019ncov_immuneviewer/.

RESULTS

T cell- and B cell-mediated immune vulnerability landscape of 2019-nCoV

We used the netMHCpan suite of software (*12*, *13*) to predict the MHC class I and class II binding peptides of all 2019-nCoV proteins, which could elicit $CD8^+$ and $CD4^+$ T cell responses for viral clearance (**Fig. 1a**). We found that the number of MHC class I and class II binders, weighted by the HLA allele frequency in the Chinese population, are not spatially uniform across the viral genome. And there are a small number of genomic regions that showed high peaks of immunogenicity corresponding to a large number of MHC binders in a small neighborhood, which could be better potential vaccine targets (**Sup. Table 1**). The MHC binding peptide profiles of a different racial population (*i.e.* European ancestry) are shown in **Sup. Fig. 1**. Interestingly, the T cell epitope intensities (number of binders weighted by allele frequency) are higher overall in the European population than the Chinese population, suggesting that the Chinese population may be more vulnerable to 2019-nCoV infection. Individual HLA alleles are examined in **Sup. Fig. 2**. The above analyses are conducted for the 2019-nCoV reference genome. However, the viral strains that have been sampled and sequenced so far are highly similar to each other (a segment of multiple alignment shown in **Fig. 1b**).

We also examined the potentials of the viral proteins to encode linear epitopes that can elicit antibody responses, by using the BepiPred 1.0 software (*14*). For the current analyses, we focused on linear epitopes, rather than conformational epitopes, because linear epitopes are more suitable for vaccine design (*15*, *16*). Similarly scanning through all 2019-nCoV proteins of the reference genome (**Fig. 1c**), we found that the viral genome is also not uniformly enriched for B cell epitopes. In particular, one small segment of the Orf1 protein and the N protein are enriched for predicted B cell epitopes (**Sup. Table 2**). Lastly, we focused on the receptor-binding motif of the 2019-nCoV S protein, which attaches to the ACE-2 protein for entry into the human cell (*17*). We blasted the motif binding domain sequence of the 2019-nCoV S protein with that of SARS, and found there is a poor conservation between the two S proteins (**Fig. 1d**), which suggests that prior antibodies developed for SARS may not work for 2019-nCOV.

For comparison, we also computed the immune vulnerability maps of SARS (**Fig. 1e**) and MERS (**Fig. 1f**), which are the two most aggressive coronaviruses, together with 2019-nCoV. We found that the B cell epitope profiles seem to be more consistent among the three viruses, while the T cell epitope profiles are more distinct. This suggests that the T cell biology of 2019-nCoV could be different from that of SARS and MERS.

Potential mutations in 2019-nCoV could affect immune vulnerability and vaccine design

Coronaviruses are all RNA virus (*18*), and RNA viruses generally have very high mutation rates (*19*). We showed the mutational rates in the viral genomes for 2019-nCoV, SARS and MERS (**Fig. 2a-c**). 2019-nCoV only emerged a very short time ago, which likely explains the lack of significant amount of genetic variations (**Fig. 2a**). In comparison, SARS (**Fig. 2b**) and MERS (**Fig. 2c**) have both accumulated significant variation between the different strains, probably due to their much longer contagion history in humans. Interestingly, in 2019-nCoV, the genomic regions with higher mutational rates and lower mutational rates can already be discerned (**Fig. 2a**), and they seem to be rather conserved with those of SARS and MERS (**Fig. 2bc**). This indicates that we should be cautious of a similar level of genetic variation in the future, which might yield an aggressive viral strain, despite the current lack of mutations in 2019-nCoV.

Inspired by this observation, we reasoned that the evolution of the immune vulnerability maps of SARS and MERS due to genetic variation may reveal insight into 2019-nCoV. In **Fig. 2d**, we showed the relative mutational rates of the viral genomic regions that are enriched for abundant CD4 T cell epitopes, CD8 T cell epitopes, and linear antibody epitopes, in each of the three viruses. It can be seen from **Fig. 2d** that the mutational rates in 2019-nCoV are still much lower than those of SARS and MERS in these epitope-enriched regions, as expected. But interestingly, the mutational rates of the CD8 epitope-enriched regions are higher than those of the CD4 epitope-enriched regions in both SARS and MERS. This may be due to selection pressure inflicted by the $CD8^+$ T cells, which are the major cytotoxic T cell population. The same may happen to the 2019-nCoV virus as well, but this remains to be proved.

A continuously updated database of the immune vulnerability of 2019-nCoV strains

We created the 2019-nCoV Immune Viewer (**Fig. 3**) to openly share the virus immunogenicity data we created for 2019-nCoV, and also SARS and MERS. In the Viewer, we provided user-friendly visualization functionality for researchers to examine immunogenicity strength of different genomic regions of each of the three viruses (**Fig. 3a**), where the users can either zoom in or zoom out. To facilitate the examination of how genetic variations could impact the immune vulnerability landscape of the viruses, we also showed the mutational rates of the viral genome along with the immunogenicity maps. Furthermore, the Viewer also displays a phylogenetic tree with annotations of the strains overlaid. The tree allows the users to highlight the strains of virus according to the annotations (**Fig. 3b**). Overall, we believe that the 2019-nCoV Immune Viewer will be a valuable resource for the research community, and will facilitate immunological research into this virus.

DISCUSSION

In this report, we characterized the immune vulnerability landscape of the 2019-nCoV, and compared it with that of SARS and MERS. Our work should be broadly useful for researchers who study the interaction between this virus and the host immune system. In particular, we discovered focused regions of this virus that encode a high density of T cell epitopes and B cell

epitopes, which could be more suitable for peptide vaccine and anti-viral antibody development. We also found that the S protein receptor-binding motifs are poorly conserved between SARS and 2019-nCoV. To facilitate wide adoption of our research outcome, we created a publicly accessible database, for the researchers to easily explore and download our results. The database is under continuous development, and will be updated timely when new strains of 2019-nCoV are made available.

Genetic variations can modify the immunogenicity landscape of the virus, and impact its survival fitness. The selection of good vaccination epitopes should focus on parts of viral proteins with good potentials of generating immunogenic epitopes, and with less chance of mutations. The low level of genetic variation in 2019-nCoV could merely be a sampling issue due to the short contagion history. However, the domains of genomes that are highly mutated in SARS and MERS are already more highly mutated in 2019-nCoV. We should remain cautious about the possible genetic variations to happen in 2019-nCoV, and immunological studies should consider their possible impact, knowing where the mutations are likely going to happen.

Overall, our work provides a window into the immunological features of 2019-nCoV, and we hope our work could aid therapeutic development against this virus to stop this pandemic earlier and to aid the vaccine development to prevent future breakouts.

MATERIALS AND METHODS

Acquisition of the viral genome sequences

The 2019-nCoV complete genome sequences and meta data were downloaded from the https://bigd.big.ac.cn/ncov database, before the data lock of Feb 5th, 2020. The reference genome annotation was acquired from NCBI: https://www.ncbi.nlm.nih.gov/nuccore/MN908947. The complete genome SARS and MERS sequences are also downloaded from NCBI: https://www.ncbi.nlm.nih.gov/nuccore/?term=txid694009%5BOrganism%3Anoexp%5D+and+complete+genome and https://www.ncbi.nlm.nih.gov/nuccore/?term=txid1335626%5BOrganism%3Anoexp%5D+and+complete+genome

Prediction of T cell and B cell epitopes

NetMHCpan (v4.0) (*20*) and NetMHCIIpan (v3.2) (*13*) with default threshold options were used to predict peptides, from the viral proteins, that bind to human MHC class I and II proteins for all the available HLA alleles. Only strong binders (<0.5% percentile rank) were retained. The HLA allele population frequency for the Chinese population was acquired from Kwok *et al* (*21*) and population frequency for the European population was from Mack *et al (22)*. The B cell epitope predictions were made by the BepiPred 1.0 software (*14*) with default parameters. Amino acids with B cell epitope prediction scores >0.6 are regarded as having good likelihood of generating linear antibodies.

DNA and protein sequence alignment

The command-line version of MUSCLE (v3.8.31) (*23*, *24*) was used to perform multiple genome sequence alignment with diagonal optimization (-diags). The default number of iteration and the default maximum number of new trees were applied during the alignment. The protein sequence alignment between the S proteins, YP_009724390.1 (2019-nCoV) and NP_828851.1 (SARS), was performed using EMBOSS needle (*25*) with the BLOSUM62 scoring matrix.

Website development

The Immune Viewer is a dynamic website. It is developed using the HTML (HyperText Markup Language), JavaScript and CSS (Cascading Style Sheets). Specifically, we used the D3.js library to allow users to interactively explore the mutation rates or immunogenic scores across the viral genomic regions. We also used the D3.phylogram.js to visualize the phylogenetic tree and the Select2 library to facilitate users' query for different 2019-nCoV strains across multiple geographic regions.

Statistical analyses

All computations and statistical analyses were carried out in the R computing environment. For all boxplots appearing in this study, box boundaries represent interquartile ranges, whiskers extend to the most extreme data point which is no more than 1.5 times the interquartile range, and the line in the middle of the box represents the median. For the line plots, the viral genomes were binned by every 60 nucleotide, and the number of T cell and B cell epitopes falling into each window is calculated. For T cell epitopes, a sum of the number of epitopes weighted by the corresponding ethnic population's HLA allele (A, B, C, and DRB1) frequency is calculated to form the T cell and B cell immunogenicity strength for that population. The genetic variation rate at each nucleotide is calculated by examining all viral strains and counting the proportion of strains with a different nucleotide or with an insertion/deletion, with respect to the reference genome. The genetic variation rates are also binned by the same length of windows and averaged.

Data availability

The 2019-nCoV Immune Viewer is available at:
https://qbrc.swmed.edu/projects/2019ncov_immuneviewer/.

FIGURE AND TABLE LEGENDS

Fig. 1 T cell- and B cell-mediated immune vulnerability landscape of 2019-nCoV in the Chinese population. (a) The $CD4^+$ and $CD8^+$ T cell epitope profiles of 2019-nCoV. The Y axis shows the immunogenicity intensity as described in the method section. (b) Part of the multiple alignment of all the 2019-nCoV strains. (c) The B cell epitope profiles of the 2019-nCoV. The Y axis shows the predicted B cell epitope score (d) The BLAST between the motif binding domain of

the 2019-nCoV S protein and the SARS S protein. (e) The T cell and B cell epitope profiles of SARS. (f) The T cell and B cell epitope profiles of MERS.

Fig. 2 Potential mutations in 2019-nCoV could affect immune vulnerability and vaccine design. The relative mutational rate profiles of the whole genomes of the three viruses: (a) 2019-nCoV, (b) SARS, and (c) MERS. The semi-transparent boxes mark the regions of high mutational rates due to artefacts of incomplete sequencing. (d) The relative mutational rates of the CD8 epitope-enriched regions, CD4 epitope-enriched regions, and B cell epitope-enriched regions of the three viruses. For all bins of the viral genome, we selected the top 50 CD8 and 50 CD4 epitope-enriched regions, and we selected the B cell epitope regions with average B cell epitope prediction score >0.6.

Fig. 3 A continuously updated database of the immune vulnerability of 2019-nCoV strains. (a) Visualization functionality to examine immunogenicity strength of the selected genomic region of the viruses (the Chinese population). Each line plot is divided into two panels stacked vertically together. At the bottom panel, the user can drag and set a region to zoom in, and the top panel zooms in and shows the details of that selected region. (b) Phylogenetic tree of viral strains, which allows subsetting based on annotations by going through a series of drop-down boxes.

Sup. Fig. 1 The T cell epitope profiles of the European population. (a) 2019-nCoV, (b) SARS, and (c) MERS.

Sup. Fig. 2 The variation of T cell epitope profiles for 2019-nCoV, SARS and MERS across populations. The heatmap represents the number of immunogenic binding epitopes across the binned genomes (500bp) of (a) nCoV, (b) SARS and (c) MERS for the major HLA-A alleles shown as examples (allele frequency larger than 1%) in the European American (EA) and Hongkong Chinese (HK) populations. These major alleles are colored in black, blue or red if they are common to both EA and HK population, unique to EA population, or unique to HK population, respectively. On the right, the band of strength represents the cumulative number of immunogenic peptides, and the bands of EA and HK represent the HLA allele frequency of EA and HK populations, respectively.

Sup. Table 1 Genomics regions of 2019-nCoV that are T cell epitope-enriched

Sup. Table 2 Genomics regions of 2019-nCoV that are B cell epitope-enriched

ACKNOWLEDGEMENTS

We acknowledge the patients who contributed the viral islets, the medical staff and researchers for performing islet purification and sequencing, and the NGDC (https://bigd.big.ac.cn/ncov) database for timely sharing of the viral genome sequences.

AUTHOR CONTRIBUTIONS

J.Z. and X.Z. performed statistical analyses of the immune vulnerability landscape of the viruses. J.K., T.W., X.X, and X.Z. retrieved and curated the virus sequence genomes. J.K. and J.Z. performed T cell and B cell epitope predictions. X.X. and J.K carried out DNA and protein sequence alignment. Y.W., D.L., R.C., and X.Z. created the Immune Viewer. H.Z. and X.Z. calculated the phylogenetic trees for the website. G.X. and L.X. provided significant input on the scientific direction of the project. T.W. and Y.X. supervised the study. J.Z., J.K., X.X., Y.W., X.Z., T.W., and Y.X wrote the manuscript.

COMPETING INTERESTS

The authors declare no conflicts of interest.

Bibliography

1. Q. Li *et al.*, Early Transmission Dynamics in Wuhan, China, of Novel Coronavirus-Infected Pneumonia. *N. Engl. J. Med.* (2020), doi:10.1056/NEJMoa2001316.
2. E. de Wit, N. van Doremalen, D. Falzarano, V. J. Munster, SARS and MERS: recent insights into emerging coronaviruses. *Nat. Rev. Microbiol.* **14**, 523–534 (2016).
3. C. Huang *et al.*, Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. *Lancet* (2020), doi:10.1016/S0140-6736(20)30183-5.
4. M. L. Holshue *et al.*, First case of 2019 novel coronavirus in the united states. *N. Engl. J. Med.* (2020), doi:10.1056/NEJMoa2001191.
5. R. Lu *et al.*, Genomic characterisation and epidemiology of 2019 novel coronavirus: implications for virus origins and receptor binding. *Lancet* (2020), doi:10.1016/S0140-6736(20)30251-8.
6. P. Zhou *et al.*, A pneumonia outbreak associated with a new coronavirus of probable bat origin. *Nature* (2020), doi:10.1038/s41586-020-2012-7.
7. E. R. Pfefferkorn, L. C. Pfefferkorn, Arabinosyl nucleosides inhibit Toxoplasma gondii and allow the selection of resistant mutants. *J. Parasitol.* **62**, 993–999 (1976).
8. H.-L. Janice Oh, S. Ken-En Gan, A. Bertoletti, Y.-J. Tan, Understanding the T cell immune response in SARS coronavirus infection. *Emerg. Microbes Infect.* **1**, e23 (2012).
9. J. Chen *et al.*, Cellular immune responses to severe acute respiratory syndrome coronavirus (SARS-CoV) infection in senescent BALB/c mice: CD4+ T cells are important in control of SARS-CoV infection. *J. Virol.* **84**, 1289–1301 (2010).
10. J. S. Testa *et al.*, MHC class I-presented T cell epitopes identified by immunoproteomics analysis are targets for a cross reactive influenza-specific T cell response. *PLoS ONE*. **7**, e48484 (2012).
11. C. Zhou, L. Zhou, Y.-H. Chen, Immunization with high epitope density of M2e derived from 2009 pandemic H1N1 elicits protective immunity in mice. *Vaccine*. **30**, 3463–3469 (2012).
12. M. Nielsen, M. Andreatta, NetMHCpan-3.0; improved prediction of binding to MHC class I

molecules integrating information from multiple receptor and peptide length datasets. *Genome Med.* **8**, 33 (2016).

13. K. K. Jensen *et al.*, Improved methods for predicting peptide binding affinity to MHC class II molecules. *Immunology*. **154**, 394–406 (2018).
14. J. E. P. Larsen, O. Lund, M. Nielsen, Improved method for predicting linear B-cell epitopes. *Immunome Res.* **2**, 2 (2006).
15. R. E. Soria-Guerra, R. Nieto-Gomez, D. O. Govea-Alonso, S. Rosales-Mendoza, An overview of bioinformatics tools for epitope prediction: implications on vaccine development. *J. Biomed. Inform.* **53**, 405–414 (2015).
16. J. L. Sanchez-Trincado, M. Gomez-Perosanz, P. A. Reche, Fundamentals and Methods for T- and B-Cell Epitope Prediction. *J. Immunol. Res.* **2017**, 2680160 (2017).
17. Y. Wan, J. Shang, R. Graham, R. S. Baric, F. Li, Receptor recognition by novel coronavirus from Wuhan: An analysis based on decade-long structural studies of SARS. *J. Virol.* (2020), doi:10.1128/JVI.00127-20.
18. A. R. Fehr, S. Perlman, Coronaviruses: an overview of their replication and pathogenesis. *Methods Mol. Biol.* **1282**, 1–23 (2015).
19. S. Duffy, Why are RNA virus mutation rates so damn high? *PLoS Biol.* **16**, e3000003 (2018).
20. V. Jurtz *et al.*, NetMHCpan-4.0: Improved Peptide-MHC Class I Interaction Predictions Integrating Eluted Ligand and Peptide Binding Affinity Data. *J. Immunol.* **199**, 3360–3368 (2017).
21. J. Kwok *et al.*, HLA-A, -B, -C, and -DRB1 genotyping and haplotype frequencies for a Hong Kong Chinese population of 7595 individuals. *Hum. Immunol.* **77**, 1111–1112 (2016).
22. S. J. Mack *et al.*, HLA-A, -B, -C, and -DRB1 allele and haplotype frequencies distinguish Eastern European Americans from the general European American population. *Tissue Antigens*. **73**, 17–32 (2009).
23. R. C. Edgar, MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics*. **5**, 113 (2004).
24. R. C. Edgar, MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* **32**, 1792–1797 (2004).
25. P. Rice, I. Longden, A. Bleasby, EMBOSS: the european molecular biology open software suite. *Trends Genet.* **16**, 276–277 (2000).

Fig. 1

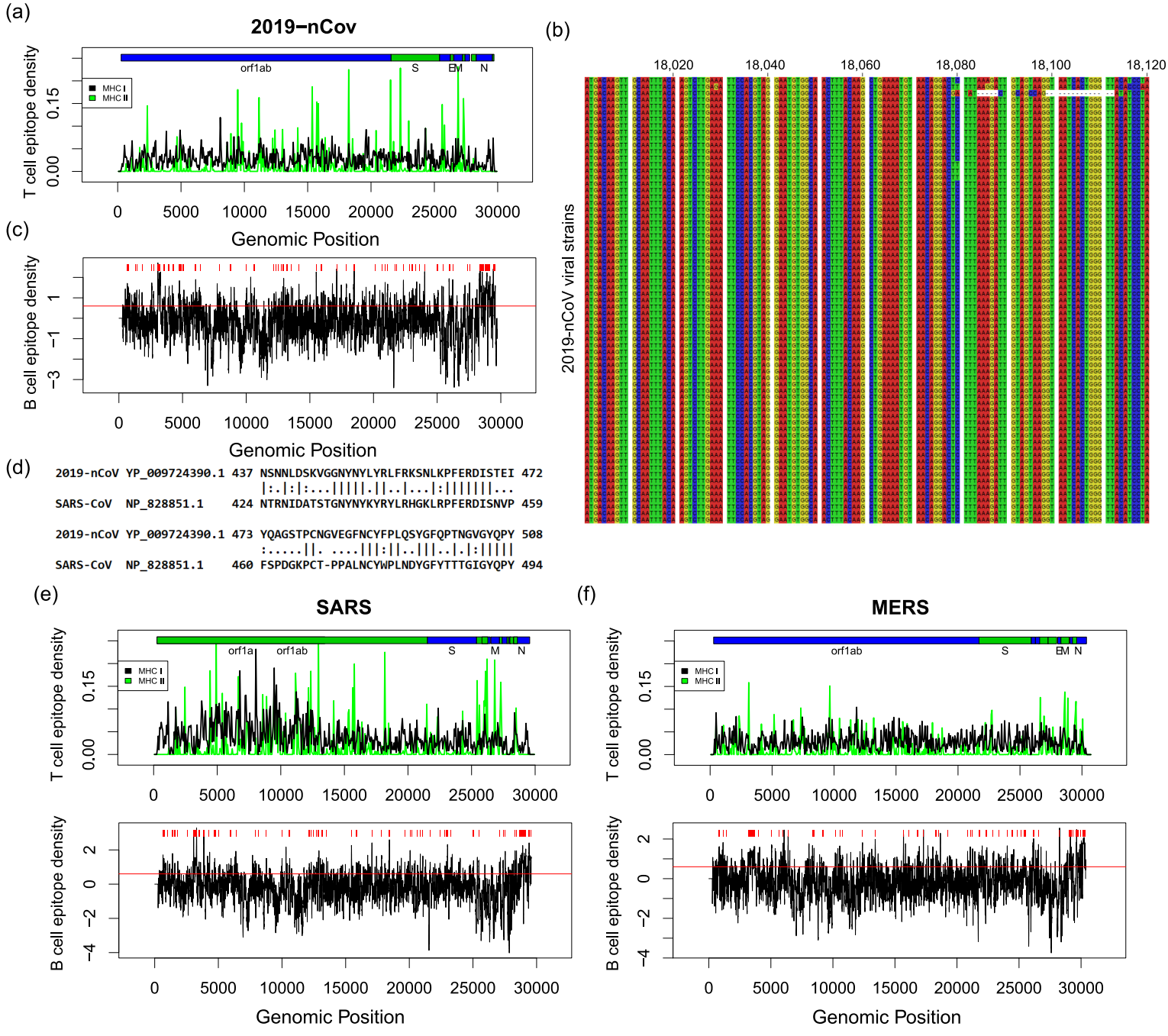

(a)
2019−nCov
orf1ab
S
E M
N
MHC I
MHC II
T cell epitope density
0.00
0.15
0
5000
10000
15000
20000
25000
30000
Genomic Position
(b)
18,020
18,040
18,060
18,080
18,100
18,120
2019-nCoV viral strains
(c)
B cell epitope density
1
−1
−3
Genomic Position
(d)
2019-nCoV YP_009724390.1 437 NSNNLDSKVGGNYNYLYRLFRKSNLKPFERDISTEI 472
SARS-CoV NP_828851.1 424 NTRNIDATSTGNYNYKYRYLRHGKLRPFERDISNVP 459
2019-nCoV YP_009724390.1 473 YQAGSTPCNGVEGFNCYFPLQSYGFQPTNGVGYQPY 508
SARS-CoV NP_828851.1 460 FSPDGKPCT-PPALNCYWPLNDYGFYTTTGIGYQPY 494
(e)
SARS
orf1a
orf1ab
S
M
N
MHC I
MHC II
T cell epitope density
0.00
0.15
B cell epitope density
2
0
−2
−4
Genomic Position
(f)
MERS
orf1ab
S
E M
N
MHC I
MHC II
T cell epitope density
0.00
0.15
B cell epitope density
2
0
−2
−4
Genomic Position

Fig. 2

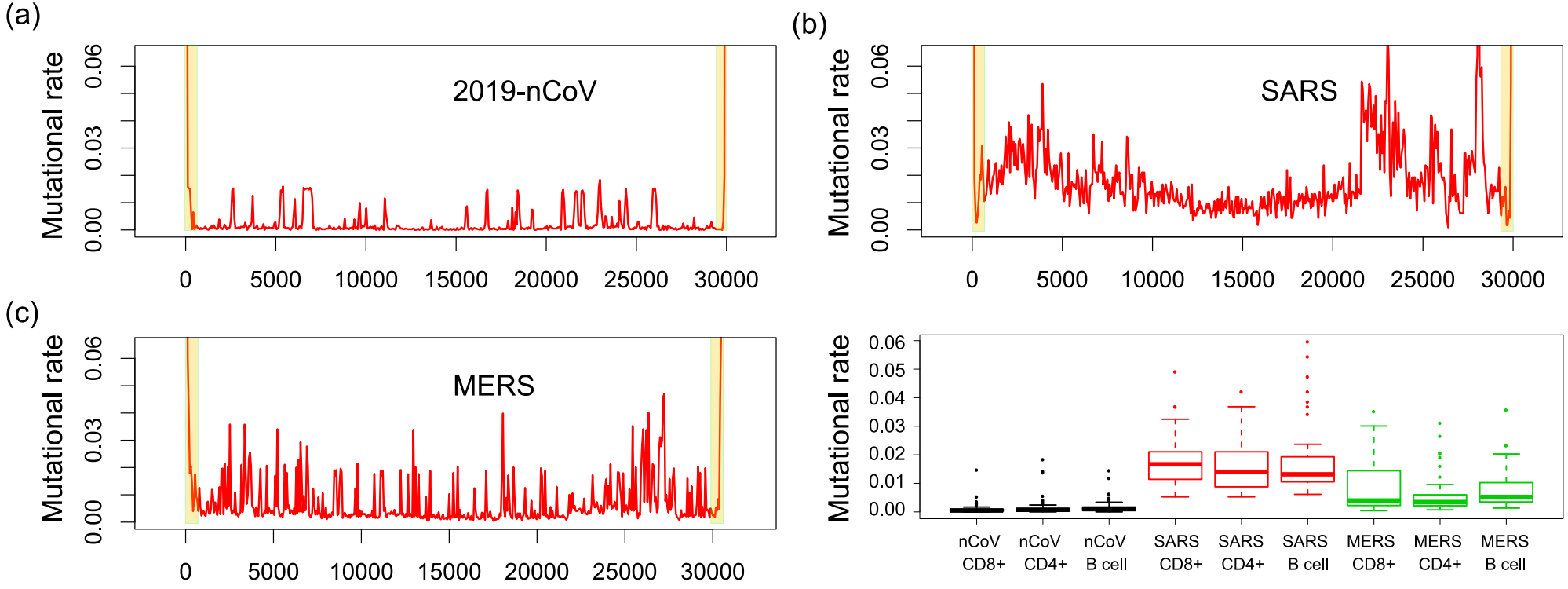

(a)
2019-nCoV
Mutational rate
0.00
0.03
0.06
0
5000
10000
15000
20000
25000
30000
(b)
SARS
(c)
MERS
0.01
0.02
0.04
0.05
nCoV CD8+
nCoV CD4+
nCoV B cell
SARS CD8+
SARS CD4+
SARS B cell
MERS CD8+
MERS CD4+
MERS B cell

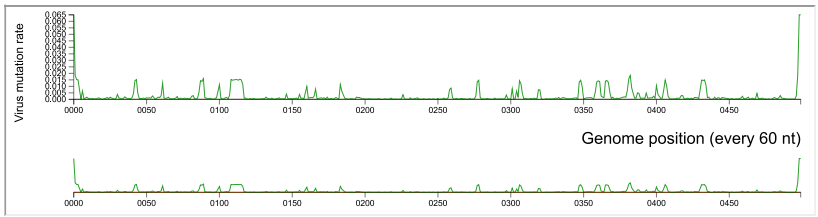
Fig. 3

(a)

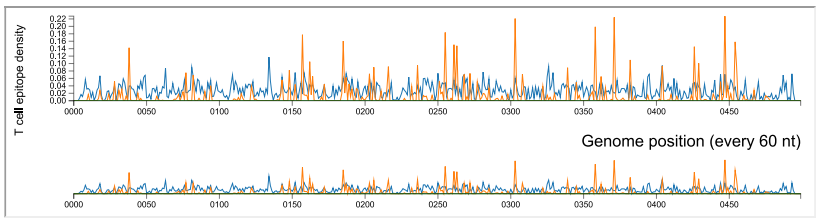
QBRC
Quantitative Biomedical Research Center

NCOV SARS MERS PHYLOGENETIC TREE MEMBER

2019-ncov - Mutation rate

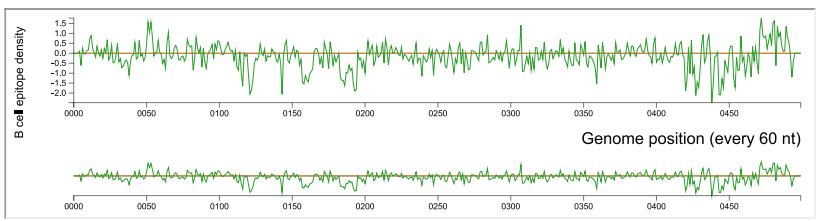


2019-ncov - T cell epitope density



T Cell Type: CD4 CD8

2019-ncov - B cell epitope density



(b)

Phylogram

Wuhan