The immune vulnerability landscape of the 2019 Novel Coronavirus, SARS-CoV-2

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], John W. Schoggins[3], Xiaowei Zhan[1,4], Tao Wang[1,4,+], Yang Xie[1,2+]

1. Quantitative Biomedical Research Center, Department of Population and Data Sciences, University of Texas Southwestern Medical Center, Dallas, TX, USA, 75390.
2. Department of Bioinformatics, University of Texas Southwestern Medical Center, Dallas, TX, USA, 75390.
3. Department of Microbiology, University of Texas Southwestern Medical Center, Dallas, TX, USA, 75390.
4. 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

Short title: The immunogenicity landscape of SARS-CoV-2

Keywords: SARS-CoV-2, T cell, B cell, immunogenicity, mutation

ABSTRACT

The outbreak of the 2019 Novel Coronavirus (SARS-CoV-2) rapidly spread from Wuhan, China to more than 150 countries, areas or territories, causing staggering number of infections and deaths. A systematic profiling of the immune vulnerability landscape of SARS-CoV-2, which can bring critical insights into the immune clearance mechanism, peptide vaccine development, and antiviral antibody development, is lacking. In this study, we investigated the potential of the SARS-CoV-2 viral proteins to induce class I and II MHC presentation and to form linear antibody epitopes. We created an online database to broadly share the predictions as a resource for the research community. Using this resource, we showed that genetic variations in SARS-CoV-2, though still few for the moment, already follow the pattern of mutations in related coronaviruses, and could alter the immune vulnerability landscape of this virus. Importantly, we discovered evidence that SARS-CoV-2, along with related coronaviruses, used mutations to evade attack from the human immune system. Overall, we present an immunological resource for SARS-CoV-2 that could promote both therapeutic development and mechanistic research.

Background

In December 2019, an outbreak of a novel coronavirus (SARS-CoV-2) was reported in Wuhan, China (1). SARS-CoV-2 rapidly spread to other regions of China along with more than 150 other countries, at a speed that is much higher than the Severe Acute Respiratory Syndrome coronavirus (SARS-CoV) and the Middle East Respiratory Syndrome coronavirus (MERS-CoV) (2). Despite the lower mortality rate of SARS-CoV-2 compared with SARS-CoV and MERS-CoV, the scale of the SARS-CoV-2 contagion has already caused more casualties than either of the previous outbreaks. Early research into SARS-CoV-2 has mostly described its epidemiological features (1, 3), case reports (4), structural characterizations (5), basic genomics features (6, 7), *etc*.

Scant works have reported on the immunological features of SARS-CoV-2, which could have significant bearing on our understanding of how this virus interacts with its host. Such analyses could also inform antiviral 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 interfering with binding to receptors, blocking uptake into cells, *etc*. For SARS-CoV, the human ACE-2 protein is the functional receptor, and anti-ACE2 antibody can block viral replication (8). On the other hand, previous studies have indicated a crucial role of both $CD8^+$ and $CD4^+$ T cells in SARS-CoV clearance (9, 10), while Janice Oh *et al* also observed that development of SARS-CoV specific neutralizing antibodies requires $CD4^+$ T helper cells (9). In fact, there are examples of vaccines for influenza that contain both antibody and T cell inducing components (11, 12).

In this work, we performed a bioinformatics profiling of the class I and class II MHC binding potentials of the SARS-CoV-2 proteins, and also a profiling of the potential for linear epitopes of the viral proteins to induce antibodies. We correlated the immune vulnerability map of the

SARS-CoV-2 proteins with its genomic mutations, compared against those of SARS-CoV and MERS-CoV, and generated several interesting observations. We made the analyses publicly available as a resource to the research community, in the form of the SARS-CoV-2 Immune Viewer: https://qbrc.swmed.edu/projects/2019ncov_immuneviewer/.

Results and Discussion

To explore the immune vulnerability landscape of SARS-CoV-2, we used the netMHCpan software (13, 14) to predict the MHC class I and class II binding peptides of all SARS-CoV-2 proteins, which could elicit $CD8^+$ and $CD4^+$ T cell responses for viral clearance (**Fig. 1a**). The counts of the MHC binders were weighted by the allele frequencies of the Chinese population (15), since most, if not all, samples analyzed in this study were obtained from Chinese patients. We found that 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 (16), are shown in **Sup. Fig. 1**. Interestingly, the total T cell epitope intensities, calculated as the sum of the number of binders weighted by allele frequency, are slightly higher in the European population than the Chinese population (**Fig. 1b** and individual alleles shown in **Sup. Fig. 2**), which indicates that different susceptibility to this virus may exist between differential populations. The above analyses were performed for the SARS-CoV-2 reference genome. The viral strains that have been sampled and sequenced so far are highly similar to each other overall, with a segment of multiple alignment shown in **Sup. Fig. 3**.

We also examined the potential of the viral proteins to encode linear epitopes that can elicit antibody responses using BepiPred 2.0 (17). We focused on linear epitopes, rather than conformational epitopes, because linear epitopes are more suitable for vaccine design (18, 19) and bioinformatics predictions are more feasible. Similar to T cell epitopes, contiguous stretches of >10 epitope-encoding residues, which are more representative of the usual length of B cell epitopes, are located in small genomic regions (red bars in **Fig. 1c**, and **Sup. Table 2**). We also focused on the receptor-binding motif (RBM) of the SARS-CoV-2 S protein, which attaches to the ACE-2 protein for entry into the human cell (20). We used BLAST to align the RBM sequence of the SARS-CoV-2 S protein with that of SARS-CoV, and found there is a relatively poor conservation between the two S proteins (**Fig. 1d**). This was supported by Wrapp *et al* (5), which reports no binding to the SARS-CoV-2 RBM by any of the three SARS-CoV RBM-directed antibodies. Therefore, antibody-based therapeutic design should start with the SARS-CoV-2 S protein *de novo*.

For comparison, we also computed the immune vulnerability maps of SARS-CoV (**Fig. 1e**) and MERS-CoV (**Fig. 1f**), which are the two other coronaviruses known to have caused past outbreaks. 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.

Coronaviruses are RNA viruses (21), which generally have very high mutation rates (22). Therefore, we examined the effect of host immune pressure on genetic drifts in SARS-CoV-2. We showed the mutational rates in the viral genomes for SARS-CoV-2, SARS-CoV and MERS-CoV (**Fig. 2a-c**). SARS-CoV-2 only emerged a short time ago, which likely explains the lack of significant amount of genetic variation (**Fig. 2a,** n=114). In comparison, SARS-CoV (**Fig. 2b**, n=19) and MERS-CoV (**Fig. 2c**, n=519) have both accumulated significant variation among the different strains. However, in SARS-CoV-2, the genomic regions with higher mutational rates can already be discerned (**Fig. 2a**), which are roughly 0-9kb and 21kb-30kb, and they largely overlap with the highly mutated regions of SARS-CoV and MERS-CoV (**Fig. 2bc**). This indicates that we may expect a similar level of genetic variation for this virus in the future, which might also occur in these same regions.

Inspired by this observation, we examined whether we can detect any correlation between host immune pressure and genetic drifts of SARS-CoV-2, even at the early stage of viral evolution. We examined individual mutations in each SARV-CoV-2 isolate with respect to the reference genome. We predicted the T cell epitopes and B cell epitopes that would be lost and gained due to the mutations in each strain. For T cell epitopes, we take a sum of the lost/gained epitopes weighted by the Chinese HLA allele frequency, to generate an overall "lost" immunogenicity score, and an overall "gained" immunogenicity score. **Fig. 2d** ($CD8^+$ T epitopes) and **Fig. 2e** ($CD4^+$ T epitopes) show that the gained and lost immunogenicity scores are very comparable for most strains, which is just an effect of random change in the epitope profiles due to mutations. But strikingly, the strains that showed a large difference in the lost and gained immunogenicity almost all have dramatic net loss, rather than net gain, of immunogenicity (**Fig. 2d and 2e**). Similarly, we calculated the number of amino acid residues that could encode B cell epitopes predicted from the protein sequences of each SARS-CoV-2 strain (**Fig. 2f**), and also found that there is an overall decrease of the number of B cell epitopes in the isolates compared with the reference genome (Pval=0.001). We limited this analysis to amino acid residues in stretches of >10 epitope-encoding residues, and made the same observation (Pval=0.038, **Sup. Fig. 4a**). Furthermore, we grouped the SARS-CoV-2 strains by their collection times in Dec/2019, Jan/2020, and Feb/2020, and compared the changes in T cell and B cell immunogenicity (**Sup. Fig. 5**). We confirmed the latter appearing strains are indeed less immunogenic than earlier strains. Taken together, these observations show that the genetic evolution of SARS-CoV-2 in humans is subject to immune pressure, and SARS-CoV-2 may be using this mechanism to evade immune surveillance by the host.

SARS-CoV-2 is still in the early stages of genetic drifts, which has brought difficulty to capturing the correlation between immune pressure and genetic drifts, given the low signal/noise ratio and low granularity of data. To further corroborate our observation, we performed the same analyses for SARS-CoV (**Fig. 2g-i** and **Sup. Fig. 4b**) and MERS-CoV (**Fig. 2j-l** and **Sup. Fig. 4c**). We observed similar phenomena that the viral strains with large immunogenicity changes almost all have net loss, rather than net gain, of immunogenicity, for both T cell epitopes and B

cell epitopes. This suggests that the immune pressure-induced genetic drifts are common to coronaviruses during human circulation, and further supports our conjecture that future genetic drifts can happen in SARS-CoV-2 under this mechanism.

To facilitate immunological studies of SARS-CoV-2, we created the SARS-CoV-2 Immune Viewer (**Sup. Fig. 6**) to openly share the viral immunogenicity data of SARS-CoV-2, and also SARS-CoV and MERS-CoV. The Viewer displays a phylogenetic tree with annotations of the strains overlaid (**Sup. Fig. 6a**). The tree allows the users to highlight the strains of virus according to the annotations. Upon clicking the node of each tree, the user will be able to download a small data packet for that strain of SARS-CoV-2, which can be directly visualized on the desktop JBrowse software (23) (https://jbrowse.org/blog/). The JBrowse visualization demonstrates the genomic sequence, protein annotations, and T cell and B cell epitopes of each viral isolate (**Sup. Fig. 6b**). In addition to downloading of the JBrowse data, we also provided user-friendly visualization functionality for researchers to examine immunogenicity strength of different genomic regions of each of the three viruses online (**Sup. Fig. 6c**). Users can either zoom in or zoom out to focus on specific genomic regions or examine the global pattern of T cell/B cell epitope profiles. We also showed the mutational rates of the viral genome along with the immunogenicity maps.

In summary, we characterized the immune vulnerability landscape of SARS-CoV-2, and compared it with those of SARS-CoV and MERS-CoV. We reported the viral genomic regions that encode high density T cell epitopes and B cell epitopes in each strain of SARS-CoV-2, which could be more suitable for peptide vaccine and anti-viral antibody development. Similar to a previous report (17), we also found that the amino acid sequences of the S protein RBMs of SARS-CoV and SARS-CoV-2 diverge in a manner that may reflect unique immune epitopes. To disseminate our research, we created a publicly accessible database, the SARV-CoV-2 Immune Viewer, for easy exploration and downloading of results. The database is under continuous development, and will be updated when new strains of SARS-CoV-2 are made available.

We found evidence that the mutations in SARS-CoV-2 are more than merely random genetic drifts, by showing some strains of SARS-CoV-2 have immunogenicity loss for T cell-based and/or B cell-based host immune surveillance, and the latter appearing strains are less immunogenic than earlier strains. We found the same phenomenon in SARS-CoV and MERS-CoV, which further supports the validity of this observation. Curiously, for all three coronaviruses, more viral strains showed large immunogenicity loss in class II MHC presentation, than class I MHC presentation. This suggests that the $CD4^+$ T cell-class II MHC axis, compared with the $CD8^+$ T cell-class I MHC axis, may be more important for viral recognition and clearance by the host immune system, in alignment with recent research showing the critical antiviral roles of $CD4^+$ T cells (24–27).

Genetic variations can modify the immunogenicity landscape of the virus, and impact its survival fitness. The selection of effective vaccination epitopes should focus on parts of viral proteins

with high potential for generating immunogenic epitopes, and with less chance of mutation. The low level of genetic variation in SARS-CoV-2 could merely be a sampling issue due to the short time this virus has circulated in humans. However, the domains of genomes that are highly mutated in SARS-CoV and MERS-CoV are already more highly mutated in SARS-CoV-2, which means that sustained mutations could happen in these same regions of SARS-CoV-2 in the future.

Conclusions

Overall, our work provides a window into the immunological features of SARS-CoV-2, and have yielded curious insights into the evolution of this virus. We hope our work could aid therapeutic development against this virus to stop this pandemic earlier and to prevent future outbreaks.

Methods

Acquisition of the viral genome sequences

The SARS-CoV-2 complete genome sequences and meta data were downloaded from the https://bigd.big.ac.cn/ncov database, before the data lock of March 6th, 2020. The reference genome was acquired from NCBI, which is one of the first few isolates of SARS-CoV-2 collected in late December of 2019: https://www.ncbi.nlm.nih.gov/nuccore/MN908947. The complete genome SARS-CoV and MERS-CoV 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.

Each isolate's genomic sequence was aligned to the reference genome sequence using EMBOSS needle with the gap opening penalty of 20 and the gap extension penalty of 0. The sequence differences were identified as genomic variants and they were annotated using Annomen. In **Fig. 2a-c**, the semi-transparent boxes mark the regions of high mutational rates due to artefacts of incomplete sequencing. These regions are shielded from calculations of genomic variants. The protein sequences of the isolates were determined based on the sequence differences and the reference gene annotations. The isolate protein sequences were then aligned to the reference protein sequences using EMBOSS needle and the strains whose protein alignments do not cover >90% of the reference protein sequence were ignored.

Prediction of T cell and B cell epitopes

NetMHCpan (v4.0) (28) and NetMHCIIpan (v3.2) (14) with default threshold options were used to predict T cell peptides from the viral proteins that bind to human MHC class I and II proteins for all the available HLA alleles. The HLA allele population frequency for the Chinese population was acquired from Kwok *et al* (15) and population frequency for the European

population was from Mack *et al (16)*. The B cell epitope predictions were made by the BepiPred 2.0 software (17), with default parameters. Amino acids with B cell epitope prediction scores >0.6 are regarded as having high likelihoods of generating linear antibodies.

DNA and protein sequence alignment

The command-line version of MUSCLE (v3.8.31) (29, 30) 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 (SARS-CoV-2) and NP_828851.1 (SARS-CoV), was performed using EMBOSS needle (31) with the BLOSUM62 scoring matrix.

Website development

The Immune Viewer is a dynamic website. It is developed using 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 SARS-CoV-2 strains across multiple geographic regions. The data packet downloaded for each viral strain can be directly visualized in the desktop JBrowse software, which can be downloaded from https://jbrowse.org/blog.

Statistical analyses

All computations and statistical analyses were carried out in the R and Python 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 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.

FIGURE AND TABLE LEGENDS

Fig. 1 T cell- and B cell-mediated immune vulnerability landscape of SARS-CoV-2 in the Chinese population. (a) The $CD4^+$ and $CD8^+$ T cell epitope profiles of SARS-CoV-2. The Y axis shows the immunogenicity intensity as described in the method section. (b) The overall T cell and B cell immunogenicity strengths of SARS-CoV-2, SARS-CoV, and MERS-CoV in the

Chinese and European populations. (c) The B cell epitope profiles of the SARS-CoV-2. The Y axis shows the predicted B cell epitope score. Only showing residues with predicted epitope score >0.6. (d) BLASTing the motif binding domain of the SARS-CoV-2 S protein and the SARS-CoV S protein. (e) The T cell and B cell epitope profiles of SARS-CoV. (f) The T cell and B cell epitope profiles of MERS-CoV.

Fig. 2 Genetic drifts in SARS-CoV-2 are influenced by host immune pressure. (a-c) The relative mutational rate profiles of the three coronaviruses: (a) SARS-CoV-2, (b) SARS-CoV, and (c) MERS-CoV. The relative mutational rates are calculated by the percentage of nucleotides (nts) from all strains of each virus (n=114 for SARS-CoV-2, 19 for SARS-CoV-, and 519 for MERS-CoV) in each 60-nt bin that are different from the reference genome. The semi-transparent boxes mark the regions of high mutational rates due to artefacts of incomplete sequencing. These regions are shielded from calculations of gain and loss of immunogenicity. (d-f) The "lost" immunogenicity and "gained" immunogenicity due to mutations in each SARS-CoV-2 isolate compared with the reference genome. (d) $CD8^+$ T epitopes, (e) $CD4^+$ T epitopes, and (f) B cell epitopes. The red line indicates the total number of predicted B cell epitope-encoding amino acids in the reference sequence. (g-i) The same immunogenicity change analyses as in (d-f), but for SARS-CoV. (j-l) The same analysis for MERS-CoV.

Sup. Fig. 1 The T cell epitope profiles of the European population. (a) SARS-CoV-2, (b) SARS-CoV, and (c) MERS-CoV.

Sup. Fig. 2 The variation of T cell epitope profiles for SARS-CoV-2, SARS-CoV and MERS-CoV across populations. The heatmap represents the number of immunogenic binding epitopes across the binned genomes (500bp) of (a) SARS-CoV-2, (b) SARS-CoV and (c) MERS-CoV 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. Fig. 3 Part of the multiple alignment results of all the SARS-CoV-2 strains.

Sup. Fig. 4 B cell epitope-encoding amino acids in contiguous stretches of >10 epitope-encoding amino acids. (a) SARS-CoV-2, (b) SARS-CoV, and (c) MERS-CoV. Red lines denote the total numbers of epitope-encoding amino acids of the reference sequences.

Sup. Fig. 5 The trend of immunogenicity to SARS-CoV-2 over time. (a) "Gained/lost" T cell immunogenicity (class I). (b) "Gained/lost" T cell immunogenicity (class II). (c) Changes in B cell immunogenicity (all epitope-encoding residues). (d) Changes in B cell immunogenicity (epitope-encoding residues in stretches of >10 epitope-encoding residues). The strains were grouped according to their collection time of Dec/2019, Jan/2020, and Feb/2020. Medians for

each group were listed below the plots; median=2875 residues for all groups in (d). One-way Wilcoxon rank-sum test was used to compare Feb/2020 vs. Dec/2019 groups in (a) (Pval=0.048) and (c) (Pval=0.0004).

Sup. Fig. 6 A continuously updated database of the immune vulnerability landscape of SARS-CoV-2. (a) Phylogenetic tree of viral strains, which allows subsetting based on annotations by going through a series of drop-down boxes. (b) Upon clicking each node, the user will be able to download a small data packet for the corresponding viral strain, which can be visualized in JBrowse desktop (https://jbrowse.org/blog). A screenshot of an example JBrowse session is shown. (c) Additional visualization functionality to examine immunogenicity profiles 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.

Sup. Table 1 Genomics regions of SARS-CoV-2 that are T cell epitope-enriched

Sup. Table 2 Genomics regions of SARS-CoV-2 that are B cell epitope-enriched

Declarations

ETHICS APPROVAL AND CONSENT TO PARTICIPATE

Not applicable

CONSENT FOR PUBLICATION

Not applicable

DATA AVAILABILITY

The SARS-CoV-2 Immune Viewer is publicly accessible, with the reference genome, protein annotation, T/B cell epitope prediction results freely downloadable for each strain: https://qbrc.swmed.edu/projects/2019ncov_immuneviewer/.

COMPETING INTERESTS

The authors declare no conflicts of interest related to this work.

FUNDING

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

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. J.S., 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.

Bibliography

1. Li,Q., Guan,X., Wu,P., Wang,X., Zhou,L., Tong,Y., Ren,R., Leung,K.S.M., Lau,E.H.Y., Wong,J.Y., *et al.* (2020) Early Transmission Dynamics in Wuhan, China, of Novel Coronavirus-Infected Pneumonia. *N. Engl. J. Med.*, 10.1056/NEJMoa2001316.

2. de Wit,E., van Doremalen,N., Falzarano,D. and Munster,V.J. (2016) SARS and MERS: recent insights into emerging coronaviruses. *Nat. Rev. Microbiol.*, **14**, 523–534.

3. Huang,C., Wang,Y., Li,X., Ren,L., Zhao,J., Hu,Y., Zhang,L., Fan,G., Xu,J., Gu,X., *et al.* (2020) Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. *Lancet*, **395**, 497–506.

4. Holshue,M.L., DeBolt,C., Lindquist,S., Lofy,K.H., Wiesman,J., Bruce,H., Spitters,C., Ericson,K., Wilkerson,S., Tural,A., *et al.* (2020) First case of 2019 novel coronavirus in the united states. *N. Engl. J. Med.*, 10.1056/NEJMoa2001191.

5. Wrapp,D., Wang,N., Corbett,K.S., Goldsmith,J.A., Hsieh,C.-L., Abiona,O., Graham,B.S. and McLellan,J.S. (2020) Cryo-EM structure of the 2019-nCoV spike in the prefusion conformation. *Science*, 10.1126/science.abb2507.

6. Lu,R., Zhao,X., Li,J., Niu,P., Yang,B., Wu,H., Wang,W., Song,H., Huang,B., Zhu,N., *et al.* (2020) Genomic characterisation and epidemiology of 2019 novel coronavirus: implications for virus origins and receptor binding. *Lancet*, **395**, 565–574.

7. Zhou,P., Yang,X.-L., Wang,X.-G., Hu,B., Zhang,L., Zhang,W., Si,H.-R., Zhu,Y., Li,B., Huang,C.-L., *et al.* (2020) A pneumonia outbreak associated with a new coronavirus of probable bat origin. *Nature*, 10.1038/s41586-020-2012-7.

8. Pfefferkorn,E.R. and Pfefferkorn,L.C. (1976) Arabinosyl nucleosides inhibit Toxoplasma gondii and allow the selection of resistant mutants. *J. Parasitol.*, **62**, 993–999.

9. Janice Oh,H.-L., Ken-En Gan,S., Bertoletti,A. and Tan,Y.-J. (2012) Understanding the T cell immune response in SARS coronavirus infection. *Emerg. Microbes Infect.*, **1**, e23.

10. Chen,J., Lau,Y.F., Lamirande,E.W., Paddock,C.D., Bartlett,J.H., Zaki,S.R. and Subbarao,K.

(2010) 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.

11. Testa,J.S., Shetty,V., Hafner,J., Nickens,Z., Kamal,S., Sinnathamby,G. and Philip,R. (2012) 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.

12. Zhou,C., Zhou,L. and Chen,Y.-H. (2012) Immunization with high epitope density of M2e derived from 2009 pandemic H1N1 elicits protective immunity in mice. *Vaccine*, **30**, 3463–3469.

13. Nielsen,M. and Andreatta,M. (2016) 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.

14. Jensen,K.K., Andreatta,M., Marcatili,P., Buus,S., Greenbaum,J.A., Yan,Z., Sette,A., Peters,B. and Nielsen,M. (2018) Improved methods for predicting peptide binding affinity to MHC class II molecules. *Immunology*, **154**, 394–406.

15. Kwok,J., Guo,M., Yang,W., Lee,C.K., Ho,J., Tang,W.H., Chan,Y.S., Middleton,D., Lu,L.W. and Chan,G.C.F. (2016) HLA-A, -B, -C, and -DRB1 genotyping and haplotype frequencies for a Hong Kong Chinese population of 7595 individuals. *Hum. Immunol.*, **77**, 1111–1112.

16. Mack,S.J., Tu,B., Lazaro,A., Yang,R., Lancaster,A.K., Cao,K., Ng,J. and Hurley,C.K. (2009) 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.

17. Jespersen,M.C., Peters,B., Nielsen,M. and Marcatili,P. (2017) BepiPred-2.0: improving sequence-based B-cell epitope prediction using conformational epitopes. *Nucleic Acids Res.*, **45**, W24–W29.

18. Soria-Guerra,R.E., Nieto-Gomez,R., Govea-Alonso,D.O. and Rosales-Mendoza,S. (2015) An overview of bioinformatics tools for epitope prediction: implications on vaccine development. *J. Biomed. Inform.*, **53**, 405–414.

19. Sanchez-Trincado,J.L., Gomez-Perosanz,M. and Reche,P.A. (2017) Fundamentals and Methods for T- and B-Cell Epitope Prediction. *J. Immunol. Res.*, **2017**, 2680160.

20. Wan,Y., Shang,J., Graham,R., Baric,R.S. and Li,F. (2020) Receptor recognition by novel coronavirus from Wuhan: An analysis based on decade-long structural studies of SARS. *J. Virol.*, 10.1128/JVI.00127-20.

21. Fehr,A.R. and Perlman,S. (2015) Coronaviruses: an overview of their replication and pathogenesis. *Methods Mol. Biol.*, **1282**, 1–23.

22. Duffy,S. (2018) Why are RNA virus mutation rates so damn high? *PLoS Biol.*, **16**, e3000003.

23. Buels,R., Yao,E., Diesh,C.M., Hayes,R.D., Munoz-Torres,M., Helt,G., Goodstein,D.M., Elsik,C.G., Lewis,S.E., Stein,L., *et al.* (2016) JBrowse: a dynamic web platform for genome visualization and analysis. *Genome Biol.*, **17**, 66.

24. Phares,T.W., Stohlman,S.A., Hwang,M., Min,B., Hinton,D.R. and Bergmann,C.C. (2012)

CD4 T cells promote CD8 T cell immunity at the priming and effector site during viral encephalitis. *J. Virol.*, **86**, 2416–2427.

25. Sant,A.J. and McMichael,A. (2012) Revealing the role of CD4(+) T cells in viral immunity. *J. Exp. Med.*, **209**, 1391–1395.

26. Cullen,J.G., McQuilten,H.A., Quinn,K.M., Olshansky,M., Russ,B.E., Morey,A., Wei,S., Prier,J.E., La Gruta,N.L., Doherty,P.C., *et al.* (2019) CD4+ T help promotes influenza virus-specific CD8+ T cell memory by limiting metabolic dysfunction. *Proc Natl Acad Sci USA*, **116**, 4481–4488.

27. Phares,T.W., Stohlman,S.A., Hinton,D.R. and Bergmann,C.C. (2012) Enhanced CD8 T-cell anti-viral function and clinical disease in B7-H1-deficient mice requires CD4 T cells during encephalomyelitis. *J. Neuroinflammation*, **9**, 269.

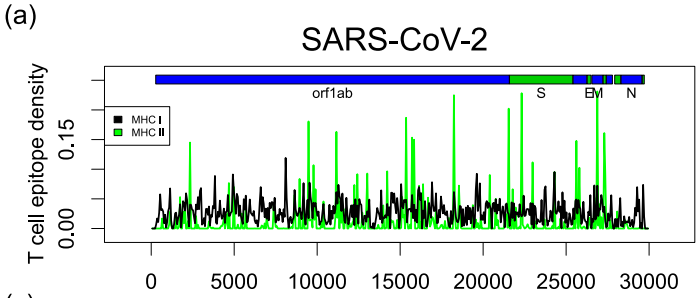
28. Jurtz,V., Paul,S., Andreatta,M., Marcatili,P., Peters,B. and Nielsen,M. (2017) NetMHCpan-4.0: Improved Peptide-MHC Class I Interaction Predictions Integrating Eluted Ligand and Peptide Binding Affinity Data. *J. Immunol.*, **199**, 3360–3368.

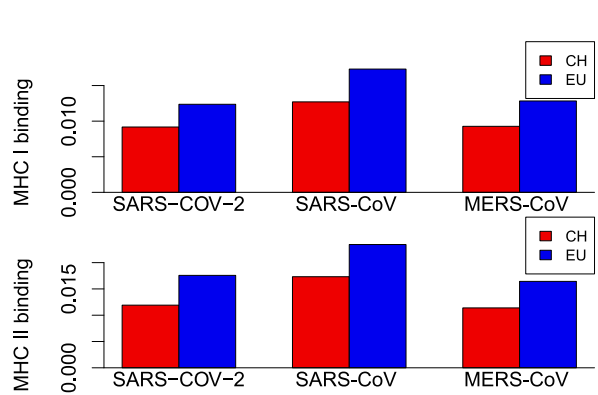
29. Edgar,R.C. (2004) MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics*, **5**, 113.

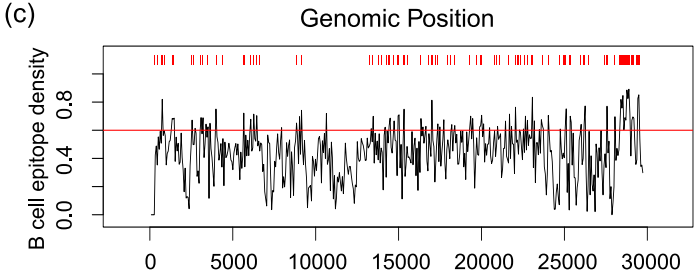
30. Edgar,R.C. (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.*, **32**, 1792–1797.

31. Rice,P., Longden,I. and Bleasby,A. (2000) EMBOSS: the european molecular biology open software suite. *Trends Genet.*, **16**, 276–277.

Fig. 1



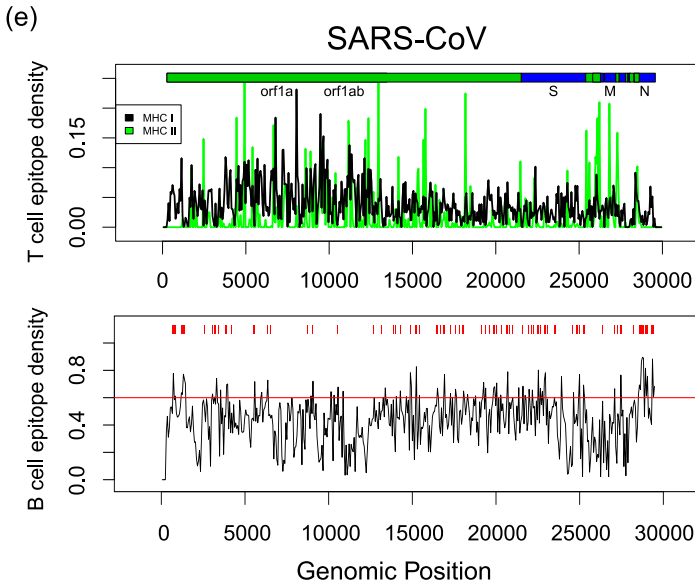




(d)

```
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
```



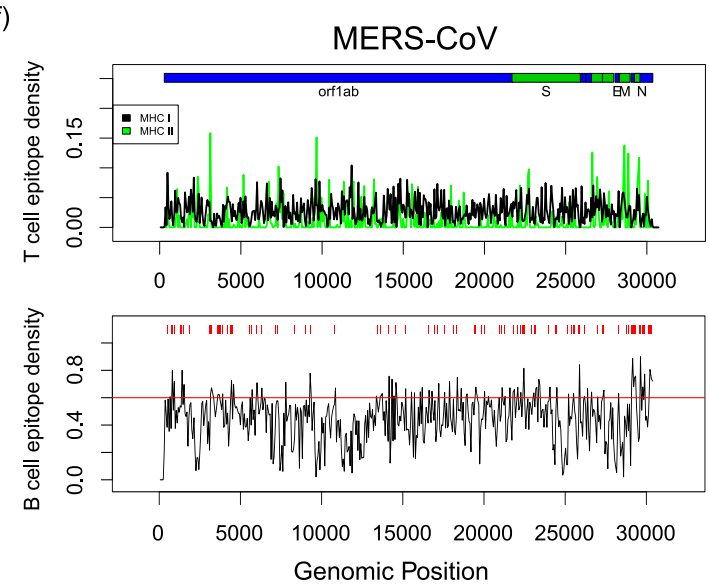
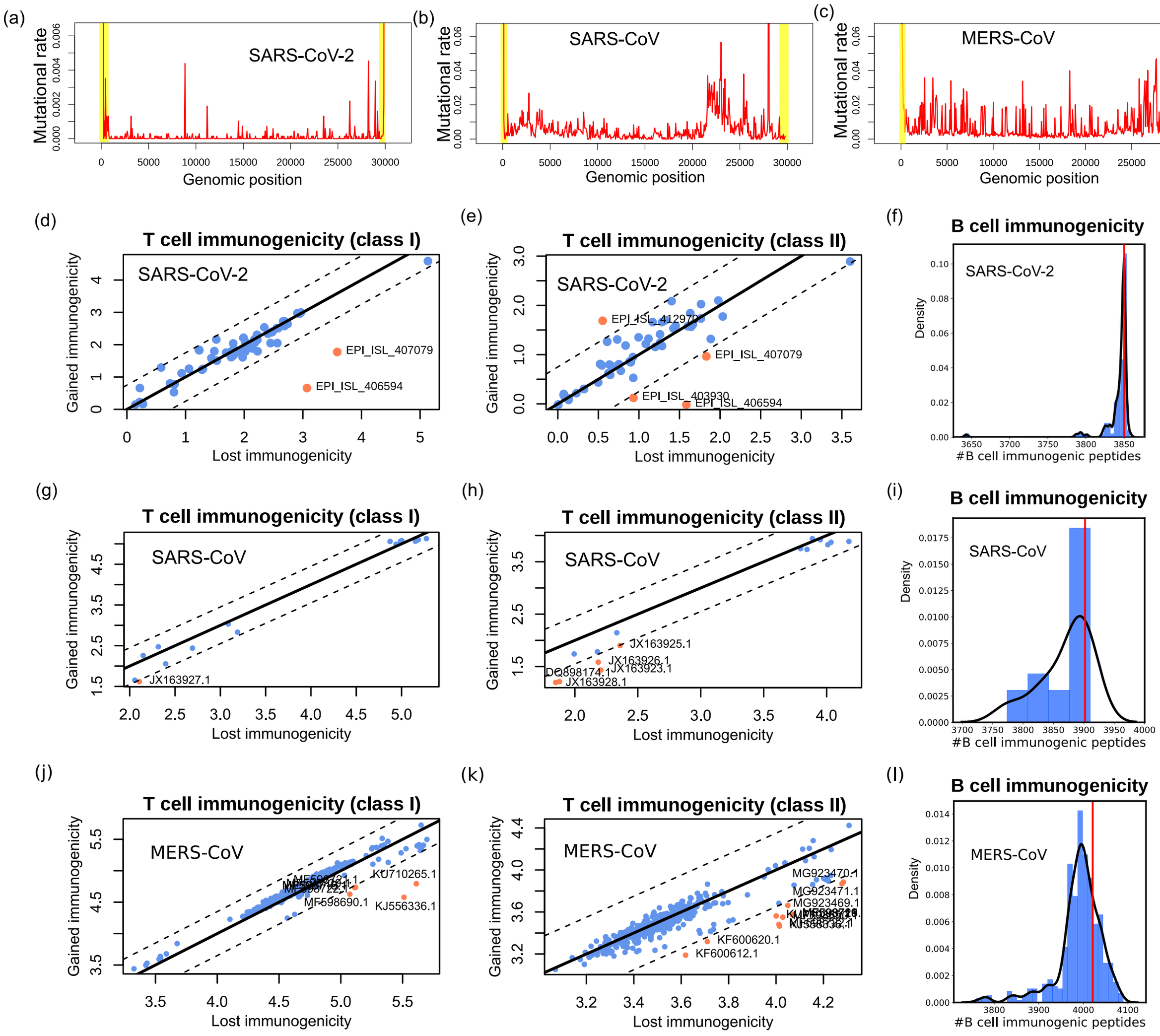


Fig. 2


(a)
SARS-CoV-2
Mutational rate
Genomic position
(b)
SARS-CoV
Mutational rate
Genomic position
(c)
MERS-CoV
Mutational rate
Genomic position
(d)
T cell immunogenicity (class I)
SARS-CoV-2
EPI_ISL_407079
EPI_ISL_406594
Gained immunogenicity
Lost immunogenicity
(e)
T cell immunogenicity (class II)
SARS-CoV-2
EPI_ISL_407079
EPI_ISL_403930
EPI_ISL_406594
Gained immunogenicity
Lost immunogenicity
(f)
B cell immunogenicity
SARS-CoV-2
Density
#B cell immunogenic peptides
(g)
T cell immunogenicity (class I)
SARS-CoV
JX163927.1
Gained immunogenicity
Lost immunogenicity
(h)
T cell immunogenicity (class II)
SARS-CoV
JX163925.1
JX163926.1
JX163923.1
DQ898174.1
JX163928.1
Gained immunogenicity
Lost immunogenicity
(i)
B cell immunogenicity
SARS-CoV
Density
#B cell immunogenic peptides
(j)
T cell immunogenicity (class I)
MERS-CoV
KU710265.1
MF598690.1
KJ556336.1
Gained immunogenicity
Lost immunogenicity
(k)
T cell immunogenicity (class II)
MERS-CoV
MG923470.1
MG923471.1
MG923469.1
KF600620.1
KF600612.1
Gained immunogenicity
Lost immunogenicity
(l)
B cell immunogenicity
MERS-CoV
Density
#B cell immunogenic peptides