Potential T-cell and B-cell Epitopes of 2019-nCoV

Ethan Fast
Nash
ethanfast@cs.stanford.edu

Binbin Chen
Stanford Medicine
bchen45@stanford.edu

Abstract

As of Feb 16th 2020, 2019-nCoV has infected more than 51,857 people across 26 countries and claimed 1666 lives. 2019-nCoV is a novel form of coronavirus that causes COVID-19 and has high similarity with SARS-CoV. No approved vaccine yet exists for 2019-nCoV or any form of coronavirus. Here we use computational tools from structural biology and machine learning to identify 2019-nCoV T-cell and B-cell epitopes based on viral protein antigen presentation and antibody binding properties. These epitopes can be used to develop more effective vaccines and identify neutralizing antibodies. We identified 405 viral peptides with good antigen presentation scores for both human MHC-I and MHC-II alleles, and two potential neutralizing B-cell epitopes near the 2019-nCoV spike protein receptor binding domain (440-460 and 494-506). Analyzing mutation profiles of 68 viral genomes from four continents, we identified 96 coding-change mutations. These mutations are more likely to occur in regions with good MHC-I presentation scores (p=0.02). No mutations are present near the spike protein receptor binding domain. We validated our computational pipeline with SARS-CoV experimental data.

1 Introduction

Coronaviruses (CoV) first became widely known after the emergence of severe acute respiratory syndrome (SARS) in 2002. They are are positive, single-strand RNA viruses that often infect birds and a variety of mammals and sometimes migrate to humans [1]. While coronaviruses that infect humans usually present mild symptoms (e.g. 15% of cases involving the common cold are caused by a coronavirus), several previous outbreaks, such as SARS and Middle East respiratory syndrome (MERS) [2], have caused significant fatalities in infected populations [3].

In December 2019, several patients with a connection in Wuhan, China developed symptoms of viral pneumonia. Sequencing of viral RNA determined that these cases were caused by a novel coronavirus named 2019-nCoV or SARS-CoV-2 [4, 5]. The virus has infected more than 51,857 people across 26 countries as of February 16th 2020 according to a World Health Organization report [6]. COVID-19, the disease caused by 2019-nCoV, has led to death of 1666 patients and is clearly transmissible between humans [3].

The deployment of a vaccine against 2019-nCoV would arrest its infection rate and help protect vulnerable populations. However, no vaccine has yet passed beyond clinical trails for any coronavirus, whether in humans or other animals [7, 8]. In animal models, inactivated whole virus vaccines without focused immune epitopes offer incomplete protection [8]. Vaccines for SARS-CoV and MERS-CoV have also resulted in hypersensitivity and immunopathologic responses in mice [9, 7]. The diversity among coronavirus strains presents another challenge [9, 1], as there are at least 6 distinct subgroups in the coronavirus genus. And while SARS-CoV and 2019-nCoV share the same subgroup (2b), their genomes are only 77% identical [4]. Understanding the shared and unique epitopes of 2019-nCoV will help the field design better vaccines or diagnostic tests for patients.

Both humoral immunity provided by B-cell antibodies and cellular immunity provided by T-cells are essential for effective vaccines [10, 11]. Though humans can usually mount an antibody response

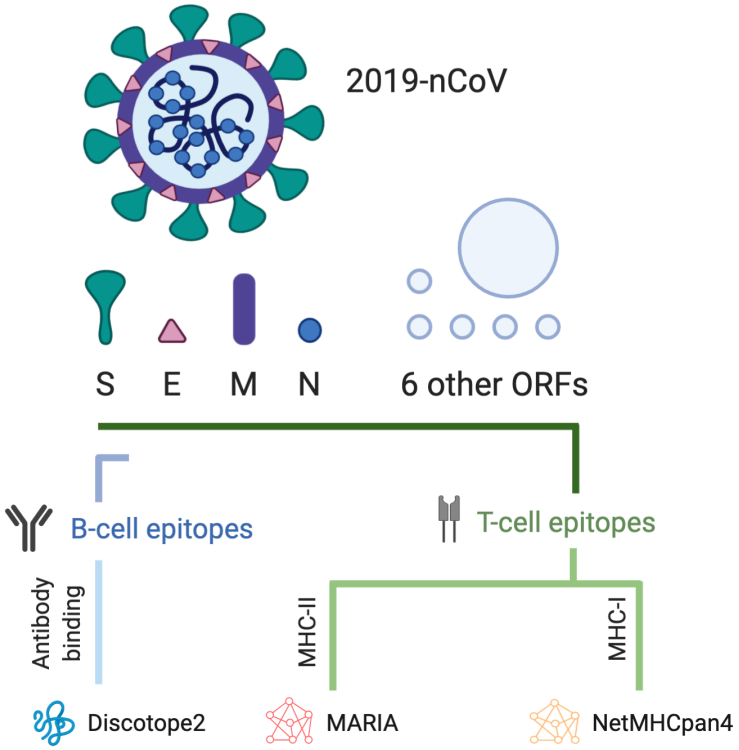

Figure 1: Our computational pipeline to identify T-cell and B-cell epitopes in 2019-nCoV. The 2019-nCoV genome codes for Spike (S), Membrane (M), Envelope (E), Nucleocapsid (N) and at least 6 other open reading frames (ORFs). The S protein is a target for antibodies and we model its 3D structure with Discotope2 to identify likely binding sites (B-cell epitopes). We scan peptide sequences from individual viral protein with NetMHCPan4 and MARIA for MHC-I and MHC-II presentation respectively to identify potential T-cell epitopes. Our MHC-I analysis include common alleles for HLA-A, HLA-B, and HLA-C, and our MHC-II analysis include common alleles of HLA-DR.

against viruses, only neutralizing antibodies can fully block the entry of viruses into human cells [12]. The location of antibody binding sites on a viral protein strongly affects the body's ability to produce neutralizing antibodies; for example, HIV has buried surface proteins that protect themselves from human antibodies [13]. For this reason, it would be valuable to understand whether 2019-nCoV has potential antibody binding sites (B-cell epitopes) near their interacting surface with its known human entry receptor Angiotensin-Converting Enzyme 2 (ACE2).

Beyond neutralizing antibodies, human bodies also rely upon cytotoxic CD8 T-cells and helper CD4 T-cells to fully clear viruses. The presentation of viral peptides by human Major Histocompatibility Complex (MHC or HLA) class I and class II is essential for anti-viral T-cell responses [14]. In contrast to B-cell epitopes, T-cell epitopes can be located anywhere in a viral protein since human cells can process and present both intracellular and extracellular viral peptides [15].

These immunology principles guide the design of our current study (**Fig. 1**). First we focus on T-cell epitopes. 2019-nCoV carries 4 major structural proteins—spike (S), membrane (M), envelope (E) and nucleocapsid (N)—and at least 6 other open reading frames (ORFs) [1, 16]. All protein fragments have the potential to be presented by MHC-I or MHC-II and recognized by T-cells. We apply NetMHCpan4 [17] and MARIA [15], two artificial neural network algorithms, to predict antigen presentation and identify potential T-cell epitopes.

Next we focus on B-cell epitopes. The spike (S) protein is the main trans-membrane glycoprotein expressed on the surface of 2019-nCoV and is responsible for receptor binding and virion entry to cells [1, 18]. We narrow our B-cell epitope search to S protein since it is the most likely target of human neutralizing antibodies. Discopte2 uses a combination of amino acid sequences and protein

surface properties to predict antibody binding sites given a protein structure [19]. We use this tool to identify potential 2019-nCoV B-cell epitopes as the co-crystal structure of 2019-nCoV's spike protein and antibody is yet available.

Finally, we examine the viral mutation pattern. Thanks to global sequencing efforts [20], we are able to obtain a cohort of 68 2019-nCoV genomes from four continents. Based on the resulting viral mutation profile, we examine whether mutations are driven by the immune selection pressure. Highly mutated regions might be excluded from the vaccine candidate pool.

2 Method

2.1 Viral sequences

We obtained 2019-nCoV (SARS-CoV-2019) and SARS-CoV reference sequence data from NCBI GeneBank (NC_045512 and NC_004718) [16, 4]. We then extracted the 2019-nCoV protein sequences of ORF1AB, S, ORF3A, E, M, ORF6, ORF7A, ORF7B, ORF8, N, and ORF10 based on the reference genome. We obtained viral sequences associated with 68 patients from GISAID on Feb 1st 2020 (**Supplementary Table 1**) [20]. Genomes with single protein are not included in the analysis.

2.2 MHC antigen presentation prediction

We broke each gene sequence in 2019-nCoV into sliding windows of length 9, the median length of MHC-I ligands, and 15, the median length of MHC-II ligands. We used netMHCpan4 [17] and MARIA [15] to predict MHC-I and MHC-II presentation scores, respectively. We used 32 MHC alleles common in the Chinese population (>4% allele frequency, 7 HLA-A, 8 HLA-B, 9 HLA-C, 8 HLA-DRB1), as determined by an analysis of human populations [21]. The complete list of common MHC alleles is included in **Supplementary Table 2**.

Both MARIA and netMHCpan4 return percentiles that characterize a peptide's likelihood of presentation relative to preset distribution of random human peptide scores. Prior work recommends thresholds of 98% for NetMHCpan4 and 95% for MARIA to determine reasonable presenters. We also applied our analysis with a more stringent 99.5% threshold for both NetMHCpan4 and MARIA. We used gene expression values of 50 TPM when running MARIA to reflect the high expression values of viral genes in human cells.

When aggregating alleles across MHC-I and MHC-II to report overall coverage, we marked a peptide sequence as covered if it is presented by more then 33% of common alleles. We chose 33% as a cut-off because it suggests a high (>90%) probability that at least one allele can present this peptide assuming the patient carries six MHC alleles (e.g. 2 As, 2 Bs, and 2 Cs) and the distribution of common MHC alleles in the population is random. We ranked potential T-cell epitopes based on their MHC-I and MHC-II presentation coverage across alleles.

2.3 T-cell epitope validation

We applied our methodology to known SARS T-cell epitopes and non-epitope SARS peptides to estimate our ability to predict 2019-nCoV epitopes. For MHC-I, we curated 17 experimentally determined HLA-A*02:01 associated CD8 T-cell epitopes and 1236 non-epitope 9mer sliding windows on SARS S protein [22, 23, 24]. For MHC-II, we curated 3 experimentally determined CD4 T-cell epitopes and 246 non-epitopes on SARS S protein [25]. No specific HLA-DR alleles were reported in the original study, so we used HLA-DRB1*09:01 and 15:01 (common alleles) to run MARIA. To calculate sensitivity and specificity for this validation set, we labeled any peptide sequence above the 98th (MHC-I) or 95th (MHC-II) percentile as a positive epitope prediction. Any sequence below that threshold we labeled as a negative prediction. We calculated AUC scores (AUROC) to estimate the overall performance of our methodology.

2.4 Homology modeling of 2019-nCoV S protein

We obtained an approximate 3D structure of 2019-nCoV S protein by homology modeling SARS S protein (PDB: 6ACC) with SWISS-MODEL [26, 27]. S proteins from 2019-nCoV and SARS share 93% similarity. The modeled structure of 2019-nCoV S protein has a QMEAN of -3.63

Gene	Length	HLA-A		HLA-B		HLA-C		HLA-DR	
ORF1ab	7096	370	5.21%	516	7.27%	775***	10.9%	918	12.9%
S	1273	59	4.63%	91	7.15%	132***	10.4%	95	7.46%
ORF3a	275	23	8.36%	16	5.82%	34**	12.4%	11	4.00%
E	75	11	14.7%	8	10.7%	12	16.0%	3	4.00%
M	222	12	5.41%	18	8.11%	27*	12.2%	14	6.31%
ORF6	61	6	9.84%	5	8.20%	8	13.1%	1	1.64%
ORF7a	121	7	5.79%	8	6.61%	11	9.09%	3	2.48%
ORF7b	43	4	9.30%	2	4.65%	3	6.98%	0	0.00%
ORF8	121	6	4.96%	6	4.96%	9	7.44%	2	1.65%
N	419	9	2.15%	19	4.53%	24	5.73%	24	5.73%
ORF10	38	2	5.26%	3	7.89%	8	21.1%	1	2.63%

Table 1: Summary of potential 2019-nCoV T-cell epitope candidates based on HLA antigen presentation scores. The HLA-A, HLA-B, HLA-C, and HLA-DR columns refer to the number of peptide sequences that were predicted to present across more than one third of common alleles among the Chinese population. Length refers to the number of amino acids in a given gene. HLA-C are more likely to present 2019-nCoV antigens across one third of alleles compared to HLA-A or B. Fisher's exact test P-value levels are indicated with *** (< 0.001), ** (< 0.01), and * (< 0.05).

(**Supplementary File 1**). We used a computational docking algorithm, ClusPro2 [28], to estimate the interactive surface of 2019-nCoV with its known human receptor, ACE2 [5] (**Supplementary File 2**). Graphic rendering and analysis was performed with PyMOL [29].

2.5 B-cell epitope prediction and validation

We predicted likely human antibody binding sites (B-cell epitopes) on SARS and 2019-nCoV S protein with Disctope2 [19]. Our analysis focused on neutralizing binding sites by only examining residues 1-600. The full prediction results can be found in **Supplementary Table 3, 4**. To validate our B-cell epitope predictions, we compared our top 3 B-cell epitopes for SARS S protein with previously experimentally identified epitopes from three independent studies [30, 31, 32].

2.6 Mutation identification

We aimed to determine whether there was any statistical relationship between regions of mutation and presentation. We translated viral genome sequences into protein sequences using BioPython [33] and indexes from the NCBI reference genome. 2019-nCoV nucleotide position 13468 contains a -1 frame shift signal and we adjusted accordingly to generate ORF1ab. We compared protein sequences from 68 patients to the reference sequence to identify mutations with edit distance analysis. Positions with poor quality reads (e.g. W or Y) were excluded from the analysis. The full sequence and mutation profiles can be found in **Supplementary Table 9, 10**. We compared positions of point mutations with MHC-I or MHC-II presentable regions in the protein with Fisher's exact test. Specifically, we use Fisher's exact test to compare the proportion of 9mers covered by >33% HLA-C alleles to the proportion covered by >33% of HLA-A and HLA-B alleles.

2.7 Statistical analysis

We computed Fisher's exact test with scipy [34] and AUROC with scikit-learn [35].

3 Results

3.1 T-cell epitopes of 2019-nCoV based on MHC presentation

We scanned 11 2019-nCoV genes with NetMHCpan4 and MARIA to identify regions highly presentable by HLA complexes (human MHC). A epitope is labeled positive if presentable by more than one third of common alleles among the Chinese population. We display a summary of our

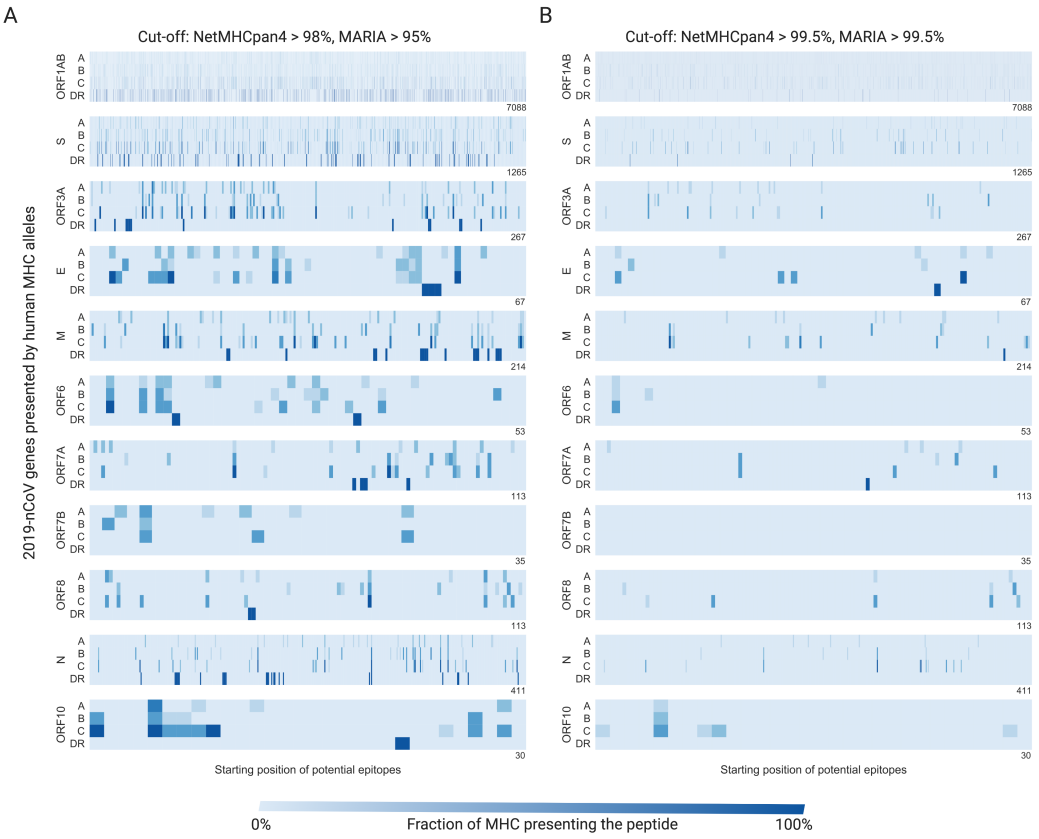

Figure 2: MHC presentation regions for 11 2019-nCoV genes across HLA-A, HLA-B, HLA-C, and HLA-DR. We plot results at both medium **A**) and high **B**) cut-offs for presentation. The X-axis indicates the position of each viral protein, and the y-axis indicates each human HLA gene family. Each blue stripe indicates the fraction of HLA alleles that can present a 9mer or 15mer viral peptide starting from a given position.

T-cell epitope findings in **Table 1**. In **Fig. 2** we show regions of strong MHC presentation across HLA-A, HLA-B, HLA-C, and HLA-DR alleles, both at a medium threshold (**Fig. 2A**), and at a high threshold (**Fig. 2B**). Consistent with the similar finding from SARS-CoV, proteins ORF1AB, S, and E contain high numbers of presentable antigen sequences across both MHC-I and MHC-II [8]. These highly presentable peptides on functional viral proteins provide a candidate pool for T-cell epitope identifications and epitope-focused vaccine design. The presentation scores of individual antigen peptides can be found in **Supplementary Table 5, 6**.

NetMHCpan4 predicts that a larger number of 2019-nCoV protein sequences can be presented by HLA-C alleles across all genes (Fisher's exact test, $p = 4.2e^{-59}$) compared to HLA-A or HLA-B (**Table 1**). Specifically, ORF1ab ($p = 6.7e^{-54}$), S ($p = 1.4e^{-9}$), ORF3a ($p = 0.01$), and M ($p = 0.025$) can be better presented by HLA-C alleles after Bonferroni correction. This might be due to relatively conserved HLA-C binding motif on the 9th position (the main anchor residue) where Leucine, Methionine, Phenylalanine, Tyrosine are commonly favored across alleles [36]. HLA-C*07:02, HLA-C*01:02, and HLA-C*06:02 all have >10% allele frequency in the Chinese population [21]. This suggests the future epitope-based vaccines should include HLA-C related epitopes beyond the common HLA-A*02:01 and HLA-B*40:01 approach.

3.2 T-cell epitope validation

To estimate the ability of antigen presentation scores to identify 2019-nCoV T-cell epitopes, we performed a validation study with known SARS-CoV CD8 and CD4 T-cell epitopes (**Fig. 3**). From 4

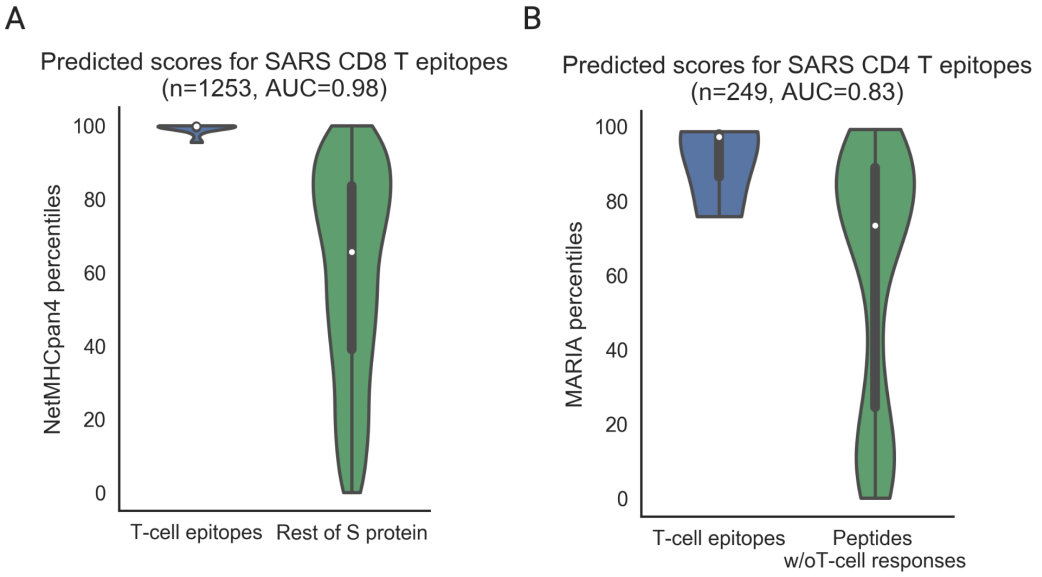

Figure 3: Validation of NetMHCpan and MARIA for T-cell epitope identification with SARS-CoV S protein epitopes. NetMHCpan4 and MARIA presentation percentiles were used to differentiate experimentally identified CD8 T-cell epitopes (MHC-I, n=17)[22, 23, 24] and CD4 T-cell epitopes (MHC-II, n=3)[25] from negative peptides (n=1236 and 246). Negative peptides for CD8 T-cell epitopes are all 9mer sliding windows of S protein without a reported epitope. Negative peptides for CD4 T-cell epitopes are experimentally tested. **A)** NetMHCpan4 has a sensitivity of 82.3% and specificity of 97.0% with a 98th presentation percentile cut-off, and an overall AUC of 0.98. **B)** MARIA has a sensitivity of 66.6% and specificity of 91.1% with a 95th presentation percentile cut-off, and an overall AUC of 0.83.

independent experimental studies [22, 23, 24, 25], we identified known CD8 T-cell epitopes (MHC-I, n=17) and known CD4 T-cell epitopes (MHC-II, n=3). We also obtained known non-epitopes (n=1236 and 246) for CD4, while for CD8 we include all 9mer sliding windows of S protein without a known epitope as non-epitopes. MHC-I presentation (with recommended cut-off 98% [17]) has a sensitivity of 82.3%, specificity of 97.0% and an AUC of 0.98 (**Fig. 3A**). MHC-II presentation (recommended cut-off 95% [15]) has a sensitivity of 66.6%, specificity of 91.1% and an AUC of 0.83 (**Fig. 3B**).

Given the high similarity between SARS-CoV and 2019-nCoV proteins, we expect our analysis on 2019-nCoV to perform similarly against future experimentally validated T-cell epitopes. The detailed scores and sequences of this analysis can be found in **Supplementary Table 7 and 8**.

3.3 Correlation between viral mutation and antigen presentation

Human immune selection pressure has been shown to drive viral mutations which evade immune surveillance (e.g. low MHC presentation). We hypothesize that a similar phenomenon can occur in 2019-nCoV. We curated a cohort of 68 viral genomes across four continents and identified 93 point mutations, 2 nonsense mutation and 1 deletion mutation compared to the published reference genome [5]. We plot point mutations against regions of MHC-I or MHC-II presentation in **Fig. 4**. The full protein sequence and mutation information can be found in **Supplementary Table 9, 10**.

Mutations occur in most genes with the exception of E, ORF6, ORF7b, and N. We identify a recurrent mutation L84S in ORF8 in 20 samples. Mutations are more likely to occur in regions with MHC-I presentation (Fischer's exact test, $p = 0.02$). No relationship was observed with MHC-II ($p = 0.58$) or an aggregate of MHC-I and MHC-II ($p = 0.14$). This is consistent with the role of CD8 T-cell in clearing infected cells and mutations supporting the evasion of immune surveillance [37].

We identified two nonsense mutations (ORF1AB T15372A and ORF7A C184T) in two viral samples from Shenzhen, China (EPI_ISL_406592 and EPI_ISL_406594). ORF7A C184U causes an immediate stop gain and a truncation of the last 60 amino acids from the ORF7A gene, which can

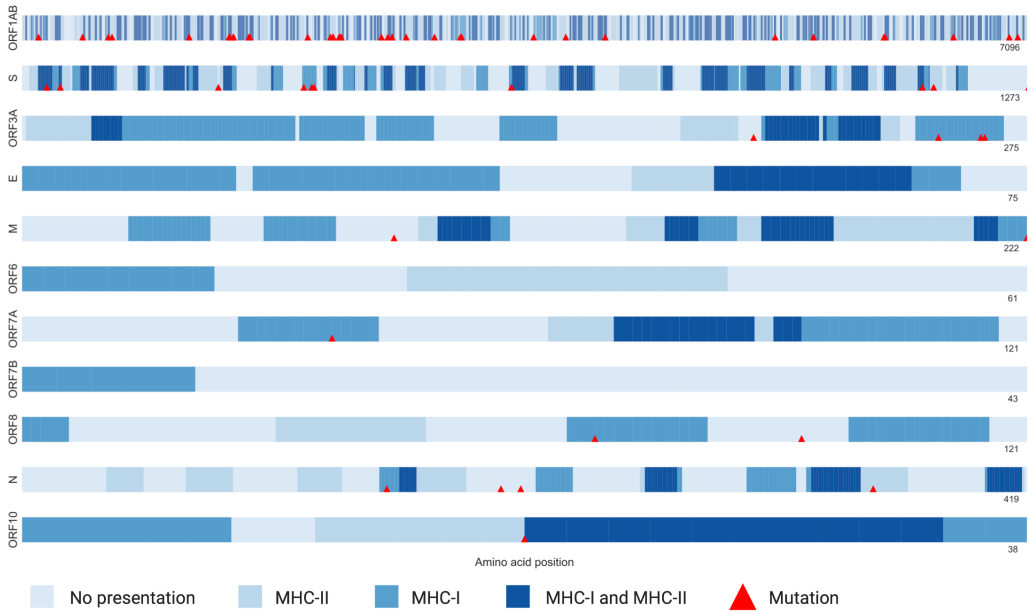

Figure 4: 2019-nCoV point mutations correlates with MHC-I presentation regions. 93 point mutations at 59 unique positions (red triangles) are present in a cohort of 68 2019-nCoV cohort. Each row represents a viral protein and the shades of blue indicate whether MHC-I or MHC-II or both can present an antigen from this region. A region is labeled as presented by MHC-I or MHC-II if more than one third of the corresponding alleles can present the peptide sequences. Mutations are more likely be located in regions with MHC-I presentation ($p = 0.02$). No relationship was observed with MHC-II ($p = 0.58$) or MHC-I and MHC-II double positive region (dark blue, $p = 0.14$).

be presented by multiple MHC alleles as indicated in our heatmap (**Fig. 4**). In a sample from Japan (EPI_ISL_407084), the viral gene ORF1AB loses nucleotides 94-118, which code for GDSVEEVL. An associated antigen FGDSVEEVL is predicted to be good presented by multiple MHC-I alleles (e.g. 99.9% for HLA-C*08:01, 99.2% for HLA-C*01:02, and 99.0% for HLA-B*39:01).

In summary, we observed that 2019-nCoV mutations often occur among or delete presentable peptides, a potential form of immune evasion. Patient MHC allele information and T-cell profiling are needed to better assess this phenomenon. However, patients typically carry fewer than 3 mutations, so a vaccine against multiple epitopes will not be nullified by such a small number of mutations.

3.4 Potential B-cell epitopes of 2019-nCoV

We obtained the likely 3D structure of 2019-nCoV's spike protein (**Supplementary File 1**) by homology modeling with SARS-CoV's spike protein (PDB: 6ACC). We scanned both spike protein structures with Discotope2 to identify potential antibody binding sites on the protein surface (**Fig. 5**). For the SARS-CoV S protein, we identified a strong cluster of antibody sites (>30 residues, pink or red) on the receptor binding domain (RBD, the ACE2 interacting surface). Three independent studies [32, 30, 31] discovered B-cell epitopes in this region (blue) with a combination of in vitro and ex vivo approaches. Additionally, Discotope2 identifies residue 541-555 as another binding site, which was supported by two independent studies (**Fig. 5A**, blue) [30, 31]. This gives us some confidence in the ability of Discotope2 to predict B-cell epitopes if given a unknown protein structure.

For the 2019-nCoV S protein, Discotope2 identified a similar antibody binding site on S protein potential RBD, but with fewer residues (17 residues, **Fig. 5B**). With computational protein-protein docking, we predicted one potential interacting conformation between the 2019-nCoV S protein and the human ACE2 entry receptor (**Fig. 5C**). The main antibody binding site substantially overlaps with the interacting surface where ACE2 binds to S protein, and an antibody binding to this surface is likely to block viral entry into cells. Discotope2 identified anther antibody binding site (residue 246-257) different from SARS-CoV but with lower scores (>-2.5 raw scores).

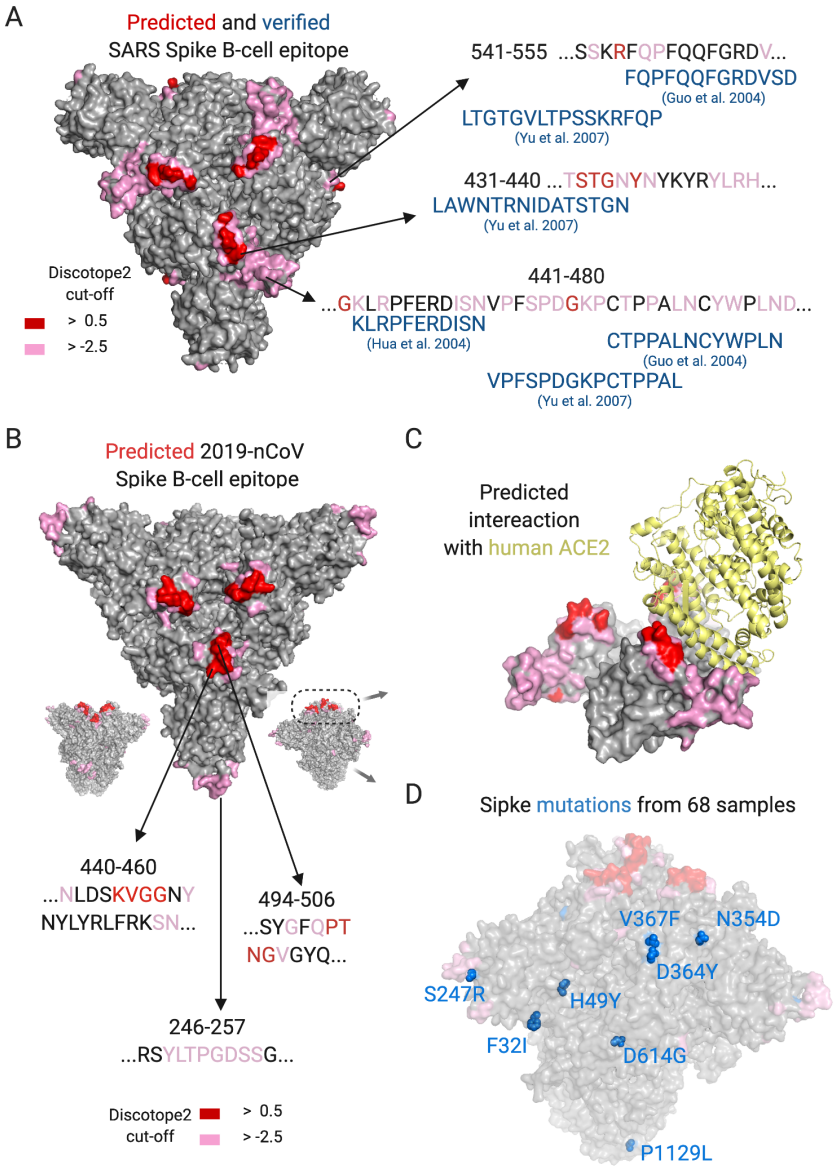

Figure 5: Predicted B-cell epitopes on SARS-CoV and 2019-nCoV spike (S) protein. 3D structures of both spike protein were scanned with Discotope2 to assess potential antibody binding sites (B-cell epitopes). Red indicates residues with score > 0.5 (cut-off for specificity of 90%), and pink indicates residues with score > -2.5 (cut-off for specificity of 80%). **A**) The receptor binding domain (RBD) of SARS spike protein has concentrated high score residues and is comprised of two linear epitopes. One additional binding site is predicted to be around residue 541-555. Three independent experimental studies [32, 30, 31] identify patient antibodies recognize these three linear epitopes (blue). **B**) 2019-nCoV spike protein is predicted to have a similar strong antibody binding site near RBD and an additional site around residue 246-257. **C**) The predicted antibody binding site on 2019-nCov RBD potentially overlaps with the interacting surface for the known human entry receptor ACE2 (yellow). **D**) Eight 2019-nCoV spike protein point mutations are present in a cohort of 68 viral samples. No mutations are near the ACE2 interacting surface. One mutation (S247R) occurs near one of the predicted antibody binding sites (residue 246-257).

Gene	Sequence	Position	MHC-I Cov.		MHC-II Cov.		Antibody
S	SYGFQPTNGVGYQPY	494	Yes	52%	Yes	100%	Yes
	SQSIIAYTMSLGAEN	689	Yes	74%	Yes	100%	No
	IPTNFTISVTTEILP	714	Yes	70%	Yes	100%	No
	AAAYYVGYLQPRTFL	262	Yes	65%	Yes	100%	No
	APHGVVFLHVTYVPA	1056	Yes	65%	Yes	100%	?
ORF1ab	DGEVITFDNLKTLLS	1547	Yes	83%	Yes	100%	No
	EVRTIKVFTTVDNIN	1564	Yes	78%	Yes	100%	No
	IINLVQMAPISAMVR	2368	Yes	78%	Yes	100%	No
	NPTTFHLDGEVITFD	1540	Yes	74%	Yes	100%	No
	VAAIFYLITPVHVMS	2783	Yes	74%	Yes	100%	No
M	IASFRLFARTRSMWS	97	Yes	65%	Yes	100%	?
N	ATKAYNVTQAFGRRG	264	Yes	74%	Yes	100%	?
E	VKPSFYVYSRVKNLN	52	Yes	74%	Yes	100%	?

Table 2: Top potential epitopes for key 2019-nCoV proteins. We ranked epitopes based on their likely coverage of presentation by MHC-I and MHC-II alleles. S protein 494-508 is highly ranked based on MHC presentation and is also one of the predicted top B-cell epitopes, localized near the S protein receptor binding domain (**Figure 5**). MHC-I coverage is calculated by the 9mer with the highest MHC-I coverage for each epitope (highlighted in orange). All candidates are likely to be presented by both MHC-I and MHC-II. A question mark (?) under the antibody column indicates that one or more SARS homolog of this peptide is a known B-cell epitope

We identified 8 point mutation sites on S protein from a cohort of 68 2019-nCoV samples (**Fig. 5D**). All mutations are apart from the protein S RBD, and one mutation (S247R) occurs near one of the predicted minor antibody binding site (residue 246-257).

In summary, we observed major structural and B-cell epitope similarity between SARS-CoV and 2019-nCoV spike proteins. A recent Cryo-EM study of 2019-nCoV S protein [18] supports such structural similarity, and it would be informative to rerun our analysis once the Cryo-EM crystal structure becomes publicly available. RBDs in both proteins seem to be important B-cell epitopes for neutralizing antibodies, however, SARS-CoV appears to have larger attack surface than 2019-nCoV (**Fig. 5A, 5B**). Fortunately, we have not observed any mutations altering binding sites for this major B-cell epitope in 2019-nCoV.

3.5 Promising candidates for 2019-nCoV vaccines

After understanding the landscape of T-cell and B-cell epitopes of 2019-nCoV, we summarize our results in **Table 2** for promising candidates for 2019-nCoV vaccines. Ideal antigens should be presented by multiple MHC-I and MHC-II alleles in a general population and contain linear B-cell epitopes associated with neutralizing antibodies. Specifically, we first identified regions with high coverage of MHC-II, then ranked them based on their MHC-I coverage. We further search these top candidates with 90% sequence similarity on IEDB [38] to assess whether any candidate has been previously reported for being a B-cell epitope or T-cell epitope.

Around 1000 peptide 15mers can be presented by a wide range of MHC-II (>33%), which is consistent with our knowledge about promiscuity of MHC-II presentation [15, 39]. When further requiring a peptide to be presented by more than 33% or 66% of common MHC-I alleles, we narrowed the results to 405 or 44 promising T-cell epitopes, respectively (**Supplementary Table 5, 6**). The top 13 candidates in key viral proteins can be presented by 52-83% of MHC-I alleles (**Table 2**). Most notably, S protein 494-508 is not only a good MHC presenter but also a predicted B-cell epitope near the S protein receptor binding domain (**Fig. 5B**), which indicates a good vaccine candidate.

When searching these 13 candidates on IEDB, we observed several of their homologous peptides in SARS are T-cell and B-cell epitopes. SARS peptide SIVAYTMSL is similar to 2019-nCoV peptide SQSIIAYMSLGAEN and elicits T-cell responses in chromium cytotoxicity assays [40]. Previously

discovered SARS B-cell epitopes LMSFPQAAPHGVVFLHV [41], ASFRLFARTRSMWSF[42], KRTATKQYNVTQAFGRR [43], and SRVKNLNSSEGVPDLLV [44] share high homology with 2019-nCoV S protein 1056-1060, M 97-111, N 264-278, and E 52-66. These search results support validity of our computational pipeline.

4 Discussion

In summary, we analyzed the 2019-nCoV viral genome for epitope candidates and found 405 likely T-Cell epitopes, with strong MHC-I and MHC-II presentation scores, and 2 potential neutralizing B-Cell epitopes on S protein. We validated our methodology by running a similar analysis on SARS-CoV against known SARS T-cell and B-cell epitopes. Inspecting 68 samples, we found that viral mutations have a tendency to occur in regions with strong MHC-I presentation, which can be a form of immune evasion. However, no mutations are present near the spike protein receptor binding domain which is critical for viral entrance to cells and neutralizing antibody binding.

Similar to SARS-CoV [1, 8], the 2019-nCoV spike protein is likely to be immunogenic as it carries a number of both T-cell and B-cell epitopes. Recombinant surface protein vaccines have demonstrated strong clinical utilities in hepatitis B and human papillomavirus (HPV) vaccines [11, 45]. A recombinant 2019-nCoV spike protein vaccine combined with other top epitopes and an appropriate adjuvant could be a reasonable next step for 2019-nCoV vaccine development.

From an economic perspective, the development of coronavirus vaccines faces significant financial hurdles as most such viruses do not cause endemic infections. After an epidemic episode (e.g. SARS) there is little financial incentive for private companies to develop vaccines against a specific strain of virus [10, 2, 8]. To address these incentive issues, governments might provide greater funding support to create vaccines against past and future outbreaks.

Our pipeline provides a framework to identify strong epitope-based vaccine candidates—beyond 2019-nCoV—and might be applied against any unknown pathogens. Previous animal studies do show antigens with high MHC presentation scores are more likely to elicit strong T-cell responses, but the correlation between vaccine efficacy and T-cell responses is relatively weak [14, 46, 7]. When combined with future clinical data, our work can help the field untangle the relationship between antigen presentation scores and vaccine efficacy.

5 Supplementary Tables and Files

Supplementary Table 1. Full details and acknowledgements for the viral genome data used.

Supplementary Table 2. Common HLA alleles among the Chinese population

Supplementary Table 3. Discotope2 prediction for SARS-nCoV spike structure chain A.

Supplementary Table 4. Discotope2 prediction for 2019-nCoV spike model chain A.

Supplementary Table 5. NetMHCpan4 percentiles for all possible 9mer peptide sequences for 2019-nCoV protein.

Supplementary Table 6. MARIA percentiles for all possible 15mer peptide sequences for 2019-nCoV protein and peptides with high coverage for both MHC-I and MHC-II alleles

Supplementary Table 7. Validation CD8 T-cell epitopes for SARS spike protein

Supplementary Table 8. Validation CD4 T-cell epitopes for SARS spike protein.

Supplementary Table 9. Protein sequence and mutation profiles of 68 2019-nCoV samples.

Supplementary Table 10. DNA sequences of 68 2019-nCoV samples.

Supplementary File 1. PDB file for 2019-nCoV spike protein homology modeling.

Supplementary File 2. PDB file for 2019-nCoV spike protein and human ACE2 protein-protein docking.

6 Conflicts of interest

We declared no conflicts of interest.

7 Acknowledgement

We thank Stanford Schor and Allen Lin for the discussion about coronaviruses. We thank all labs contributing to the global efforts of sequencing 2019-nCoV samples and we attached the full acknowledgment list in **Supplementary Table 1**.

References

[1] Perlman, S. & Netland, J. Coronaviruses post-sars: update on replication and pathogenesis. *Nature reviews microbiology* **7**, 439–450 (2009).

[2] Group, W. M.-C. R. *et al.* State of knowledge and data gaps of middle east respiratory syndrome coronavirus (mers-cov) in humans. *PLoS currents* **5** (2013).

[3] Huang, C. *et al.* Clinical features of patients infected with 2019 novel coronavirus in wuhan, china. *The Lancet* (2020).

[4] Wu, F. *et al.* A new coronavirus associated with human respiratory disease in china. *Nature* 1–8 (2020).

[5] Zhou, P. *et al.* A pneumonia outbreak associated with a new coronavirus of probable bat origin. *Nature* 1–4 (2020).

[6] World Health Organization report. `https://www.who.int/docs/default-source/coronaviruse/situation-reports/20200216-sitrep-27-covid-19.pdf` (Accessed February 2020).

[7] Yong, C. Y., Ong, H. K., Yeap, S. K., Ho, K. L. & Tan, W. S. Recent advances in the vaccine development against middle east respiratory syndrome-coronavirus. *Frontiers in Microbiology* **10**, 1781 (2019). URL `https://www.frontiersin.org/article/10.3389/fmicb.2019.01781`.

[8] Coleman, C. M. *et al.* Purified coronavirus spike protein nanoparticles induce coronavirus neutralizing antibodies in mice. *Vaccine* **32**, 3169–3174 (2014).

[9] Tseng, C.-T. *et al.* Immunization with sars coronavirus vaccines leads to pulmonary immunopathology on challenge with the sars virus. *PloS one* **7** (2012).

[10] Rappuoli, R., Black, S. & Bloom, D. E. Vaccines and global health: In search of a sustainable model for vaccine development and delivery. *Science Translational Medicine* **11**, eaaw2888 (2019).

[11] Olsson, S.-E. *et al.* Induction of immune memory following administration of a prophylactic quadrivalent human papillomavirus (hpv) types 6/11/16/18 l1 virus-like particle (vlp) vaccine. *Vaccine* **25**, 4931–4939 (2007).

[12] Suarez, D. & Schultz-Cherry, S. Immunology of avian influenza virus: a review. *Developmental & Comparative Immunology* **24**, 269–283 (2000).

[13] Briney, B. *et al.* Tailored immunogens direct affinity maturation toward hiv neutralizing antibodies. *Cell* **166**, 1459–1470 (2016).

[14] Pedersen, S. R. *et al.* Immunogenicity of hla class i and ii double restricted influenza a-derived peptides. *PloS one* **11** (2016).

[15] Chen, B. *et al.* Predicting hla class ii antigen presentation through integrated deep learning. *Nature biotechnology* **37**, 1332–1343 (2019).

[16] Lu, R. *et al.* Genomic characterisation and epidemiology of 2019 novel coronavirus: implications for virus origins and receptor binding. *The Lancet* (2020).

[17] Jurtz, V. *et al.* Netmhcpan-4.0: improved peptide–mhc class i interaction predictions integrating eluted ligand and peptide binding affinity data. *The Journal of Immunology* **199**, 3360–3368 (2017).

[18] Wrapp, D. *et al.* Cryo-em structure of the 2019-ncov spike in the prefusion conformation. *bioRxiv* (2020). URL `https://www.biorxiv.org/content/early/2020/02/15/2020.02.11.944462. https://www.biorxiv.org/content/early/2020/02/15/2020.02.11.944462.full.pdf`.

[19] Kringelum, J. V., Lundegaard, C., Lund, O. & Nielsen, M. Reliable b cell epitope predictions: impacts of method development and improved benchmarking. *PLoS computational biology* **8** (2012).

[20] Shu, Y. & McCauley, J. Gisaid: Global initiative on sharing all influenza data–from vision to reality. *Eurosurveillance* **22** (2017).

[21] Zhou, F. *et al.* Deep sequencing of the mhc region in the chinese population contributes to studies of complex disease. *Nature genetics* **48**, 740–746 (2016).

[22] Chen, H. *et al.* Response of memory cd8+ t cells to severe acute respiratory syndrome (sars) coronavirus in recovered sars patients and healthy individuals. *The Journal of Immunology* **175**, 591–598 (2005).

[23] Zhou, M. *et al.* Screening and identification of severe acute respiratory syndrome-associated coronavirus-specific ctl epitopes. *The Journal of Immunology* **177**, 2138–2145 (2006).

[24] Tsao, Y.-P. *et al.* Hla-a 0201 t-cell epitopes in severe acute respiratory syndrome (sars) coronavirus nucleocapsid and spike proteins. *Biochemical and biophysical research communications* **344**, 63–71 (2006).

[25] Ng, O.-W. *et al.* Memory t cell responses targeting the sars coronavirus persist up to 11 years post-infection. *Vaccine* **34**, 2008–2014 (2016).

[26] Schwede, T., Kopp, J., Guex, N. & Peitsch, M. C. Swiss-model: an automated protein homology-modeling server. *Nucleic acids research* **31**, 3381–3385 (2003).

[27] Song, W., Gui, M., Wang, X. & Xiang, Y. Cryo-em structure of the sars coronavirus spike glycoprotein in complex with its host cell receptor ace2. *PLoS pathogens* **14**, e1007236 (2018).

[28] Comeau, S. R., Gatchell, D. W., Vajda, S. & Camacho, C. J. Cluspro: a fully automated algorithm for protein–protein docking. *Nucleic acids research* **32**, W96–W99 (2004).

[29] DeLano, W. L. *et al.* Pymol: An open-source molecular graphics tool. *CCP4 Newsletter on protein crystallography* **40**, 82–92.

[30] Yu, H. *et al.* Selection of sars-coronavirus-specific b cell epitopes by phage peptide library screening and evaluation of the immunological effect of epitope-based peptides on mice. *Virology* **359**, 264–274 (2007).

[31] Guo, J.-P., Petric, M., Campbell, W. & McGeer, P. L. Sars corona virus peptides recognized by antibodies in the sera of convalescent cases. *Virology* **324**, 251–256 (2004).

[32] Hua, R., Zhou, Y., Wang, Y., Hua, Y. & Tong, G. Identification of two antigenic epitopes on sars-cov spike protein. *Biochemical and biophysical research communications* **319**, 929–935 (2004).

[33] Cock, P. J. *et al.* Biopython: freely available python tools for computational molecular biology and bioinformatics. *Bioinformatics* **25**, 1422–1423 (2009).

[34] Virtanen, P. *et al.* Scipy 1.0: fundamental algorithms for scientific computing in python. *Nature Methods* 1–12 (2020).

[35] Pedregosa, F. *et al.* Scikit-learn: Machine learning in python. *Journal of machine learning research* **12**, 2825–2830 (2011).

[36] Rasmussen, M. *et al.* Uncovering the peptide-binding specificities of hla-c: a general strategy to determine the specificity of any mhc class i molecule. *The Journal of Immunology* **193**, 4790–4802 (2014).

[37] Channappanavar, R., Fett, C., Zhao, J., Meyerholz, D. K. & Perlman, S. Virus-specific memory cd8 t cells provide substantial protection from lethal severe acute respiratory syndrome coronavirus infection. *Journal of virology* **88**, 11034–11044 (2014).

[38] Vita, R. *et al.* The immune epitope database (iedb) 3.0. *Nucleic acids research* **43**, D405–D412 (2015).

[39] Hu, Y. *et al.* Immunologic hierarchy, class ii mhc promiscuity, and epitope spreading of a melanoma helper peptide vaccine. *Cancer Immunology, Immunotherapy* **63**, 779–786 (2014).

[40] Lv, Y., Ruan, Z., Wang, L., Ni, B. & Wu, Y. Identification of a novel conserved hla-a* 0201-restricted epitope from the spike protein of sars-cov. *BMC immunology* **10**, 61 (2009).

[41] He, Y. *et al.* Identification of immunodominant sites on the spike protein of severe acute respiratory syndrome (sars) coronavirus: implication for developing sars diagnostics and vaccines. *The Journal of Immunology* **173**, 4050–4057 (2004).

[42] He, Y., Zhou, Y., Siddiqui, P., Niu, J. & Jiang, S. Identification of immunodominant epitopes on the membrane protein of the severe acute respiratory syndrome-associated coronavirus. *Journal of clinical microbiology* **43**, 3718–3726 (2005).

[43] He, Y. *et al.* Mapping of antigenic sites on the nucleocapsid protein of the severe acute respiratory syndrome coronavirus. *Journal of clinical microbiology* **42**, 5309–5314 (2004).

[44] Chow, S. C. *et al.* Specific epitopes of the structural and hypothetical proteins elicit variable humoral responses in sars patients. *Journal of clinical pathology* **59**, 468–476 (2006).

[45] Tajiri, K. *et al.* Analysis of the epitope and neutralizing capacity of human monoclonal antibodies induced by hepatitis b vaccine. *Antiviral research* **87**, 40–49 (2010).

[46] Bettencourt, P. *et al.* Identification of antigens presented by mhc for vaccines against tuberculosis. *npj Vaccines* **5**, 1–14 (2020).