



OPEN

Artificial intelligence predicts the immunogenic landscape of SARS-CoV-2 leading to universal blueprints for vaccine designs

Brandon Malone^{2,3}, Boris Simovski^{1,3}, Clément Moliné^{1,3}, Jun Cheng², Marius Gheorghe¹, Hugues Fontenelle¹, Ioannis Vardaxis¹, Simen Tennøe¹, Jenny-Ann Malmberg¹, Richard Stratford¹ & Trevor Clancy¹✉

The global population is at present suffering from a pandemic of Coronavirus disease 2019 (COVID-19), caused by the novel coronavirus Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2). The goal of this study was to use artificial intelligence (AI) to predict blueprints for designing universal vaccines against SARS-CoV-2, that contain a sufficiently broad repertoire of T-cell epitopes capable of providing coverage and protection across the global population. To help achieve these aims, we profiled the entire SARS-CoV-2 proteome across the most frequent 100 HLA-A, HLA-B and HLA-DR alleles in the human population, using host-infected cell surface antigen presentation and immunogenicity predictors from the *NEC Immune Profiler* suite of tools, and generated comprehensive epitope maps. We then used these epitope maps as input for a Monte Carlo simulation designed to identify statistically significant “epitope hotspot” regions in the virus that are most likely to be immunogenic across a broad spectrum of HLA types. We then removed epitope hotspots that shared significant homology with proteins in the human proteome to reduce the chance of inducing off-target autoimmune responses. We also analyzed the antigen presentation and immunogenic landscape of all the nonsynonymous mutations across 3,400 different sequences of the virus, to identify a trend whereby SARS-CoV-2 mutations are predicted to have reduced potential to be presented by host-infected cells, and consequently detected by the host immune system. A sequence conservation analysis then removed epitope hotspots that occurred in less-conserved regions of the viral proteome. Finally, we used a database of the HLA haplotypes of approximately 22,000 individuals to develop a “digital twin” type simulation to model how effective different combinations of hotspots would work in a diverse human population; the approach identified an optimal constellation of epitope hotspots that could provide maximum coverage in the global population. By combining the antigen presentation to the infected-host cell surface and immunogenicity predictions of the *NEC Immune Profiler* with a robust Monte Carlo and digital twin simulation, we have profiled the entire SARS-CoV-2 proteome and identified a subset of epitope hotspots that could be harnessed in a vaccine formulation to provide a broad coverage across the global population.

The outbreak of Coronavirus disease 2019 (COVID-19) and its rapid worldwide transmission resulted in the World Health Organization (WHO) declaring COVID-19 as a pandemic and global health emergency¹. COVID-19 is caused by the novel coronavirus Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2)². Like all *Coronaviridae*, SARS-CoV-2 is a positive-sense RNA virus encapsulated by an envelope, and characterized by an exposed spike glycoprotein (S-protein) that is projected from the viral surface³. Although the main structural proteins on *Coronaviridae*, such as the S-protein, are reasonably well studied, many of the other proteins are less well characterized. Correcting this gap may be important to improve the design of therapeutic interventions⁴. This particular gap in knowledge is very relevant from the perspective of finding immunogenic targets across the entire virus proteome, in order to guide the design of effective vaccines. The SARS-CoV-2 virus is closely related

¹NEC Oncolmmunity AS, Ullernchaussen 64/66, 0379 Oslo, Norway. ²NEC Laboratories Europe GmbH, Kurfuersten-Anlage 36, 69115 Heidelberg, Germany. ³These authors contributed equally: Brandon Malone, Boris Simovski, Clément Moliné. ✉email: trevor@oncoimmunity.com

in sequence identity and receptor binding to SARS-CoV^{5,6}, and therefore it has been purported that one may borrow from this similarity to validate targets in potential vaccines^{7,8}. Much of the emphasis on *Coronaviridae* vaccines to date has focused on antibody responses against the S-protein, which is the most “antibody exposed” structural protein in the virus. Although demonstrated to be effective with short-lived responses in a mouse study⁹, the immune response against the S-protein of SARS-CoV is associated with low neutralizing antibody titers and short-lived memory B cell responses in recovered patients^{10,11}. Additionally, potential harmful effects of vaccines based on the antibody response to S-protein in SARS-CoV have raised possible safety concerns regarding this approach. For example, in a macaque model, it was observed that anti-S-protein antibodies caused severe acute lung injury¹², and sera from SARS-CoV patients also revealed that elevated anti-S-protein antibodies were observed in those patients that succumbed to the infection¹². When considering antibody responses to the S protein, it is also important to consider the possibility that antibody-dependent enhancement (ADE) may occur, whereby antibodies facilitate viral entry into host cells and enhance the infection and inflammation pathology of the virus^{13–15}. Considering the potential for ADE, the reported short-lived protective antibody response reported for SARS-CoV^{10,11,16}, and that of the seasonal *Coronaviridae*^{17,18} and SARS-CoV-2^{17,19,20}, coupled with the pathological consequence of S-protein specific antibodies in certain animal models; it is worth considering diverse strategies for vaccine development that also drive T-cell responses from targets other than the S-protein when designing *Coronaviridae* vaccines^{17,21,22}.

In senior citizens, T cell mediated immunity has been shown to a more reliable correlate of protective vaccination²³. In cases of weakened or a waning antibody response it has been clearly demonstrated that a vaccine strategy that induces both neutralizing antibodies and T cell mediated immunity provides the optimal protection²⁴. Additionally, it is possible, due to the waning immune response generated by “antibody alone” driven vaccines, that the first round of immunized populations may also benefit from 2nd generation vaccines that may also explicitly incorporate a broad T cell response²¹.

Although T cells cannot prevent the initial entry of a virus into host cells, they can provide protection by recognizing viral peptides presented by human leukocyte antigens (HLAs) on the surface of host-infected cells or antigen presenting cells (APCs). Several studies have demonstrated in SARS-CoV that virus-specific CD8 T cells are required for mounting an effective immune response and viral clearance^{10,25–29}. A vaccine design that confers optimal protection may also need to involve the generation of memory T cell responses³⁰. It has been shown that the activation of memory T cells specific for a conserved epitope shared by SARS-CoV and MERS-CoV is a potential strategy for developing coronavirus vaccines³¹. In addition, levels of memory T cell responses to SARS-CoV against peptides from its structural proteins were detected in a proportion of SARS-recovered patients, years after infection^{26,32–34}, and most recently detected up to as much as 17 years post infection in 100% of tested patients³⁵. Studies of the cross-reactive T cell epitopes in the common cold *Coronaviridae* speculated durable and protective T cell memory responses against SARS-CoV-2^{36,37}. For SARS-CoV-2 specific responses, a study of the T cell kinetics revealed that virus specific T cells are present early and increase over time³⁸, and these SARS-CoV-2 specific T cell responses are overwhelmingly beneficial Th1 based responses³⁹. SARS-CoV-2 specific T cells are clearly important in the elimination of the virus and controlling COVID-19 progression, which lends support to including T cell responses in the design of COVID-19 vaccines^{17,22,39}.

However, a T cell response, in isolation, may not be sufficient to combat SARS-CoV-2 infections²². The importance of an accompanied neutralizing antibody response with SARS-CoV-2 specific T cell responses has already been demonstrated in cohorts of convalescent patients⁴⁰. In a study of 128 recovered SARS-CoV patients, the immune correlates of protection were investigated and broad CD8, CD4 and neutralizing antibody response were all shown to contribute to protection⁴¹. The CD4 T cell responses mainly clustered in the S-protein, presumably as B cell antibody responses to the S-protein require the help of CD4 T cells specific to the same protein⁴². Given that in the before-mentioned study⁴², neutralizing antibody responses correlated with CD8 T cell responses against a broad set of CD8 T cell epitopes in the S-protein, a vaccine design that centers on the S-protein or any other viral protein will need to stimulate a broad CD8 response⁴³. In the previous study⁴³, robust T cell responses correlated significantly with higher neutralizing antibody activity, consistent with the hypothesis that T cells play an important role in the generation of antibody responses in recovered SARS-CoV patients⁴¹.

The required CD4 T cell help for a both protective antibody response and protective CD8 T cell activation has been previously well described⁴⁴. The importance of an integrative antibody, CD8 and CD4 T cell response in mounting a successful immune response against the present SARS-CoV-2 threat was well established in a case study during an early stage of the COVID-19 pandemic⁴⁵. This important correlation of protective antibody and T cell response was later confirmed in larger cohorts of convalescent COVID-19 patients^{40,46}. Indeed, it is now well established that activated CD4 T cells and CD8 T cells in concert with antibodies against SARS-CoV-2 are recruited in successful protective immune responses against the SARS-CoV-2 virus^{41,45,46}.

Many of the previous SARS-CoV studies have found promising CD8 targets^{10,25,28,41}, including sustainable memory T cell responses^{10,25–27,30–33} that recognize epitopes in proteins across the entire spectrum of the virus, although the S-protein has been reported to be enriched for dominant CD8 T cell responses⁴¹. Vaccines that incorporate full length proteins, or attenuated viruses may allow for the patient’s own T cell immunity to sample a broad spectrum of epitopes in a natural manner. However, with whole protein or attenuated viruses, one loses the specificity offered by a targeted T cell epitope-based approach. By predicting virus specific epitopes in a vaccine, highly specific T cell responses can be induced. The computational prediction of specific T cell epitopes minimizes the risk of antigenic competition, the unwanted inclusion of inhibitory epitopes, and is generally considered safer⁴⁷.

Taken together, this supports the approach taken in this study, which is to computationally map a broad epitope landscape across the global viral SARS-CoV-2 proteome, which includes integrated CD8 and CD4 T cell targets in the modeling. There has been some preliminary efforts recently that describe SARS-CoV-2 epitope maps^{48–51}, however it appears that the emphasis in those approaches were based mostly on HLA binding. It is

important to profile, as in this study, not only the candidates that may bind to HLA but also those CD8 epitopes that are naturally processed and presented by the cell's antigen processing (AP) machinery and presented on the surface of the infected host cells. Layered on top of the antigen presentation predictions in the host infected cells, we also make predictions across the entire viral proteome that measure the likelihood that the peptides presented on the host infected cells are capable of being recognized by T cells that are not yet tolerized or deleted from a patient's T cell repertoire. Cross reactivity to the seasonal *Coronaviridae* should also be taken into consideration in this regard in future studies, when considering the pre-existing patient's T cell repertoire^{36,52}. The subsequent immunogenic landscape of the SARS-CoV-2 that we present here is taken further to analyze the immunogenicity of all the non-synonymous variations across approximately 3400 different SARS-CoV-2 sequences, to map the trajectory of differential immunogenic potential between all the currently sequenced viral strains.

Any viable vaccine to tackle SARS-CoV-2 that incorporates T cell epitopes in its design would need to contain a constellation of overlapping epitopes that protect the vast majority of the human HLA population against the virus. In this study, we demonstrate that the SARS-CoV-2 immunogenic landscape clusters into distinct groups across the spectrum of HLA alleles in the human population. Our predicted immunogenic landscape of the SARS-CoV-2 virus is then processed through a robust comprehensive statistical Monte Carlo simulation, incorporating the integrative immune parameters, to identify epitope hotspots for a broad adaptive immune response across the most common HLA genetic makeup in the human population. The central question that the Monte Carlo simulation attempted to answer is whether specific regions in the viral proteins are enriched with higher immunogenic scores with respect to a large set of HLA alleles in the human population, more than expected by chance. In addition, epitope hotspots containing viral epitopes that have high similarity with human peptides, especially those expressed in critical organs were removed. The resulting epitope hotspots we identified represent areas in the viral proteome that are likely to be viable vaccine targets and represent blueprints for vaccine design.

In order to rank-prioritize these potential universal epitope hotspots, and the peptides that underlie them at high resolution, the baseline peptide predictions are then taken through a graph-based "digital twin" type simulation³³, to prioritize hotspots and the specific overlapping peptides that they comprise at a patient-specific and population-specific level. In this context, the digital twin information is the precise HLA haplotype of an individual, and many virtual individuals are considered within a given population being analyzed. The HLA haplotype is a key determinant of the immune response that specific individuals can mount against SARS-CoV-2 infections^{51,54,55}. This HLA genetic background is an important factor for determining whether a vaccine is effective in establishing immunity for the specific individual and a broader population (consisting of multiple diverse individuals). The candidate sequence targets that emerge from this computational analysis represent blueprints for potential vaccine designs modeled across the global human population.

Results

The immunogenic landscape of SARS-CoV-2 reveals diversity among the different HLA groups in the human population.

We carried out an epitope mapping of the entire SARS-CoV-2 virus proteome using cell-surface antigen presentation and immunogenicity predictors from the *NEC Immune Profiler* suite of tools. Antigen presentation (AP) was predicted from an ensemble machine-learning model that integrates information from several HLA binding predictors (in this case three distinct HLA binding predictors trained on IC50nm binding affinity data) and 13 different predictors of antigen processing (all trained on mass spectrometry data, see section "Generation of global epitope maps and amino acid scores" in Materials and Methods). The outputted AP score ranges from 0 to 1, and that was used as input to compute immune presentation (IP) across the epitope map. The IP score penalizes those presented peptides that have degrees of "similarity to human" when compared against the human proteome, and awards peptides that are less similar. The resulting IP score represents those HLA-presented peptides that are likely to be recognized by circulating T cells in the periphery, that is, T cells that have not been deleted or tolerized, and therefore most likely to be immunogenic. Both the AP and the IP epitope prediction models are "pan" HLA, or HLA-agnostic, and can be carried out for any allele in the human population; however for the purpose of this study we limited the analysis to the 100 most frequent HLA-A, HLA-B and HLA-DR alleles in the human population (as documented in the Allele Frequency Net Database⁵⁶). Class II HLA binding predictions were also incorporated into the large scale epitope screen from the IEDB consensus of tools⁵⁷. The resulting epitope maps allowed for the identification of regions in the viral proteome that are most likely to be presented by host-infected cells using the most frequent HLA-A, HLA-B and HLA-DR alleles in the global human population (defined conceptually for the purpose of this study as the 100 most frequent HLA-A, HLA-B and HLA-DR alleles in the human population documented in the Allele Frequency Net Database⁵⁶). Epitope maps were created for all of the viral proteins, and an example based on the IP scores for the S-protein is depicted in Fig. 1A and for AP in Fig. 1B; it illustrates distinct regions of the S-protein that contain candidate CD8 and CD4 epitopes for the 100 most frequent human HLA-A, HLA-B and HLA-DR alleles. It is clear from Fig. 1A,B that different HLA alleles have different Class I AP (and IP) and Class II binding properties. This strongly suggests, as one might anticipate, that the SARS-CoV-2 antigen presentation (and immune presentation) landscape clusters into distinct population groups across the spectrum of different human HLA alleles. This trend is further illustrated in the hierarchical-clustering map presented in Fig. 2 after the AP scores have been binarized. Figure 2 clearly demonstrates that some allelic clusters present many viral targets to the human immune system, while others only present a few targets, and some are unable to present any. This implies that different groups in the human population with different HLAs will respond differentially to a T cell driven vaccine composed of viral peptides. Therefore, in order to design the optimal vaccine that leverages the benefits of T cell immunity across a broad human population, we need to predict "epitope hotspots" in viral proteome. These hotspots are regions of the virus that are enriched for overlapping epitopes, and/or epitopes in close spatial proximity, that can be recognized by multiple HLA types across the human population.

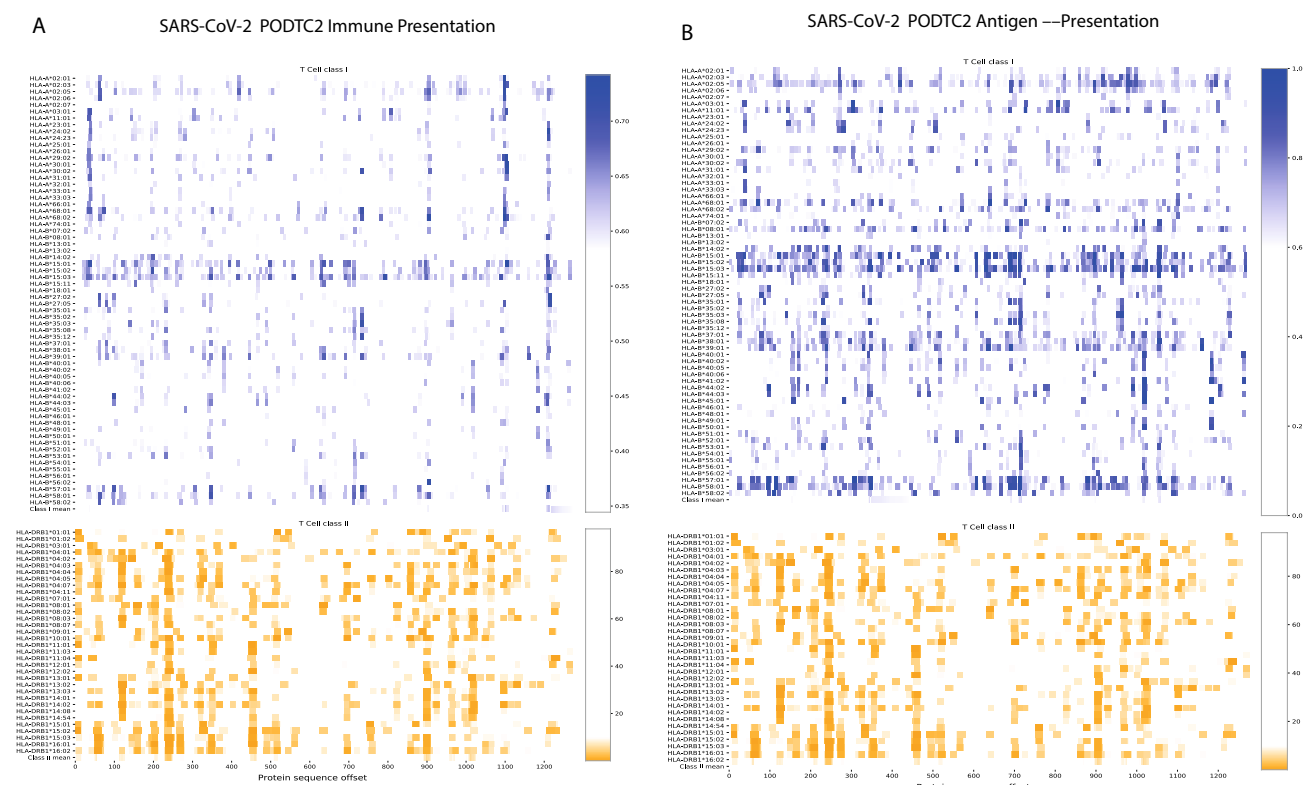


Figure 1. Epitope map of the S-Protein (PODTC2) across the most frequent HLA-A, HLA-B and HLA-DRB alleles in the human population. Data is transformed such that a positive results for CD8 relates to 0.7 or above, and 0.1 or below for Class II. Broad coverage for CD8 and CD4 is demonstrated. PODTC2 is the Uniprot accession ID for the S protein (spike glycoprotein of SARS-CoV-2).

Prior to applying the *NEC Immune Profiler* suite of tools to map the SARS-CoV-2 viral proteome, it was important to first validate, to the extent that is possible from the limited number of validated SARS-CoV viral epitopes, that the T cell based AP and IP scores are predicting viable targets. We identified Class I epitopes from the original SARS-CoV virus (that first emerged in the Guangdong province in China in 2002) that shared $\geq 90\%$ sequence identity with the current SARS-CoV-2. Unfortunately, many of the published epitopes were identified using ELISPOT on PBMCs from convalescent patients and/or healthy donors (or humanized mouse models) where the restricting HLA was not explicitly deconvoluted. In order to circumvent this problem, we identified a subset of 8 epitopes where the minimal epitopes and HLA restriction had been identified using tetramers^{7,36}. In this survey, 7 out of the 8 epitopes tested were identified as positive, i.e., had an IP score of above 0.5 (see Table 1), demonstrating an accuracy of 87%. The IP score ranges between 0 and 1, where 1 indicates the largest probability of an immune response among the candidates.

Although this was a very small test dataset, this provides a degree of confidence that the *NEC Immune Profiler* prediction pipeline can accurately identify good immunogenic candidates and that the epitope hotspots identified by this analysis and subsequent analyses represent interesting targets for vaccine development.

A robust statistical analysis identifies epitope hotspots for a broad T cell response. In order to identify epitope hotspots that have the potential to be viable immunogenic targets for the vast majority of the human population, we first carried out a Monte Carlo random sampling procedure, on the epitope maps generated previously (for the Wuhan reference sequence exemplified in Fig. 1 for the S-protein) to identify specific areas of the SARS-CoV-2 proteome that have the highest probability of being epitope hotspots (see Material and Methods). Three bin sizes were investigated for potential epitope hotspots: 27, 50 and 100. A statistic was calculated for each defined subset region of the protein (bin) from the set of 100 HLAs; the statistic accounts for individual epitope scores and epitope lengths. The Monte Carlo simulation method was then used to estimate the p-values for each bin, whereby each bin represented a candidate epitope hotspot. The epitope hotspots are the statistically significant bins, that is, those below a 5% false discovery rate (FDR) according to the Monte Carlo simulation; these hotspots represent regions that are most likely to contain viable T cell driven vaccine targets that can be recognized by multiple HLA types across the human population. A summary of the epitope hotspots identified across the entire viral proteome for AP is depicted in Fig. 3. It reveals that the most immunogenic regions of the virus, that target the most frequent HLA alleles in the global population, are found in several of the viral proteins above and beyond the antibody exposed structural proteins, such as the S-protein.

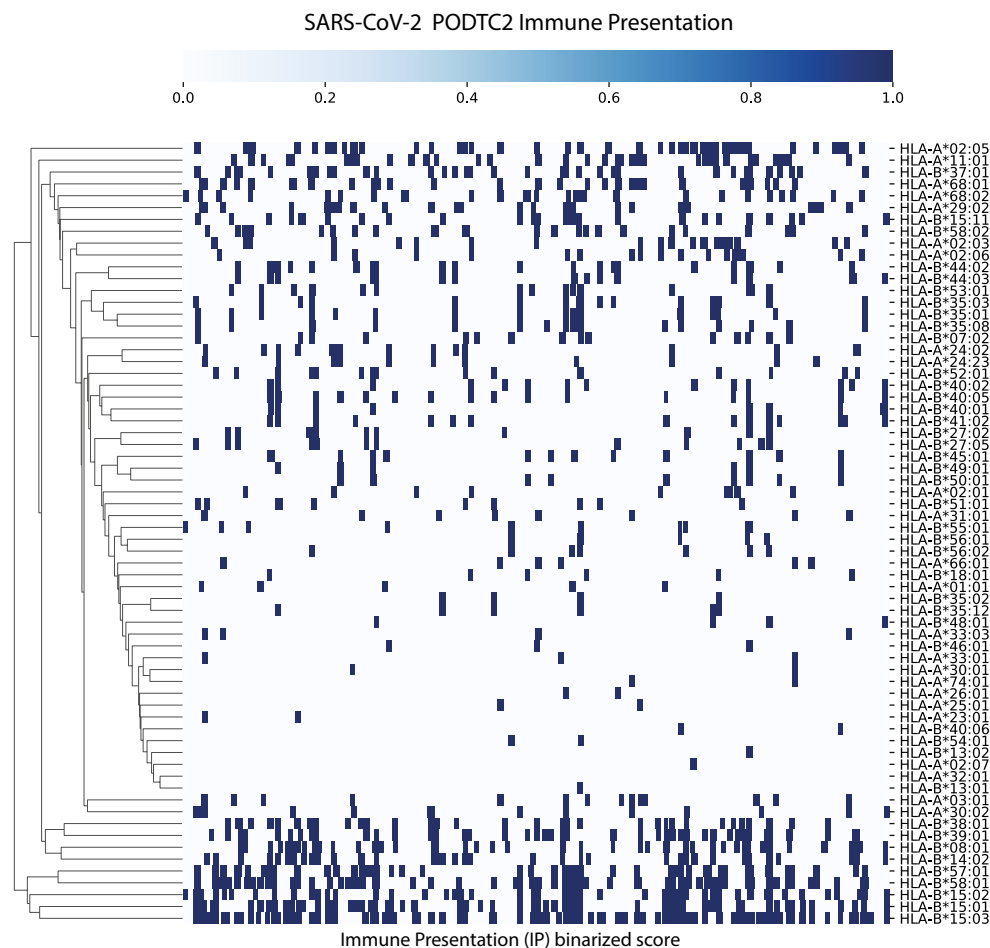


Figure 2. Hierarchical clustering of binary transformation of the epitope maps. Illustrated here for Class I CD8 epitopes in HLA-A and HLA-B alleles. Predictions for Class I greater than 0.7 are transformed to 1. Illustrated for the S-Protein here for demonstration purposes.

Conservation analysis identifies robust epitope hotspots in SARS-CoV-2. Although an reports in demonstrated in a few sequences that the SARS-CoV-2 genome has a lower mutation rate and genetic diver-

Peptide	Sequence similarity (%)	Parental protein	IP score	Correct prediction
FIAGLIAIV	100	Spike	0.54	Yes
MEVTPSGTWL	100	Nucleoprotein	0.61	Yes
RLNEVAKNL	100	Spike	0.39	No
TLACFVLA AV	100	Membrane	0.54	Yes
KLPDDFTGCV	90	Spike	0.58	Yes
GMSRIGMEV	100	Nucleoprotein	0.59	Yes
LLLDRLNQL	100	Nucleoprotein	0.57	Yes
VVFLHVTYV	100	Spike precursor	0.53	Yes

Table 1. Immune Presentation (IP) scores for validated SARS-CoV peptides with high similarity to SARS-CoV-2.

sity compared to that of SARS-CoV⁵⁸, another study has demonstrated that there are evolving genetic patterns emerging in different strains of SARS-CoV-2 in diverse geographic locations⁵⁹. A universal vaccine blueprint should ideally protect all populations against different emerging clades of the SARS-CoV-2 virus. Motivated by the necessity to design universal vaccine blueprints that are robust to the mutating nature of SARS-CoV-2, we compared the AP potential of approximately 3,400 virus sequences in the GISAID database against the AP potential of the Wuhan Genbank reference sequence (see Materials and Methods). The outcome of that com-

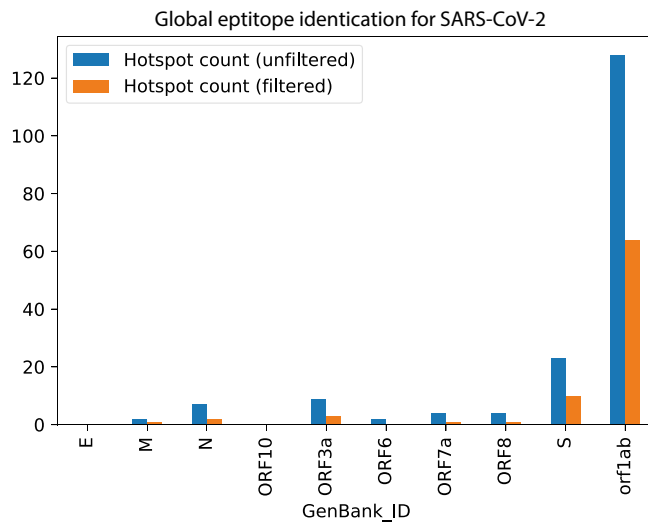


Figure 3. Epitope hotspots from the Monte Carlo analysis are captured across the majority of the entire viral proteome using filtering procedures for conserved and human self-peptides. The most abundant signal for hotspots is in the orf1ab polyprotein.

Changes in the Antigen Presentation of peptide variants from 3400 SARS-CoV-2 sequences

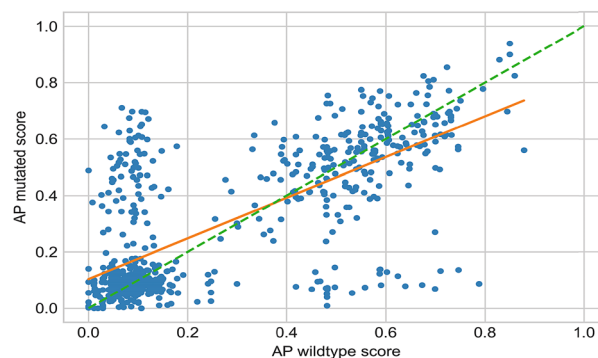


Figure 4. The scatter plot shows the mutated AP score (Y-axis) against its wildtype AP score (X-axis) for each peptide variant. Each data point (blue circles) represents a peptide variant identified in a SARS-COV-2 sequence when compared to the Genbank (Wuhan) reference sequence. There was a total of 1122 peptide variants identified in such a comparison and illustrated on the plot. The orange line is a least square fit with slope = 0.72. Peptide variants on the green dashed line of slope 1 would show no change in AP after introduction of the peptide variant.

parison is illustrated in Fig. 4, and although only illustrated for one HLA allele, hints at a trend whereby SARS-COV-2 peptide variants seem to reduce their potential to be presented and, consequently, detected by the host immune system. Similar trends have been observed in chronic infections such as HPV⁶⁰ and HIV⁶¹.

On the other hand, a few mutations significantly increase the presentation potential. However, a robust vaccine should ideally target a robust set of hotspots which consistently appear in populations from all geographical regions.

In order to assess if these epitope hotspots are sufficiently robust across sequenced and mutating strains of SARS-CoV-2, we next used the AP based epitope hotspot Monte Carlo statistical framework, and analyzed 10 sequences of the virus from among the 10 most mutated viral sequences from different geographical regions⁶². The vast majority of the hotspots were present in all of the sequenced viruses, however occasionally hotspots were eliminated and/or new hotspots emerged in these divergent strains as shown in Fig. 5.

Although the identified hotspots seem to be maintained across different viral strains, in order to design the most robust vaccine blueprint that will provide the broadest protection possible against new emerging clades of the SARS-CoV-2 virus, the epitope hotspots were subject to a sequence conservation analysis. The goal of this analysis was to identify hotspots that appear to be less prone to mutation across thousands of viral sequences, for inclusion into the optimal universal HLA vaccine blueprints, applied to the most conserved regions of the virus. We calculated a conservation score for each hotspot based on the consensus sequence of a protein (see Materials and Methods). Figure 6 shows conservation scores for the hotspots identified based on IP using different bin sizes. Only the epitope hotspots with a conservation score higher than the median conservation score were kept

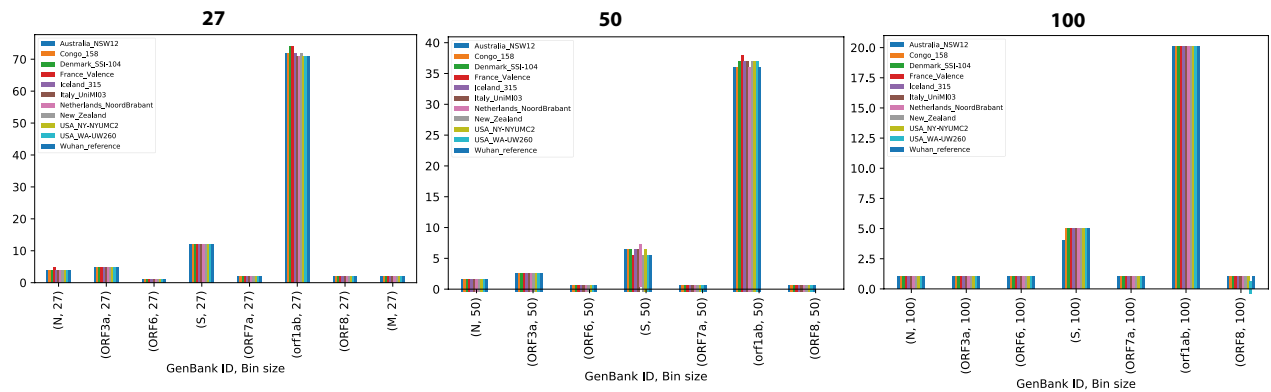


Figure 5. Application of the Antigen presentation (AP) based Monte Carlo epitope hotspot prediction method to mutated viral sequences from 10 different geographical locations, as well as the Wuhan reference sequence. Each group of bars shows the count of epitope hotspots found in the respective viral protein in each location. The epitope hotspot counts are shown for three different bin lengths: 27 (left), 50 (centre) and 100 (right). It is clear that the epitope hotspots are robust across mutation sequences. Only those hotspots that were identified for AP in Fig. 3 are illustrated in the plot.

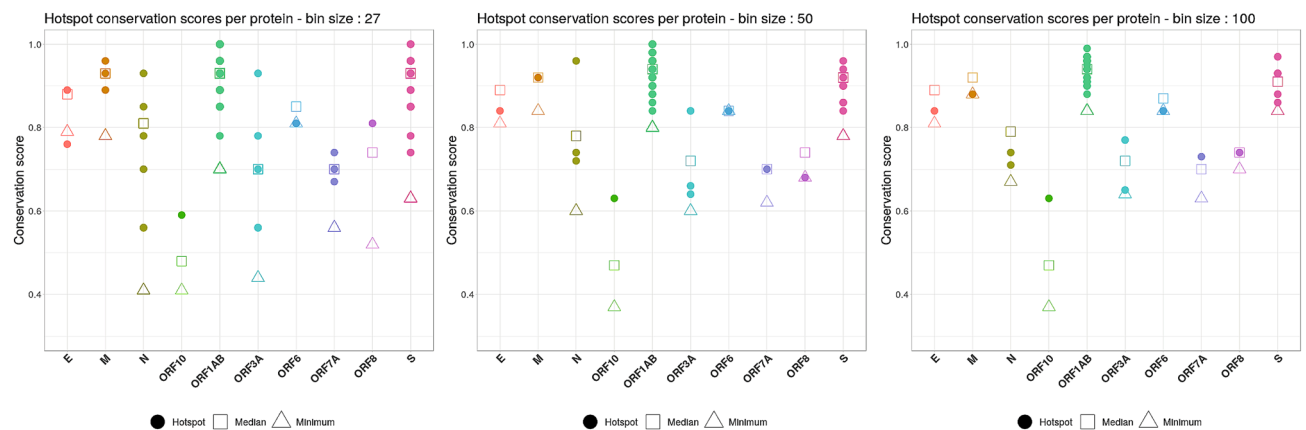


Figure 6. The scatter plots show the distribution of the hotspot conservation scores (Y-axis) for proteins in the viral genome (X-axis). Each plot contains the conservation scores of the hotspots identified based on immune presentation (IP), using different bin sizes: 27 (left), 50 (center), and 100 (right). A hotspot is represented by a filled coloured circle, while the median conservation score for a given protein is depicted by a hollow square. As a reference, upwards facing triangles show the minimum conservation score for that protein. Only hotspots with a conservation score above the median were taken for further consideration.

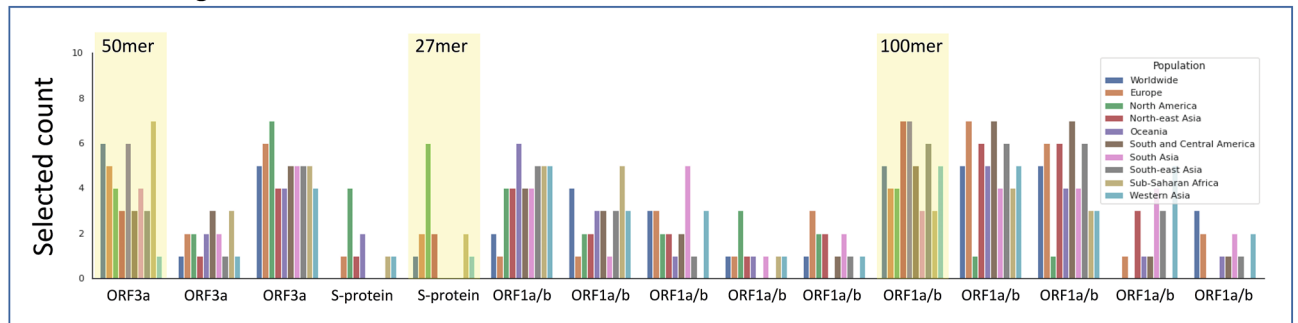
for further analysis. This allowed us to filter out a significant amount of less conserved epitope hotspots, that although have high immunogenic scores, harbor a higher degree of potential sequence variation.

In addition, to reduce the potential for off-target autoimmune responses against host tissue, we removed hotspots that contained exact sequence matches for all epitope lengths analyzed to proteins in the human proteome.

A graph based "digital twin" optimization prioritizes epitopes hotspots to select universal blueprints for vaccine design.

The Monte Carlo simulation identified well over 100 different hotspots of length 27, 50 or 100 amino acids, for both AP and IP. Even after filtering for conservation and self-similarity, we were left with over 50 different hotspots for both the AP and IP based analyses. In order to develop a blueprint for viable universal vaccine against SARS-CoV-2, it is necessary to (1) cover with fidelity a broad proportion of the human population, and (2) prioritize the selection to even fewer regions (the exact number may depend on the size of the bin and the vaccine platform under consideration). Consequently, we need to identify the optimal constellation of hotspots, or relevant viral segments, that can provide broad coverage in the human population with a limited and targeted vaccine "payload". In order to achieve this aim, we developed and applied (see Materials and Methods) a "digital twin" method, which models the specific HLA haplotypes of different geographical populations. A graph-based mathematical optimization approach is then used to select the optimal combination of immunogenic epitope hotspots that will induce immunity in the broad human population. The results of this analysis are shown in Fig. 7. The output clearly identified a subset of hotspots that may be combined to stimulate a robust immune response in a broad global population. An example hotspot for the ORF3a100-150 region is

IP filtered: All lengths



AP filtered: All lengths

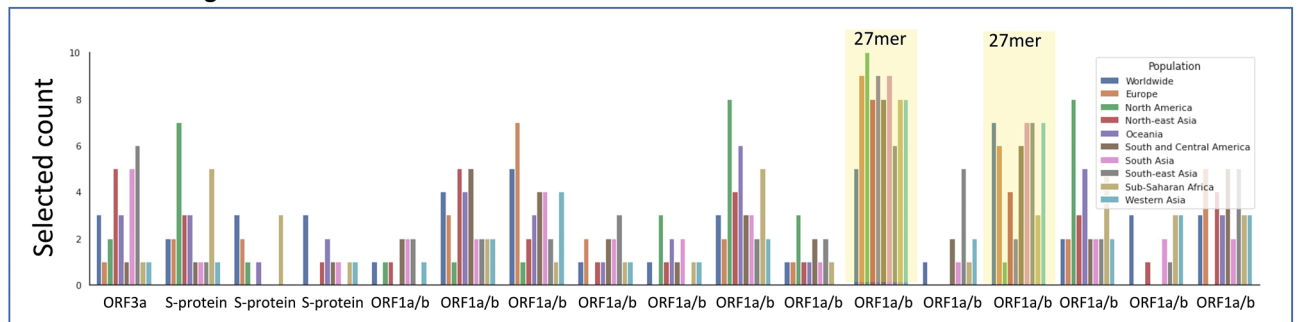


Figure 7. The plot illustrates a set of digital twin simulation experiments to identify effective epitope hotspots based on immune presentation (IP) and antigen presentation (AP), which broadly cover the population. The aim of the analysis was to select an optimum set of hotspots (respecting a given budget) such that the likelihood that each citizen (in a given modelled population) has a positive response is maximized (or, equivalently, that the log likelihood of no response for each citizen is minimized). Ten simulations were run for each region illustrated, where each simulation consisted of 10,000 digital twins (see supplementary file for more details). This plot shows the number of times each hotspot was selected for use in one of the simulations. Each hotspot selected at least 10 times is shown on the x-axis, and the y-axis shows the selected counts per region. Each bar corresponds to a different region-specific simulation setting. A subset of 5 hotspots were selected based on their profile across the AP and IP digital twin analyses, which could theoretically provide coverage of >90% across a global population (highlighted in yellow).

provided in supplementary Table 1, which shows the amino acid sequence and its component Class I and Class II epitopes.

Conclusions

In order to effectively combat the SARS-CoV-2 pandemic, a vaccine will need to protect the vast majority of the human population and stimulate diverse T cell responses against multiple viral targets including, but not limited to, the S-protein. To help achieve this ambitious aim, we have profiled the entire SARS-CoV-2 proteome across the most frequent 100 HLA-A, HLA-B and HLA-DR alleles in the human population and generated comprehensive epitope maps. We subsequently used these epitope maps as the basis for modeling the specific HLA haplotype of individual persons in a diverse set of different human populations using the most significant CD8 and CD4 T cell “epitope hotspots” in the virus. To the best of our knowledge, this is the first computational approach that generates comprehensive vaccine design blueprints from large-scale epitope maps of SARS-CoV-2 in a manner that optimizes for diverse T cell immune responses across the global population. Underlying this approach are two novel methods that, when integrated together, result in a solution that is uniquely suited to achieving the objective of the study, i.e., designing blueprints for universal vaccines. Firstly, a framework that leverages Monte Carlo simulations was developed to identify statistically significant epitope hotspot regions in the virus that are most likely to be immunogenic across a broad spectrum of HLA types. Secondly, a novel person-specific or “digital twin” type simulation, based on the actual HLA haplotypes of approximately 22,000 individuals, prioritizes these epitope hotspots, to identify the optimal constellation of vaccine hotspots in the SARS-CoV-2 proteome that are most likely to promote a robust T cell immune response in the global population.

Importantly, the CD8 epitope maps that underlie these optimized epitope hotspots are based on our AP predictions of peptides presented on the surface of host-infected cells, and visible to the host’s CD8 T cells. Additionally, these antigen presented peptides are subject to our IP predictions that infer those specific epitopes that are most likely to activate a T cell in a host’s repertoire that has not been deleted or tolerized. These features confer unique properties to the epitope maps that underlie our epitope hotspot predictions and digital twin optimization. These properties differ from the SARS-CoV-2 epitope maps that have been reported in recent preprints since the outbreak of this virus, which mainly utilize predictions based on HLA binding^{48–51}.

A genomic analysis of approximately 3400 SARS-CoV-2 sequences revealed that the epitope hotspots that we predict are robust across different evolving clades of the virus, which may be important in the design of robust vaccine blueprints that are applied to conserved sequences in the virus, and universal to the HLA haplotypes of the global human population. However, on average, mutations in the virus that cause amino acid changes in peptides seem to reduce their potential to be presented on the cell surface and consequently detected by the host immune system. We therefore apply sequence-based filters on the vaccine blueprints that discard less conserved hotspots, and hotspots that harbor peptides that have an exact match in the human proteome, before performing our digital twin simulation.

It is important to note that the predictions derived in this study are *in silico* based antigen prediction and should be subject to further experimental assay scrutiny⁶³, before final definitive selection into a COVID-19 vaccine design. Further research that characterizes potential protective Th1 versus harmful Th2 CD4 T cell responses is critically needed to determine the optimal vaccine design for T cell based vaccines⁶⁴.

The findings described in this study highlight the potential of looking beyond the S-protein and mining the whole viral proteome in order to identify optimal constellations of epitopes that can be used to develop efficacious and universal T-cell vaccines. The novel integrated methodological approaches described in this study may result in the design of diverse T cell driven vaccines that may help combat the SARS-CoV-2 pandemic and bring much needed relief to the suffering global human population.

Materials and methods

Generation of global epitope maps and amino acid scores. For a given HLA allele, the score allocated to an amino acid corresponds to the best score obtained by an epitope prediction overlapping with this amino acid. For Class I HLA alleles, the epitope lengths are 8, 9, 10 and 11, and predicted for antigen presentation (AP) or immune presentation (IP) of the viral peptide to host-infected cell surface, generated using the NEC Immune Profiler software. These Class I scores range between 0 and 1, where by 1 is the best score, i.e., higher likelihood of being naturally presented on the cell surface (AP) or being recognized by a T cell (IP). For Class II HLA alleles, we only consider 15mers. The Class II predictions were percentile rank binding affinity scores (not antigen presentation), so the lower scores are best (the scores range from 0 to 100, with 0 being the best score).

The NEC Immune Profiler is software based on machine learning algorithms, which predict which antigens have the required features of HLA-binding, processing, presentation to the cell surface, and the potential to be recognized by T cells to be good clinical targets for immunotherapy. The main machine-learning components of the NEC Immune Profiler will be further detailed in the following sections.

1. HLA Binding: The ability of an antigen to bind HLA represents the most important step in determining immunogenicity, as only HLA-bound peptides can be detected by circulating T-cells. The NEC Immune Profiler HLA binding module predicts binding affinity of the peptide to the inputted HLA allele(s). The binding affinity predictions are measured by IC₅₀ (nM) scores. The lower the IC₅₀ score, the stronger the peptide binds to the HLA molecule. The module is composed of three different binding affinity predictors.
2. Processing: In order to have an opportunity to bind HLA and be subsequently presented at the surface, an antigen must be generated by proteasomal cleavage of its parental polypeptide/protein in the cytosol and be subsequently transported into the endoplasmic reticulum by the TAP transporters. The processing module consists of a series of Support Vector Machines (SVM), trained on large databases of mass spectrometry immunopeptidome data, that are incorporated into the NEC Immune Profiler Software and operate in an ensemble machine learning layer of 13 processing models to predict which antigens have the right physico-chemical features to be efficiently processed by the processing apparatus. The different algorithms work in concert to produce a consensus score that ranges between 0 and 1. A consensus score of 1 means that the antigen is predicted to be efficiently processed while conversely a score of 0 means that that the antigen is predicted to be poorly processed.
3. Antigen Presentation (AP) and Immune Presentation (IP): In order to stimulate a T-cell a candidate antigen must be presented at the surface of the tumor complexed with HLA. The most important variables that determines whether an antigen will be efficiently presented are: (1) the binding affinity between the candidate antigen and a specific HLA molecule, (2) its potential to be efficiently processed by the antigen processing machinery, (3) the level of expression of the protein containing the mutation and (4) the ability of the source protein to contribute component peptides to the antigen processing pathway. The immune presentation (IP) method generates a distance measure that determines the relative uniqueness of the candidate antigens and can be used in combination with the antigen presentation score to generate an immune presentation (IP) score. Both the AP and IP scores range between 0 and 1. For AP, a score of 1 indicates the largest chance of presentation on the host infected cell surface, and scores > 0.7 are generally considered acceptable. For IP, a score of 1 indicates the largest chance of an immune response among the peptide candidates, and scores > 0.5 are generally considered acceptable. These thresholds were identified as acceptable scores in benchmarking of the NEC Immune Profiler on numerous clinical datasets, whereby the optimal trade-off between specificity and sensitivity in terms of identifying immunogenic epitopes was identified (manuscript in preparation).

Statistical framework for the detection of epitope hotspot epitope regions in different HLA populations. *Input data.* The data sets inputted into the statistical framework are epitope maps generated for each amino-acid position in all the proteins in the SARS-CoV-2 proteome, for all of the studied 100 HLA alleles (SARS-CoV-2 sequences from the GenBank Wuhan reference, MN908947.3, were downloaded April 15th 2020). A score for any given amino acid was determined as the maximum AP or IP score that a peptide overlap-

ping that amino acid holds in the epitope map. All peptide lengths of size 8–11 amino acids for Class I, and 15 for Class II were processed, generating one HLA dataset per viral protein. Each row in the dataset represents the amino acid epitope scores predicted for one HLA type.

Statistical framework. The central question that the statistical framework attempts to answer is: “are specific regions in a given viral protein enriched with higher immunogenic scores, with respect to a given set of HLA types, more than expected by chance?” To answer the question, we implemented a hypothesis-testing framework inspired by work done in statistical genomics^{65,66}.

HLA tracks. The raw input datasets are first transformed into binary tracks. For each Class I HLA dataset, the epitope scores are transformed to binary (0 and 1) values, such that amino-acid positions with predicted epitope scores larger than 0.7 for AP, or larger than 0.5 for IP, are assigned the value 1 (positively predicted epitope), and the rest are assigned the value 0. Similarly, for Class II HLA datasets, amino-acid positions with predicted epitope scores smaller than 10 are assigned the value 1, otherwise 0. Each binary track can effectively be presented as a list of segments, intervals of consecutive ones, and gaps, intervals of consecutive zeros. For a description of the AP and IP scores please refer to; “[Generation of global epitope maps and amino acid scores](#)”, in Material and Methods.

Test statistic. For a group of K HLA binary tracks, a test statistic S_i is calculated for each bin b_i of given size m , dividing the protein in n bins (e.g. $m=100$ amino-acids for the larger proteins). For a single HLA track k , a test statistic $s_{i,k}$ is calculated for each bin b_i as follows.

$$s_{i,k} = \sum_{j=1}^m b_{i,j} * weight_k,$$

where $weight_k$ is by default 1.0.

Then, for each bin $i=1\dots n$,

$$S_i = \frac{\sum_{k=1}^K s_{i,k}}{K}$$

which is the average number of amino acids predicted to be included in an epitope (epitope enrichment) of the bin b_i normalized by the number of selected HLAs.

Null model. As with genomic tracks⁶⁶, analytical approaches to estimate the statistical significance of the observed enrichment with predicted epitopes, of a region, across multiple HLA tracks, are intractable. An effective alternative to this problem is Monte Carlo-based simulation. A null model is defined as the generative model of the HLA tracks, if they were generated by chance. From the null model, through sampling, arises the null distribution of the test statistic $s_{i,k}$. Amino acids that are positively predicted as epitopes will clump together in segments with minimal length of eight, which is the shortest peptide length for which epitope scores are predicted, and often form longer segments when the source peptides overlap each other. Similarly, non-epitope amino acids will form gaps, with a minimal possible length of one amino acid. To preserve these features of the observed HLA tracks in the null model, as a sampling strategy we selected to shuffle the order of the segments and the gaps, respectively, within an HLA track.

P-value estimation. To sample from the null model, all of the K HLA tracks are divided in segments and gaps, which are then shuffled to produce a randomized HLA track. This is repeated 10,000 times, to produce 10,000 samples of $s_{i,k}$ statistic for each bin. For each bin, the p-value is estimated as the proportion of the samples that are equal or larger than the truly observed enrichment. Further, the generated p-values are adjusted for multiple testing with the Benjamini–Yekutieli procedure to control for a false discovery rate (FDR) of 0.05.

Graph-based optimization in digital twin simulations of the epitope hotspots. We consider a population as a set C of “digital twin” citizens c , and a vaccine as a set V of vaccine elements v . We denote the likelihood that all citizens have a positive response to a vaccine as $P(R = +|C, V)$. Our goal is to design a vaccine, that is, select a set of vaccine elements, to maximize this probability:

$$\max_V P(R = +|V, C)$$

In this setting, maximizing the probability of positive response is the same as minimizing the probability of no response. Thus, we approach vaccine design by minimizing the probability of no response for the citizen who has the highest probability of no response $P(R = -|V, c_j)$:

$$\max_V P(R = +|V, C) := \min_V \max_{c_j \in C} \{P(R = -|V, c_j)\}$$

We consider that a vaccine causes a positive response if at least one of its elements causes a positive response. That is, the probability of no response is the joint likelihood that all elements fail. For a particular citizen c_j , this probability is given as follows.

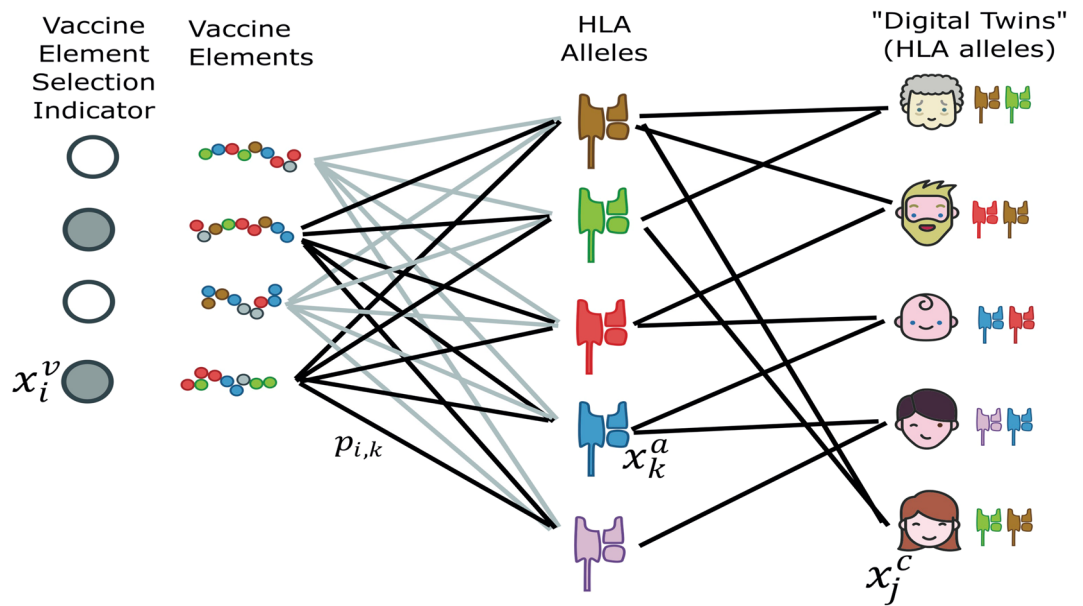


Figure 8. Schematic of the problem setting. The vaccines elements were the significant epitope hotspots that emerged from the statistical hotspot detection framework.

$$P(R = -|V, c_j) = \prod_{v_i \in V} P(R = -|v_i, c_j, V)$$

The original optimization problem can then be expressed as:

$$\minmax_V \prod_{c_j \in C} \prod_{v_i \in V} P(R = -|v_i, c_j, V)$$

Since the logarithm function is monotonic, the value of V which minimizes the logarithm of the function also minimizes the original function.

$$\minmax_V \sum_{c_j \in C} \sum_{v_i \in V} \log P(R = -|v_i, c_j, V)$$

Further, we consider each citizen as a HLA haplotype, and we assume that each vaccine element v_i may result in a response for each HLA allele independently; we refer to the alleles for citizen c_j as $A(c_j)$. Thus, our final objective is as follows.

$$\minmax_V \sum_{c \in C} \sum_{v_i \in V} \sum_{a_k \in A(c)} \log P(R = -|v_i, k, V)$$

We approach this minimax problem as a type of network flow problem, with one set of nodes corresponding to vaccine elements, one set corresponding to HLA alleles, and one set corresponding to citizens. The goal is to select the set of vaccine elements such that the likelihood of no response is minimized for each citizen. Figure 8 gives an overview of the problem setting. Supplementary table 2 provides an overview of the haplotypes used in the simulations, including the number of unique haplotypes in each geographical region, and supplementary table 3 provides the full details of the complete haplotypes used for each individual in the digital twin simulations.

Vaccine design process. Concretely, we approach the vaccine design process in four steps:

1. Select a set of candidate vaccine elements for inclusion in the vaccine. The epitope hotspots are the candidate vaccine elements.
2. Create a set of “digital twin” citizens for a population of interest, where a digital twin is an HLA haplotype.
3. Create a tripartite graph in which the nodes correspond to vaccine elements, HLA alleles, and citizens; edges correspond to relevant biological terms described in the supplementary methods.
4. Select a set of vaccine elements (respecting a given budget) such that the likelihood that each citizen has a positive response is maximized (or, equivalently, that the log likelihood of no response for each citizen is minimized).

Each step is described in more detail in the supplementary methods.

Variant immunogenic potential across the mutating sequences of SARS-CoV-2. We downloaded all the strains available in the GISAID database⁶² as of 31.03.2020, and ran them through the Nexstrain/Augur software suite with default parameters⁶⁷. We parsed the resulting phylogenetic tree to obtain all protein variants. For each peptide variant, we computed a wildtype score and a mutated Antigen Presentation (AP) score for HLA-A*02:01. One HLA allele was chosen to illustrate the robustness across mutating sequences of the virus, illustrated in Fig. 4. The mutated score is the maximum AP score among point mutations in the nine possible 9-mer peptides that include the variant. The wildtype score is the maximum AP score for the 9-mers at the same positions in the reference (Wuhan) strain.

Epitope hotspot conservation scores. For each protein within the viral genome, the set of unique amino acid sequences was compiled from all the strains available in the GISAID database⁶² as of 29.03.2020. These sets were individually processed using the Clustal Omega (v1.2.4)⁶⁸ software via the command line interface with default parameter settings. The software outputs a consensus sequence that contains conservation information for each amino acid within the protein sequence. An amino acid depicted as an “*” at position *i* within the consensus sequence translates to that amino acid being conserved (exact match in sequence similarity) at position *i* among all the input sequences⁶⁸.

The hotspot offsets were then used to extract their respective consensus sub-sequence. For each hotspot, the conservation score was calculated as the ratio of “*”s within its consensus sub-sequence to the total length of the sub-sequence. Accordingly, each hotspot was assigned a conservation score between 0 and 1, with 1 representing a perfect conservation across all available strains.

The median conservation score was calculated by sampling 1,000 sub-sequences equal to the hotspot size from the entire consensus sequence of a protein. Each sample was assigned a conservation score, and the median value from all 1,000 conservation scores was calculated. The minimum conservation score was calculated using a sliding window approach, with the window size being equal to the hotspot size. For each increment, a conservation score was calculated, and the resulting minimum conservation score was kept.

Data availability

The data associated with this study is also available at www.Synapse.org (accession: syn22216071), and the method is also downloadable at Protocol Exchange (<https://protocolexchange.researchsquare.com/>) under the same title and authorship.

Received: 21 June 2020; Accepted: 30 November 2020

Published online: 23 December 2020

References

1. WHO. *World Health Organization. WHO Director-General's opening remarks at the media briefing on COVID-19—11 March 2020.*
2. Coronaviridae Study Group of the International Committee on Taxonomy of V. The species Severe acute respiratory syndrome-related coronavirus: classifying 2019-nCoV and naming it SARS-CoV-2. *Nat. Microbiol.* **5**, 536–544. <https://doi.org/10.1038/s41564-020-0695-z> (2020).
3. Barcena, M. *et al.* Cryo-electron tomography of mouse hepatitis virus: Insights into the structure of the coronavirus. *Proc. Natl. Acad. Sci. USA* **106**, 582–587. <https://doi.org/10.1073/pnas.0805270106> (2009).
4. Fehr, A. R. & Perlman, S. Coronaviruses: an overview of their replication and pathogenesis. *Methods Mol. Biol.* **1282**, 1–23. https://doi.org/10.1007/978-1-4939-2438-7_1 (2015).
5. Zhou, P. *et al.* A pneumonia outbreak associated with a new coronavirus of probable bat origin. *Nature* **579**, 270–273. <https://doi.org/10.1038/s41586-020-2012-7> (2020).
6. Lu, R. *et al.* Genomic characterisation and epidemiology of 2019 novel coronavirus: implications for virus origins and receptor binding. *Lancet* **395**, 565–574. [https://doi.org/10.1016/S0140-6736\(20\)30251-8](https://doi.org/10.1016/S0140-6736(20)30251-8) (2020).
7. Grifoni, A. *et al.* A sequence homology and bioinformatic approach can predict candidate targets for immune responses to SARS-CoV-2. *Cell Host Microbe* <https://doi.org/10.1016/j.chom.2020.03.002> (2020).
8. Ahmed, S. F., Quadeer, A. A. & McKay, M. R. Preliminary identification of potential vaccine targets for the COVID-19 coronavirus (SARS-CoV-2) based on SARS-CoV immunological studies. *Viruses* <https://doi.org/10.3390/v12030254> (2020).
9. Yang, Z. Y. *et al.* A DNA vaccine induces SARS coronavirus neutralization and protective immunity in mice. *Nature* **428**, 561–564. <https://doi.org/10.1038/nature02463> (2004).
10. Channappanavar, R., Zhao, J. & Perlman, S. T cell-mediated immune response to respiratory coronaviruses. *Immunol. Res.* **59**, 118–128. <https://doi.org/10.1007/s12026-014-8534-z> (2014).
11. Liu, W. *et al.* Two-year prospective study of the humoral immune response of patients with severe acute respiratory syndrome. *J. Infect. Dis.* **193**, 792–795. <https://doi.org/10.1086/500469> (2006).
12. Liu, L. *et al.* Anti-spike IgG causes severe acute lung injury by skewing macrophage responses during acute SARS-CoV infection. *JCI Insight* <https://doi.org/10.1172/jci.insight.123158> (2019).
13. Tirado, S. M. & Yoon, K. J. Antibody-dependent enhancement of virus infection and disease. *Viral Immunol.* **16**, 69–86. <https://doi.org/10.1089/088282403763635465> (2003).
14. Wan, Y. *et al.* Molecular mechanism for antibody-dependent enhancement of coronavirus entry. *J. Virol.* <https://doi.org/10.1128/JVI.02015-19> (2020).
15. Tetro, J. A. Is COVID-19 receiving ADE from other coronaviruses?. *Microbes Infect.* **22**, 72–73. <https://doi.org/10.1016/j.micinf.2020.02.006> (2020).
16. Cao, W. C., Liu, W., Zhang, P. H., Zhang, F. & Richardus, J. H. Disappearance of antibodies to SARS-associated coronavirus after recovery. *N. Engl. J. Med.* **357**, 1162–1163. <https://doi.org/10.1056/NEJMc070348> (2007).
17. Sariol, A. & Perlman, S. Lessons for COVID-19 immunity from other coronavirus infections. *Immunity* **53**, 248–263. <https://doi.org/10.1016/j.immuni.2020.07.005> (2020).
18. Edridge, A. W. D. *et al.* Seasonal coronavirus protective immunity is short-lasting. *Nat. Med.* **26**, 1691–1693. <https://doi.org/10.1038/s41591-020-1083-1> (2020).
19. Seow, J. *et al.* Longitudinal observation and decline of neutralizing antibody responses in the three months following SARS-CoV-2 infection in humans. *Nat. Microbiol.* **5**, 1598–1607. <https://doi.org/10.1038/s41564-020-00813-8> (2020).

20. Long, Q. X. *et al.* Clinical and immunological assessment of asymptomatic SARS-CoV-2 infections. *Nat. Med.* **26**, 1200–1204. <https://doi.org/10.1038/s41591-020-0965-6> (2020).
21. Jeyanathan, M. *et al.* Immunological considerations for COVID-19 vaccine strategies. *Nat. Rev. Immunol.* **20**, 615–632. <https://doi.org/10.1038/s41577-020-00434-6> (2020).
22. Tay, M. Z., Poh, C. M., Renia, L., MacAry, P. A. & Ng, L. F. P. The trinity of COVID-19: immunity, inflammation and intervention. *Nat. Rev. Immunol.* <https://doi.org/10.1038/s41577-020-0311-8> (2020).
23. Haq, K. & McElhaney, J. E. Immunosenescence: influenza vaccination and the elderly. *Curr. Opin. Immunol.* **29**, 38–42. <https://doi.org/10.1016/j.coi.2014.03.008> (2014).
24. Arunachalam, P. S. *et al.* T cell-inducing vaccine durably prevents mucosal SHIV infection even with lower neutralizing antibody titers. *Nat. Med.* **26**, 932–940. <https://doi.org/10.1038/s41591-020-0858-8> (2020).
25. 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. *J. Virol.* **88**, 11034–11044. <https://doi.org/10.1128/JVI.01505-14> (2014).
26. Yang, L. T. *et al.* Long-lived effector/central memory T-cell responses to severe acute respiratory syndrome coronavirus (SARS-CoV) S antigen in recovered SARS patients. *Clinical Immunol.* **120**, 171–178. <https://doi.org/10.1016/j.clim.2006.05.002> (2006).
27. Yang, L. *et al.* Persistent memory CD4+ and CD8+ T-cell responses in recovered severe acute respiratory syndrome (SARS) patients to SARS coronavirus M antigen. *J. General Virol.* **88**, 2740–2748. <https://doi.org/10.1099/vir.0.82839-0> (2007).
28. Chen, J. *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. <https://doi.org/10.1128/JVI.01281-09> (2010).
29. Janice Oh, H. L., Ken-En Gan, S., Bertoletti, A. & Tan, Y. J. Understanding the T cell immune response in SARS coronavirus infection. *Emerg. Microbes Infect.* **1**, e23. <https://doi.org/10.1038/emi.2012.26> (2012).
30. Wherry, E. J. & Ahmed, R. Memory CD8 T-cell differentiation during viral infection. *J. Virol.* **78**, 5535–5545. <https://doi.org/10.1128/JVI.78.11.5535-5545.2004> (2004).
31. Zhao, J. *et al.* Airway memory CD4(+) T cells mediate protective immunity against emerging respiratory coronaviruses. *Immunity* **44**, 1379–1391. <https://doi.org/10.1016/j.immuni.2016.05.006> (2016).
32. Fan, Y. Y. *et al.* Characterization of SARS-CoV-specific memory T cells from recovered individuals 4 years after infection. *Adv. Virol.* **154**, 1093–1099. <https://doi.org/10.1007/s00705-009-0409-6> (2009).
33. Ng, O. W. *et al.* Memory T cell responses targeting the SARS coronavirus persist up to 11 years post-infection. *Vaccine* **34**, 2008–2014. <https://doi.org/10.1016/j.vaccine.2016.02.063> (2016).
34. Libraty, D. H., O’Neil, K. M., Baker, L. M., Acosta, L. P. & Olveda, R. M. Human CD4(+) memory T-lymphocyte responses to SARS coronavirus infection. *Virology* **368**, 317–321. <https://doi.org/10.1016/j.virol.2007.07.015> (2007).
35. Le Bert, N. *et al.* SARS-CoV-2-specific T cell immunity in cases of COVID-19 and SARS, and uninfected controls. *Nature* **584**, 457–462. <https://doi.org/10.1038/s41586-020-2550-z> (2020).
36. Mateus, J. *et al.* Selective and cross-reactive SARS-CoV-2 T cell epitopes in unexposed humans. *Science* **370**, 89–94. <https://doi.org/10.1126/science.abd3871> (2020).
37. Sette, A. & Crotty, S. Pre-existing immunity to SARS-CoV-2: the knowns and unknowns. *Nat. Rev. Immunol.* **20**, 457–458. <https://doi.org/10.1038/s41577-020-0389-z> (2020).
38. Weiskopf, D. *et al.* Phenotype and kinetics of SARS-CoV-2-specific T cells in COVID-19 patients with acute respiratory distress syndrome. *Sci. Immunol.* <https://doi.org/10.1126/sciimmunol.abd2071> (2020).
39. Altmann, D. M. & Boyton, R. J. SARS-CoV-2 T cell immunity: specificity, function, durability, and role in protection. *Sci. Immunol.* <https://doi.org/10.1126/sciimmunol.abd6160> (2020).
40. Ni, L. *et al.* Detection of SARS-CoV-2-specific humoral and cellular immunity in COVID-19 convalescent individuals. *Immunity* **52**, 971–977 e973. <https://doi.org/10.1016/j.immuni.2020.04.023> (2020).
41. Li, C. K. *et al.* T cell responses to whole SARS coronavirus in humans. *J. Immunol.* **181**, 5490–5500. <https://doi.org/10.4049/jimmunol.181.8.5490> (2008).
42. Mitchison, N. A. T-cell-B-cell cooperation. *Nat. Rev. Immunol.* **4**, 308–312. <https://doi.org/10.1038/nri1334> (2004).
43. Herst, C. V. *et al.* An effective CTL peptide vaccine for Ebola Zaire Based on Survivors’ CD8+ targeting of a particular nucleocapsid protein epitope with potential implications for COVID-19 vaccine design. *Vaccine* **38**, 4464–4475. <https://doi.org/10.1016/j.vaccine.2020.04.034> (2020).
44. Chen, K. & Kolls, J. K. T cell-mediated host immune defenses in the lung. *Annu. Rev. Immunol.* **31**, 605–633. <https://doi.org/10.1146/annurev-immunol-032712-100019> (2013).
45. Thevarajan, I. *et al.* Breadth of concomitant immune responses prior to patient recovery: a case report of non-severe COVID-19. *Nat. Med.* <https://doi.org/10.1038/s41591-020-0819-2> (2020).
46. Grifoni, A. *et al.* Targets of T cell responses to SARS-CoV-2 coronavirus in humans with COVID-19 disease and unexposed individuals. *Cell* **181**, 1489–1501 e1415. <https://doi.org/10.1016/j.cell.2020.05.015> (2020).
47. Panagioti, E., Klenerman, P., Lee, L. N., van der Burg, S. H. & Arens, R. Features of effective T cell-inducing vaccines against chronic viral infections. *Front. Immunol.* **9**, 276. <https://doi.org/10.3389/fimmu.2018.00276> (2018).
48. Campbell, K. M., Steiner, G., Wells, D. K., Ribas, A. & Kalbasi, A. Prediction of SARS-CoV-2 epitopes across 9360 HLA class I alleles. *bioRxiv* (2020).
49. Nguyen, A. *et al.* Human leukocyte antigen susceptibility map for SARS-CoV-2. *medRxiv* (2020).
50. Poran, A. *et al.* Sequence-based prediction of vaccine targets for inducing T cell responses to SARS-CoV-2 utilizing the bioinformatics predictor RECON. *bioRxiv* (2020).
51. Nguyen, A. *et al.* Human leukocyte antigen susceptibility map for severe acute respiratory syndrome coronavirus 2. *J. Virol.* <https://doi.org/10.1128/JVI.00510-20> (2020).
52. Pacheco-Olvera, D. L. *et al.* Bioinformatic analysis of shared B and T cell epitopes amongst relevant coronaviruses to human health: Is there cross-protection? *bioRxiv* (2020).
53. Björnsson, B. *et al.* Digital twins to personalize medicine. *Genome Med.* **12**, 4. <https://doi.org/10.1186/s13073-019-0701-3> (2019).
54. Zahn, L. M. HLA genetics and COVID-19. *Science* **368**, 841–841. <https://doi.org/10.1126/science.368.6493.841-b> (2020).
55. Barquera, R. *et al.* Binding affinities of 438 HLA proteins to complete proteomes of seven pandemic viruses and distributions of strongest and weakest HLA peptide binders in populations worldwide. *HLA* **96**, 277–298. <https://doi.org/10.1111/tan.13956> (2020).
56. Gonzalez-Galarza, F. F. *et al.* Allele frequency net database (AFND) 2020 update: gold-standard data classification, open access genotype data and new query tools. *Nucleic Acids Res.* **48**, D783–D788. <https://doi.org/10.1093/nar/gkz1029> (2020).
57. Dhanda, S. K. *et al.* IEDB-AR: immune epitope database-analysis resource in 2019. *Nucleic Acids Res.* **47**, W502–W506. <https://doi.org/10.1093/nar/gkz452> (2019).
58. Jia, Y. *et al.* Analysis of the mutation dynamics of SARS-CoV-2 reveals the spread history and emergence of RBD mutant with lower ACE2 binding affinity. *bioRxiv* <https://doi.org/10.1101/2020.04.09.034942> (2020).
59. Pachetti, M. *et al.* Emerging SARS-CoV-2 mutation hot spots include a novel RNA-dependent-RNA polymerase variant. *J. Transl. Med.* **18**, 179. <https://doi.org/10.1186/s12967-020-02344-6> (2020).
60. Rosenberg, W. Mechanisms of immune escape in viral hepatitis. *Gut* **44**, 759–764. <https://doi.org/10.1136/gut.44.5.759> (1999).

61. Batorsky, R., Sergeev, R. A. & Rouzine, I. M. The route of HIV escape from immune response targeting multiple sites is determined by the cost-benefit tradeoff of escape mutations. *PLoS Comput. Biol.* **10**, e1003878. <https://doi.org/10.1371/journal.pcbi.1003878> (2014).
62. Shu, Y. & McCauley, J. GISAIID: global initiative on sharing all influenza data—from vision to reality. *Eur. Commun. Disease Bull.* <https://doi.org/10.2807/1560-7917.ES.2017.22.13.30494> (2017).
63. Paul, S. *et al.* Benchmarking predictions of MHC class I restricted T cell epitopes in a comprehensively studied model system. *PLoS Comput. Biol.* **16**, e1007757. <https://doi.org/10.1371/journal.pcbi.1007757> (2020).
64. Lurie, N., Saville, M., Hatchett, R. & Halton, J. Developing covid-19 vaccines at pandemic speed. *N. Engl. J. Med.* **382**, 1969–1973. <https://doi.org/10.1056/NEJMp2005630> (2020).
65. Simovski, B. *et al.* GSuite HyperBrowser: integrative analysis of dataset collections across the genome and epigenome. *GigaScience* **6**, 1–12. <https://doi.org/10.1093/gigascience/gix032> (2017).
66. Sandve, G. K. *et al.* The Genomic HyperBrowser: inferential genomics at the sequence level. *Genome Biol.* **11**, R121. <https://doi.org/10.1186/gb-2010-11-12-r121> (2010).
67. Hadfield, J. *et al.* Nextstrain: real-time tracking of pathogen evolution. *Bioinformatics* **34**, 4121–4123. <https://doi.org/10.1093/bioinformatics/bty407> (2018).
68. Sievers, F. & Higgins, D. G. Clustal omega for making accurate alignments of many protein sequences. *Protein Sci. Public. Protein Soc.* **27**, 135–145. <https://doi.org/10.1002/pro.3290> (2018).

Author contributions

B.S., C.M., M.G., H.F., I.V., S.T., J.M., R.S. and T.C. are employees of NEC OncoImmunity, a subsidiary of NEC Corporation. BM and JC are employees of NEC Laboratories Europe.

Competing interests

BS, CM, MG, HF, IV, ST, JM, RS and TC are employees of NEC OncoImmunity, a subsidiary of NEC Corporation. BM and JC are employees of NEC Laboratories Europe. The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at <https://doi.org/10.1038/s41598-020-78758-5>.

Correspondence and requests for materials should be addressed to T.C.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2020