Epitopes targeted by T cells in convalescent COVID-19 patients

Ahmed A. Quadeer[1,*], Syed Faraz Ahmed[1], and Matthew R. McKay[1,2,*]

[1]Department of Electronic and Computer Engineering, The Hong Kong University of Science and Technology, Hong Kong, China

[2]Department of Chemical and Biological Engineering, The Hong Kong University of Science and Technology, Hong Kong, China

*Correspondence should be addressed to eeaaquadeer@ust.hk and m.mckay@ust.hk

Abstract

There is growing evidence pointing to the protective role of T cells against COVID-19. Vaccines eliciting targeted T cell responses have the potential to provide robust, long-lasting immunity. However, their design requires knowledge of the SARS-CoV-2-specific epitopes that can elicit a T cell response and confer protection across a wide population. Here, we provide a unified description of emerging data of SARS-CoV-2 T cell epitopes compiled from results of 8 independent studies of convalescent COVID-19 patients. We describe features of these epitopes relevant for vaccine design, while indicating knowledge gaps that can, in part, be augmented using prior immunological data from SARS-CoV. The landscape of SARS-CoV-2 T cell epitopes that we describe can help guide SARS-CoV-2 vaccine development as well as future immunological studies. A web-based platform has also been developed to complement these efforts.

Introduction

Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), the causative agent of COVID-19, has led to a global public health crisis and resulted in more than 820,000 deaths. Unprecedented efforts are underway to produce vaccines to prevent COVID-19 (Callaway, 2020). However, uncertainty remains about the effectiveness of different vaccine candidates in eliciting long-term protective immunity needed to control the pandemic.

Vaccine development is aided by an understanding of the natural protective immune responses against SARS-CoV-2. This includes both humoral and cellular immunity, mediated by antibodies and T cells respectively. Significant attention has focused on understanding antibody responses against SARS-CoV-2 (Yong *et al.*, 2020; Zohar and Alter, 2020; Wu *et al.*, 2020); though, studies informing the role of T cells in conferring protection have started to emerge.

There is growing evidence to suggest that vaccines designed to produce targeted T cell responses may be particularly important. Preliminary studies have reported an inverse relation between the intensity of T cell responses and disease severity (Snyder *et al.*, 2020; Nelde *et al.*, 2020; Zhang *et al.*, 2020), while the opposite has been observed for antibody titers (Long, Liu, *et al.*, 2020; Snyder *et al.*, 2020). Further investigations on larger cohorts are required, however these initial results suggest a protective role of T cells against COVID-19. Studies of convalescent COVID-19 patients have also detected SARS-CoV-2-specific T cells 3 months post infection (Snyder *et al.*, 2020), giving promising signs for the potential of T cells to provide lasting protection. Longevity may be a concern for antibodies, which have been reported to be undetectable within just a few months post-infection (Snyder *et al.*, 2020; Long, Tang, *et al.*, 2020; Seow *et al.*, 2020). Consistent observations have also been reported for the most closely related human coronavirus, SARS-CoV, for which T cells were shown to persist up to 11 years post infection (Ng *et al.*, 2016), whereas antibody responses waned after a few years (Tang *et al.*, 2011).

In developing T-cell-based vaccines, it is important to determine the SARS-CoV-2-specific T cell epitopes that can induce a protective immune response. During the very early phase of the pandemic, bioinformatics-based studies were used to predict these epitopes (Ahmed *et al.*, 2020c; Grifoni *et al.*, 2020). More recently, SARS-CoV-2-specific epitopes have been identified experimentally by analysing blood samples from convalescent COVID-19 patients, often guided by the earlier predictions. These studies, performed using a range of experimental procedures, lend new insight into the specific T cell responses that may confer protection against SARS-CoV-2, and provide important clues for vaccine development.

Here, we present a summary of emerging data (as of 7 August 2020) about experimentally-identified SARS-CoV-2 T cell epitopes compiled from 1,259 convalescent COVID-19 patients across 8 independent studies. This includes published studies as well as preprints. We report distinctive features of the epitopes reported to be targeted in these patients, including their positioning within the SARS-CoV-2 proteome, genetic conservation, immunodominance, and associated human leukocyte antigen (HLA) alleles. We also describe some knowledge gaps, due in part to a limited diversity of patient samples that have been studied, and in some cases, to limitations of the experimental procedures themselves. Most notable is the HLA association of T cell epitopes, an important factor for vaccine design, since it is used to estimate what fraction of the population may respond to a given set of epitopes. For most of the identified T cell epitopes, only partial HLA information is currently available, or no information at all. In seeking to partly resolve the knowledge gaps, we propose protocols to augment the existing T cell epitope data by leveraging prior experimental immunological studies of SARS-CoV.

The set of T cell epitopes that we compile and describe can aid the design of T-cell-based vaccines with broad population coverage, and inform related experimental immunological studies aimed at understanding T cell responses against SARS-CoV-2 and elucidating their protective role against COVID-19. We report a web dashboard that we have developed to support these on-going scientific efforts.

Results

SARS-CoV-2 T cell epitopes identified by studies of convalescent COVID-19 patients

Results compiled from 8 experimental studies of T cell responses in a total of 1,259 convalescent COVID-19 patients (Le Bert *et al.*, 2020; Chour *et al.*, 2020; Peng *et al.*, 2020; Nelde *et al.*, 2020; Poran *et al.*, 2020; Shomuradova *et al.*, 2020; Ferretti *et al.*, 2020; Snyder *et al.*, 2020) revealed 303 unique T cell epitopes. Of these, 71 are $CD8^+$ (HLA class I restricted), 21 are $CD4^+$ (HLA class II restricted), while 211 epitopes had no HLA allele information reported (Figure 1A). Epitopes were identified for each protein of SARS-CoV-2 (Figure 1B), with the majority falling within the spike (S) and nucleocapsid (N)—the most exposed and the most abundantly expressed SARS-CoV-2 protein respectively. All identified epitopes have high genetic conservation (>95%) among the 42,815 SARS-CoV-2 sequences (as of 7 August

2020), except for two epitopes: $_{77}$GNYTVSCLPFTI$_{88}$ in the ORF8 protein (conservation = 0.89) and $_{312}$FPPTSFGPL$_{320}$ in the orf1b protein (conservation = 0.28) (Supplementary Table 1).

A study of the immunodominance of each epitope, estimated by the standard response frequency (RF) metric (Dhanda *et al.*, 2018; Tian *et al.*, 2019; Weiskopf *et al.*, 2013) (Methods), revealed that 23 of the reported epitopes induced T cell responses in over three quarters of the subjects tested (Figure 1C). Confidence in the RF score is high for some epitopes, such as $_{361}$KTFPPTEKP$_{369}$ in N, $_{176}$LSYYKLGASQRVAGD$_{190}$ in membrane (M), and $_{1637}$TTDPSFLGRY$_{1646}$ in orf1a, whereas for others there is uncertainty due to the limited number of subjects tested. Specific epitopes with high RF, such as $_{1637}$TTDPSFLGRY$_{1646}$ in orf1a, $_{134}$ATEGALNTPK$_{143}$ in N, and $_{1000}$RLQSLQTYV$_{1008}$ in S, were observed in multiple independent studies.

Mapping the epitopes of the immunodominant S and N proteins onto the primary structure (Methods) revealed their potential functional or structural significance (Supplementary Figure 2). Of the 98 S protein epitopes, 19 lie within the receptor-binding domain (330-583), which facilitates binding of the S protein to the ACE2 receptor on human cells (Shang *et al.*, 2020). In addition, 48 epitopes lie within the S2 subunit, including an epitope ($_{821}$LLFNKVTLA$_{829}$) that lies within the fusion peptide region (816-855) (Tang *et al.*, 2020). Similarly, of the 60 N protein epitopes, 24 lie within the RNA-binding domain (41-186), while 20 lie within the dimerization region (258-361) (UniProt, 2020).

For most of the experimental studies, bioinformatics methods were used to select an initial set of potential SARS-CoV-2 epitopes for which to test a response against. This includes methods that use peptide-MHC binding prediction algorithms, such as NetMHCpan (Jurtz *et al.*, 2017), or methods designed to exploit the genetic similarity between SARS-CoV-2 and SARS-CoV, along with knowledge of experimentally-determined SARS-CoV epitopes (Ahmed *et al.*, 2020c, 2020b). Of the 303 unique SARS-CoV-2 T cell epitopes (Supplementary Figure 1), 59 (~19%) matched exactly with known SARS-CoV epitopes.

Augmenting SARS-CoV-2 T cell epitope information using experimental data for SARS-CoV

T-cell-based vaccines should be designed to induce a protective response in a large fraction of the population. The population coverage of a given T cell epitope is tied to its HLA association. For the SARS-CoV-2 T cell epitopes discovered so far however, the reported number of HLA alleles is typically limited to one or two, or in many cases, to none at all (Figure 1).

The experiments that report at least partial HLA information were designed to analyse specific synthetic peptides using functional assays (IFN-γ ELISPOT/ELISA) or peptide-MHC (p-MHC) multimers (Figure 2A). Only a single associated HLA allele was reported for the majority of identified epitopes, with strong biases towards the most globally prevalent HLA alleles (Figure 2B). This is due, in part, to the limited number of convalescent COVID-19 patients that have been studied. Experimental complications in epitope identification from blood samples, such as challenges in generating specific p-MHC multimers (Mallone and Nepom, 2004), also contribute to the data limitations.

We proposed methods to augment the SARS-CoV-2 T cell epitope data (Methods) with additional HLA information by using prior knowledge of experimentally-identified T cell epitopes for SARS-CoV and their corresponding HLA associations. Among the SARS-CoV-2 T cell epitopes having at least partial HLA allele information (92 unique) (Figure 2A), 20 were genetically identical to known SARS-CoV epitopes (Supplementary Figure 3A; the remaining epitopes are listed in Supplementary Figure 3B). The HLA alleles determined experimentally for these 20 SARS-CoV-2 epitopes matched exactly with those that had been determined for the corresponding SARS-CoV epitopes (Supplementary Figure 3A), suggesting that an epitope-HLA combination associated with a T cell response for one virus is highly likely to be associated with a response for the other (Košmrlj *et al.*, 2008; Ngono and Shresta, 2019). Hence, the comparatively more extensive HLA associations known for SARS-CoV epitopes (Ahmed *et al.*, 2020b, 2020c) could complement the limited existing information for SARS-CoV-2. Based on this rationale, of the 20 genetically-identical SARS-CoV epitopes, we identified additional HLA alleles for 12 epitopes. This also increased the overall diversity of HLA alleles associated with at least one epitope (Figure 2C).

The experiments that do not report HLA information identified epitopes by testing responses against synthetic peptide pool libraries using functional assays, or molecular assays such as MIRA (Figure 2D). For the SARS-CoV-2 epitopes identified with these methods (211 unique), in addition to having a lack of HLA associations, the reported epitopes have varying lengths, ranging from 9 amino acids to long immunogenic peptides comprising more than 18 amino acids (Figure 2E). That is, for around half of the reported epitopes, it is likely that the precise epitopes against which the T cell responses were mounted were not exactly resolved. The uncertainty is most significant for $CD8^+$ T cell epitopes, which are relatively short (~9-12 amino acids long (Serwold *et al.*, 2002)).

To facilitate exact identification, these 211 reported SARS-CoV-2 epitopes were first scanned to determine if they had an identical genetic match with known SARS-CoV epitopes, or if any fragment had an identical match (Methods). This

procedure revealed 125 unique SARS-CoV epitopes that exactly matched with, or were contained within, 71 of the reported SARS-CoV-2 epitopes (Supplementary Figure 3C). These matching epitopes, comprised of 9 to 11 amino acids, likely correspond to the actual T cell epitopes targeted within the convalescent COVID-19 patients. The epitopes were associated with a diverse set of HLA class I alleles (Figure 2F).

Population coverage of the T cell epitopes

The population coverage of the T cell epitopes was estimated from the corresponding HLA associations. For the epitopes with partial experimental HLA information (Figures 2A and 2B), the median coverage of the global population was estimated to be 21%, with a few epitopes having close to 50% coverage (Figure 3A, left panel). For the 12 SARS-CoV-2 epitopes having additional HLA associations after augmenting with SARS-CoV data, the median population coverage was 44%, significantly increasing from 17% prior to data augmentation (Figure 3A, right panel). Specific SARS-CoV-2 epitopes, such as $_{66}$FPRGQGVPI$_{74}$ in N and $_{171}$ATSRTLSYYK$_{180}$ in M, were observed to have quite broad global population coverage (~60%) after data augmentation (Supplementary Figure 3A).

For the SARS-CoV-2 epitopes reported without any HLA information (Figure 2D), the precise epitopes, predicted with SARS-CoV data, had a high estimated median global population coverage of ~40% (Figure 3B). One particular epitope in orf1b, $_{1708}$VLWAHGFEL$_{1716}$, had a very high estimated coverage of 85% (Supplementary Figure 3C). This epitope may also have functional significance. It is located in the 3'-to-5' exonuclease nsp14 protein, suggested to play a possible role in providing proofreading support to the RNA polymerase of SARS-CoV-2 (Romano *et al.*, 2020).

Web-dashboard for reporting SARS-CoV-2 T cell epitope data

The SARS-CoV-2 T cell epitope data that we compile and report here has been integrated into a web-based dashboard (Ahmed *et al.*, 2020a). This dashboard provides exportable data tables listing the SARS-CoV-2 epitopes, as well as information gained by augmenting with SARS-CoV epitope data. Graphical displays are used to summarize different characteristics of the SARS-CoV-2 epitopes, including aggregate information as well as specific details of the individual epitopes. We plan to update the dashboard with new experimental information as it becomes available, with the goal of aiding further research in understanding T cell responses against SARS-CoV-2 in humans, and in guiding studies related to COVID-19 vaccines.

DISCUSSION

It is becoming increasingly apparent that T cells, one of the main protagonists of the adaptive immune system, may play a significant role in providing protective immunity against COVID-19. Understanding the repertoire of SARS-CoV-2-specific epitopes that can trigger a robust T cell response in humans is important for rational vaccine design. Knowledge of these epitopes can be used in designing peptide or nucleic acid based vaccines against SARS-CoV-2. For peptide-based vaccine designs, epitope sequences can be directly incorporated into the vaccine formulations (Herst *et al.*, 2020; Atsmon *et al.*, 2012), whereas nucleic acid sequences encoding for epitopes, along with the appropriate signal peptides and linkers, can be included in DNA or RNA vaccine constructs (Sahin *et al.*, 2017).

In specifying epitopes as potential targets for a universal COVID-19 vaccine, epitope sets may be selected to maximize accumulated global population coverage, possibly incorporating other factors such as the number of responding subjects and the response frequency. A simple analysis of the SARS-CoV-2 epitopes chosen to maximize accumulated population coverage (Dahirel *et al.*, 2011; Quadeer *et al.*, 2014; Ahmed *et al.*, 2019) revealed a group of 12 epitopes (9 HLA class I restricted, 3 HLA class II restricted), providing an estimated global coverage of 99.7% (Supplementary Figure 4). Considering only three epitopes ($_{1708}$VLWAHGFEL$_{1716}$ in orf1b, $_{171}$ATSRTLSYY$_{180}$ in M, and $_{311}$ASAFFGMSRIGMEVT$_{325}$ in N) already provided an estimated global coverage of ~97.5%, due to the association of these epitopes with multiple distinct HLA alleles. Hence, a vaccine designed to activate T cells that specifically target such epitopes (Rosendahl Huber *et al.*, 2014) might help to elicit a protective immune response against SARS-CoV-2 in a large percentage of the global population.

Our focus was on SARS-CoV-2-specific T cell responses recorded in convalescent patients, however knowledge of T cell epitopes targeted in animal studies is also informative (Corbett *et al.*, 2020; Smith *et al.*, 2020). SARS-CoV-2 epitopes that have been reported in animal models may inform potential immune targets in COVID-19-infected individuals, and can help in guiding further immunological experiments that seek to probe T cell responses that arise due to natural infection, or those elicited by vaccination. Of particular note is a preclinical study (Smith *et al.*, 2020) that measured T cell responses in mice after immunization with a synthetic DNA-based vaccine candidate targeting the SARS-CoV-2 S protein. The reported T cell immunogenic peptides were shown to comprise multiple previously known SARS-CoV human epitopes, providing a further indication of their potential immunogenicity in humans. The SARS-CoV-2 epitopes reported to be targeted in animal models have been incorporated into the web dashboard.

In augmenting the current T cell epitope data for SARS-CoV-2, we reasoned that if an epitope from SARS-CoV is known to elicit a T cell response in infected persons with a specific HLA allele, then it is highly likely that a genetically-identical epitope from SARS-CoV-2 will also induce a T cell response in infected persons with the same HLA allele. This is because binding of T cells to epitopes is independent of the pathogen from which the epitope is derived (Košmrlj *et al.*, 2008; Butler *et al.*, 2013). Such cross-reactive T cell epitopes have been reported previously in genetically similar viruses, including dengue and zika viruses (Ngono and Shresta, 2019). Nonetheless, for the specific epitope-HLA combinations predicted based on SARS-CoV epitope data, further experiments are still required to confirm their immunogenicity for SARS-CoV-2.

Overall, the data that we describe based on recent experimental studies demonstrates an impressive and diverse list of SARS-CoV-2 T cell epitopes targeted by convalescent COVID-19 patients. Subsets of these epitopes exhibit desirable properties, including high genetic conservation and response frequency, and broad global population coverage. Vaccines designed to target epitopes with these properties have the potential to induce widespread protective T-cell immunity in the human population, and could play a role in curbing the spread of the COVID-19 pandemic.

METHODS

T cell epitope data derived from convalescent COVID-19 patients

Data was compiled from 8 immunological studies (as of 7 August 2020) reporting SARS-CoV-2 T cell epitopes in convalescent COVID-19 patients. More specifically, epitope data was extracted from Figure 3 in (Chour *et al.*, 2020), Table 1 in (Shomuradova *et al.*, 2020), Extended Data Tables 2 and 3 in (Nelde *et al.*, 2020), Figure 3 in (Poran *et al.*, 2020), Table 1 in (Ferretti *et al.*, 2020), Table 1 in (Le Bert *et al.*, 2020), Tables 1 and 2 in (Peng *et al.*, 2020), and data deposited by authors of (Snyder *et al.*, 2020) in the Immune Epitope Database (IEDB).

Sequence data, epitope conservation and coverage

SARS-CoV-2 genomic sequences were obtained from the GISAID database (https://www.gisaid.org/) on 7 August 2020. We downloaded only the complete (full-genome) sequences with high coverage using the options provided on the GISAID database. All the downloaded sequences were aligned to the SARS-CoV-2 reference genome (GenBank ID: NC_045512.2) using MAFFT (Katoh and Standley, 2013). The genomic MSA was translated using an in-house code to obtain the protein MSAs. The positions of the open reading frames provided with the reference sequence were used to identify the respective protein regions of the full-genome.

The conservation of each SARS-CoV-2 T cell epitope was calculated as the fraction of SARS-CoV-2 sequences that encompassed the precise epitope sequence. The coverage of SARS-CoV-2 T cell epitopes at any position of a protein was calculated by counting the number of epitopes that included that position.

Response frequency (RF)

RF score (Dhanda *et al.*, 2018) was used to quantify the immunodominance of the SARS-CoV-2 epitopes reported to be recognized by T cells in convalescent COVID-19 patients. The RF score of an epitope is defined as follows:

$$\mathrm{RF} = \frac{\sum_{i=1}^{S} r_i}{\sum_{i=1}^{S} t_i}$$

where r_i is the number of subjects responding to the epitope in study i, t_i is the number of subjects tested for a response against the epitope in study i, and S is the total number of studies. An RF score calculated using a large number of tested subjects would be more reliable than one calculated using relatively few subjects. To account for this, we computed the 95% confidence interval for the RF score of each epitope using the binomial cumulative distribution function (Dhanda *et al.*, 2018).

Augmenting SARS-CoV-2 T cell epitope data using immunological information of SARS-CoV

SARS-CoV T cell epitopes, determined using positive MHC binding assay or/and T cell assay, were obtained from the Virus Pathogen Database and Analysis Resource (ViPR; www.viprbrc.org) on 7 August 2020.

For augmenting the SARS-COV-2 epitope data with additional HLA information, we compared each SARS-CoV-2 epitope against the set of SARS-CoV epitopes. For any SARS-CoV-2 epitope whose sequence matched exactly to a

SARS-CoV epitope sequence, any additional HLA associations of the SARS-CoV epitope were appended to those of the SARS-CoV-2 epitope.

For the specific SARS-CoV-2 epitopes with some uncertainty about their exact epitope sequence (corresponding to those epitopes with no HLA information reported), SARS-CoV epitope data was used to help pinpoint the exact sequence identity. This was achieved by identifying those SARS-CoV epitopes whose sequence had an exact match to any fragment of any of the SARS-CoV-2 epitope sequences. The HLA associations of these matched SARS-CoV epitopes were also recorded.

Estimating global population coverage of epitopes

The global population coverage of an epitope refers to the percentage of individuals in the world population that is expected to mount a T cell response against that epitope. We calculated the population coverage of a T cell epitope based on the HLA alleles associated with it. The population coverage was estimated using the tool downloaded from the IEDB Analysis Resource (http://tools.iedb.org/population/download/). This tool employs global HLA allele frequency data obtained from the Allele Frequency database (http://www.allelefrequencies.net/).

Data and code availability

Compiled data of the SARS-CoV-2 T cell epitopes is available to download from the web dashboard (Ahmed *et al.*, 2020a). All data and scripts for reproducing the figures in this manuscript are available at https://github.com/ahmedaq/SARS2TcellEpitopes.

Author contributions

A.A.Q. and M.R.M. conceptualized the study and wrote the manuscript. A.A.Q. compiled the data from the literature. A.A.Q., S.F.A., and M.R.M. analyzed the data. A.A.Q. and S.F.A performed the computations. S.F.A. generated the figures and developed the web-dashboard.

Acknowledgements

The authors were supported by the General Research Fund of the Hong Kong Research Grants Council (RGC) [Grant No. 16204519 and 16201620]. S.F.A. was additionally supported by the Hong Kong Ph.D. Fellowship Scheme (HKPFS).

Competing interests

The authors declare no competing interests.

REFERENCES

Ahmed,S.F. *et al.* (2020a) A web-based dashboard for reporting epitopes targeted by T cells in convalescent COVID-19 patients. https://www.mckayspcb.com/SARS2TcellEpitopes/.

Ahmed,S.F. *et al.* (2020b) COVIDep: a web-based platform for real-time reporting of vaccine target recommendations for SARS-CoV-2. *Nat. Protoc.*, **15**, 2141–2142.

Ahmed,S.F. *et al.* (2020c) Preliminary identification of potential vaccine targets for the COVID-19 coronavirus (SARS-CoV-2) based on SARS-CoV immunological studies. *Viruses*, **12**, 254.

Ahmed,S.F. *et al.* (2019) Sub-dominant principal components inform new vaccine targets for HIV Gag. *Bioinformatics*, **35**, 3884–3889.

Atsmon,J. *et al.* (2012) Safety and immunogenicity of multimeric-001—a novel universal influenza vaccine. *J. Clin. Immunol.*, **32**, 595–603.

Le Bert,N. *et al.* (2020) SARS-CoV-2-specific T cell immunity in cases of COVID-19 and SARS, and uninfected controls. *Nature*, **584**, 457–462.

Butler,T.C. *et al.* (2013) Quorum sensing allows T cells to discriminate between self and nonself. *Proc. Natl. Acad. Sci. U. S. A.*, **110**, 11833–11838.

Callaway,E. (2020) The race for coronavirus vaccines: a graphical guide. *Nature*, **580**, 576–577.

Chour,W. *et al.* (2020) Shared antigen-specific CD8+ T cell responses against the SARS-COV-2 spike protein in HLA A*02:01 COVID-19 participants. *medRxiv*, 2020.05.04.20085779.

Corbett,K.S. *et al.* (2020) SARS-CoV-2 mRNA vaccine design enabled by prototype pathogen preparedness. *Nature*.

Dahirel,V. *et al.* (2011) Coordinate linkage of HIV evolution reveals regions of immunological vulnerability. *Proc. Natl. Acad. Sci.*, **108**, 11530–11535.

Dhanda,S.K. *et al.* (2018) ImmunomeBrowser: a tool to aggregate and visualize complex and heterogeneous epitopes in reference proteins. *Bioinformatics*, **34**, 3931–3933.

Ferretti,A.P. *et al.* (2020) COVID-19 patients form memory CD8+ T cells that recognize a small set of shared immunodominant epitopes in SARS-CoV-2. *SSRN Electron. J.*

Grifoni,A. *et al.* (2020) A sequence homology and bioinformatic approach can predict candidate targets for immune responses to SARS-CoV-2. *Cell Host Microbe*, **27**, 1–10.

Herst,C.V. *et al.* (2020) 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.

Jurtz,V. *et al.* (2017) NetMHCpan-4.0: Improved peptide–MHC class I interaction predictions integrating eluted ligand and peptide binding affinity data. *J. Immunol.*, **199**, 3360–3368.

Katoh,K. and Standley,D.M. (2013) MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. *Mol. Biol. Evol.*, **30**, 772–780.

Košmrlj,A. *et al.* (2008) How the thymus designs antigen-specific and self-tolerant T cell receptor sequences. *Proc. Natl. Acad. Sci. U. S. A.*, **105**, 16671–16676.

Long,Q.-X., Liu,B.-Z., *et al.* (2020) Antibody responses to SARS-CoV-2 in patients with COVID-19. *Nat. Med.*, **26**, 845–848.

Long,Q.-X., Tang,X.-J., *et al.* (2020) Clinical and immunological assessment of asymptomatic SARS-CoV-2 infections. *Nat. Med.*

Mallone,R. and Nepom,G.T. (2004) MHC Class II tetramers and the pursuit of antigen-specific T cells: define, deviate, delete. *Clin. Immunol.*, **110**, 232–242.

Nelde,A. *et al.* (2020) SARS-CoV-2 T-cell epitopes define heterologous and COVID-19-induced T-cell recognition. *Res. Sq.*

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

Ngono,A.E. and Shresta,S. (2019) Cross-reactive T cell immunity to dengue and zika viruses: New insights into vaccine development. *Front. Immunol.*, **10**, 1–9.

Peng,Y. *et al.* (2020) Broad and strong memory CD4+ and CD8+ T cells induced by SARS-CoV-2 in UK convalescent COVID-19 patients. *bioRxiv*, 2020.06.05.134551.

Poran,A. *et al.* (2020) Sequence-based prediction of SARS-CoV-2 vaccine targets using a mass spectrometry-based bioinformatics predictor identifies immunogenic T cell epitopes. *Genome Med.*, **12**, 70.

Quadeer,A.A. *et al.* (2014) Statistical linkage analysis of substitutions in patient-derived sequences of genotype 1a hepatitis C virus nonstructural protein 3 exposes targets for immunogen design. *J. Virol.*, **88**, 7628–7644.

Romano,M. *et al.* (2020) A structural view of SARS-CoV-2 RNA replication machinery: RNA synthesis, proofreading and final capping. *Cells*, **9**, 1267.

Rosendahl Huber,S. *et al.* (2014) T cell responses to viral infections - opportunities for peptide vaccination. *Front. Immunol.*, **5**, 1–12.

Sahin,U. *et al.* (2017) Personalized RNA mutanome vaccines mobilize poly-specific therapeutic immunity against cancer. *Nature*, **547**, 222–226.

Seow,J. *et al.* (2020) Longitudinal evaluation and decline of antibody responses in SARS-CoV-2 infection. *medRxiv*, 2020.07.09.20148429.

Serwold,T. *et al.* (2002) ERAAP customizes peptides for MHC class I molecules in the endoplasmic reticulum. *Nature*, **419**, 480–483.

Shang,J. *et al.* (2020) Cell entry mechanisms of SARS-CoV-2. *Proc. Natl. Acad. Sci.*, **117**, 11727–11734.

Shomuradova,A.S. *et al.* (2020) SARS-CoV-2 epitopes are recognized by a public and diverse repertoire of human T-cell receptors. *SSRN Electron. J.*

Smith,T.R.F. *et al.* (2020) Immunogenicity of a DNA vaccine candidate for COVID-19. *Nat. Commun.*, **11**, 2601.

Snyder,T.M. *et al.* (2020) Magnitude and dynamics of the T-Cell response to SARS-CoV-2 infection at both individual and population levels. *medRxiv*, 2020.07.31.20165647.

Tang,F. *et al.* (2011) Lack of peripheral memory B cell responses in recovered patients with severe acute respiratory syndrome: A six-year follow-up study. *J. Immunol.*, **186**, 7264–7268.

Tang,T. *et al.* (2020) Coronavirus membrane fusion mechanism offers a potential target for antiviral development. *Antiviral Res.*, **178**, 104792.
Tian,Y. *et al.* (2019) Human T cell response to dengue virus infection. *Front. Immunol.*, **10**, 1–9.
UniProt (2020) SARS-CoV-2 nucleoprotein. https://www.uniprot.org/uniprot/P0DTC9.
Weiskopf,D. *et al.* (2013) Comprehensive analysis of dengue virus-specific responses supports an HLA-linked protective role for CD8+ T cells. *Proc. Natl. Acad. Sci. U. S. A.*, **110**, E2046-53.
Wu,F. *et al.* (2020) Neutralizing antibody responses to SARS-CoV-2 in a COVID-19 recovered patient cohort and their implications. *medRxiv*, 2020.03.30.20047365.
Yong,S.E.F. *et al.* (2020) Connecting clusters of COVID-19: an epidemiological and serological investigation. *Lancet Infect. Dis.*, **20**, 809–815.
Zhang,Xiaonan *et al.* (2020) Viral and host factors related to the clinical outcome of COVID-19. *Nature*, **583**, 437–440.
Zohar,T. and Alter,G. (2020) Dissecting antibody-mediated protection against SARS-CoV-2. *Nat. Rev. Immunol.*, **20**, 392–394.

Figures

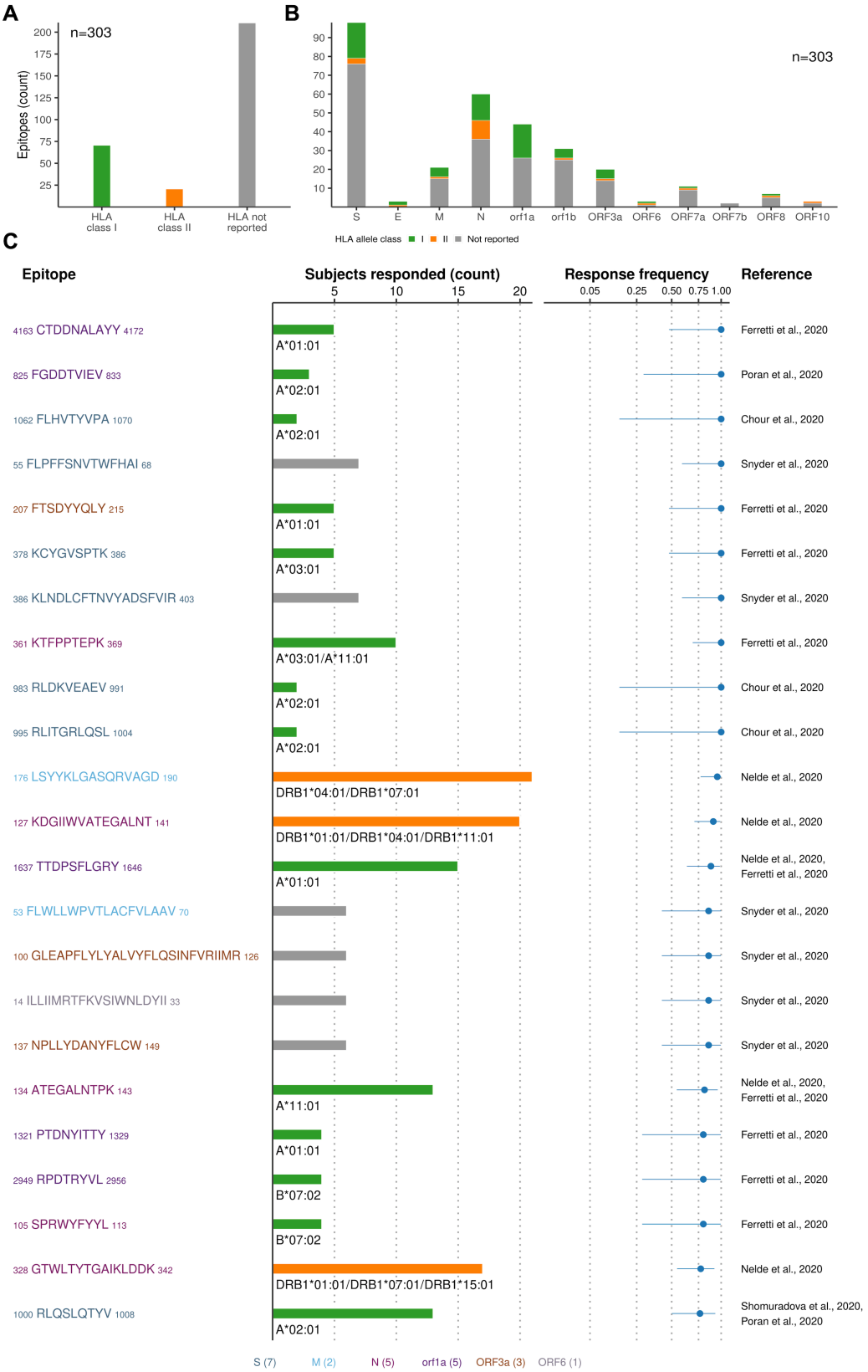


Figure 1. Summary of unique SARS-CoV-2-specific T cell epitopes reported based on blood samples of convalescent COVID-19 patients. (**A**) The number of epitopes (n=303) according to HLA class restriction. (**B**) The number of epitopes (n=303) in specific SARS-CoV-2 proteins. The bars are shaded according to the HLA class restriction in (A). (**C**) List of SARS-CoV-2-specific T cell epitopes and the corresponding reported number of responding patients, the reported associated HLA alleles (where available), and the computed response frequency (RF) along with the 95% confidence interval (CI) (see Methods for details). The studies reporting each epitope are shown on the right. Epitopes are arranged in descending order according to RF and colored according to the protein from which they were derived (legend shown at the bottom). The subscript numbers given before and after the epitope indicate its location within the protein. The bars are shaded according to HLA class restriction in (A). For brevity, only the epitopes (n=23) with RF > 0.75 are displayed (see Supplementary Figure 1 and Supplementary Table 1 for the complete list).

Summary of identified T cell epitopes with experimentally-determined HLA allele information

A

Epitope identification procedure*	Identified epitopes	References
IFN-γ ELISA/ELISPOT using specific synthetic peptides	71	Nelde et al. 2020
MHC multimer staining using specific synthetic peptides	37	Poran et al. 2020; Shomuradova et al., 2020; Ferretti et al., 2020; Chour et al., 2020
IFN-γ ELISA using specific synthetic peptides	26	Ferretti et al., 2020
Total	***92** (unique)*	

*IFN = Interferon; ELISA = enzyme-linked immunosorbent assay; ELISPOT = enzyme-linked immune absorbent spot.

Summary of identified T cell epitopes without experimentally-determined HLA allele information

D

Epitope identification procedure*	Identified epitopes	References
MIRA using synthetic peptide pool library	176	Snyder et al., 2020
IFN-γ ELISPOT using synthetic peptide pool library	51	Le Bert et al., 2020; Peng et al., 2020
Total	***211** (unique)*	

*MIRA = multiplex identification of T-cell receptor antigen specificity.

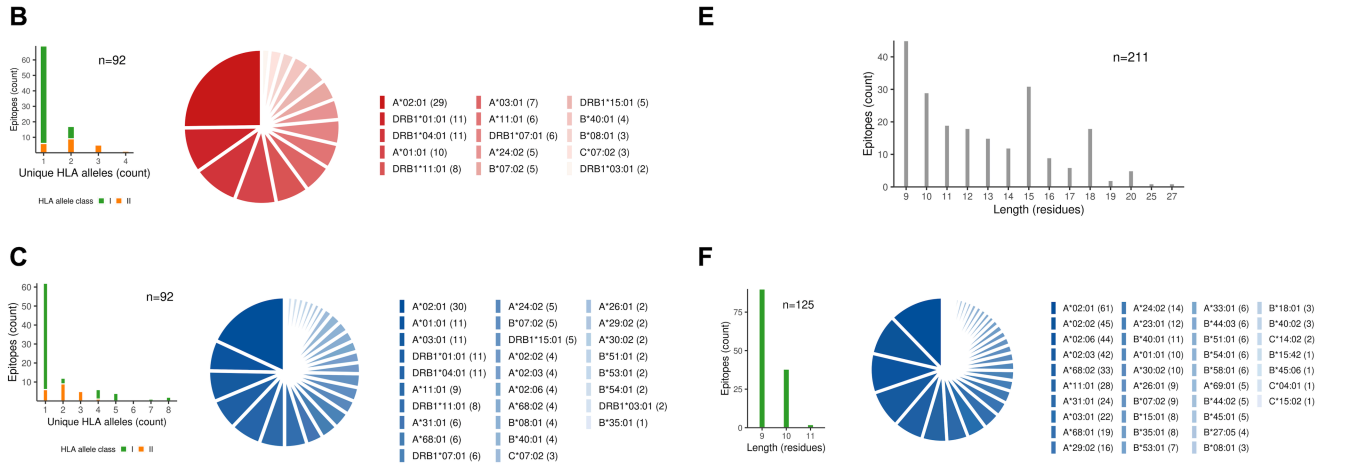


Figure 2. Experimental procedures used for identifying the SARS-CoV-2-specific T cell epitopes, and data augmentation using T cell epitope information for SARS-CoV. (**A**) Summary of studies reporting epitopes (n=92) with experimentally-determined HLA allele information. (**B**) Diversity of HLA associations reported for SARS-CoV-2 epitopes (n=92). The number of epitopes associated with a specific number of unique HLA alleles (left panel) and the number of epitopes associated with a particular HLA allele (right panel). (**C**) Results analogous to (B) after augmenting with the HLA information of genetically-identical SARS-CoV epitopes. (**D**) Summary of studies reporting epitopes (n=211) without experimentally-determined HLA allele information. (**E**) The length of epitopes (n=211) reported without HLA associations. (**F**) Summary statistics of the epitopes (n=125) identified after augmenting with SARS-CoV epitope data, including epitope length (left panel) and diversity of HLA associations (right panel).

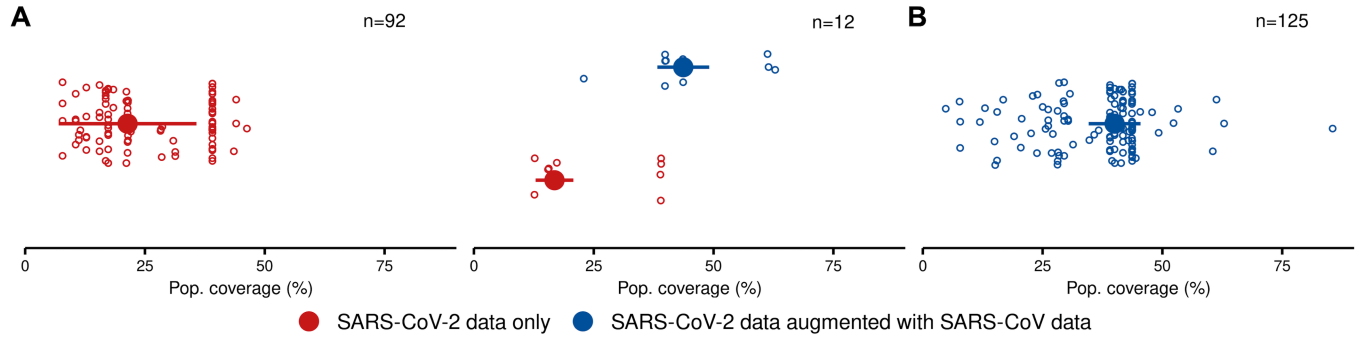


Figure 3. Population coverage of the SARS-CoV-2 T cell epitopes and gain after augmenting with SARS-CoV epitope data. (**A**) Estimated global population coverage of the SARS-CoV-2 epitopes (n=92) that were reported with at least partial HLA allele information (left panel), and the gain in estimated coverage after data augmentation of those epitopes having an identical genetic match to SARS-CoV epitopes (n=12) (right panel). (**B**) Estimated global population coverage of the SARS-CoV-2 epitopes (n=125) that were reported with no HLA allele information, after augmenting with SARS-CoV epitope data.

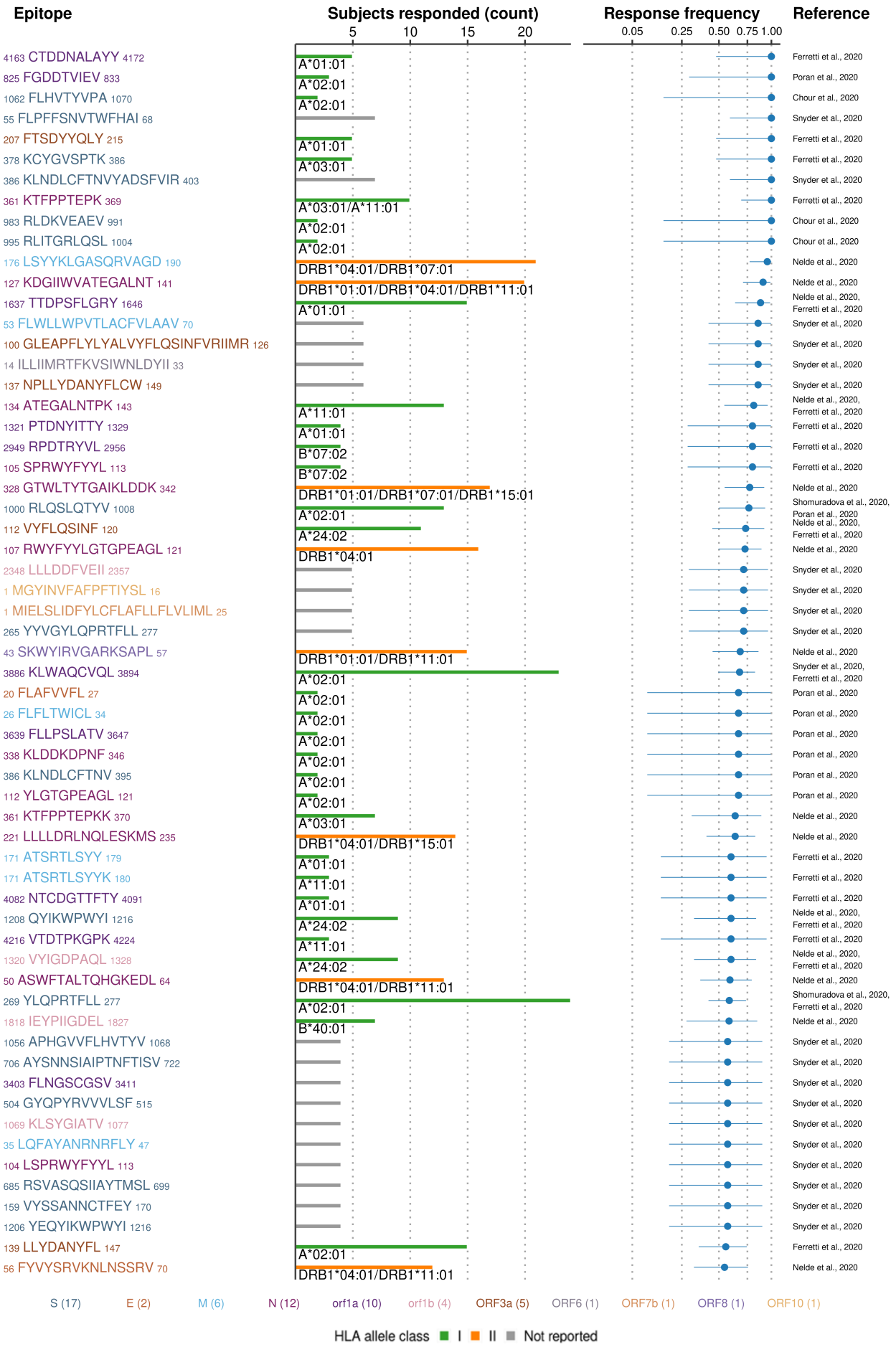
Supplementary Information

Supplementary Tables

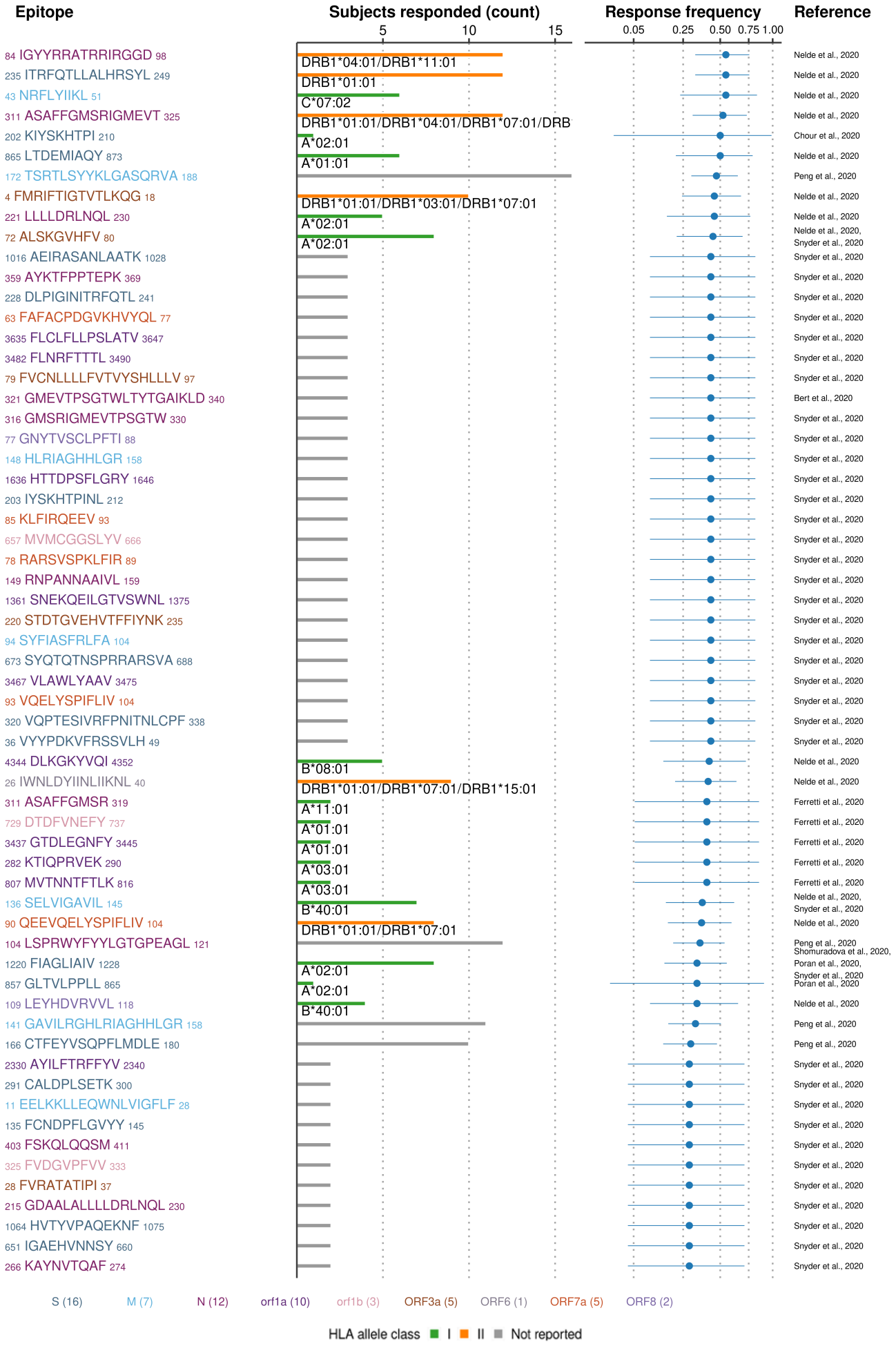
Supplementary Table 1 (included as a separate file)**. Complete list of unique SARS-CoV-2-specific T cell epitopes (n=303) identified experimentally from blood samples of convalescent COVID-19 patients.** Location within the proteome, genetic conservation, associated HLA alleles, number of responding subjects, and response frequency are listed for each epitope. This list of epitopes is compiled from 8 experimental studies, which includes both published studies and non-peer-reviewed studies available so far only on preprint servers. Results from non-peer-reviewed studies should be interpreted with caution.

Supplementary Figures

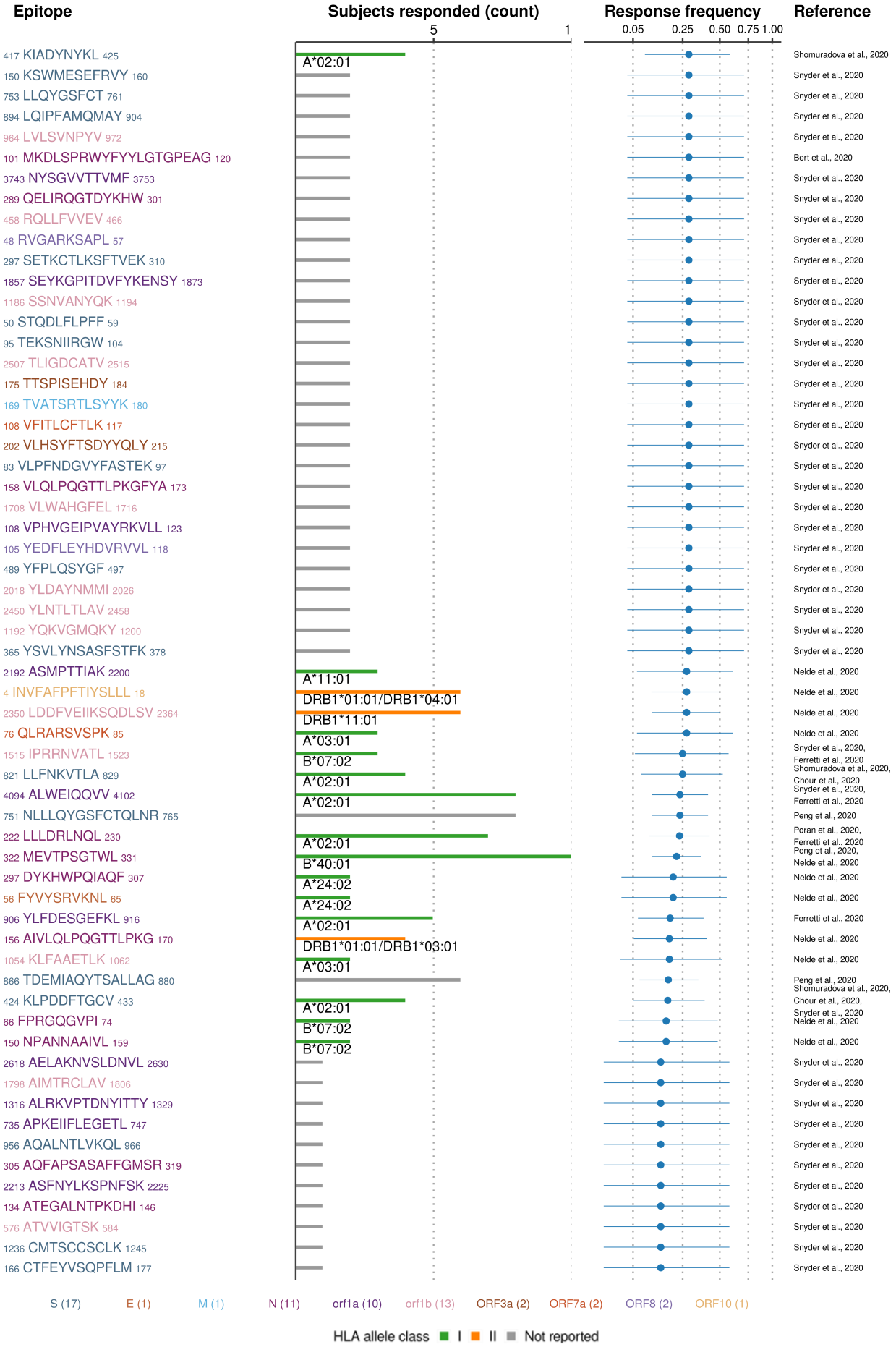
Supplementary Figure 1



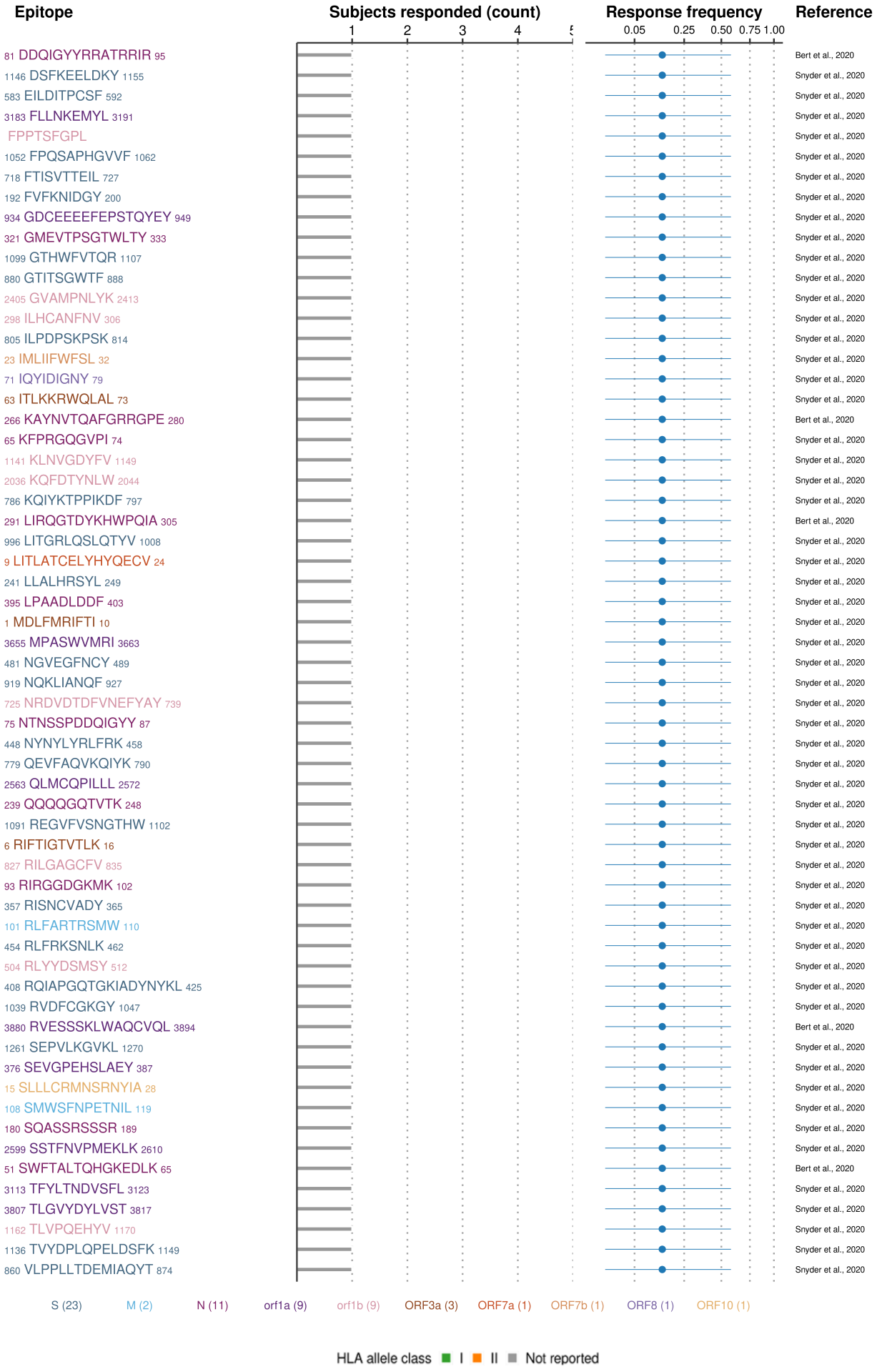
Supplementary Figure 1 (cont'd)



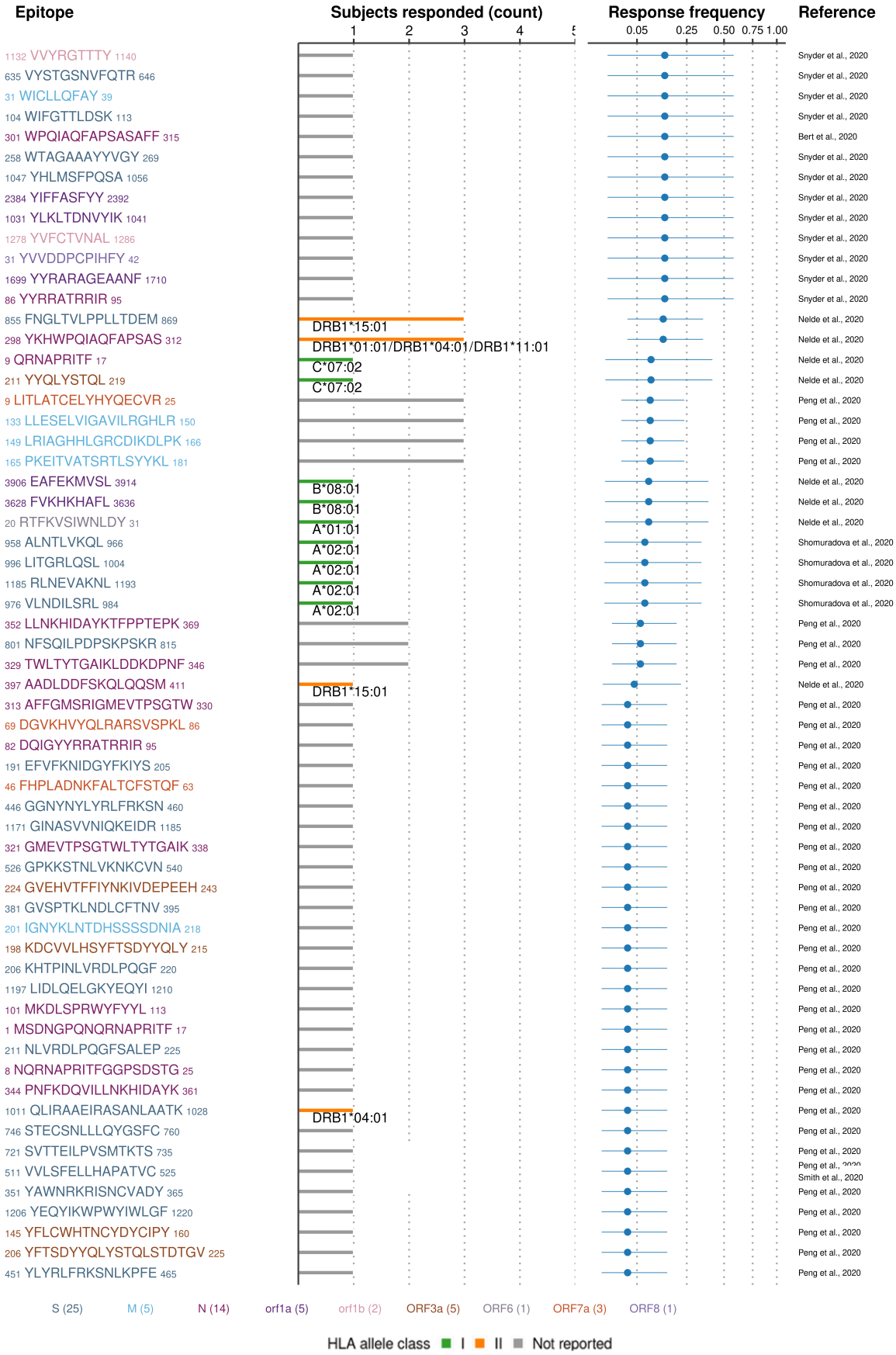
Supplementary Figure 1 (cont'd)



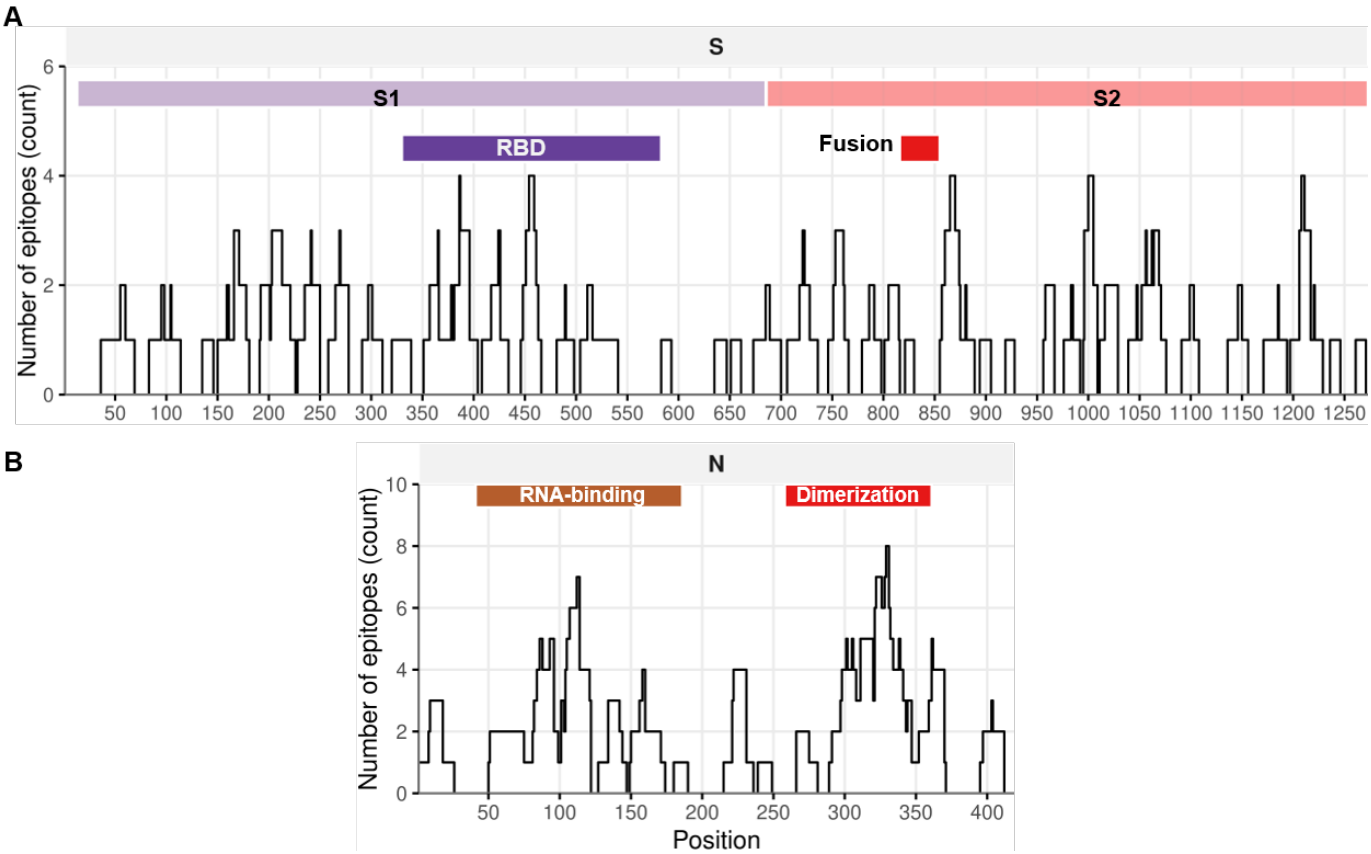
Supplementary Figure 1 (cont'd)



Supplementary Figure 1 (cont'd)



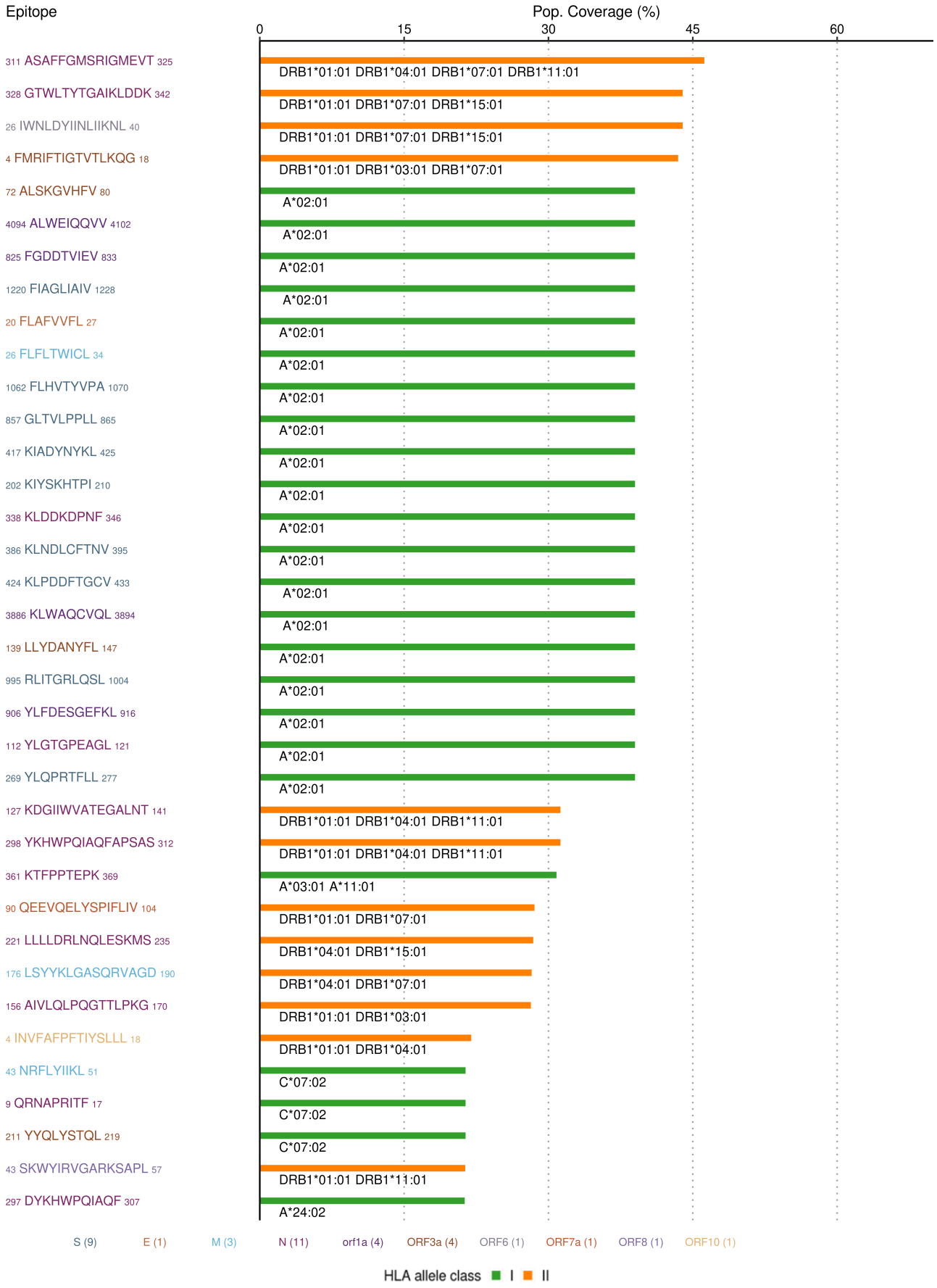
Supplementary Figure 1. Details of the unique SARS-CoV-2-specific T cell epitopes (n=303) reported from studies of convalescent COVID-19 patients. List of epitopes and the corresponding reported number of responding patients, the reported associated HLA alleles (where available), and the computed RF along with the 95% CI (see Methods for details). The studies reporting each epitope are shown on the right. Epitopes are arranged in descending order according to RF and colored according to the protein from which they were derived (legend shown at the bottom). The subscript numbers before and after the epitope indicate its location within the protein. The bars are shaded according to HLA class restriction.



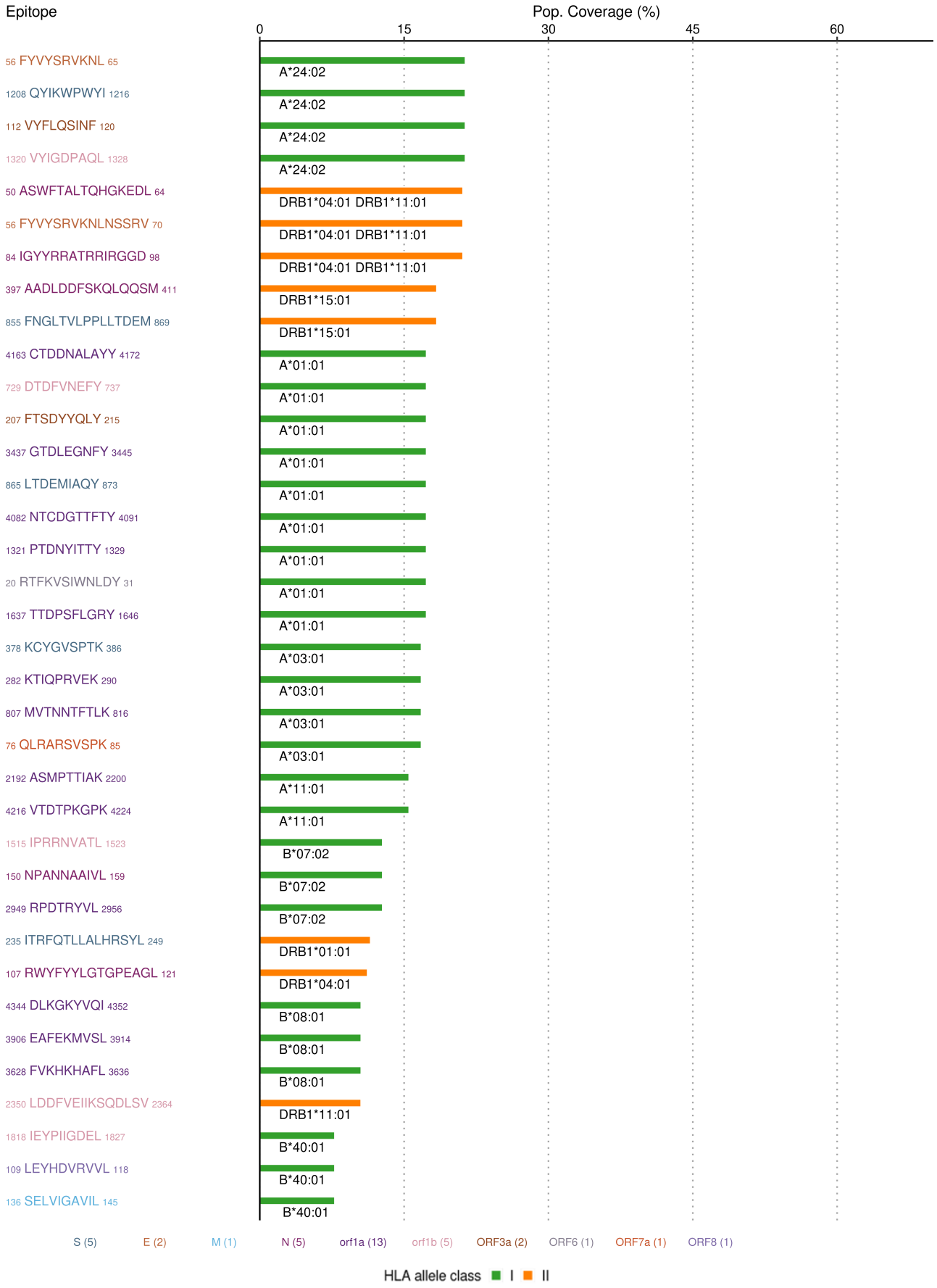
Supplementary Figure 2. Epitope coverage of the spike (S) and nucleocapsid (N) proteins. (**A**) Number of epitopes that cover the S protein are indicated along its primary structure. Receptor-binding domain (RBD) (330-583) within the S1 subunit, and fusion peptide region (816-855) within the S2 subunit are highlighted. (**B**) Number of epitopes that cover the N protein are indicated along its primary structure. RNA-binding domain (41-186), and dimerization region (258-361) are highlighted.

Supplementary Figure 3A

Supplementary Figure 3B

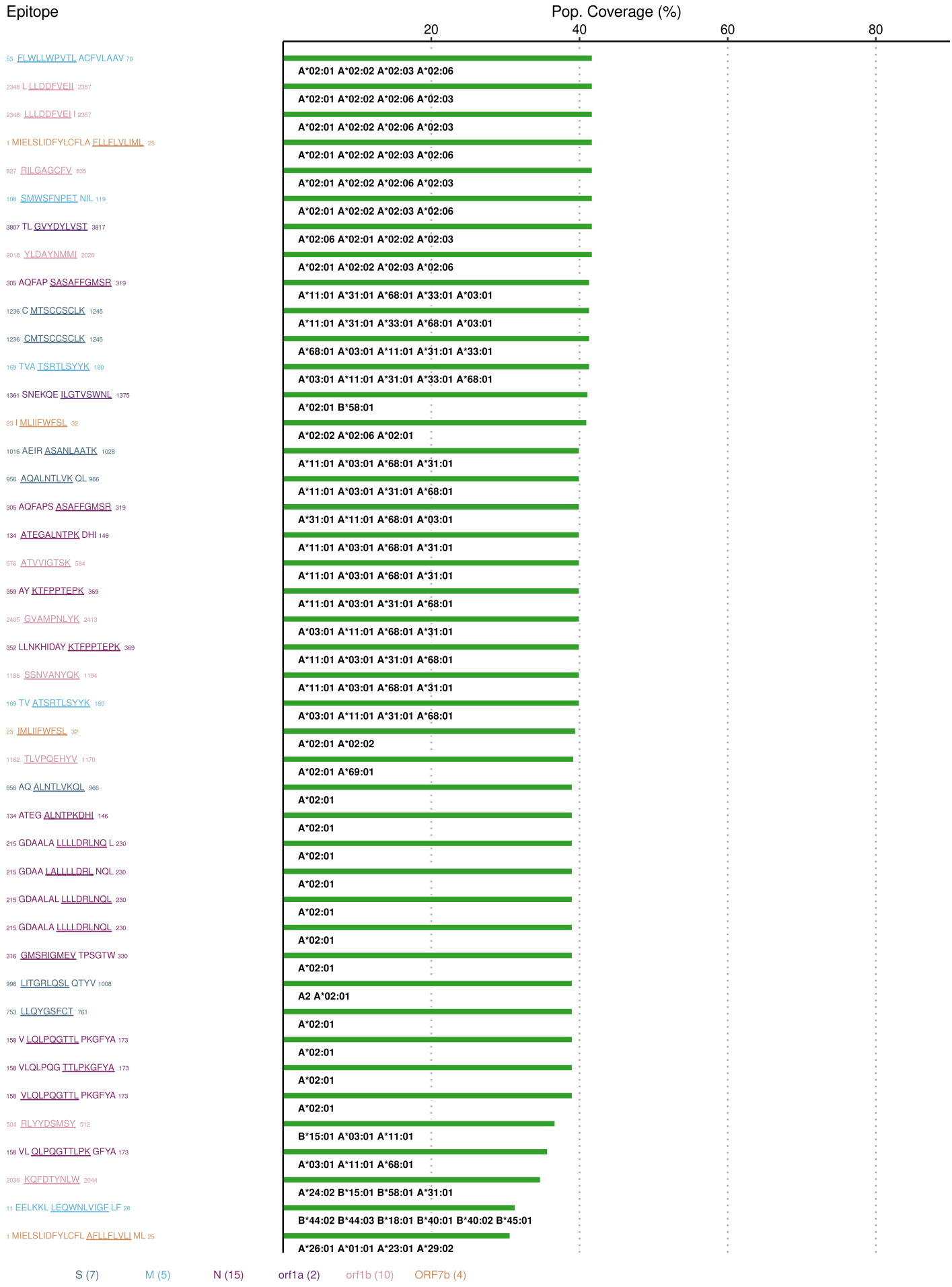


Supplementary Figure 3B (cont'd)



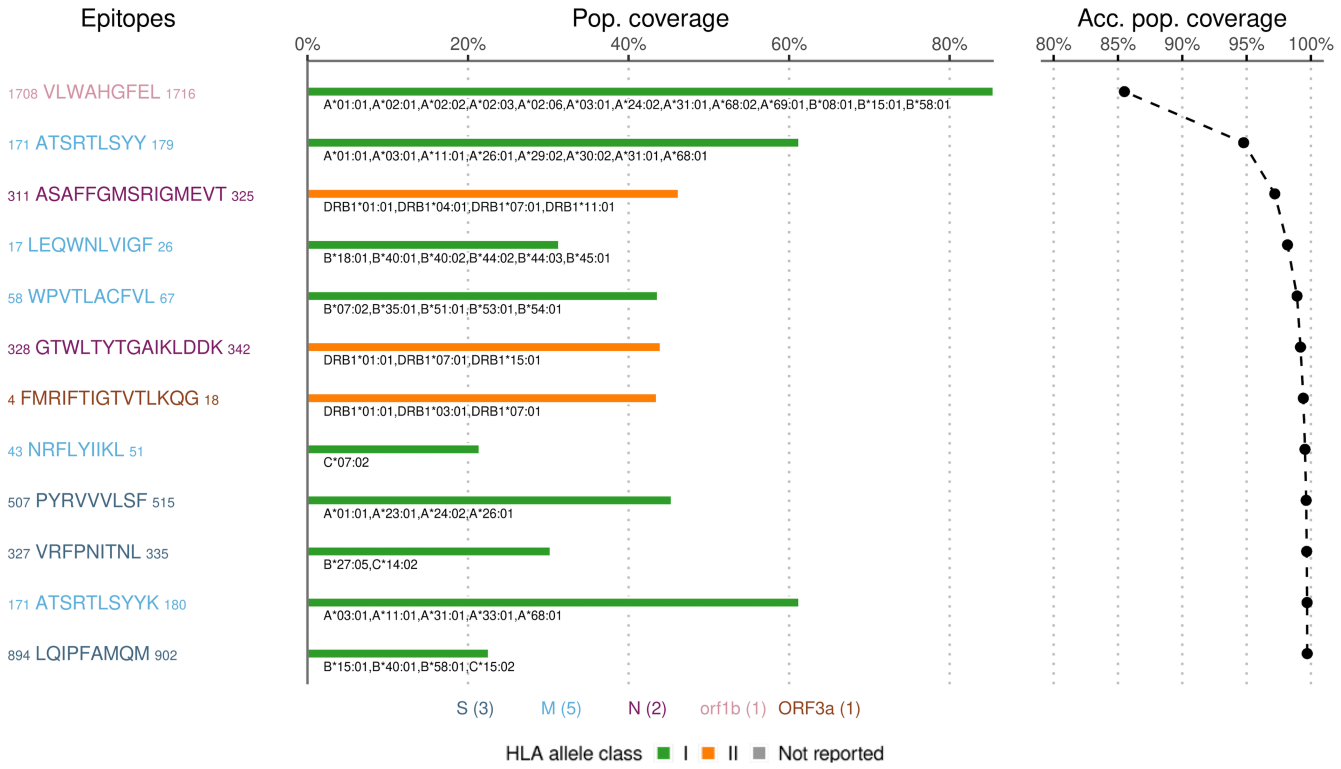
Supplementary Figure 3C

Supplementary Figure 3C (cont'd)



Supplementary Figure 3C (cont'd)

Supplementary Figure 3. Estimated global population coverage of SARS-CoV-2-specific T cell epitopes. (**A**) Estimated global population coverage of the SARS-CoV-2 epitopes that were reported with at least partial HLA allele information and that are genetically identical to known SARS-CoV epitopes (n=20). (**B**) SARS-CoV-2 epitopes that were reported with at least partial HLA allele information and that are not genetically identical to any known SARS-CoV epitope (n=72). (**C**) SARS-CoV-2 epitopes that were reported without HLA allele information and that encompass one or more known SARS-CoV epitopes (n=125). The HLA alleles shown in normal font (A, B) were reported in SARS-CoV-2 studies, while those in bold font (A-C) are obtained from SARS-CoV T cell epitope data. In (C), the known SARS-CoV epitopes encompassed by the SARS-CoV-2 epitopes are underlined. In (A-C), epitopes are arranged in decreasing order of estimated global population coverage, and are colored according to the protein from which they were derived (legend shown at the bottom). The subscript numbers before and after the epitope indicate its location within the protein. The bars indicating estimated population coverage are colored according to HLA class restriction.



Supplementary Figure 4. Combination of epitopes selected to maximize accumulated global population coverage. List of epitopes (left panel) chosen to collectively cover a maximum global population. The global population coverage corresponding to each epitope estimated based on the information of the associated HLA alleles (middle panel), and the accumulated global population coverage as epitopes are progressively added in the group (right panel). The accumulated population coverage for a group of epitopes is defined as the percentage of individuals in the world population that is expected to induce a T cell response against at least one epitope in the group. Given the complete set of SARS-CoV-2 epitopes (except the single orf1b epitope with very low genetic conservation (0.28) among the SARS-CoV-2 sequences), this selected subset of epitopes was chosen to maximize accumulated global population coverage using a greedy procedure. First, the epitope with the highest individual population coverage was selected to initialize the group. Second, of the remaining epitopes, the epitope maximizing the accumulated global population coverage (i.e., the global population coverage of the two epitopes considered together) was added to the group. Third, this procedure of adding epitopes one-by-one was continued until a group of epitopes was obtained for which the accumulated global population coverage saturated to a maximum value (99.71%). Epitopes are colored according to the protein from which they were derived (shown at the bottom), while the number of epitopes within the group derived from each protein is shown within parentheses at the bottom. Location of each epitope within the protein is indicated using subscripts.