The mutational landscape of SARS-CoV-2 variants diversifies T cell targets in an HLA supertype-dependent manner

AUTHORS

David J. Hamelin[1], Dominique Fournelle[2], Jean-Christophe Grenier[2], Jana Schockaert[3], Kevin Kovalchik[1], Peter Kubiniok[1], Fatima Mostefai[2], Jérôme D. Duquette[1], Frederic Saab[1], Isabelle Sirois[1], Martin A. Smith[1,4], Sofie Pattijn[3], Hugo Soudeyns[1,5,6], Hélène Decaluwe[1,6], Julie Hussin[2,4]*, Etienne Caron[1,7,8]*

AFFILIATIONS

[1]CHU Sainte-Justine Research Center, Montreal, QC, Canada

[2]Montreal Heart Institute, Department of Medicine, Université de Montréal, Montréal, QC, Canada

[3]ImmunXperts, a Nexelis Group Company, 6041 Gosselies, Belgium

[4]Department of Biochemistry and Molecular Medicine, Faculty of Medicine, Université de Montréal, QC, Canada

[5]Department of Microbiology, Infectiology and Immunology, Faculty of Medicine, Université de Montréal, Montréal, QC, Canada

[6]Department of Pediatrics, Faculty of Medicine, Université de Montréal, Montréal, QC, Canada

[7]Department of Pathology and Cellular Biology, Faculty of Medicine, Université de Montréal, Montreal, QC, Canada

[8]Lead Contact

*Corresponding author: Julie Hussin (julie.hussin@umontreal.ca) and Etienne Caron (etienne.caron@umontreal.ca)

SUMMARY

The rapid, global dispersion of SARS-CoV-2 since its initial identification in December 2019 has led to the emergence of a diverse range of variants. The initial concerns regarding the virus were quickly compounded with concerns relating to the impact of its mutated forms on viral infectivity, pathogenicity and immunogenicity. To address the latter, we seek to understand how the mutational landscape of SARS-CoV-2 has shaped HLA-restricted T cell immunity at the population level during the first year of the pandemic, before mass vaccination. We analyzed a total of 330,246 high quality SARS-CoV-2 genome assemblies sampled across 143 countries and all major continents. Strikingly, we found that specific mutational patterns in SARS-CoV-2 diversify T cell epitopes in an HLA supertype-dependent manner. In fact, we observed that proline residues are preferentially removed from the proteome of prevalent mutants, leading to a predicted global loss of SARS-CoV-2 T cell epitopes in individuals expressing HLA-B alleles of the B7 supertype family. In addition, we show that this predicted global loss of epitopes is largely driven by a dominant C-to-U mutation type at the RNA level. These results indicate that B7 supertype-associated epitopes, including the most immunodominant ones, were more likely to escape CD8+ T cell immunosurveillance during the first year of the pandemic. Together, our study lays the foundation to help understand how SARS-CoV-2 mutants shape the repertoire of T cell targets and T cell immunity across human populations. The proposed theoretical framework has implications in viral evolution, disease severity, vaccine resistance and herd immunity.

INTRODUCTION

As of May 2021, the COVID-19 pandemic, caused by the novel Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2), has led to upwards 3.4 million deaths and 165 million confirmed cases worldwide (https://coronavirus.jhu.edu/map.html), making vaccine development and deployment an urgent necessity (Callaway, 2020). As a result of unprecedent efforts, vaccines have been developed and licensed within a 1-year timeframe and are currently being widely distributed for mass vaccination (Krammer, 2020).

A clear understanding of the natural protective immune response against SARS-CoV-2 is essential for the development of vaccines that can trigger lifelong immunologic memory to prevent COVID-19 (Sette and Crotty, 2021; Stephens and McElrath, 2020). Since the start of the pandemic, numerous studies have investigated the association between COVID-19 clinical outcomes and SARS-CoV-2 specific antibodies and T cell immunity (Altmann and Boyton, 2020; Bert et al., 2020; Braun et al., 2020; Grifoni et al., 2020a; Long et al., 2020a, 2020b; Meckiff et al., 2020; Moderbacher et al., 2020; Sekine et al., 2020; Weiskopf et al., 2020). Memory may be a concern for SARS-CoV-2 specific antibodies, as they were recently shown to be present in convalescent COVID-19 patients in a highly heterogenous manner (Dan et al., 2021) and, in some cases, observed to be undetectable just a few months post-infection (Seow et al., 2020). In contrast, an increasing number of studies point CD4+ and CD8+ T cells as key regulators of disease severity (Liao et al., 2020; Moderbacher et al., 2020; Schub et al., 2020; Weiskopf et al., 2020; Zhou et al., 2020). Studies of convalescent COVID-19 patients have also shown broad and strong CD4+ and CD8+ memory T cells induced by SARS-COV-2, suggesting that T cells may provide robust and long-term protection (Dan et al., 2021; Peng et al., 2020). Similar observations have been made for the most closely related human coronavirus, SARS-CoV, for which T cells have been detected

11 years (Ng et al., 2016) and 17 years (Bert et al., 2020) after the initial infection, whereas antibodies were noted to be undetectable after 2-3 years (Liu et al., 2006; Tang et al., 2011; Wu et al., 2007). Thus, vaccines designed to produce robust T cell responses are likely to be important for eliciting lifelong immunity against COVID-19 in the general population.

To investigate how T cells could contribute to long-term vaccine effectiveness, precise knowledge about SARS-CoV-2 T cell-specific epitopes is of paramount importance (Liu et al., 2020). To this end, bioinformatics tools were developed to predict T cell-specific epitopes during the early phase of the pandemic (Grifoni et al., 2020b). A comprehensive map of epitopes recognized by CD4+ and CD8+ T cell responses across the entire SARS-CoV-2 viral proteome was also recently reported (Tarke et al., 2020). Notably, the structural proteins Spike (S), Nucleocapsid (N) and Membrane (M) were shown to be rich sources of immunodominant HLA-associated epitopes, accounting for a large proportion of the total CD4+ and CD8+ T cell response in the context of a broad set of HLA alleles (Tarke et al., 2021). To date (May 2021), ~700 HLA class I-restricted SARS-CoV-2-derived epitopes have been experimentally validated (https://www.mckayspcb.com/SARS2TcellEpitopes/) (Quadeer et al., 2020).

T cell epitopes that have been mapped across the entire SARS-CoV-2 viral proteome are reference peptides that are unmutated because they have been predicted from the sequence of the original SARS-CoV-2 that emerged from Wuhan, China (Grifoni et al., 2020b). However, analyses of unprecedented numbers of SARS-CoV-2 genome assemblies available from large-scale efforts have shown that SARS-CoV-2 is accumulating an array of mutations across the world, leading to the circulation and transmission of thousands of variants around the globe at various frequencies, and hence, contributing to the global genomic diversification of SARS-CoV-2 (Dorp et al., 2020a; Korber et al., 2020; Laamarti et al., 2020; Mercatelli and Giorgi, 2020; Mercatelli et al., 2020;

Popa et al., 2020). In this regard, recent data indicate that most recurrent mutations appear to be evolutionary neutral with no evidence for increased transmissibility (Dorp et al., 2020a). Nonetheless, it is important to highlight that those neutral mutations are associated with a remarkably high proportion of cytidine-to-uridine (C-to-U) changes that were hypothesized to be induced by members of the APOBEC RNA-editing enzyme family (Dorp et al., 2020a; Giorgio et al., 2020; Klimczak et al., 2020; Kosuge et al., 2020; Li et al., 2020; Matyášek and Kovařík, 2020; Rice et al., 2020; Simmonds, 2020; Wang et al., 2020). Since shown for other viruses (Grant and Larijani, 2017; Monajemi et al., 2014), we reasoned that the putative action of such host enzymes during the first year of the pandemic could lead to the large-scale escape from immunodominant and protective SARS-CoV-2-specific T cell responses, thereby potentially compromising their effectiveness to control the virus at the population-scale.

In this study, we report a comprehensive study of the global genetic diversity of SARS-CoV-2 to expose the impact of mutation bias on epitope presentation and HLA-restricted T cell response within the first year of the pandemic, from December 2019 to December 2020. More specifically, we asked the following questions: 1) What are the impact of SARS-CoV-2 prevalent mutations detected across the global human population on the repertoire of validated SARS-CoV-2 T cell targets, with specific emphasis on CD8+ T cell epitopes? and 2) Are mutational patterns in the genomic and proteomic composition of SARS-CoV-2 indicative of disrupted (or enhanced) epitope presentation and T cell immunity in human populations? By answering these questions, we provide a theoretical framework to understand how SARS-CoV-2 mutants have shaped T cell immunity to evade effective T cell immune responses at the population level during the first year of the pandemic, i.e. without mass vaccination-induced immune pressure on viral evolution and adaptation.

RESULTS

The global diversity of SARS-CoV-2 genomes influences the repertoire of T cell targets

As of May 2021, nearly 1.7M complete SARS-CoV-2 genome assemblies are publicly available via the Global Initiative on Sharing All Influenza Data (GISAID) repository. In the context of this large-scale effort, we performed a global analysis of SARS-CoV-2 genomes to assess whether mutations that emerged during the first year of the pandemic could disrupt HLA binding of clinically relevant SARS-CoV2 CD8+ T cell epitopes. First, we identified missense mutations by aligning 330,246 high-quality consensus SARS-CoV-2 genomic sequences (GISAID; December 31st 2020, prior to mass vaccination) to the reference sequence, Wuhan-1 SARS-CoV-2 genome (**Figure S1**). We found a total of 13,780 mutations identified in at least 4 SARS-CoV-2 genomes/individuals from GISAID, including 1,721 unique amino acid mutations in the S protein, with D614G as the most frequent one (94%) (Korber et al., 2020) (**Table S1** and **Table S2**). Next, we implemented a bioinformatics pipeline to assess the impact of these mutations on HLA binding for 620 unique SARS-CoV-2 HLA class I epitopes that were recently reported to trigger a CD8+ T cell response in acute or convalescent COVID-19 patients (Quadeer et al., 2020; Tarke et al., 2020) (see Methods). On average, we found that the predicted binding affinity of 181 of these SARS-CoV-2 epitopes (30%) for common HLA-I alleles was reduced by ~100-fold (**Table S3** and **Figure S1**). It is also apparent that mutations negatively impacted the HLA binding affinity of 56 (31%) and 19 (10%) CD8+ T cell epitopes located in the immunodominant S and N proteins, respectively (**Figure 1A,B)**. Notably, a gap in the N protein, composed of a serine-rich region, is associated with higher mutation rate and a marked lack of predicted T cell epitopes and response

(**Figure 1B**). Epitopes located in the RBD vaccine locus were also impacted by mutations (**Figure 1C**).

Loss of epitope binding for commonly expressed HLA class I molecules was validated *in vitro* for a subset of representative SARS-CoV-2 epitopes (**Figure S2**). Of relevance, we found that the common D614G mutation in the S protein is linked to a 15-fold decrease in the binding affinity for the mutated HLA-A*02:01 epitope YQGVNCTEV when compared to the reference/unmutated epitope YQDVNCTEV (**Figure S2A,B**). Interestingly, our analysis also identified a mutation in the HLA-B*07:02-restricted N105 epitope SPRWYFYYL, which is one of the most immunodominant SARS-CoV-2 epitope (Ferretti et al., 2020; Kared et al., 2021; Saini et al., 2021; Schulien et al., 2021; Sekine et al., 2020; Tarke et al., 2021). Although relatively rare (found in only two genomes), the mutation in the N105 epitope consists of P→S at anchor residue position P2 (P106S: SPRWYFYYL → SSRWYFYYL) (**Figure 1B**) and is predicted to decrease HLA epitope binding by 47-fold (**Figure 3D**), thereby likely reducing the breadth of the immune response in B*07:02 individuals carrying this mutation. Moreover, our global analysis validated the presence of two previously reported CD8+ T cell mutated epitopes (i.e. GLMWLSYFI → GFMWLSYFI, found in 38 genomes; and MEVTPSGTWL → MKVTPSGTWL, found in 23 genomes), which were shown to lose binding to HLA-A*02:01 and -B*40:01, respectively, in addition to disrupt epitope-specific CD8+ T cell response in COVID-19 patients (**Figure S3**) (Agerer et al., 2021). Together, these results demonstrate that mutations driving the global genomic diversity of SARS-CoV-2 can drastically disrupt HLA binding of clinically relevant CD8+ T cell epitopes, including epitopes encoded by the immunodominant S and N antigens, therefore affecting epitope-specific T cell responses in COVID-19 patients.

In addition to mutations leading to a loss of HLA epitope binding, we identified a significant number of mutations predicted to enhance the presentation of peptides by their respective HLA molecules, leading to a 'Gain' of binding (**Figure S4**). Because the unmutated epitopes are predicted to be non-HLA binders, these mutations were not searched against the list of known validated epitopes, which consist of strong-HLA binding reference epitopes. Whether SARS-CoV-2 mutations predicted to increase HLA epitope binding can enhance T cell responses to control the virus in COVID-19 patients remains to be determined experimentally.

Amino acid mutational biases shape the global diversity of SARS-CoV-2 proteomes

While analysing the impact of the mutational landscape of SARS-CoV-2 on validated CD8+ T-cell epitopes, we observed that specific mutation types were over-represented while others were under-represented (**Figure S2C,D**). For instance, we found that 31% of the mutated epitopes were represented by a removal of proline residue (**Figure S2C,D**), leading to the hypothesis that such biases could originate from biases in the proteome of SARS-CoV-2 mutants. To further investigate whether specific amino acid mutational biases could be observed globally in the proteome of SARS-CoV-2 mutants, we asked whether certain amino acid residues were preferentially removed from, or introduced into the global proteomic diversity of SARS-CoV-2, thereby potentially diversifying CD8+ T cell epitopes in a systematic manner.

To test this, we computed all residue substitutions (amino acid removed and introduced) found in SARS-CoV-2 proteomes and calculated Global Residue Substitution Output (GRSO) values, i.e. the % difference in overall amino acid composition for individual amino acids (see Methods for details). GRSO values were computed for mutations found at various frequencies in GISAID (i.e. found in only 1 genome, 2 to 100 genomes, 100 to 1000 genomes and > 1000

genomes) (**Figure 2**). Interestingly, distinct mutational patterns at the amino acid level were observed amongst mutations detected in more than 100 genomes/individuals (**Figure 2),** referred in this study to as 'prevalent mutations' (see Methods and **Table S2**). Amongst those mutations, the amino acids alanine (A), proline (P) and threonine (T) were preferentially removed by 10.2% ($p = 1.2x10^{-13}$), 9.1% ($p = 1.6x10^{-15}$), and 10.5% ($p = 1.3x10^{-14}$), respectively. In contrast, phenylalanine (F), isoleucine (I), leucine (L) and tyrosine (Y) were preferentially introduced by 13.4% ($p = 2.0x10^{-17}$), 15.2% ($p = 2.4x10^{-17}$), 4.3% ($p = 6.3x10^{-11}$) and 5.0% ($p = 7.0x10^{-14}$), respectively (**Figure 2**). Statistical significance of these GRSO values was assessed by generating simulated samples of 1000 SARS-CoV-2 genomes evolving under neutrality (N = 10 replicates) using the SANTA-SIM algorithm (Jariani et al., 2019) (see Methods for details). Of note, mutations that were detected in 2 to 100 individuals appeared significantly more neutral, with none of the mutational patterns enriched above the selected cut-off values (fold change > 4; p-value < $1x10^{-11}$). Thus, our results show that specific amino acid residues were preferentially removed or introduced in the proteome of SARS-CoV-2 mainly by prevalent mutations. Therefore, we introduce the notion that the global diversity of SARS-CoV-2 proteomes is shaped by specific amino acid mutational biases. Such biased amino acid composition generated by prevalent mutations may have a systematic impact on epitope processing and presentation to shape SARS-CoV-2 T cell immunity in human populations. To address this systematic impact, all downstream analyses described in this study were performed from the set of 1,933 prevalent mutations (>100 genomes) listed in **Table S2**.

Prominent removal of proline residues leads to a predicted global loss of epitopes presented by HLA-B7 supertype molecules

The association of peptides with the binding groove of HLA molecules largely relies on the presence of anchor residues, also known as peptide binding motifs (Falk et al., 1991). Hundreds of different peptide binding motifs have been reported over the last decades (Gfeller and Bassani-Sternberg, 2018). Overlapping binding motifs are qualified as "HLA supertypes" on the basis of their main anchor specificity (Greenbaum et al., 2011; Sidney et al., 2008). Of relevance here, proline acts as a critical anchor residue at position P2 for epitopes presented by HLA-B7 (B7) supertype molecules, which include a wide range of commonly expressed HLA-B alleles in humans, i.e. HLA-B*07, -B*15, -B*35, -B*42, -B*51, -B*53, -B*54, -B*55, -B*56, -B*67 and B*78 (Sidney et al., 2008). In fact, the B7 supertype covers ~35% of the human population (Francisco et al., 2015). Hence, we reasoned that the global removal of proline residues observed in the proteome of prevalent SARS-CoV-2 mutants (**Figure 2**) could drastically compromise T cell epitope binding to B7 supertype molecules, thereby potentially interfering with SARS-CoV-2 T cell immunity in a relatively large proportion of the human population.

Due to the preferential removal of proline by prevalent mutations, we investigated the extent at which proline residues were substituted at anchor binding position P2 and, consequently, resulted in loss of epitopes presented by B7 supertype molecules. To answer this, we performed the following four steps: (i) We applied NetMHCpan 4.1 (Reynisson et al., 2020) using the reference and mutated SARS-CoV-2 genomes to generate a list of all possible reference/mutated peptide pairs (8-11 mers) predicted to bind 16 common HLA-B types that belong to the B7 supertype family (**Figure S5B**). (ii) We analyzed all reference/mutated peptide pairs, along with their differential predicted binding affinities to quantitatively identify HLA strong binder (SB) to non-binder (NB) transitions [(SB) NetMHCpan %rank < 0.5 to (NB) NetMHCpan %rank >2]. (iii) We categorized all peptide pairs based on the mutation type (amino acid X → amino acid Y) and

the position of the mutation within the peptide sequence. (iv) Lastly, we quantified the number of reference/mutated peptide pairs and the associated fold-change in predicted binding affinity for each category. Our results show that prevalent mutations predicted to impact the presentation of peptides by the B7 supertype are dominated by P→L ($p = 8.6x10^{-35}$) and P→S ($p = 3.4x10^{-24}$) substitutions at anchor residue position P2 (**Figure 3A,B**). Reference/mutated peptide pairs from these categories were the most abundant, with > 250 mutated peptides per category (**Figure 3C**). P→L and P→S mutations resulted, on average, in a 61-fold reduction in predicted HLA binding affinity for a representative set of clinically validated CD8+ T cell epitopes (**Figure 3D**).

In addition to the dominant P→S/L substitution type, other P→X substitutions were observed. Interestingly, analysis of mutations found in the Pangolin B.1.1.7 variant (January 2021) showed that the P681H mutation found in the Spike protein led to disrupted association of the reference epitope SPRRARSVA for several HLA-B7 types. In fact, the P-to-H substitution resulted in a strong loss of epitope binding predicted for 7/16 HLA-B types tested. Thus, our results strongly suggest that biased substitutions of proline residues in the proteome of SARS-CoV-2 shapes the repertoire of epitopes presented by B7 supertype, including epitopes encoded by the genome of the B.1.1.7 variant. This finding let us to propose that mutation biases found in SARS-CoV-2 may contribute to CD8+ T cell epitope escape in a B7 supertype-dependent manner.

The mutational landscape of SARS-CoV-2 enables disruption or enhancement of epitope presentation in an HLA supertype-dependent manner

We found that specific amino acid residues were preferentially removed (proline, alanine and threonine) or introduced (isoleucine, phenylalanine, leucine and tyrosine) in SARS-CoV-2 proteomes (**Figure 2**). Importantly, most of these amino acids act as key epitope anchor residues

for multiple HLA class I supertypes (**Figure S5**). For instance, phenylalanine and tyrosine are key anchor residues for all known A*24 alleles of the A24 supertype family, whereas proline is known to play a critical role in the anchoring of epitopes to alleles of the B7 supertype family (**Figure 4**). Therefore, one would expect the introduction of phenylalanine and tyrosine in SARS-CoV-2 proteomes to facilitate peptide presentation by A24, whereas the removal of proline would disrupt peptide presentation by B7. With this concept in mind, we hypothesized that the distinct amino acid mutational biases found throughout prevalent SARS-CoV-2 mutations could systematically mold epitope presentation in an HLA supertype-dependent manner.

In order to compare supertypes to each other, we generated a 'Gain/Loss plot' for each supertype assessed (**Figure 4C**). Gain/Loss plot were generated by computing the number of mutations that resulted in 'Gain' or 'Loss' of epitopes for representative class I alleles selected for each supertype (see methods for details). 'Gain' was assigned for mutated epitopes that were predicted to transit from non-HLA binders (NetMHCpan %rank > 2) to strong HLA binders (NetMHCpan %rank < 0.5), whereas 'Loss' was assigned for mutated epitopes that were predicted to transit from strong HLA binders to non-HLA binders. Surprisingly, our analysis shows that most supertypes preferentially gain new epitopes as a result of SARS-CoV-2 mutations: A1 ($p = 4.5x10^{-11}$), A2 ($p = 0.001$), A24 ($p = 1.0x10^{-26}$), B8 ($p = 2.4x10^{-14}$), B27 ($p = 2.5x10^{-6}$). Interestingly, preferential loss of epitopes was only shown to be statistically significant for B7 supertype ($p = 0.0012$). Note that we explain the relatively low statistical value obtained for B7 supertype by the presence of isoleucine and phenylalanine (preferentially introduced in SARS-CoV-2 proteomes; see Figure 2) at anchor residue P9 for certain HLA types (namely HLAB*51:01 and HLA-B*53:01) (**Figure 4A**). In fact, omitting motifs containing isoleucine or phenylalanine increased the significance of epitope lost *versus* gained ($p = 2.6x10^{-7}$) (**Figure 4C**). Together, our

283 results show that the amino acid mutational biases that feature the global diversity of SARS-CoV-
284 2 proteomes can positively or negatively affect binding affinities of mutated epitopes for a wide
285 range of HLA class I molecules in a supertype-dependent manner.
286
287 **The C-to-U point mutation bias largely drives diversification of SARS-CoV-2 T cell epitopes**
288 Next, we sought to better understand the genetic determinants that drive the association between
289 epitope presentation and the amino acid mutational biases found in the SARS-CoV-2 population.
290 To this end, we analyzed the abundance of all the possible nucleotide mutation types (i.e. A-to-C,
291 A-to-G, A-to-U, C-to-A, C-to-G, C-to-U, etc.). This analysis indicates that C-to-U is the most
292 common mutation type (43%), followed by G-to-U (28%), as well as A-to-G, G-to-A and U-to-C
293 (from 9.7% to 11.6%) (**Figure S6A**), in line with observations made by others (Giorgio et al.,
294 2020; Klimczak et al., 2020; Kosuge et al., 2020; Li et al., 2020; Matyášek and Kovařík, 2020;
295 Rice et al., 2020; Simmonds, 2020; Wang et al., 2020).
296 Next, we aimed to determine the contribution of these different nucleic acid mutation types
297 to the global mutational pattern observed at the amino acid level in Figure 2. To do so, we
298 generated simulated population samples of 1000 SARS-CoV-2 genomes using SANTA-SIM
299 (Jariani et al., 2019), applying various extents of mutational biases corresponding to the two most
300 common mutation types observed (i.e. C-to-U and G-to-U). The resulting simulated viral
301 populations were then analyzed to elucidate the global amino acid mutational pattern engendered
302 by these simulated nucleic acid point mutation biases, and whether they recapitulate the observed
303 patterns. Indeed, our data show that the mutational pattern resulting from the simulated C-to-U
304 bias very closely mimicked the mutational pattern observed in the real-life dataset (**Figure 5A**).
305 Namely, the *in silico* introduction of a C-to-U mutation bias resulted in the preferential removal

of alanine, proline, and threonine, by 6.7% ($p = 5.1x10^{-11}$), 6.9% ($p = 1.2x10^{-11}$) and 8% ($p = 4.8x10^{-12}$), respectively, as well as the introduction of isoleucine and phenylalanine by 8.2% ($p = 1.3x10^{-8}$) and 5.2% ($p = 4.3x10^{-11}$), respectively (**Figure 5A**). The G-to-U mutation bias also contributed to the introduction of isoleucine and phenylalanine (**Figure S6**). Together, these results show that the predominant C-to-U point mutations largely contribute to shaping the global proteomic diversity of SARS-CoV-2.

Given the significant impact of the C-to-U point mutation bias on the amino acid content of SARS-CoV-2 proteomes, we reasoned that C-to-U could be the main driver shaping the repertoire and diversification of SARS-CoV-2 T cell targets in human populations, including targets presented by the particularly interesting B7 supertype molecules. To investigate this, we used all the SARS-CoV-2 CD8+ T cell epitopes that were experimentally validated using peripheral blood mononuclear cells (PBMC) of acute and convalescent COVID-19 patients (Quadeer et al., 2020; Tarke et al., 2020) and matched them with their corresponding nucleic acid sequence found in reference/mutated genome pairs. We then calculated the frequency of the various mutation types (i.e. A-to-C, A-to-G, A-to-U, C-to-A, C-to-G, C-to-U, etc.) coding for the mutated form of those clinically validated CD8+ T cell epitopes. Importantly, we found that C-to-U and G-to-U were the two main mutation types leading to mutated epitopes, both accounting for 37% of all mutation types amongst prevalent mutations (>100 individuals) (**Figure 5B**). Most strikingly, 62% of the prevalent mutations predicted to disrupt the presentation of epitopes by HLA alleles for the B7 supertype were found to derive from the C-to-U mutation type (**Figure 5B**). These results strongly suggest that the dominant C-to-U point mutation bias found amongst prevalent SARS-CoV-2 mutants has the potential to significantly contribute to shaping the repertoire of SARS-CoV-2 T cell epitopes in B7 supertype individuals across human populations.

Collectively, our study lets us to propose the model that C-to-U editing enzymes play a fundamental role in shaping the mutational landscape dynamics of SARS-CoV-2 CD8+ T cell targets in humans (**Figure 5C**), and hence, may contribute to molding T cell immunity against COVID-19 at the population level.

DISCUSSION

Mutations contribute to the genetic diversity of SARS-CoV-2 and shape the progression of the COVID-19 pandemic (Dorp et al., 2020b, 2020a; Popa et al., 2020). T cells are key players controlling COVID-19 disease severity. Therefore, determining whether and how the mutational landscape of SARS-CoV-2 shapes or is shaped by HLA-restricted T cell response is fundamentally important. Traditionally, most studies have investigated how viral mutations are shaped by T cell response in the context of HLA-typed cohort patients. This type of approach sought to determine the evolutionary relationship between HLA genotypes and variants of long-standing viruses such as HIV-1 (Brumme et al., 2007; Kawashima et al., 2009) and influenza (Woolthuis et al., 2016). In the case of novel virus such as SARS-CoV-2, such a relationship remains to be established and does not constitute the scope of our work. Here, we rationalized that an alternative approach to interrogating SARS-CoV-2 epitope-associated variants is by investigating the global genomic and proteomic diversity of SARS-CoV-2 for any outstanding mutational biases, and then, assessing the relationship between such biases and epitope presentation for a broad set of HLA alleles. In other words, in this study, we did not seek to understand how viral mutations are shaped by T cell immunity, but rather to understand how mutational biases in SARS-CoV-2 may have shaped T cell immunity at the population level during the first year of the pandemic. This approach was possible thanks to an unprecedented number of SARS-CoV-2 genome sequences available for

352 downstream analysis. Our approach is universal and could be applied to other epidemic or
353 pandemic viruses in the future, given the development of distinct, prevalent mutational biases.
354 Importantly, our global approach has led to several striking conclusions to help understand how
355 the increasing genomic diversity of SARS-CoV-2 may shape T cell immunity in human
356 populations. Our findings have important implications that are discussed below in the context of
357 disease severity, viral evolution and vaccine resistance.

358 In this study, we found that prevalent SARS-CoV-2 mutations are governed by defined
359 mutational patterns, with C-to-U being a predominant mutation type, as previously shown by
360 others (Giorgio et al., 2020; Klimczak et al., 2020; Kosuge et al., 2020; Li et al., 2020; Matyášek
361 and Kovařík, 2020; Rice et al., 2020; Simmonds, 2020; Wang et al., 2020). In fact, we show that
362 the C-to-U mutation bias in SARS-CoV-2 genomes has a remarkably intimate relationship with
363 the observed amino acid mutational biases, indicating that C-to-U mutations largely contribute to
364 the global proteomic diversity of SARS-CoV-2. Most importantly, we show that this mutational
365 bias leads to the preferential substitution of proline residues with leucine or serine residues in the
366 P2 anchor position of SARS-CoV-2 CD8+ T cell epitopes, and hence, drastically compromise
367 epitope binding to B7 supertype molecules, which represent ~35% of the human population
368 (Francisco et al., 2015). Therefore, the C-to-U mutational bias observed amongst prevalent
369 mutants may partially disrupt SARS-CoV-2 T cell immunity in a very significant proportion of the
370 human population. Noteworthy, this impact of C-to-U mutations on B7-depedent epitope escape
371 was somehow predictable. In fact, proline residues originate from codons that are highly rich in C
372 whereas serine and leucine residues originate from codons that are rich in both C and U. One could
373 therefore predict, at least to some extent, that a strong C-to-U bias would lead to proline-to-leucine
374 or proline-to-serine substitutions. Thus, this study highlights the impact of viral mutational biases

and codon usage in shaping the diversity of CD8+ T cell targets. This being said, it is important to realize that we do not make the claim that the presence of proline-to-leucine or proline-to-serine mutations in the SARS-CoV-2 proteomes depend on patients being B7 supertype-positive, or that the B7 supertype drives the evolution of proline-to-leucine/serine mutations. We do, however, demonstrate that the prevalent mutations currently in circulation are enriched for proline-to-leucine/serine, and our *in silico* predictions suggest that the high occurrence of this mutation type leads to widespread hinderance of epitope presentation in B7 supertype-positive individuals.

A key question to address is to what extent does the C-to-U bias drives SARS-CoV-2 evolution and adaptation over the course of the ongoing pandemic. As proposed by others, the most likely explanation for the observed C-to-U bias is the action of the host-mediated RNA-editing APOBEC enzymes, a family of cytidine deaminases that catalyze deamination of cytidine to uridine in RNA (Dorp et al., 2020a; Giorgio et al., 2020; Kosuge et al., 2020; Olson et al., 2018; Salter et al., 2016). In this regard, APOBEC activity has been shown to broadly drive viral evolution and diversity, including in human immunodeficiency virus (HIV) (Albin et al., 2010; Cuevas et al., 2015; Haché et al., 2008; Jern et al., 2009; Peretti et al., 2018; Sadler et al., 2010; Wood et al., 2009). In fact, APOBEC-induced mutations driving the evolution and diversification of HIV-1 were shown to have an intimate relationship with T cell immunity (Kim et al., 2014; Wood et al., 2009). Notably, those studies have shown that the impact of APOBEC-induced mutations may result in either a decrease or increase of CD8+ T cell recognition, and that the direction of this response is dictated by the HLA context (Casartelli et al., 2010; Grant and Larijani, 2017; Kim et al., 2014; Monajemi et al., 2014; Squires et al., 2015; Wood et al., 2009). This is very much in line with our findings. Indeed, we showed that amino acid mutation biases in SARS-CoV-2 proteomes generally positively affect epitope binding for various HLA class I supertypes,

and most strikingly for A24, whereas B7 is the only supertype negatively affected by the mutation biases given the markable loss of proline residues in SARSCoV-2 proteomes. Together, our results raise the important hypothesis that host-mediated RNA editing systems shape the repertoire of SARS-CoV-2 T cell epitopes in a positive and negative HLA-dependant manner.

Another question is whether populations of B7 supertype individuals represent an advantageous reservoir for the virus to evolve toward more transmissible variants. As the genetic diversity of the SARS-CoV-2 population continue to increase, and as new variants emerge, our global analysis suggests that the probability for SARS-CoV-2 epitopes to escape CD8+ T cell immunosurveillance is much higher in B7 individuals compared to A24 individuals. In fact, a slower T cell response dynamic to control SARS-CoV-2 infection in B7 individuals may offer a selective advantage for the virus to evolve. In this regard, we noted that the B.1.1.7 variant lost the B7 supertype-associated epitope SP/HRRARSVA as a result of a proline-to-histidine substitution. While genomic surveillance is ongoing in different regions of the world, measuring the level of transmission of the B.1.1.7 variant within geographical regions of the world with low B7 population densities and high A24 population densities (in Asia) or the opposite trend (in Sub-Saharan Africa) (http://www.allelefrequencies.net/top10freqs.asp) may provide insights into this concern. As new variants of concern continue to emerge and as new epitope data are continuously being generated (Grifoni et al., 2021), another interesting avenue would be to study the mutational patterns of those emerging variants and assess whether and how the potential loss of B7-associated epitopes in those specific variants impact T cell response in infected patients. Understanding the impact of losing several subdominant B7-associated epitopes versus one single immunodominant epitope could also be investigated in the context of those variants. In this regard, a particular attention was allocated in our study to the B*07:02-restricted N105 epitope SPRWYFYYL. This

epitope is of high interest as its immunodominance was experimentally demonstrated in many independent studies (Ferretti et al., 2020; Kared et al., 2021; Saini et al., 2021; Schulien et al., 2021; Sekine et al., 2020; Tarke et al., 2021). Precisely, we found a rare mutation consisting of P→S at P2 of this epitope (SPRWYFYYL → SSRWYFYYL). Its occurrence was predicted to result in the complete abrogation of binding of the epitope to B*07:02, thereby likely reducing the breadth of the immune response in individuals carrying this mutation. As such, we advise the community to carefully monitor this mutation in subsequent months. Moreover, it is also possible that B7 individuals respond less efficiently to the currently available vaccines, as genetic variants promoting B7 escape might favorably emerge in the future. The B7 supertype could therefore potentially represent a biomarker of vaccine resistance.

In summary, our study shows that mutation biases in the SARS-CoV-2 population diversify the repertoire of SARS-CoV-2 T cell targets in humans in an HLA-supertype dependent manner. Hence, we provide a foundation model to help understand how SARS-CoV-2 may continue to mutate over time to shape T cell immunity at a global population scale. The proposed process will likely continue to influence the evolution and diversification of SARS-CoV-2 lineages as the virus is under tremendous pressure to adapt in response to mass vaccination.

LIMITATIONS AND FUTURE DIRECTIONS

Our analyses focused on class I molecules for which predictors are established to be more accurate in comparison with class II. HLA-C and non-classical HLA were not included in this study. Predictions were performed on the most common HLA class I alleles and rare HLA alleles were not included. Study has been performed using the GISAID dataset available in December 31st 2020, i.e. first year of the pandemic, before mass vaccination. Our epitope binding results rely on

in silico predictions using a method that has been widely benchmarked, but is designed to predict peptide presentation rather than immunogenicity. Follow up experiments would need to be performed to further validate the proposed model. Priority follow up studies are 1) to investigate T cell response to SARS-CoV-2 mutants in large cohorts of B7 supertype-positive versus negative patients, and 2) to determine the direct role of APOBEC family proteins in modulation of SARS-CoV-2-specific T cell immunity. Moreover, this study lays the foundation to understand the evolutionary dynamics of pandemic viruses with a time 0 / no vaccine-induced immune pressure start point. Employing SARS-CoV-2 as model provides an opportunity in future studies to look at the dynamic of the relationship between mutational patterns and HLA-restricted T cell immunity in real-time. Kinetic analyses using the latest GISAID datasets, which now include 1.7M SARS-CoV-2 genomes as of May 2021, may lead to additional insights in this regard.

ACKNOWLEDGMENTS

We thank Drs. Alessandro Sette, John Sidney and Alba Grifoni (La Jolla Institute for Immunology, USA) for helpful discussions. This study was supported by funding from the Fonds de recherche du Québec – Santé (FRQS), the Cole Foundation, CHU Sainte-Justine and the Charles-Bruneau Foundations, Canada Foundation for Innovation, IVADO COVID19 Rapid Response grant (CVD19-030), Montreal Heart Institute Foundation, the National Sciences and Engineering Research Council (NSERC) (#RGPIN-2020-05232) and the Canadian Institutes of Health Research (CIHR) (#174924). K.K. is a recipient of IVADO's postdoctoral scholarship (#4879287150). D.F. is a BioTalent awardee. E.C. and J.H. are FRQS Junior 1 Research Scholars.

AUTHOR CONTRIBUTIONS

Conceptualization: D.H., J.H., and E.C.; Data Curation and Bioinformatic Analysis: D.H., D.F., J-C.G., F.M, K.K., and P.K.; Formal Analysis: D.H., and D.F.; Investigation: D.H., D.F., J.S., J-C.G., K.K., J.D.D., F.S., P.K., I.S., H.D., S.P., J.H., and E.C.; Writing – Original Draft: D.H., and E.C.; Writing – Review & Editing: D.H., D.F., J.S., J.S., J-C.G., F.M., K.K., P.K., J.D.D., F.S., I.S., M.S., H.S., H.D., S.P., J.H., and E.C.; Supervision: J.H., and E.C.; Funding Acquisition: J.H., and E.C.

DECLARATION OF INTERESTS

Jana Schockaert and Sofie Pattijn are employees of ImmunXperts, a Nexelis Group Company.

FIGURE LEGENDS

Figure 1. Distribution of CD8+ T cell epitopes and their mutated variants across the immunodominant S and N antigens. (**A, B**) Lower panel: blue dots showing all mutations that occurred in at least 4 SARS-CoV-2 genomes (GISAID). Middle panel: epitope density showing the overlap of HLA class I epitopes predicted within the 1st percentile for 12 queried HLA-I molecules. Upper panel: dots showing the frequency of CD8+ T cell response as determined from multiple studies aggregated in the database https://www.mckayspcb.com/SARS2TcellEpitopes as of January 2021. Red dots are mutated epitopes wherein the mutation event led to a predicted loss of binding. Sequences of specific epitopes are shown with the mutant amino acid in red. The red box in the N protein highlights a serine-rich region associated with no T cell response, low epitope density and high mutation frequency. (**C**) 3D structure of the Spike glycoprotein (Moderna

Vaccine) and highlighted in yellow is the Receptor Binding Domain (Pfizer Vaccine). Shown in red are mutated epitopes wherein mutation events led to a predicted loss of HLA binding.

Figure 2. Global amino acid mutational biases in SARS-CoV-2 proteomes. A total of 330,246 SARS-CoV-2 genomes were translated into protein sequences and analyzed for the identification of any amino acid mutational bias. Amino acid residues (x-axis) that were removed and introduced in SARS-CoV-2 variants are presented by negative and positive %-difference in overall amino acid composition (GRSO values; y-axis), respectively. Analysis of mutational biases was performed for mutations occurring at various frequencies: 1 genome (blue line), 2 to 100 genomes (orange line), 100 to 1000 genomes (green line) and more than 1000 genomes (red line). Simulation of neutral evolution simulation (random mutations) were performed using the SANTA-SIM algorithm and serves as control for assessing the statistical significance of the observed pattern for individual amino acid residues. The dotted red lines show the cutoff values (fold change > 4; p-value < 1x10-11) that were used to define the residues that were preferentially removed or introduced (asterisk).

Figure 3. Mutation of proline at the anchor residue position for B7 supertype-associated epitopes. (**A**) (Left panel) Motif view of SARS-CoV-2 reference peptides predicted to bind B7 supertype molecules (HLA-B*07:02, -B*35:03, -B42:02 , -B*5101, -B*53:01, -B*54:01, -B*55:01, -B*56:01, -B*67:01). (Right panel) Motif view of the corresponding mutated peptides. (**B**) Heat map showing the frequency of specific amino acid substitutions between reference and mutated peptides. (**C**) Graph showing the number of mutations (upper panel; y-axis) leading to specific amino acid substitutions (x-axis) at anchor residue positions P2 (red dots) and P9 (green

511 dots) or elsewhere (black dots). Dotted red line indicate the cutoff used to define dominant
512 substitutions. The lower panel shows fold changes for individual amino acid substitutions. (**D**)
513 Representative examples of validated CD8+ T cell epitopes
514 (https://www.mckayspcb.com/SARS2TcellEpitopes as of January 2021). Effect of the P→X
515 substitutions on predicted epitope binding affinities (NetMHCpan 4.1 EL %Rank) are shown. T
516 cell response data for reference epitopes extracted from
517 https://www.mckayspcb.com/SARS2TcellEpitopes.
518
519 **Figure 4. Loss or gain of SARS-CoV-2 mutated epitopes for different HLA class I supertypes.**
520 (**A, B**) Motif views showing established epitope binding motifs for different HLA-I alleles that
521 belong to the HLA-B7 (A) and HLA-A24 (B) supertype family. Shaded squares highlight anchor
522 residues that are preferentially removed (pale green) or introduced (pale orange) in SARS-CoV-2
523 proteomes (related to Figure 2), respectively. Histograms below the binding motifs indicate the
524 number of frequent mutations (identified in at least 100 individuals) leading to the loss or gain of
525 epitopes. (**C**) 'Gain/Loss plots' showing number of mutations (y-axis) leading to a preferentially
526 loss (pale green) or gain (pale orange) of epitopes for different HLA class I supertypes. Each black
527 dot represents the number of mutations associated with gain and loss of epitopes for a given HLA-
528 I allele. Between 14 to 19 alleles per supertype (Figure S5) were used to generate the graphs and
529 p-values (*$p \leq 0.001$, **$p < 1e\text{-}5$, ***$p < 1e\text{-}10$).
530
531 **Figure 5. The C-to-U point mutation bias largely drives the diversity of SARS-CoV-2**
532 **proteomes and CD8+ T cell epitopes.** (**A**) Comparison of global amino acid mutational patterns
533 generated from real-life versus simulated SARS-COV-2 genomes. Amino acid residues (x-axis)

that were removed and introduced in real-life versus simulated SARS-CoV-2 are presented by negative and positive %-difference in overall amino acid composition (GRSO values; y-axis), respectively. Evolution of SARS-CoV-2 was simulated by introducing various extents of C-to-U biases, i.e. x1, x15 and x20 (n = 10). The red line shows the pattern obtained from mutations identified in more than 100 SARS-CoV-2 genomes, related to Figure 2. (**B**) (Top) Pie chart showing the proportion of nucleotide substitution types from the list of validated CD8+ T cell epitopes in https://www.mckayspcb.com/SARS2TcellEpitopes as of January 2021. (Bottom) Pie chart showing the proportion of nucleotide substitution types from the list of validated CD8+ T cell epitopes specific to the B7 supertype. (**C**) Schematic illustrating the C-to-U-mediated epitope escape model. The observed mutation of the immunodominant SPRWYLFYYL epitope in the N protein is shown as an example.

STAR METHODS

RESOURCE AVAILABILITY

Lead Contact

Further information and requests should be directed to the lead contact, Dr. Etienne Caron (etienne.caron@umontreal.ca)

Materials Availability

This study did not generate new unique reagents.

Data and Code Availability

All sequence data used here are available from The Initiative for Sharing All Influenza Data (GISAID), at https://gisaid.org/. The user agreement for GISAID does not permit redistribution of sequences, but researchers can register to get access to the dataset. Code to create the alignments,

to predict mutated and unmutated HLA-I peptides, and to perform the global analysis of SARS-CoV-2 proteomes are available at https://github.com/CaronLab.

METHOD DETAILS

Identification of SARS-CoV-2 mutations

All SARS-CoV-2 nucleotide sequences were acquired from the GISAID on 31/12/2021. A total of 330,246 SARS-CoV-2 sequences spanning 143 countries were acquired and analyzed. All sequences isolated from animals (including viral RNA isolated from bat, pangolin, mink, cat and tiger) were removed from the list and only high-quality sequences were further analysed. Consensus sequences were aligned to the reference sequence, Wuhan-1 (NC_045512.2) using minimap2 2.17-r974. All mapped sequences were then merged back with all others in a single alignment bam file. The variant calling was done using bcftools mpileup v1.91 in a haploid calling mode. Sequences were processed by batches of 1000 to overcome technical issues with very low-frequency variants. With the variant calling obtained for each batch, vcf-merge (from the vcftools suite) was used to merge all the variant calls across the entire dataset. A total of 24,220 variants in at least two consensus sequences were identified. Mutations appearing in only one genome were excluded as they are likely enriched for sequencing errors. A list of all missense mutations considered in our analyses is provided in **Table S1**. The 1,933 prevalent mutations observed in more than 100 genomes are also clearly shown in **Table S2**.

Prediction of mutated and reference CD8+ T-cell epitopes

Prediction of CD8+ T cell epitopes was carried out using netMHCpan 4.0 EL (Reynisson et al., 2020). For each unique missense mutation, short sequence windows consisting of 14 amino acids

on either side of the mutation site were generated, containing either the reference or mutated amino acid. Working from the resulting 29-residue sequence windows (mutation +/- 14 residues), 811mers were predicted against the 12 most frequent HLA alleles within the global population (HLA-A*01:01, HLA-A*02:01, HLA-A*03:01, HLA-A*11:01, HLA-A*23:01, HLA-A*24:02, HLA-B*07:02, HLA-B*08:01, HLA-B*35:01, HLA-B*40:01, HLA-B*44:02, and HLAB*44:03). Briefly, the NetMHCpan 4.0 EL method relies on a neural network trained on both binding affinity as well as eluted ligand data to produce a likelihood score for a peptide to be an eluted ligand for the indicated HLA types. The likelihood score consists of a percentile rank (%rank) wherein predicted (weak) binders obtain a %rank below 2.0, whereas strong binder (SB) obtain a %rank below 0.5. Using this ranking system, only mutation-containing peptides where the mutated and/or the reference peptide were ranked as SB were considered for further analyses. Mutations causing percentile ranks to transition from strong HLA-binder (SB, netMHCpan %Rank < 0.5) to HLA non-binders (NB, netMHCpan %Rank > 2.0) were considered as leading to 'Loss of binding'. Mutations causing predicted binding affinities to transition from NB to SB were considered as leading to 'Gain of binding'.

Selection of clinically validated CD8+ T-Cell epitopes

A list of validated CD8+ T Cell epitopes presented by both HLA-A and -B molecules were downloaded from https://www.mckayspcb.com/SARS2TcellEpitopes/ (as of January 2021). This database, developed by Dr. Matthew R. McKay and his team, contains compiled and catalogued validated T-cell epitope-HLA pairs from 13 studies aimed at identifying immunogenic SARSCOV-2 T-cell epitopes.

603 **In vitro HLA-peptide binding assays**

604 Peptide binding to class I HLA molecules was quantitatively measured using classical competition
605 assays based on the inhibition of binding of a high affinity radiolabeled peptide to purified HLA
606 molecules, as detailed elsewhere (Sidney et al., 2013). Briefly, HLA molecules were purified from
607 lysates of EBV transformed homozygous cell lines by affinity chromatography by repeated
608 passage over Protein A Sepharose beads conjugated with the W6/32 (anti-HLA-A, -B, -C)
609 antibody, following separation from HLA-B and -C molecules by pre-passage over a B1.23.2
610 (antiHLA B, C) column. Protein purity, concentration, and the effectiveness of depletion steps was
611 monitored by SDS-PAGE and BCA assay. Peptide affinity for respective class I molecules was
612 determined by incubating 0.1-1 nM of radiolabeled peptide at room temperature with 1 μM to 1
613 nM of purified HLA in the presence of a cocktail of protease inhibitors and 1 μM B2microglobulin.
614 Following a two-day incubation, HLA bound radioactivity was determined by capturing
615 MHC/peptide complexes on W6/32 antibody coated Lumitrac 600 plates (Greiner Bioone,
616 Frickenhausen, Germany). Bound cpm was measured using the TopCount (Packard Instrument
617 Co., Meriden, CT) microscintillation counter. The concentration of peptide yielding 50%
618 inhibition of the binding of the radiolabeled peptide was calculated. Under the conditions utilized,
619 where [label]<[MHC] and IC50 ≥ [MHC], the measured IC50 values are reasonable
620 approximations of the true Kd values. Each competitor peptide was tested at six different
621 concentrations covering a 100,000-fold dose range, and in three or more independent experiments.
622 As a positive control for inhibition, the unlabeled version of the radiolabeled probe was also tested
623 in each experiment.
624

SANTA-SIM simulations

We simulated SARS-CoV-2 genomes with SANTA-SIM, using the consensus sequence WuhanHu-1 as input sequence available at https://www.ncbi.nlm.nih.gov/nuccore/MN908947.3. Each simulation was run with a population size of 10,000 individual viral sequences evolving for 1000 generations, and analyses were conducted on random samples of 1,000 viral sequences. Following Huddelston et.al. (Huddleston et al., 2020) who used SANTA-SIM to simulate influenza A/H3N2 that has a yearly substitution rate approximately twice as high as SARS-CoV-2 [~48,824 substitutions/year (https://nextstrain.org/flu/seasonal/h3n2/ha/2y?l=clock) vs. ~24.5 substitution/year (https://nextstrain.org/ncov/global?l=clock)], we chose 400 generations/year, with the mutation rate per position per generation set to 2.04E-6 (yearly substitution rate/(generations in one year * genome size)). The transition bias was set to 3.0 for baseline simulations. To evaluate the impact of specific substitution biases, additional simulations were conducted using a substitution matrix with scores set to 1.0 of transversions, 3.0 for transitions, and biases ranging from 4.0 to 20.0 for the targeted substitution. We generated 10 replicates for all simulated scenarios, except for C-to-U where we made 100 replicates to better assess statistical significance.

Determination of amino acid mutational patterns

Mutational biases were identified by calculating the overall change in amino acid composition caused by the mutational landscape of SARS-CoV-2 for each individual amino acid, referred in the main text as 'global residue substitution output' (GRSO). For this analysis, all mutations found globally in at least 4 GISAID entries were analysed together. Preferential introduction or removal of amino acids was determined by comparing the overall amino acid composition in reference

residues vs mutated residues throughout the mutation pool, resulting in a percentile difference in amino acid composition. As such, for amino acid *X*, the % difference was calculated according to the following formula:

$$\%\ difference\ = \left(\frac{Nbr\ of\ mutations\ introducing\ X - Nbr\ of\ mutations\ removing\ X}{All\ Global\ mutations\ in\ at\ least\ 4\ GISAID\ entries}\right) \times 100$$

This analysis took into consideration the number of unique mutations. Therefore, to consider mutational biases in the context of mutation frequencies, the analysis described above was conducted separately for mutations occurring in a single GISAID entry (expected to be enriched for errors); 2-10 GISAID entries; 11-99 GISAID entries; and 100 or more GISAID entries. As a negative control, the SANTA SIM algorithm was used to simulate the neutral evolution of 1000 SARS-CoV-2 genomes (baseline simulations, N = 10 replicates). This control was used to calculate the statistical significance of the observed biases, by way of a One-Sample T-Test.

Prediction of mutation impacts on peptide presentation in the context of HLA supertypes

Reference/mutated peptide pairs for which the differential predicted binding affinities led to transitions from strong HLA binder (SB) to non-HLA binder (NB) [(SB) NetMHCpan %rank < 0.5 to (NB) NetMHCpan %rank >2] or from NB to SB, were identified, catalogued and analyzed as described above. Binding affinities were predicted for representative HLA types from several major HLA supertypes (A1, A2, A3, A24, B7, B8, B27, B44), as defined by Sydney *et al*. We then categorized all reference/mutated peptide pairs on the basis of their 1) mutation type (amino acid *X* → amino acid *Y*) and 2) the position of the mutation in the peptide sequence. Finally, we quantified the number of reference/mutated peptide pairs and the associated average fold change in predicted binding affinity for each category. P-values were generated for each category by performing a two-tailed independent T-Test between the fold changes in binding affinity

associated with mutation type *A* at position *X*, and all fold changes in binding affinity associated with position *X*.

Assessing the contribution of nucleic acid mutation types to the global amino acid mutational patterns.

To assess the contribution of various nucleic acid mutation types to the observed amino acid mutational patterns, we first determined the respective contributions of each nucleic acid mutation type to the global mutation landscape. We then selected the five most abundant mutation types [C→U (41%), G→U (18%), A→G, G→A, U→C (9.7-11.6%)] and assessed their individual impacts on amino acid mutational patterns using the simulation algorithm SANTA SIM as follows: For each mutation type, we simulated the evolution of 1000 SARS-CoV-2 genomes over 1000 generations (N = 10 replicates) with varying degrees of biases (the coefficient used to determine the extent of the biases was exploratively set to 'x4', 'x8', 'x15', and 'x20') (Figure S6A). Because the input coefficient does not have a linear relationship with the abundance of the mutation type observed in the simulation output, we used the simulations with all four parameter values (x4, x8, x15, x20) in order to identify the simulation parameter that most closely reflected observations in real-life SARS-CoV-2 data. The coefficient for the ratio of $X \rightarrow Y$ nucleic acid mutation type to all other mutation types was generated using the following formula:

$$Mutation\ Bias\ Coefficient = \frac{\left(\frac{All\ X \rightarrow Y\ mutations}{All\ X\ positions\ in\ reference\ genome}\right)}{\left(\frac{All\ mutations}{All\ positions\ in\ reference\ genome}\right)}$$

Finally, all amino acid mutations were identified for the output of each simulation, as described above. To determine statistical significances, simulated mutational biases (at the amino acid level)

692 were compared to a neutral evolution as a negative control (N = 10 replicates) by way of twotailed
693 independent T-Test.
694
695 **Statistical analysis**
696 A Two-tailed One-Sample T-Test was used to assess the statistical significance of the observed
697 mutational biases against the neutral simulations (N = 10 replicates). A Two-tailed Independent
698 T-Test assuming different variances was used to assess the statistical significances of 1) the
699 simulated biased SARS-CoV-2 evolution, 2) the gain/loss plots in the context of supertypes, and
700 3) the statistical significance associated with the average fold change in %rank associated with
701 each position-specific amino acid mutation type in the supertype analysis.
702
703 **SUPPLEMENTARY MATERIALS**
704 **SUPPLEMENTARY FIGURE LEGENDS**
705 **Figure S1. Impact of SARS-CoV-2 mutations on CD8+ T cell epitopes, Related to Figure 1**
706 **and 4.** (**A**) Bioinformatic pipeline for the prediction of SARS-CoV-2 mutated class I peptides
707 associated to 12 common HLA alleles. (**B**) Pyramidal graph showing the number of i) missense
708 mutations in SARS-CoV-2 genomes, ii) predicted class I mutated peptides, iii) predicted class I
709 peptides subject to Weak Binder (WB) to Non-Binder (NB) and Strong Binder (SB) to NB
710 transition (epitope loss category), and iv) predicted class I mutated peptides matching reference
711 CD8+ T cell epitopes that have been experimentally validated. (**C**) Representative examples of
712 predicted class I mutated peptides and the impact of the identified amino acid mutation (bold) on
713 peptide binding to a given HLA-I allele. Reference and mutated EL (eluted ligand) Rank (%)
714 generated by NetMHCpan 4.1 EL is indicated for individual predictions. Gain = NB to SB (pale

715 red); Loss = SB to NB (pale green). (**D**) Left panel: number of unique mutations leading to 'Gain'
716 or 'Loss' of class I peptides for the indicated HLA-I alleles. Right panel: number of unique
717 mutations showing no effect on peptide binding for the indicated HLA-I alleles. (**E**) Validated
718 SARS-CoV-2 CD8+ T cell epitopes (McKay Database) subjected to mutation events detected in
719 more than 4 individuals (GISAID) and predicted lead to a strong loss of HLA-epitope binding.
720 Top: number of unique missense mutations corresponding to the indicated amino acid substitution
721 type. Bottom: Predicted loss of HLA-epitope binding (NetMHCpan4.1 %Rank) corresponding to
722 the indicated residue substitution type from the list of validated CD8+ T cell epitopes in the McKay
723 Database. Each dot represents an epitope pair (mutated / reference). Color indicates HLA type
724 affected by the mutations.
725
726 **Figure S2. HLA peptide binding measurements and mutational biases in SARS-CoV-2**
727 **mutated epitopes, Related to Figure 1.** (**A**) HLA binding assay was performed to determine the
728 in vitro binding affinity (nM) of representative SARS-CoV-2 peptides for specific HLA class I
729 alleles. Peptides were selected based on 1) frequency of mutations, 2) presentation by common
730 HLA class I alleles, and 3) the mutated form was predicted to lose binding to its corresponding
731 HLA. (**B**) Plots showing raw values for the binding affinities (nM) of the reference vs mutated
732 peptides in (A). The first three amino acid residues of the reference peptides with fold change >
733 2.5 are shown. (**C**) Pie chart showing the proportion of X-to-Y substitution types from the list of
734 validated CD8+ T cell epitopes in https://www.mckayspcb.com/SARS2TcellEpitopes/ (as of
735 January 2021). (**D**) Predicted loss of HLA-epitope binding clustered by substitution type from the
736 list of validated CD8+ T cell epitopes in the McKay database. Each dot represents an epitope pair
737 (mutated / reference; NetMHCpan 4.1 %rank ratio).

738

739 **Figure S3. Identification of two SARS-CoV-2 mutated epitopes in this study that were**
740 **previously associated with decreased CD8+ T cell responses, Related to Figure 1.** (**A**) The
741 mutated epitopes GFMWLSYFI (A*02) and MKVTPSGTWL (B*40) were detected in 38 and 23
742 genomes/individuals in this study (GISAID) and their T cell immunogenicity was thoroughly
743 investigated in Agerer et al. (**B-E from Agerer et al., copyright 2021, with permission from**
744 **AAAS)** (**B**) Experimental overview. (**C**) T cells expanded with mutant peptides do not give rise to
745 wild type peptide-specific CD8+ T cell. PBMCs were isolated from HLA-A*02:01 or HLA-
746 B*40:01 positive SARS-CoV-2 patients, stimulated with wild type or mutant peptides and stained
747 with tetramers containing the wild type peptide. (**D**) Impact of mutations on CD8+ T cell response.
748 PBMCs expanded with wild type or mutant peptides as indicated, were analyzed for IFN-γ-
749 production via ICS after restimulation with wild type or mutant peptide. (**E**) Representative FACS
750 plots for (D).

751

752 **Figure S4. Impact of mutations on gain of peptide binding to various HLA class I molecules**
753 **across the immunodominant S and N antigens, Related to Figure 1.** (**A, B**) Lower panel: blue
754 dots showing all mutations that occurred in at least 4 SARS-CoV-2 genomes (GISAID). Upper
755 panel: dots showing predicted peptides subjected to a strong gain of binding (see also Figure
756 S1C,D) to one of 12 highly common HLA types queried (color coded) due to a mutation.

757

758 **Figure S5. HLA class I supertypes, Related to Figure 4.** (**A**) Epitope binding motifs for several
759 HLA class I supertypes. Anchor residues are located at P2 and P9. Pale orange and green squares
760 cover amino acid residues that are preferentially introduced (F, I, L, Y) and removed (A, P, T) in

761 SARS-CoV-2 proteomes, respectively. Representative supertypes used in this study are shown by
762 an asterisk. Epitope binding motifs were extracted from NetMHCpan Motif Viewer
763 (http://www.cbs.dtu.dk/services/NetMHCpan/logos_ps.php). (**B**) Table showing the selected
764 alleles per supertype that were used in this study to generate the 'Gain/Loss plots'.
765
766 **Figure S6. Comparison of mutation biases between real-life/observed and simulated data,**
767 **Related to Figure 5.** (**A**) Histograms showing the number of unique mutations identified for each
768 mutation type (A-to-C, A-to-G, etc.) after simulating the evolution of SARS-CoV-2 genomes
769 through the introduction of different C-to-U bias values (x4 to x20) using the SANTA-SIM
770 software. Simulated (black squares) and real-life/observed prevalent mutations found in more than
771 100 genomes (red square) at the nucleotide level are shown. (**B**) Comparison of global amino acid
772 mutational patterns generated from simulated versus real-life/observed SARS-COV-2 genomes.
773 Various extents of C-to-U (top) and G-to-U (bottom) biases were introduced to perform the
774 simulation and to generate the graphs.
775
776 **SUPPLEMENTARY TABLE LEGENDS**
777 **Table S1. SARS-CoV-2 mutations identified from 330,246 GISAID entries (December 31st**
778 **2020), Related to Figure 1.** SARS-CoV-2 mutations at the nucleic and amino acid level are
779 indicated. Number of genomes carrying mutation show the frequency of individual mutations
780 among all SARS-CoV-2 variants.
781
782 **Table S2. SARS-CoV-2 prevalent mutations identified from 330,246 GISAID entries**
783 **(December 31st 2020) and detected in at least 100 individuals, Related to Figure 1**.

Table S3. Documented SARS-CoV-2 CD8+ T cell epitopes and their matching mutated forms identified in this study, Related to Figure 1.

Table S4. List of documented SARS-CoV-2 CD8+ T cell epitopes. Epitopes were downloaded from https://www.mckayspcb.com/SARS2TcellEpitopes/ (as of January 2021). This database has effectively catalogued all SARS-CoV-2 CD8+ epitopes validated by 18 separate studies.

807 **REFERENCES**

808

809 Agerer, B., Koblischke, M., Gudipati, V., Montaño-Gutierrez, L.F., Smyth, M., Popa, A.,
810 Genger, J.-W., Endler, L., Florian, D.M., Mühlgrabner, V., et al. (2021). SARS-CoV-2 mutations
811 in MHC-I-restricted epitopes evade CD8+ T cell responses. Sci Immunol *6*, eabg6461.

812 Albin, J.S., Haché, G., Hultquist, J.F., Brown, W.L., and Harris, R.S. (2010). Long-Term
813 Restriction by APOBEC3F Selects Human Immunodeficiency Virus Type 1 Variants with
814 Restored Vif Function. J Virol *84*, 10209–10219.

815 Altmann, D.M., and Boyton, R.J. (2020). SARS-CoV-2 T cell immunity: Specificity, function,
816 durability, and role in protection. Sci Immunol *5*, eabd6160.

817 Bert, N.L., Tan, A.T., Kunasegaran, K., Tham, C.Y.L., Hafezi, M., Chia, A., Chng, M.H.Y., Lin,
818 M., Tan, N., Linster, M., et al. (2020). SARS-CoV-2-specific T cell immunity in cases of
819 COVID-19 and SARS, and uninfected controls. Nature *584*, 457–462.

820 Braun, J., Loyal, L., Frentsch, M., Wendisch, D., Georg, P., Kurth, F., Hippenstiel, S.,
821 Dingeldey, M., Kruse, B., Fauchere, F., et al. (2020). SARS-CoV-2-reactive T cells in healthy
822 donors and patients with COVID-19. Nature *587*, 270-274.

823 Brumme, Z.L., Brumme, C.J., Heckerman, D., Korber, B.T., Daniels, M., Carlson, J., Kadie, C.,
824 Bhattacharya, T., Chui, C., Szinger, J., et al. (2007). Evidence of Differential HLA Class I-
825 Mediated Viral Evolution in Functional and Accessory/Regulatory Genes of HIV-1. Plos Pathog
826 *3*, e94.

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

828 Casartelli, N., Guivel-Benhassine, F., Bouziat, R., Brandler, S., Schwartz, O., and Moris, A.
829 (2010). The antiviral factor APOBEC3G improves CTL recognition of cultured HIV-infected T
830 cells. J Exp Medicine *207*, 39–49.

831 Cuevas, J.M., Geller, R., Garijo, R., López-Aldeguer, J., and Sanjuán, R. (2015). Extremely High
832 Mutation Rate of HIV-1 In Vivo. Plos Biol *13*, e1002251.

833 Dan, J.M., Mateus, J., Kato, Y., Hastie, K.M., Yu, E.D., Faliti, C.E., Grifoni, A., Ramirez, S.I.,
834 Haupt, S., Frazier, A., et al. (2021). Immunological memory to SARS-CoV-2 assessed for up to
835 8 months after infection. Science *371*, eabf4063.

836 Dorp, L. van, Richard, D., Tan, C.C.S., Shaw, L.P., Acman, M., and Balloux, F. (2020a). No
837 evidence for increased transmissibility from recurrent mutations in SARS-CoV-2. Nat Commun
838 *11*, 5986.

839 Dorp, L. van, Acman, M., Richard, D., Shaw, L.P., Ford, C.E., Ormond, L., Owen, C.J., Pang, J.,
840 Tan, C.C.S., Boshier, F.A.T., et al. (2020b). Emergence of genomic diversity and recurrent
841 mutations in SARS-CoV-2. Infect Genetics Evol *83*, 104351.

842 Falk, K., Rötzschke, O., Stevanovic, S., Jung, G., and Rammensee, H.-G. (1991). Allele-specific
843 motifs revealed by sequencing of self-peptides eluted from MHC molecules. Nature *351*, 290-
844 296.

845 Ferretti, A.P., Kula, T., Wang, Y., Nguyen, D.M.V., Weinheimer, A., Dunlap, G.S., Xu, Q.,
846 Nabilsi, N., Perullo, C.R., Cristofaro, A.W., et al. (2020). Unbiased Screens Show CD8+ T Cells
847 of COVID-19 Patients Recognize Shared Epitopes in SARS-CoV-2 that Largely Reside outside
848 the Spike Protein. Immunity *53*, 1095-1107.e3.

849 Francisco, R. dos S., Buhler, S., Nunes, J.M., Bitarello, B.D., França, G.S., Meyer, D., and
850 Sanchez-Mazas, A. (2015). HLA supertype variation across populations: new insights into the
851 role of natural selection in the evolution of HLA-A and HLA-B polymorphisms.
852 Immunogenetics *67*, 651–663.

853 Gfeller, D., and Bassani-Sternberg, M. (2018). Predicting Antigen Presentation-What Could We
854 Learn From a Million Peptides? Front Immunol *9*, 1716.

855 Giorgio, S.D., Martignano, F., Torcia, M.G., Mattiuz, G., and Conticello, S.G. (2020). Evidence
856 for host-dependent RNA editing in the transcriptome of SARS-CoV-2. Sci Adv *6*, eabb5813.

857 Grant, M., and Larijani, M. (2017). Evasion of adaptive immunity by HIV through the action of
858 host APOBEC3G/F enzymes. Aids Res Ther *14*, 44.

859 Greenbaum, J., Sidney, J., Chung, J., Brander, C., Peters, B., and Sette, A. (2011). Functional
860 classification of class II human leukocyte antigen (HLA) molecules reveals seven different
861 supertypes and a surprising degree of repertoire sharing across supertypes. Immunogenetics *63*,
862 325-335.

863 Grifoni, A., Weiskopf, D., Ramirez, S.I., Mateus, J., Dan, J.M., Moderbacher, C.R., Rawlings,
864 S.A., Sutherland, A., Premkumar, L., Jadi, R.S., et al. (2020a). Targets of T cell responses to
865 SARS-CoV-2 coronavirus in humans with COVID-19 disease and unexposed individuals. Cell
866 *18*, 1489-1501.e15.

867 Grifoni, A., Sidney, J., Zhang, Y., Scheuermann, R.H., Peters, B., and Sette, A. (2020b). A
868 Sequence Homology and Bioinformatic Approach Can Predict Candidate Targets for Immune
869 Responses to SARS-CoV-2. Cell Host Microbe *27*, 671-680.e2.

870 Grifoni, A., Sidney, J., Vita, R., Peters, B., Crotty, S., Weiskopf, D., and Sette, A. (2021). SARS-
871 CoV-2 Human T cell Epitopes: adaptive immune response against COVID-19. Cell Host
872 Microbe. In press.

873 Haché, G., Shindo, K., Albin, J.S., and Harris, R.S. (2008). Evolution of HIV-1 Isolates that Use
874 a Novel Vif-Independent Mechanism to Resist Restriction by Human APOBEC3G. Curr Biol *18*,
875 819–824.

876 Huddleston, J., Barnes, J.R., Rowe, T., Xu, X., Kondor, R., Wentworth, D.E., Whittaker, L.,
877 Ermetal, B., Daniels, R.S., McCauley, J.W., et al. (2020). Integrating genotypes and phenotypes
878 improves long-term forecasts of seasonal influenza A/H3N2 evolution. Elife *9*, e60067.

879 Jariani, A., Warth, C., Deforche, K., Libin, P., Drummond, A.J., Rambaut, A., IV, F.A.M., and
880 Theys, K. (2019). SANTA-SIM: simulating viral sequence evolution dynamics under selection
881 and recombination. Virus Evol *5*, vez003.

882 Jern, P., Russell, R.A., Pathak, V.K., and Coffin, J.M. (2009). Likely Role of APOBEC3G-
883 Mediated G-to-A Mutations in HIV-1 Evolution and Drug Resistance. Plos Pathog *5*, e1000367.

884 Kared, H., Redd, A.D., Bloch, E.M., Bonny, T.S., Sumatoh, H.R., Kairi, F., Carbajo, D., Abel,
885 B., Newell, E.W., Bettinotti, M., et al. (2021). SARS-CoV-2-specific CD8+ T cell responses in
886 convalescent COVID-19 individuals. J Clin Invest *131*, e145476.

887 Kawashima, Y., Pfafferott, K., Frater, J., Matthews, P., Payne, R., Addo, M., Gatanaga, H.,
888 Fujiwara, M., Hachiya, A., Koizumi, H., et al. (2009). Adaptation of HIV-1 to human leukocyte
889 antigen class I. Nature *458*, 641–645.

890 Kim, E.-Y., Lorenzo-Redondo, R., Little, S.J., Chung, Y.-S., Phalora, P.K., Berry, I.M., Archer,
891 J., Penugonda, S., Fischer, W., Richman, D.D., et al. (2014). Human APOBEC3 Induced
892 Mutation of Human Immunodeficiency Virus Type-1 Contributes to Adaptation and Evolution in
893 Natural Infection. Plos Pathog *10*, e1004281.

894 Klimczak, L.J., Randall, T.A., Saini, N., Li, J.-L., and Gordenin, D.A. (2020). Similarity between
895 mutation spectra in hypermutated genomes of rubella virus and in SARS-CoV-2 genomes
896 accumulated during the COVID-19 pandemic. Plos One *15*, e0237689.

897 Korber, B., Fischer, W.M., Gnanakaran, S., Yoon, H., Theiler, J., Abfalterer, W., Hengartner, N.,
898 Giorgi, E.E., Bhattacharya, T., Foley, B., et al. (2020). Tracking changes in SARS-CoV-2 Spike:
899 evidence that D614G increases infectivity of the COVID-19 virus. Cell *182*, 812-827.e19.

900 Kosuge, M., Furusawa-Nishii, E., Ito, K., Saito, Y., and Ogasawara, K. (2020). Point mutation
901 bias in SARS-CoV-2 variants results in increased ability to stimulate inflammatory responses.
902 Sci Rep-Uk *10*, 17766.

903 Krammer, F. (2020). SARS-CoV-2 vaccines in development. Nature *586*, 516–527.

904 Laamarti, M., Alouane, T., Kartti, S., Chemao-Elfihri, M.W., Hakmi, M., Essabbar, A.,
905 Laamarti, M., Hlali, H., Bendani, H., Boumajdi, N., et al. (2020). Large scale genomic analysis
906 of 3067 SARS-CoV-2 genomes reveals a clonal geo-distribution and a rich genetic variations of
907 hotspots mutations. Plos One *15*, e0240345.

908 Li, Y., Yang, X., Wang, N., Wang, H., Yin, B., Yang, X., and Jiang, W. (2020). Mutation profile
909 of over 4500 SARS-CoV-2 isolations reveals prevalent cytosine-to-uridine deamination on viral
910 RNAs. Future Microbiol *15*, 1343–1352.

911 Liao, M., Liu, Y., Yuan, J., Wen, Y., Xu, G., Zhao, J., Cheng, L., Li, J., Wang, X., Wang, F., et
912 al. (2020). Single-cell landscape of bronchoalveolar immune cells in patients with COVID-19.
913 Nat Med *26*, 842–844.

914 Liu, G., Carter, B., Bricken, T., Jain, S., Viard, M., Carrington, M., and Gifford, D.K. (2020).
915 Computationally Optimized SARS-CoV-2 MHC Class I and II Vaccine Formulations Predicted
916 to Target Human Haplotype Distributions. Cell Syst *11*, 131-144.e.6.

917 Liu, W., Fontanet, A., Zhang, P.-H., Zhan, L., Xin, Z.-T., Baril, L., Tang, F., Lv, H., and Cao,
918 W.-C. (2006). Two-Year Prospective Study of the Humoral Immune Response of Patients with
919 Severe Acute Respiratory Syndrome. J Infect Dis *193*, 792–795.

920 Long, Q.-X., Liu, B.-Z., Deng, H.-J., Wu, G.-C., Deng, K., Chen, Y.-K., Liao, P., Qiu, J.-F., Lin,
921 Y., Cai, X.-F., et al. (2020a). Antibody responses to SARS-CoV-2 in patients with COVID-19.
922 Nat Med *26*, 845–848.

923 Long, Q.-X., Tang, X.-J., Shi, Q.-L., Li, Q., Deng, H.-J., Yuan, J., Hu, J.-L., Xu, W., Zhang, Y.,
924 Lv, F.-J., et al. (2020b). Clinical and immunological assessment of asymptomatic SARS-CoV-2
925 infections. Nat Med *26*, 1200–1204.

926 Matyášek, R., and Kovařík, A. (2020). Mutation Patterns of Human SARS-CoV-2 and Bat
927 RaTG13 Coronavirus Genomes Are Strongly Biased Towards C>U Transitions, Indicating Rapid
928 Evolution in Their Hosts. Genes-Basel *11*, 761.

929 Meckiff, B.J., Ramírez-Suástegui, C., Fajardo, V., Chee, S.J., Kusnadi, A., Simon, H.,
930 Eschweiler, S., Grifoni, A., Pelosi, E., Weiskopf, D., et al. (2020). Imbalance of regulatory and
931 cytotoxic SARS-CoV-2-reactive CD4+ T cells in COVID-19. Cell *183*, 1340-1353.e16.

932 Mercatelli, D., and Giorgi, F.M. (2020). Geographic and Genomic Distribution of SARS-CoV-2
933 Mutations. Front Microbiol *11*, 1800.

934 Mercatelli, D., Triboli, L., Fornasari, E., Ray, F., and Giorgi, F.M. (2020). coronapp: A Web
935 Application to Annotate and Monitor SARS-CoV-2 Mutations. J Med Virol *93*, 3238-3245.

936 Moderbacher, C.R., Ramirez, S.I., Dan, J.M., Grifoni, A., Hastie, K.M., Weiskopf, D., Belanger,
937 S., Abbott, R.K., Kim, C., Choi, J., et al. (2020). Antigen-specific adaptive immunity to SARS-
938 CoV-2 in acute COVID-19 and associations with age and disease severity. Cell *183*, 996-
939 1012.e19.

940 Monajemi, M., Woodworth, C.F., Zipperlen, K., Gallant, M., Grant, M.D., and Larijani, M.
941 (2014). Positioning of APOBEC3G/F Mutational Hotspots in the Human Immunodeficiency
942 Virus Genome Favors Reduced Recognition by CD8+ T Cells. Plos One *9*, e93428.

943 Ng, O.-W., Chia, A., Tan, A.T., Jadi, R.S., Leong, H.N., Bertoletti, A., and Tan, Y.-J. (2016).
944 Memory T cell responses targeting the SARS coronavirus persist up to 11 years post-infection.
945 Vaccine *34*, 2008–2014.

946 Olson, M.E., Harris, R.S., and Harki, D.A. (2018). APOBEC Enzymes as Targets for Virus and
947 Cancer Therapy. Cell Chem Biol *25*, 36–49.

948 Peng, Y., Mentzer, A.J., Liu, G., Yao, X., Yin, Z., Dong, D., Dejnirattisai, W., Rostron, T.,
949 Supasa, P., Liu, C., et al. (2020). Broad and strong memory CD4+ and CD8+ T cells induced by
950 SARS-CoV-2 in UK convalescent individuals following COVID-19. Nat Immunol *21*, 1336-
951 1345.

952 Peretti, A., Geoghegan, E.M., Pastrana, D.V., Smola, S., Feld, P., Sauter, M., Lohse, S., Ramesh,
953 M., Lim, E.S., Wang, D., et al. (2018). Characterization of BK Polyomaviruses from Kidney
954 Transplant Recipients Suggests a Role for APOBEC3 in Driving In-Host Virus Evolution. Cell
955 Host Microbe *23*, 628-635.e7.

956 Popa, A., Genger, J.-W., Nicholson, M.D., Penz, T., Schmid, D., Aberle, S.W., Agerer, B.,
957 Lercher, A., Endler, L., Colaço, H., et al. (2020). Genomic epidemiology of superspreading
958 events in Austria reveals mutational dynamics and transmission properties of SARS-CoV-2. Sci
959 Transl Med *12*, eabe2555.

960 Quadeer, A.A., Ahmed, S.F., and McKay, M.R. (2020). Epitopes targeted by T cells in
961 convalescent COVID-19 patients. BioRxiv 2020.08.26.267724.

962 Reynisson, B., Alvarez, B., Paul, S., Peters, B., and Nielsen, M. (2020). NetMHCpan-4.1 and
963 NetMHCIIpan-4.0: improved predictions of MHC antigen presentation by concurrent motif
964 deconvolution and integration of MS MHC eluted ligand data. Nucleic Acids Res *48*, W449-
965 W454.

966 Rice, A.M., Morales, A.C., Ho, A.T., Mordstein, C., Mühlhausen, S., Watson, S., Cano, L.,
967 Young, B., Kudla, G., and Hurst, L.D. (2020). Evidence for strong mutation bias towards, and
968 selection against, U content in SARS-CoV-2: implications for vaccine design. Mol Biol Evol *38*,
969 67-83.

970 Sadler, H.A., Stenglein, M.D., Harris, R.S., and Mansky, L.M. (2010). APOBEC3G Contributes
971 to HIV-1 Variation through Sublethal Mutagenesis. J Virol *84*, 7396–7404.

972 Saini, S.K., Hersby, D.S., Tamhane, T., Povlsen, H.R., Hernandez, S.P.A., Nielsen, M., Gang,
973 A.O., and Hadrup, S.R. (2021). SARS-CoV-2 genome-wide T cell epitope mapping reveals
974 immunodominance and substantial CD8+ T cell activation in COVID-19 patients. Sci Immunol
975 *6*, eabf7550.

976 Salter, J.D., Bennett, R.P., and Smith, H.C. (2016). The APOBEC Protein Family: United by
977 Structure, Divergent in Function. Trends Biochem Sci *41*, 578–594.

978 Schub, D., Klemis, V., Schneitler, S., Mihm, J., Lepper, P.M., Wilkens, H., Bals, R., Eichler, H.,
979 Gärtner, B.C., Becker, S.L., et al. (2020). High levels of SARS-CoV-2 specific T-cells with
980 restricted functionality in severe course of COVID-19. JCI Insight *5*, e142167.

981 Schulien, I., Kemming, J., Oberhardt, V., Wild, K., Seidel, L.M., Killmer, S., Sagar, Daul, F.,
982 Lago, M.S., Decker, A., et al. (2021). Characterization of pre-existing and induced SARS-CoV-
983 2-specific CD8+ T cells. Nat Med *27*, 78–85.

984 Sekine, T., Perez-Potti, A., Rivera-Ballesteros, O., Strålin, K., Gorin, J.-B., Olsson, A.,
985 Llewellyn-Lacey, S., Kamal, H., Bogdanovic, G., Muschiol, S., et al. (2020). Robust T cell
986 immunity in convalescent individuals with asymptomatic or mild COVID-19. Cell *183*, 158-
987 168.e14.

988 Seow, J., Graham, C., Merrick, B., Acors, S., Pickering, S., Steel, K.J.A., Hemmings, O.,
989 O'Byrne, A., Kouphou, N., Galao, R.P., et al. (2020). Longitudinal observation and decline of
990 neutralizing antibody responses in the three months following SARS-CoV-2 infection in humans.
991 Nat Microbiol *5*, 1598–1607.

992 Sette, A., and Crotty, S. (2021). Adaptive immunity to SARS-CoV-2 and COVID-19. Cell *184*,
993 861-880.

994 Sidney, J., Peters, B., Frahm, N., Brander, C., and Sette, A. (2008). HLA class I supertypes: a
995 revised and updated classification. BMC Immunol *9*, 1.

996 Sidney, J., Southwood, S., Moore, C., Oseroff, C., Pinilla, C., Grey, H.M., and Sette, A. (2013).
997 Measurement of MHC/Peptide Interactions by Gel Filtration or Monoclonal Antibody Capture.
998 Curr Protoc Immunol Chapter 18:Unit 18.3.

999 Simmonds, P. (2020). Rampant C→U Hypermutation in the Genomes of SARS-CoV-2 and
1000 Other Coronaviruses: Causes and Consequences for Their Short- and Long-Term Evolutionary
1001 Trajectories. mSphere *5*, e00408-20.

1002 Squires, K.D., Monajemi, M., Woodworth, C.F., Grant, M.D., and Larijani, M. (2015). Impact of
1003 APOBEC Mutations on CD8+ T Cell Recognition of HIV Epitopes Varies Depending on
1004 the Restricting HLA. Jaids J Acquir Immune Defic Syndromes *70*, 172–178.

1005 Stephens, D.S., and McElrath, M.J. (2020). COVID-19 and the Path to Immunity. Jama *324*,
1006 1279–1281.

1007 Tang, F., Quan, Y., Xin, Z.-T., Wrammert, J., Ma, M.-J., Lv, H., Wang, T.-B., Yang, H.,
1008 Richardus, J.H., Liu, W., et al. (2011). Lack of Peripheral Memory B Cell Responses in
1009 Recovered Patients with Severe Acute Respiratory Syndrome: A Six-Year Follow-Up Study. J
1010 Immunol *186*, 7264–7268.

1011 Tarke, A., Sidney, J., Kidd, C.K., Dan, J.M., Ramirez, S.I., Yu, E.D., Mateus, J., Antunes, R. da
1012 S., Moore, E., Rubiro, P., et al. (2021). Comprehensive analysis of T cell immunodominance and

1013 immunoprevalence of SARS-CoV-2 epitopes in COVID-19 cases. Cell Reports Medicine *2*,
1014 100204.

1015 Wang, R., Hozumi, Y., Zheng, Y.-H., Yin, C., and Wei, G.-W. (2020). Host Immune Response
1016 Driving SARS-CoV-2 Evolution. Viruses *12*, 1095.

1017 Weiskopf, D., Schmitz, K.S., Raadsen, M.P., Grifoni, A., Okba, N.M.A., Endeman, H., Akker,
1018 J.P.C. van den, Molenkamp, R., Koopmans, M.P.G., Gorp, E.C.M. van, et al. (2020). Phenotype
1019 and kinetics of SARS-CoV-2-specific T cells in COVID-19 patients with acute respiratory
1020 distress syndrome. Sci Immunol *5*, eabd2071.

1021 Wood, N., Bhattacharya, T., Keele, B.F., Giorgi, E., Liu, M., Gaschen, B., Daniels, M., Ferrari,
1022 G., Haynes, B.F., McMichael, A., et al. (2009). HIV Evolution in Early Infection: Selection
1023 Pressures, Patterns of Insertion and Deletion, and the Impact of APOBEC. Plos Pathog *5*,
1024 e1000414.

1025 Woolthuis, R.G., Dorp, C.H. van, Keşmir, C., Boer, R.J. de, and Boven, M. van (2016). Long-
1026 term adaptation of the influenza A virus by escaping cytotoxic T-cell recognition. Sci Rep *6*,
1027 33334.

1028 Wu, L.-P., Wang, N.-C., Chang, Y.-H., Tian, X.-Y., Na, D.-Y., Zhang, L.-Y., Zheng, L., Lan, T.,
1029 Wang, L.-F., and Liang, G.-D. (2007). Duration of Antibody Responses after Severe Acute
1030 Respiratory Syndrome. Emerg Infect Dis *13*, 1562–1564.

1031 Zhou, R., To, K.K.-W., Wong, Y.-C., Liu, L., Zhou, B., Li, X., Huang, H., Mo, Y., Luk, T.-Y.,
1032 Lau, T.T.-K., et al. (2020). Acute SARS-CoV-2 infection impairs dendritic cell and T cell
1033 responses. Immunity *53*, 864-877.e5.

1034

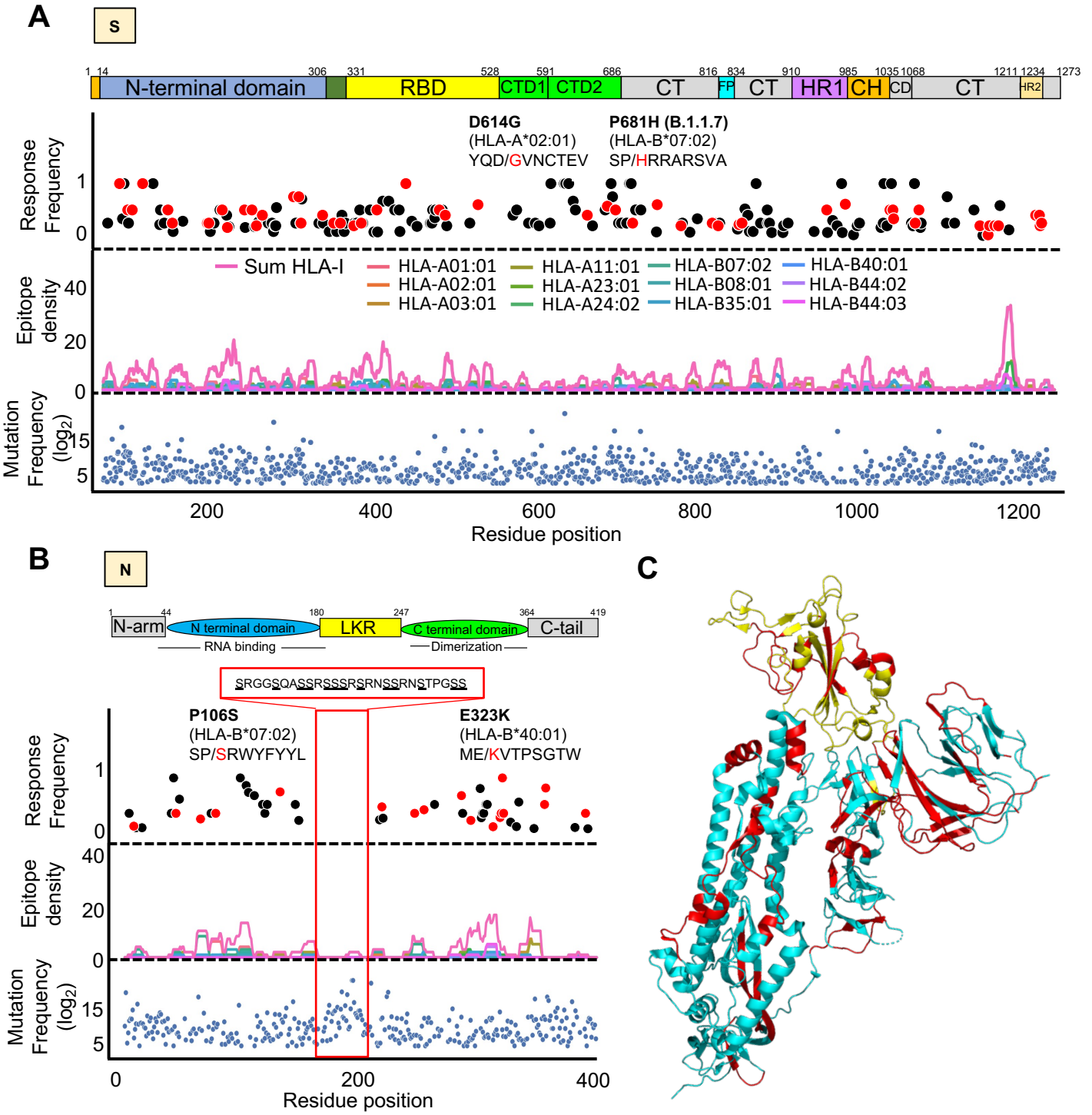


Figure 1. Distribution of CD8+ T cell epitopes and their mutated variants across the immunodominant S and N antigens. (**A, B**) Lower panel: blue dots showing all mutations that occurred in at least 4 SARS-CoV-2 genomes (GISAID). Middle panel: epitope density showing the overlap of HLA class I epitopes predicted within the 1st percentile for 12 queried HLA-I molecules. Upper panel: dots showing the frequency of CD8+ T cell response as determined from multiple studies aggregated in the database https://www.mckayspcb.com/SARS2TcellEpitopes as of January 2021. Red dots are mutated epitopes wherein the mutation event led to a predicted loss of binding. Sequences of specific epitopes are shown with the mutant amino acid in red. The red box in the N protein highlights a serine-rich region associated with no T cell response, low epitope density and high mutation frequency. (**C**) 3D structure of the Spike glycoprotein (Moderna Vaccine) and highlighted in yellow is the Receptor Binding Domain (Pfizer Vaccine). Shown in red are mutated epitopes wherein mutation events led to a predicted loss of HLA binding.

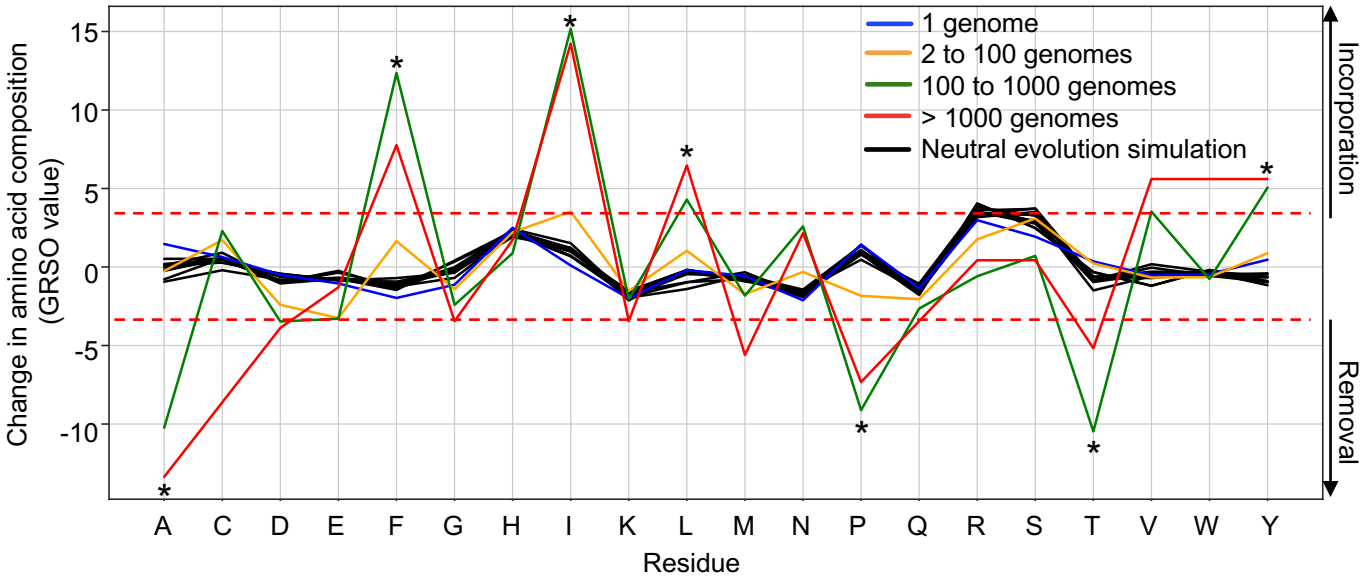
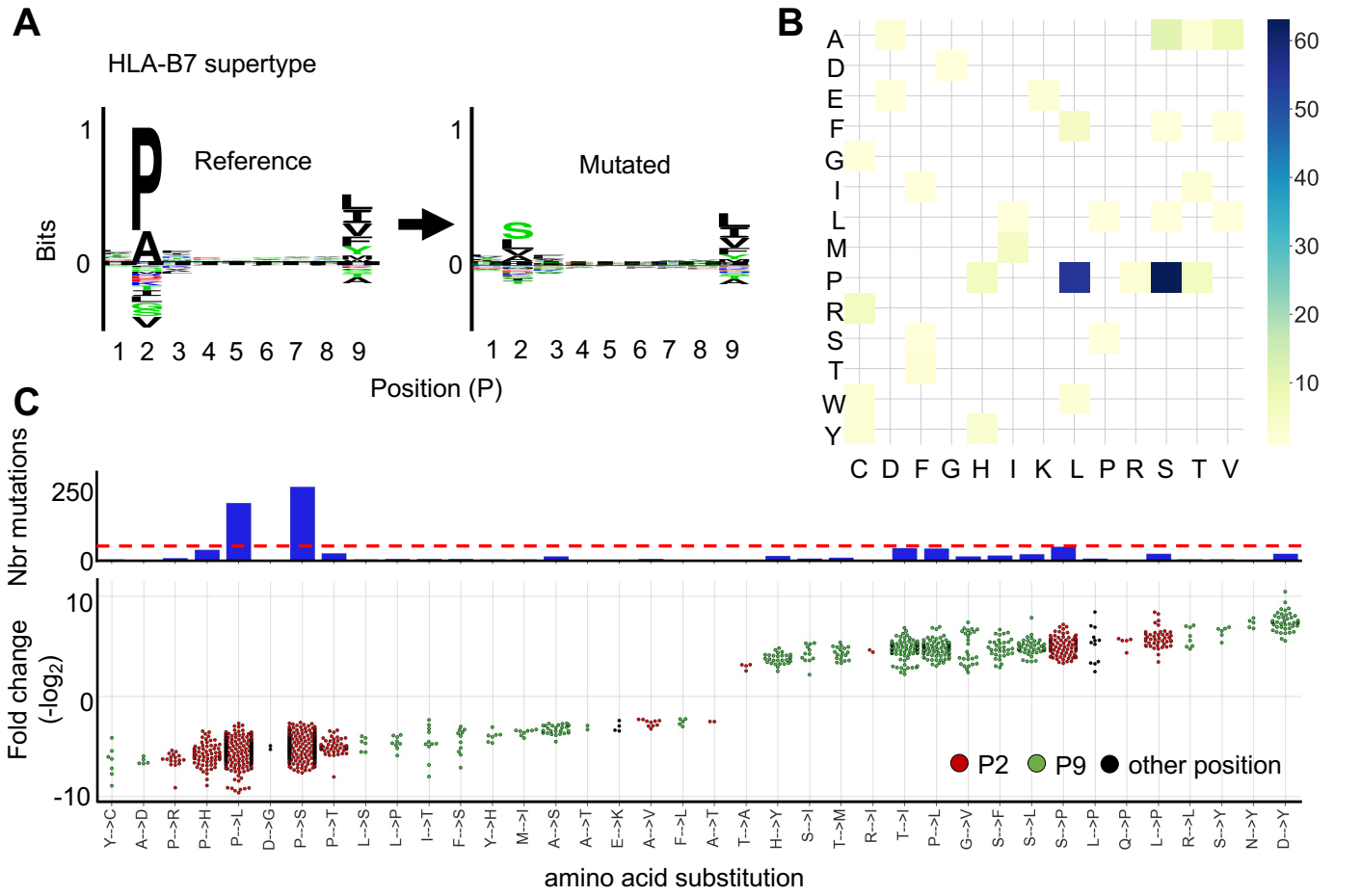


Figure 2. Global amino acid mutational biases in SARS-CoV-2 proteomes. A total of 330,246 SARS-CoV-2 genomes were translated into protein sequences and analyzed for the identification of any amino acid mutational bias. Amino acid residues (x-axis) that were removed and introduced in SARS-CoV-2 variants are presented by negative and positive %-difference in overall amino acid composition (GRSO values; y-axis), respectively. Analysis of mutational biases was performed for mutations occurring at various frequencies: 1 genome (blue line), 2 to 100 genomes (orange line), 100 to 1000 genomes (green line) and more than 1000 genomes (red line). Simulation of neutral evolution simulation (random mutations) were performed using the SANTA-SIM algorithm and serves as control for assessing the statistical significance of the observed pattern for individual amino acid residues. The dotted red lines show the cutoff values (fold change > 4; p-value < $1x10^{-11}$) that were used to define the residues that were preferentially removed or introduced (asterisk).



D

Protein	Mutation	No. of Genomes	Epitope	Substitution	HLA allele	NetMHCpan (reference)	NetMHCpan (mutated)	Magnitude of T cell response
S	P1263L	976	EP/LVLKGVKL	P → L	B*07:02	0.2	10.0	0.4
S	P681H	21,186	SP/HRRARSVA	P → H	B*07:02	0.06	5.5	0.571
N	P151L	4,488	NP/LANNAAIVL	P → L	B*67:01	0.2	6.0	0.2
N	P122H	171	LP/HYGANKDGI	P → H	B*51:02	0.3	26.0	0.333
N	P106S	2	SP/SRWYFYYL	P → S	B*07:02	0.1	4.7	0.857
N	P67S	9,791	FP/SRGQGVPI	P → S	B*55:02	0.06	4.1	0.227
ORF1a	P3613S	104	LP/SFAMGIIAM	P → S	B*56:01	0.4	11.0	0.5
S	P34S	443	LP/SPAYTNSF	P → S	B*51:01	0.1	5.6	0.2
ORF3a	P42L	171	LP/LFGWLIV	P → L	B*51:01	0.1	18.0	0.5
N	P326S	248	VP/SSGTWLTY	P → S	B*53:01	0.3	3.8	1
ORF1b	P2533L	873	NP/LIQLSSYSL	P → L	B*07:02	0.4	9.0	0.5

Figure 3. Mutation of proline at the anchor residue position for B7 supertype-associated epitopes. (**A**) (Left panel) Motif view of SARS-CoV-2 reference peptides predicted to bind B7 supertype molecules (HLA-B*07:02, -B*35:03, -B42:02 , -B*5101, -B*53:01, -B*54:01, -B*55:01, -B*56:01, -B*67:01). (Right panel) Motif view of the corresponding mutated peptides. (**B**) Heat map showing the frequency of specific amino acid substitutions between reference and mutated peptides. (**C**) Graph showing the number of mutations (upper panel; y-axis) leading to specific amino acid substitutions (x-axis) at anchor residue positions P2 (red dots) and P9 (green dots) or elsewhere (black dots). Dotted red line indicate the cutoff used to define dominant substitutions. The lower panel shows fold changes for individual amino acid substitutions. (**D**) Representative examples of validated CD8+ T cell epitopes (https://www.mckayspcb.com/SARS2TcellEpitopes as of January 2021). Effect of the P→X substitutions on predicted epitope binding affinities (NetMHCpan 4.1 EL %Rank) are shown. T cell response data for reference epitopes extracted from https://www.mckayspcb.com/SARS2TcellEpitopes.

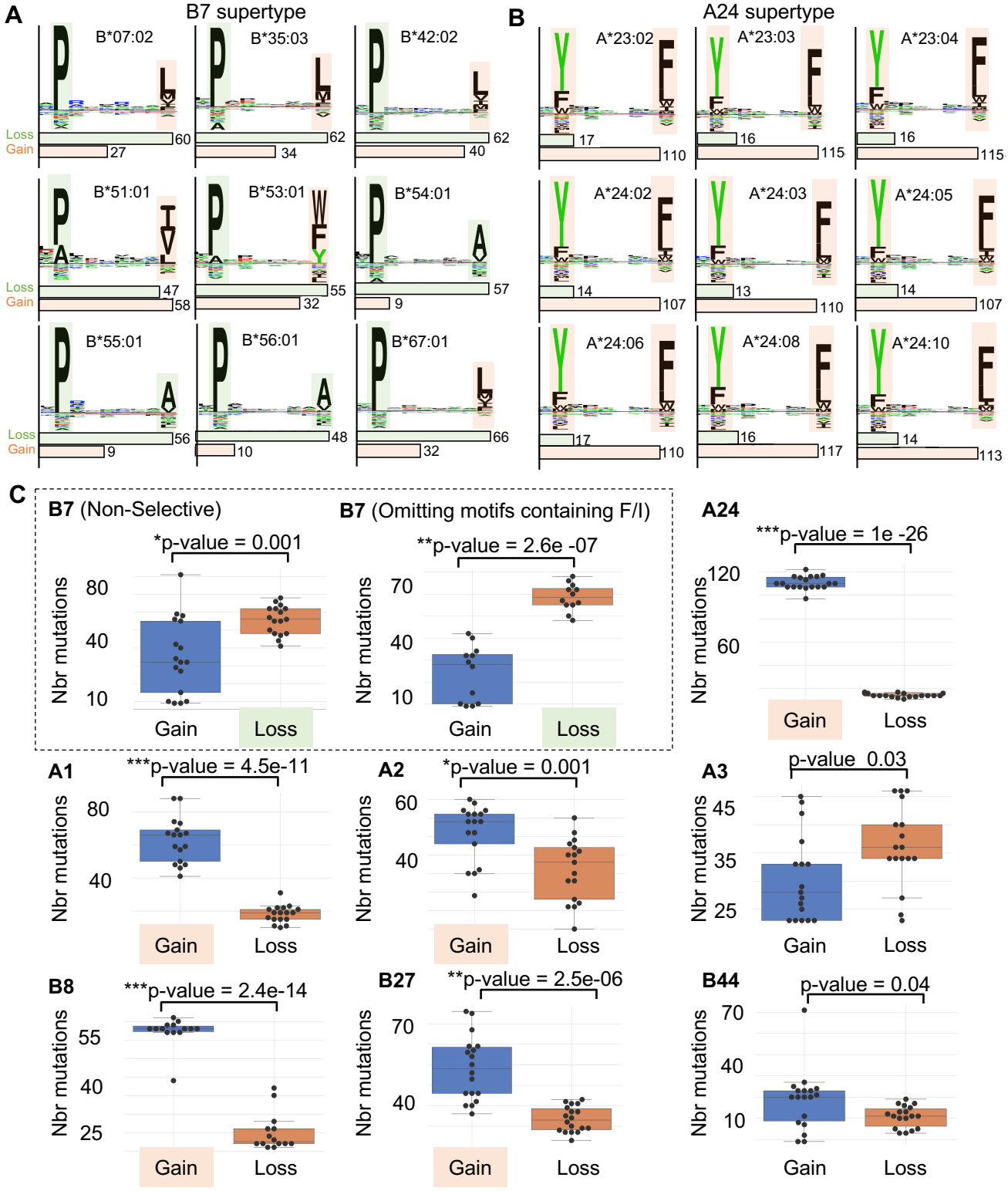


Figure 4. Loss or gain of SARS-CoV-2 mutated epitopes for different HLA class I supertypes. (**A, B**) Motif views showing established epitope binding motifs for different HLA-I alleles that belong to the HLA-B*07 (A) and HLA-A*24 (B) supertype family. Shaded squares highlight anchor residues that are preferentially removed (pale green) or introduced (pale orange) in SARS-CoV-2 proteomes (related to Figure 2), respectively. Histograms below the binding motifs indicate the number of frequent mutations (identified in at least 100 individuals) leading to the loss or gain of epitopes. (**C**) 'Gain/Loss plots' showing number of mutations (y-axis) leading to a preferentially loss (pale green) or gain (pale orange) of epitopes for different HLA class I supertypes. Each black dot represents the number of mutations associated with gain and loss of epitopes for a given HLA-I allele. Between 14 to 19 alleles per supertype (Figure S5) were used to generate the graphs and p-values (*p ≤ 0.001, **p < 1e-5, ***p < 1e-10).

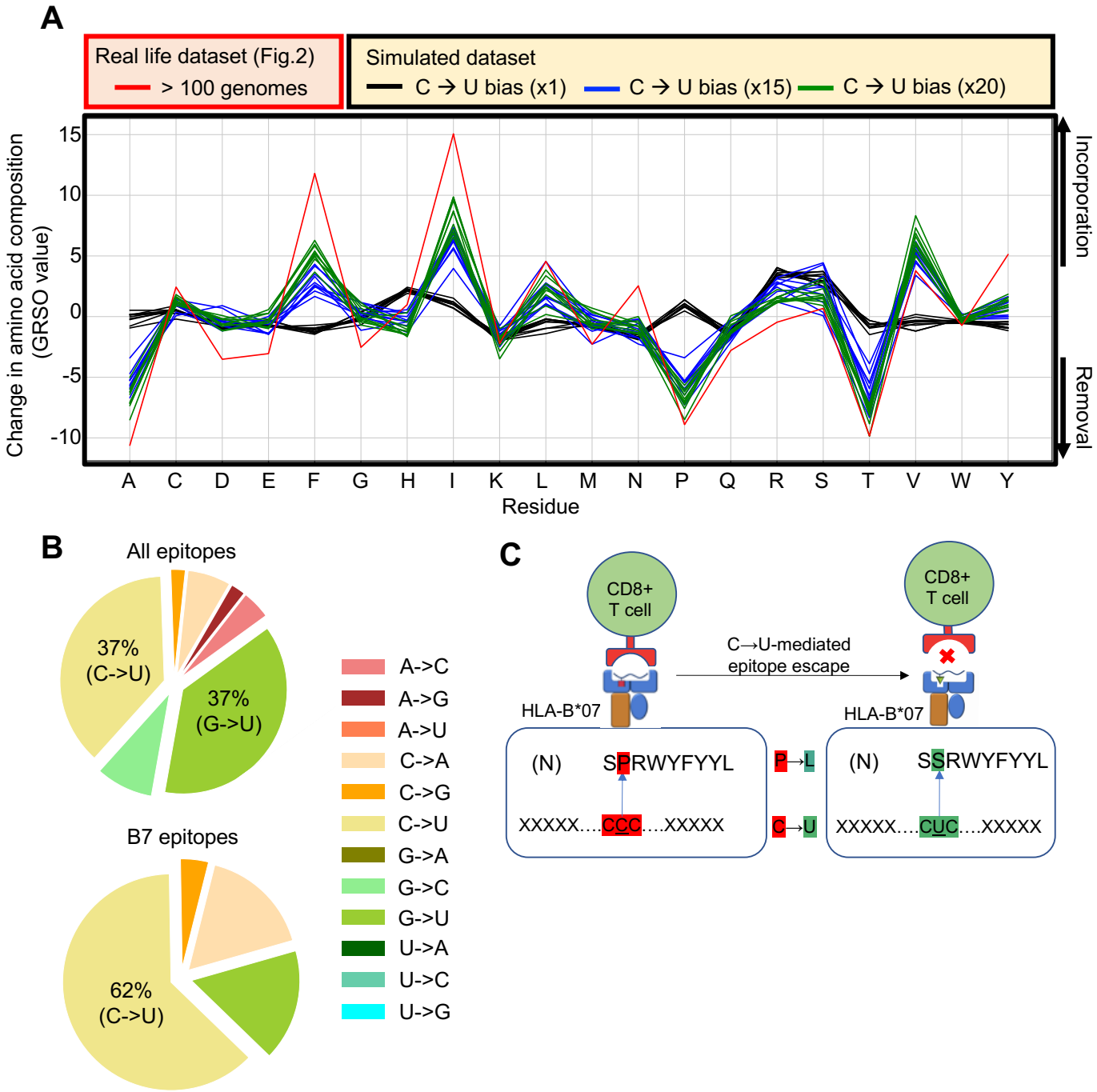
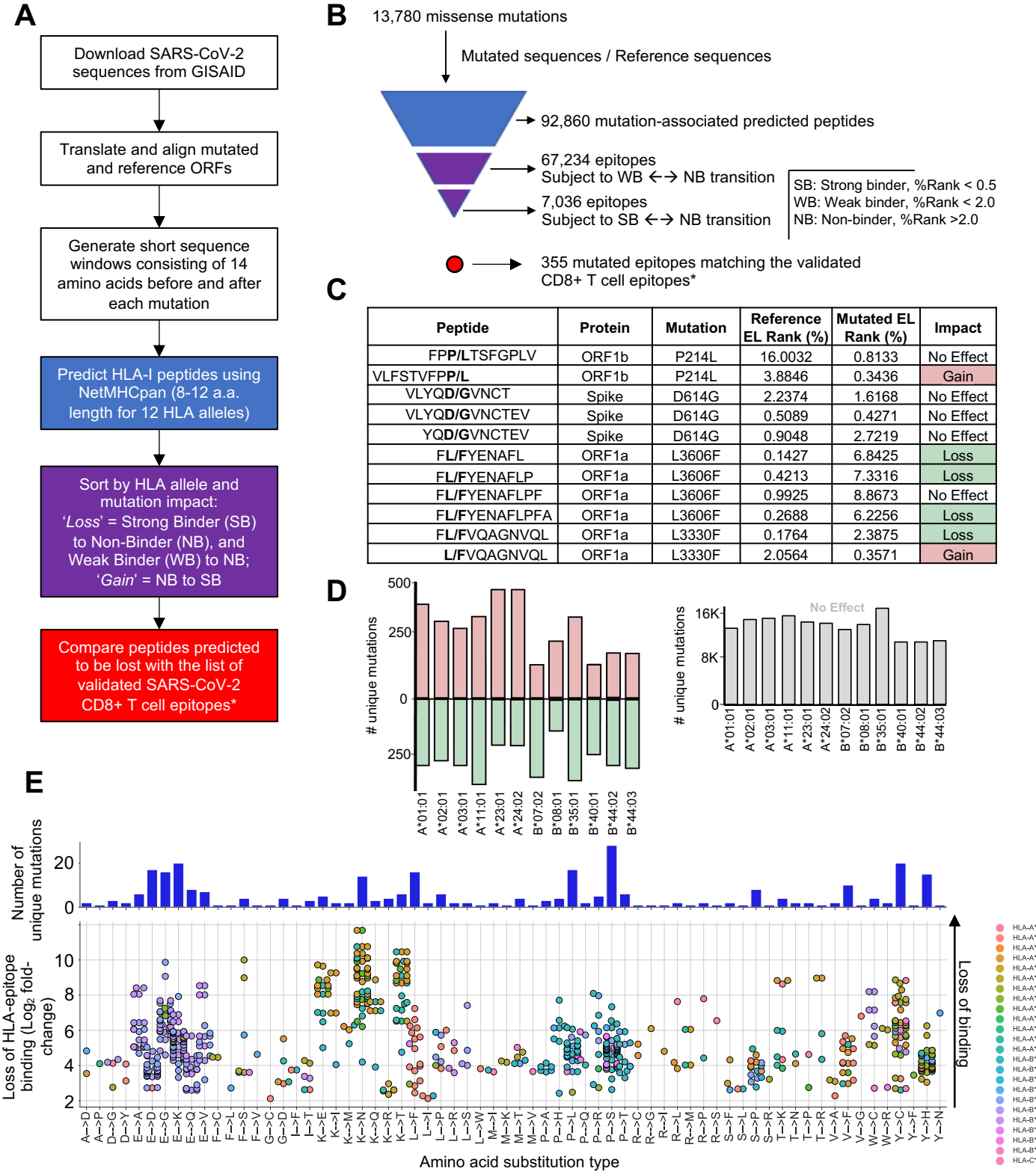


Figure 5. The C-to-U point mutation bias largely drives the diversity of SARS-CoV-2 proteomes and CD8+ T cell epitopes. (**A**) Comparison of global amino acid mutational patterns generated from real-life versus simulated SARS-COV-2 genomes. Amino acid residues (x-axis) that were removed and introduced in real-life versus simulated SARS-CoV-2 are presented by negative and positive %-difference in overall amino acid composition (GRSO values; y-axis), respectively. Evolution of SARS-CoV-2 was simulated by introducing various extents of C-to-U biases, i.e. x1, x15 and x20 (n = 10). The red line shows the pattern obtained from mutations identified in more than 100 SARS-CoV-2 genomes, related to Figure 2. (**B**) (Top) Pie chart showing the proportion of nucleotide substitution types from the list of validated CD8+ T cell epitopes in https://www.mckayspcb.com/SARS2TcellEpitopes as of January 2021. (Bottom) Pie chart showing the proportion of nucleotide substitution types from the list of validated CD8+ T cell epitopes specific to the B7 supertype. (**C**) Schematic illustrating the C-to-U-mediated epitope escape model. The observed mutation of the immunodominant SPRWYLFYYL epitope in the N protein is shown as an example.



Peptide	Protein	Mutation	Reference EL Rank (%)	Mutated EL Rank (%)	Impact
FP**P/L**TSFGPLV	ORF1b	P214L	16.0032	0.8133	No Effect
VLFSTVFP**P/L**	ORF1b	P214L	3.8846	0.3436	Gain
VLYQ**D/G**VNCT	Spike	D614G	2.2374	1.6168	No Effect
VLYQ**D/G**VNCTEV	Spike	D614G	0.5089	0.4271	No Effect
YQ**D/G**VNCTEV	Spike	D614G	0.9048	2.7219	No Effect
F**L/F**YENAFL	ORF1a	L3606F	0.1427	6.8425	Loss
F**L/F**YENAFLP	ORF1a	L3606F	0.4213	7.3316	Loss
F**L/F**YENAFLPF	ORF1a	L3606F	0.9925	8.8673	No Effect
F**L/F**YENAFLPFA	ORF1a	L3606F	0.2688	6.2256	Loss
F**L/F**VQAGNVQL	ORF1a	L3330F	0.1764	2.3875	Loss
L/FVQAGNVQL	ORF1a	L3330F	2.0564	0.3571	Gain

Figure S1. Impact of SARS-CoV-2 mutations on CD8+ T cell epitopes, Related to Figure 1 and 4. (**A**) Bioinformatic pipeline for the prediction of SARS-CoV-2 mutated class I peptides associated to 12 common HLA alleles. (**B**) Pyramidal graph showing the number of i) missense mutations in SARS-CoV-2 genomes, ii) predicted class I mutated peptides, iii) predicted class I peptides subject to Weak Binder (WB) to Non-Binder (NB) and Strong Binder (SB) to NB transition (epitope loss category), and iv) predicted class I mutated peptides matching reference CD8+ T cell epitopes that have been experimentally validated. (**C**) Representative examples of predicted class I mutated peptides and the impact of the identified amino acid mutation (bold) on peptide binding to a given HLA-I allele. Reference and mutated EL (eluted ligand) Rank (%) generated by NetMHCpan 4.1 EL is indicated for individual predictions. Gain = NB to SB (pale red); Loss = SB to NB (pale green). (**D**) Left panel: number of unique mutations leading to 'Gain' or 'Loss' of class I peptides for the indicated HLA-I alleles. Right panel: number of unique mutations showing no effect on peptide binding for the indicated HLA-I alleles. (**E**) Validated SARS-CoV-2 CD8+ T cell epitopes (McKay Database) subjected to mutation events detected in more than 4 individuals (GISAID) and predicted lead to a strong loss of HLA-epitope binding. Top: number of unique missense mutations corresponding to the indicated amino acid substitution type. Bottom: Predicted loss of HLA-epitope binding (NetMHCpan4.1 %Rank) corresponding to the indicated residue substitution type from the list of validated CD8+ T cell epitopes in the McKay Database. Each dot represents an epitope pair (mutated / reference). Color indicates HLA type affected by the mutations.

A

HLA	Genome location ref>mut	Mutation ID	number of GISAID entries carrying mutation (n = 68,031)	peptide (reference/mutated)	substitution	Reference (nM)	Mutated (nM)
A*02:01	23403A>G	S_D614G	50186	YQDVNCTEV / YQGVNCTEV	D -> G	166	2423
A*02:01	11083G>T	ORF1a_L3606F	6739	FLYENAFL / FFYENAFL	L -> F	36	1298
A*02:01	11083G>T	ORF1a_L3606F	6739	TQWSLFFFL / TQWSLFFFF	L -> F	83	9088
A*02:01	25528C>T	ORF3a_L46F	209	WLIVGVALL / WFIVGVALL	L -> F	163	15161
A*02:01	11417G>T	ORF1a_V3718F	90	WTLMNVLTLV / WTLMNVLTLF	V -> F	47	1098
A*02:01	25831C>T	ORF3a_L147F	45	NPLLYDANYFL / NPLLYDANYFF	L -> F	2221	5041
A*02:01	11417G>T	ORF1a_V3718F	90	TLMNVLTLV / TLMNVLTLF	V -> F	50	682
A*02:01	25831C>T	ORF3a_L147F	45	LLYDANYFL / LLYDANYFF	L -> F	3.6	36
A*02:06	25563G>T	ORF3a_Q57H	14981	FQSASKIITL / FHSASKIITL	Q -> H	572	749
A*11:01	24781G>T	S_K1073N	124	VTYVPAQEK / VTYVPAQEN	K -> N	28	50000
A*11:01	25593G>C	ORF3a_K67N	47	ASKIITLKK / ASKIITLKN	K -> N	23	6494
B*07:02	28881G>A	N_R203K	20893	SSRGTSPARM / SSKGTSPARM	P -> L	5445	50000
B*07:02	17747C>T	ORF1b_P1327L	1991	NPAWRKAVF / NLAWRKAVF	P -> L	11	840
B*07:02	28311C>T	N_P13L	1252	APRITFGGP / ALRITFGGP	P -> L	45	976
B*07:02	25350C>T	S_P1263L	538	SEPVLKGVKL / SELVLKGVKL	P -> L	1397	50000
B*07:02	25350C>T	S_P1263L	538	EPVLKGVKL / ELVLKGVKL	P -> L	942	50000
B*08:01	21624G>T	S_R21I	431	NLTTRTQL / NLTTITQL	R -> I	14	869

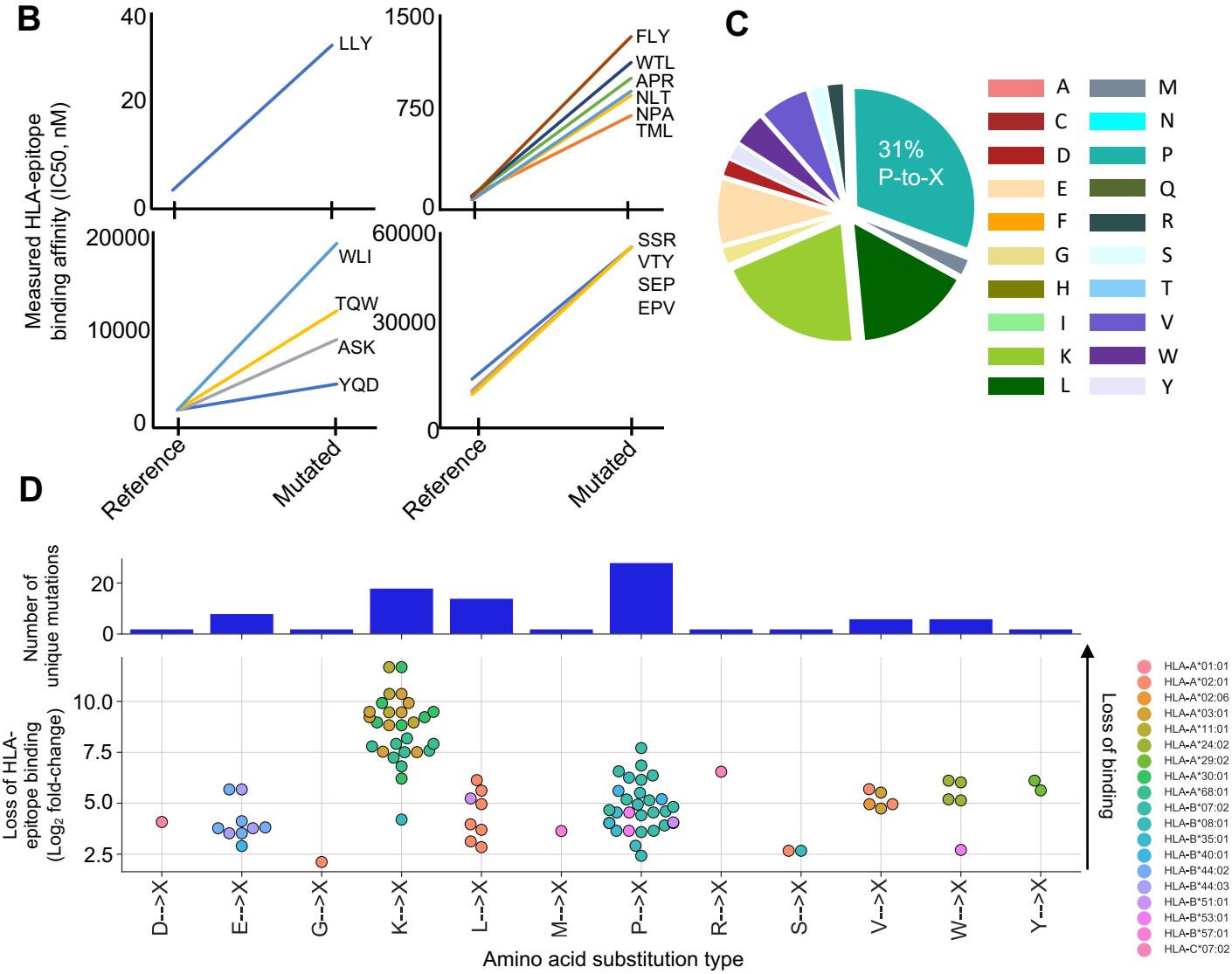


Figure S2. HLA peptide binding measurements and mutational biases in SARS-CoV-2 mutated epitopes, Related to Figure 1. (**A**) HLA binding assay was performed to determine the in vitro binding affinity (nM) of representative SARS-CoV-2 peptides for specific HLA class I alleles. Peptides were selected based on 1) frequency of mutations, 2) presentation by common HLA class I alleles, and 3) the mutated form was predicted to lose binding to its corresponding HLA. (**B**) Plots showing raw values for the binding affinities (nM) of the reference vs mutated peptides in (A). The first three amino acid residues of the reference peptides with fold change > 2.5 are shown. (**C**) Pie chart showing the proportion of X-to-Y substitution types from the list of validated CD8+ T cell epitopes in https://www.mckayspcb.com/SARS2TcellEpitopes/ (as of January 2021). (**D**) Predicted loss of HLA-epitope binding clustered by substitution type from the list of validated CD8+ T cell epitopes in the McKay database. Each dot represents an epitope pair (mutated / reference; NetMHCpan 4.1 %rank ratio).

A

Epitope	HLA	Mutation	No genomes (GISAID)	NetMHCpan %rank (reference)	NetMHCpan %rank (mutated)
GL/FMWLSYFI	A02:01	M-L90F	38	0.5	11.9
ME/KVTPSGTWL	B40:01	N-E323K	23	0.3	10.1

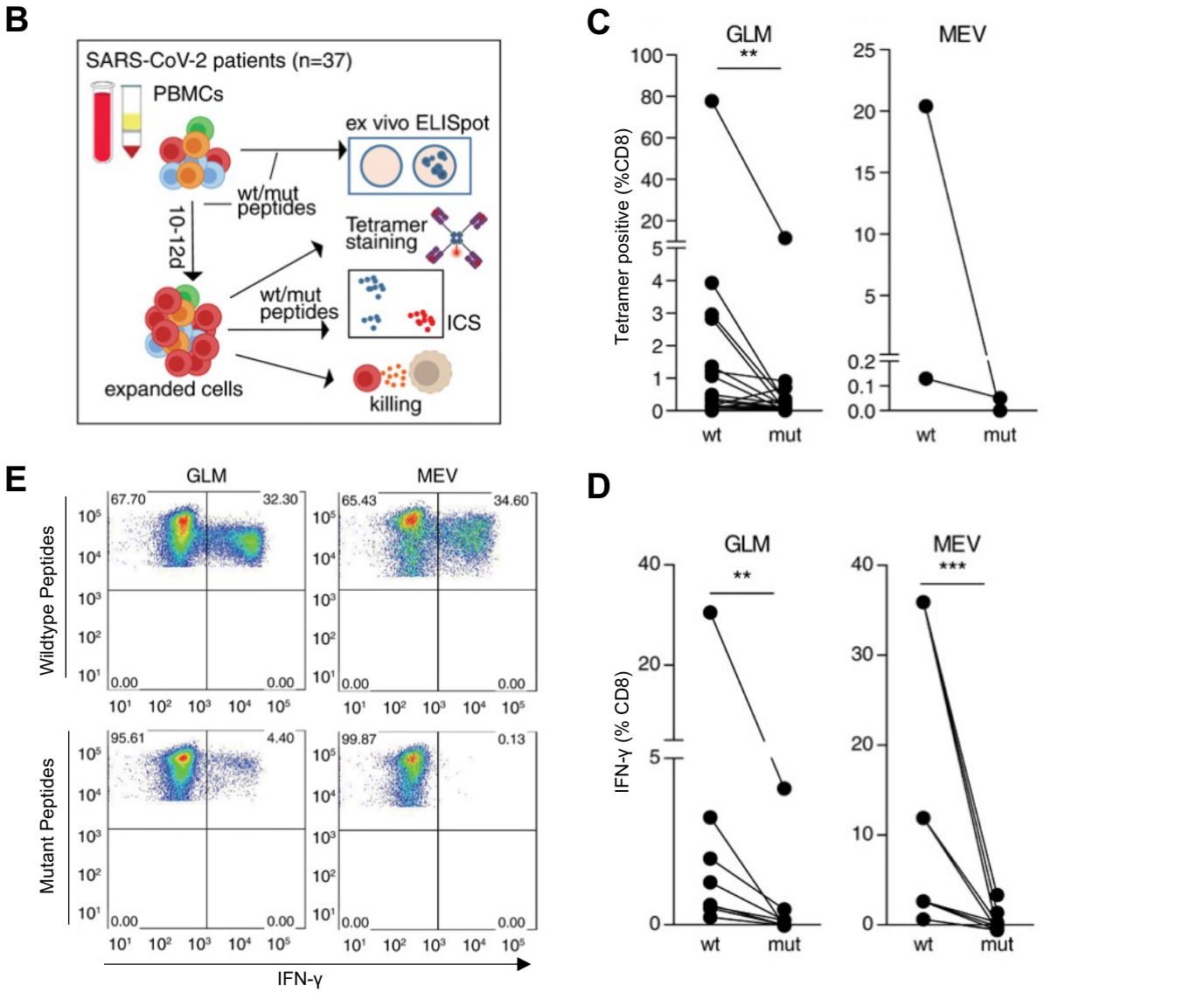


Figure S3. Identification of two SARS-CoV-2 mutated epitopes in this study that were previously associated with decreased CD8+ T cell responses, Related to Figure 1. (**A**) The mutated epitopes GFMWLSYFI (A*02) and MKVTPSGTWL (B*40) were detected in 38 and 23 genomes/individuals in this study (GISAID) and their T cell immunogenicity was thoroughly investigated in Agerer et al. (**B-E from Agerer et al., copyright 2021, with permission from AAAS)** (**B**) Experimental overview. (**C**) T cells expanded with mutant peptides do not give rise to wild type peptide-specific CD8+ T cell. PBMCs were isolated from HLA-A*02:01 or HLA-B*40:01 positive SARS-CoV-2 patients, stimulated with wild type or mutant peptides and stained with tetramers containing the wild type peptide. (**D**) Impact of mutations on CD8+ T cell response. PBMCs expanded with wild type or mutant peptides as indicated, were analyzed for IFN-γ-production via ICS after restimulation with wild type or mutant peptide. (**E**) Representative FACS plots for (D).

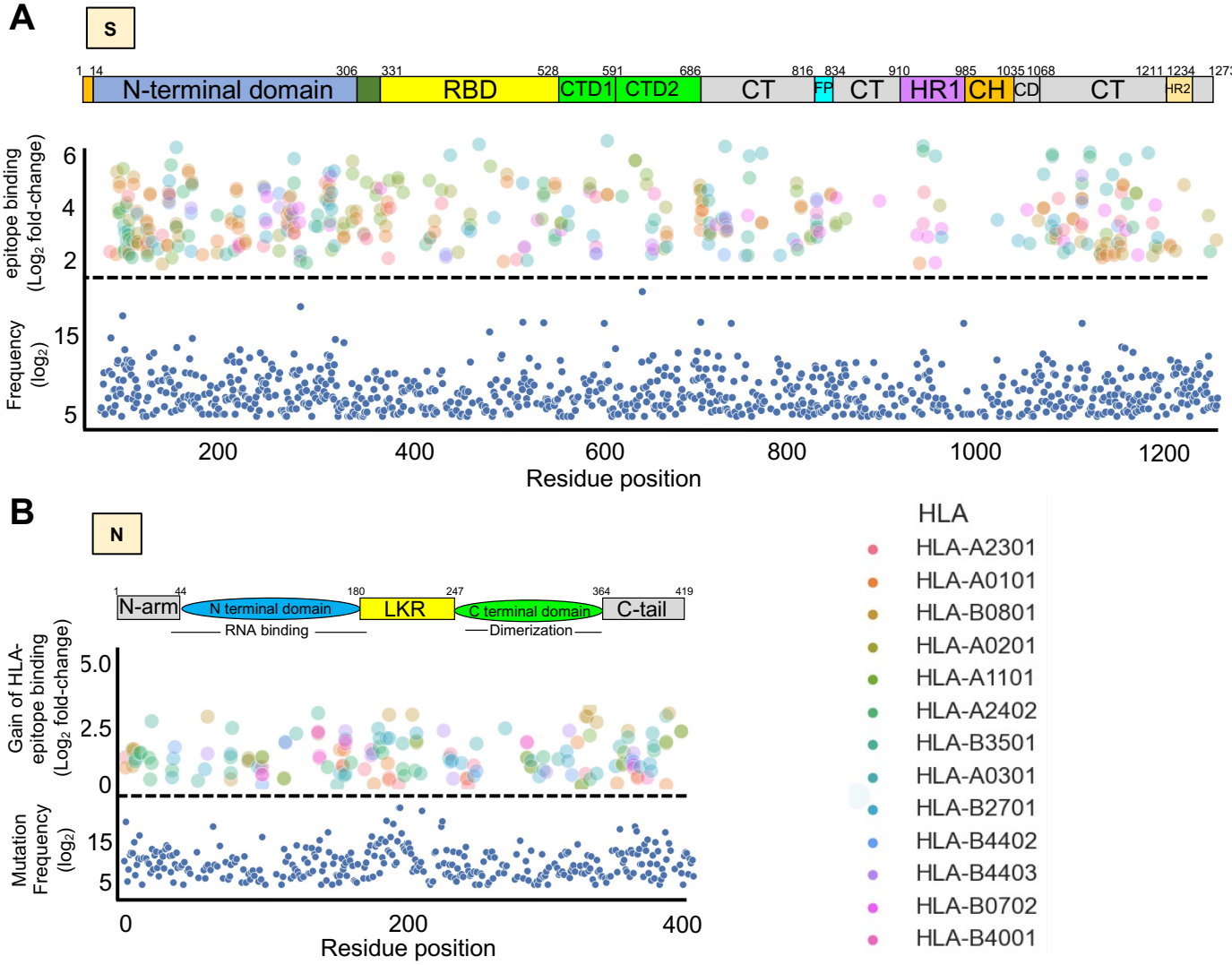
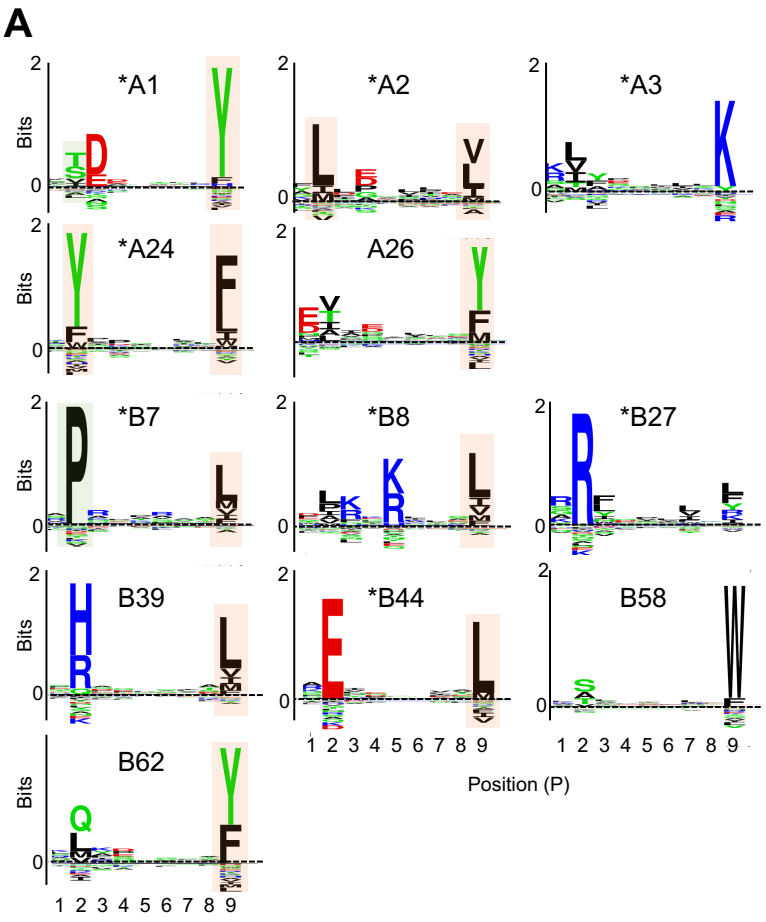


Figure S4. Impact of mutations on gain of peptide binding to various HLA class I molecules across the immunodominant S and N antigens, Related to Figure 1. (**A, B**) Lower panel: blue dots showing all mutations that occurred in at least 4 SARS-CoV-2 genomes (GISAID). Upper panel: dots showing predicted peptides subjected to a strong gain of binding (see also Figure S1C,D) to one of 12 highly common HLA types queried (color coded) due to a mutation.



B

A1		A2	
A*01:01	A*30:02	A*02:01	A*02:14
A*01:03	A*30:03	A*02:02	A*02:16
A*01:04	A*30:04	A*02:03	A*02:17
A*01:07	A*30:06	A*02:04	A*02:18
A*26:01	A*30:09	A*02:05	A*02:30
A*26:02	A*32:01	A*02:06	A*02:36
A*26:03	A*32:02	A*02:07	A*68:02
A*26:04	A*32:05	A*02:13	A*69:01
A*26:05		A*02:45	

A3		A24	
A*03:01	A*11:05	A*23:02	A*24:13
A*03:02	A*31:01	A*23:03	A*24:18
A*03:04	A*33:03	A*23:04	A*24:20
A*03:05	A*66:01	A*23:06	A*24:21
A*03:06	A*66:02	A*24:02	A*24:22
A*11:01	A*68:01	A*24:03	A*24:23
A*11:02	A*68:03	A*24:05	A*24:26
A*11:03	A*74:01	A*24:06	A*24:27
A*11:04		A*24:08	A*24:03
		A*24:10	

B7		B8	
B*07:02	B*53:01	B*08:01	B*08:20
B*07:05	B*54:01	B*08:02	B*08:21
B*07:08	B*55:01	B*08:07	B*08:22
B*15:08	B*55:02	B*08:09	B*08:23
B*35:03	B*56:01	B*08:11	B*08:24
B*42:02	B*67:01	B*08:13	B*08:25
B*51:01	B*78:01	B*08:15	
B*51:02		B*08:18	
B*51:03			

B27		B44	
B*14:02	B*27:05	B*15:53	B*40:06
B*15:03	B*27:06	B*18:01	B*40:16
B*15:09	B*27:07	B*18:03	B*44:02
B*15:10	B*27:09	B*18:05	B*44:03
B*15:18	B*38:01	B*18:06	B*44:04
B*27:01	B*39:01	B*37:01	B*44:07
B*27:02	B*39:02	B*37:04	B*44:13
B*27:03	B*39:09	B*40:01	B*45:01
B*27:04	B*48:01	B*40:02	B*45:03

Figure S5. HLA class I supertypes, Related to Figure 4. (**A**) Epitope binding motifs for several HLA class I supertypes. Anchor residues are located at P2 and P9. Pale orange and green squares cover amino acid residues that are preferentially introduced (F, I, L, Y) and removed (A, P, T) in SARS-CoV-2 proteomes, respectively. Representative supertypes used in this study are shown by an asterisk. Epitope binding motifs were extracted from NetMHCpan Motif Viewer (http://www.cbs.dtu.dk/services/NetMHCpan/logos_ps.php). (**B**) Table showing the selected alleles per supertype that were used in this study to generate the 'Gain/Loss plots'.

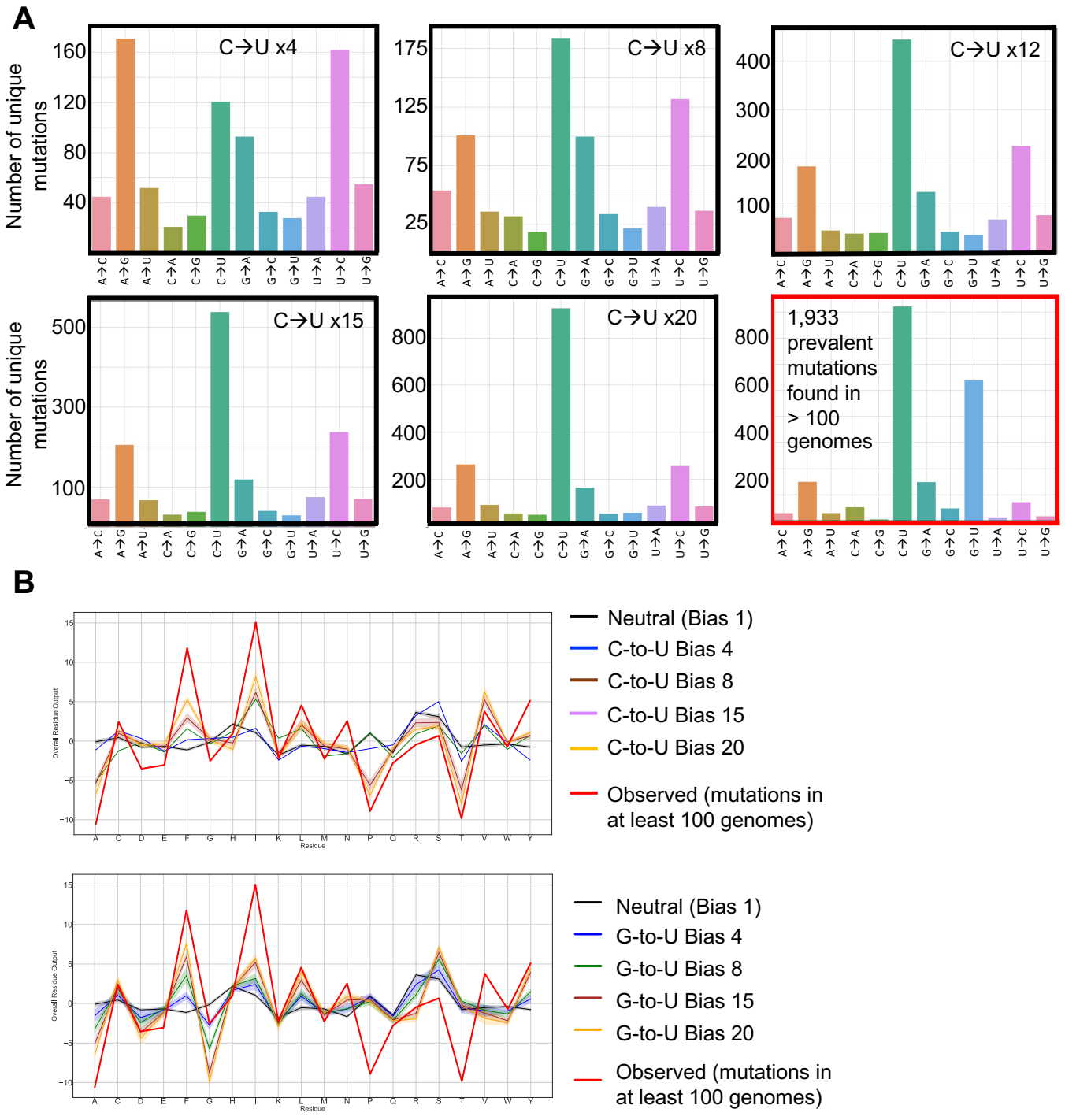


Figure S6. Comparison of mutation biases between real-life/observed and simulated data, Related to Figure 5. (**A**) Histograms showing the number of unique mutations identified for each mutation type (A-to-C, A-to-G, etc.) after simulating the evolution of SARS-CoV-2 genomes through the introduction of different C-to-U bias values (x4 to x20) using the SANTA-SIM software. Simulated (black squares) and real-life/observed prevalent mutations found in more than 100 genomes (red square) at the nucleotide level are shown. (**B**) Comparison of global amino acid mutational patterns generated from simulated versus real-life/observed SARS-COV-2 genomes. Various extents of C-to-U (top) and G-to-U (bottom) biases were introduced to perform the simulation and to generate the graphs.