Université de Montréal

# Characterizing the impact of the mutational landscape of SARS-CoV-2 on epitope presentation and CTL

Par

David Hamelin

Département de biochimie et Médecine Moléculaire

Faculté de Médecine

Submitted in Partial Fulfillment of the Requirements for the degree of

Master of Science (M.Sc.) in

Bioinformatics

December 2021

Université de Montréal

Faculté des études supérieures et postdoctorales

This dissertation, titled

# Characterizing the impact of the mutational landscape of SARS-CoV-2 on epitope presentation and CTL

Presented by

David Hamelin

Was evaluated by a jury composed of the following members:

Etienne Caron
Supervisor

Julie Hussin
Co-Supervisor

Adrian Serohijos
Président-Rapporteur

Luis Barreiro
Jury Member

**Résumé**

La pandémie actuelle de COVID-19, causée par le coronavirus 2 du syndrome respiratoire aigu sévère (SRAS-CoV-2), a entraîné plus de 6 millions de décès et près de 680 millions de cas confirmés dans le monde. Depuis l'émergence du virus en décembre 2019, beaucoup d'efforts de recherche mondiaux ont visé à étudier la relation entre le SRAS-CoV-2 et l'immunité adaptative à médiation cellulaire. La caractérisation des réponses immunitaires à base de lymphocytes T CD4+ et CD8+ contre le SRAS-CoV-2 dans le contexte de mutations virales est d'une pertinence immédiate pour l'approfondissement de nos connaissances concernant l'immunité adaptative envers un virus en évolution, ainsi que l'amélioration de vaccins. Dans cette thèse, je passerai en revue les découvertes actuelles concernant la biologie du SRAS-CoV-2 et sa relation avec le système immunitaire adaptatif humain. Je discuterai ensuite les divers mécanismes par lesquels le SRAS-CoV-2, ainsi que d'autres virus, se sont avérés échapper l'immunité adaptative humoral et cellulaire. Enfin, je présenterai mes contributions à la compréhension du paysage mutationnel global du SRAS-CoV-2 et de sa capacité à échapper à la reconnaissance par les lymphocytes T CD8+. Dans ce travail, j'ai observé que le paysage mutationnel global du SRAS-CoV-2 était régi par des biais de mutation au cours de la première année de la pandémie, le plus répandu d'entre eux conduisant à la suppression de la proline. Il a ensuite été prédit que cette élimination globale de la proline conduirait à la perte d'épitopes reconnues par les cellules T CD8+ d'une manière dépendante sur les super-types HLA, avec la perte d'épitopes survenant préférentiellement dans le contexte du super-type HLA-B7. Le modèle développé dans ce travail propose un lien entre les biais mutationnels globaux du SRAS-CoV-2, les allèles HLA et l'évasion des lymphocytes T. Ce travail crée un cadre pour anticiper l'impact des variantes existantes et émergentes du SRAS-CoV-2 envers la réponse immunitaire à base de lymphocytes T CD8+.

Mots clés : Génomique, Virologie, Cellules T, Épitopes, SRAS-CoV-2, immunopeptidomique, Evasion Immunitaire

**Abstract**

The current COVID-19 pandemic, caused by Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2), has led to upwards of 6 million deaths and nearly 680 million confirmed cases worldwide. Since the emergence of the virus in December 2019, astounding global research efforts have been aimed at investigating the relationship between SARS-CoV-2 and cell-mediated adaptive immunity. Characterizing CD4+ and CD8+ T Lymphocyte responses to SARS-CoV-2 in the context of viral mutations is of immediate relevance to understanding the breadth of a population's adaptive immunity to an evolving virus and is central to the improving existing vaccines. In this thesis, I will review all present findings pertaining to the biology of SARS-CoV-2 and its relationship with the human adaptive immune system. I will then discuss the various mechanisms by which SARS-CoV-2, along with other viruses, have been found to evade the various arms of the adaptive immune system. Finally, I will present my contributions to the understanding of the global mutational landscape of SARS-CoV-2 and its ability to evade recognition by CD8+ T lymphocytes. By investigating over 300,000 SARS-CoV-2 genomic sequences, I observed that the mutational landscape of SARS-CoV-2 was governed by mutation biases during the first year of the pandemic, with the most prevalent bias leading to the removal of proline. The observed global removal of Proline was predicted to lead to the loss of CD8+ T cell epitopes in an HLA-supertype-dependent manner, with the loss of epitopes occurring preferentially in the context of the HLA-B7 supertype. The model developed proposes a link

between SARS-CoV-2 global mutational biases, HLA alleles and T cell evasion. This work creates a framework to anticipate the population-specific impact of existing and emerging SARS-CoV-2 variants on CD8+ T cell-based immunity.

Key words: Genomics, Virology, T cells, Epitopes, SARS-CoV-2, Immunopeptidomics, Immune Escape

**Table of Contents**

**List of Tables**

**List of figures**

**Acronyms and Abbreviations**

Ab : Antibody

AIDS: Acquired Immunodeficiency Syndrome

AIM : Activation-Induced Assay

AIM-2 : Absent In Melanoma 2

APC : Antigen Presenting Cells

BA : Binding Affinity

CD (CD8; CD4): Cluster Of Differentiation

CD : Cytoplasm Domain

COVID-19 : Coronavirus Disease 2019

CTD : C-Terminal Domain

CTL : Cytotoxic T lymphocytes

DAMPs : Damage-Associated Molecular Patterns

DC : Dendritic Cells

DDA : Data-Dependant Acquisition

DIA : Data-Independant Acquisition

DNA : Desoxyribonucleic acid

EL : Eluted Ligand

ELISpot : Enzyme-Linked Immunosorbent Spots

ER : Endoplasmic Reticulum

FACS : Fluorescence-Activated Cell Sorting

GISAID : Global Initiative on Sharing Avian Influenza Data

HIV : Human Immunodeficiency Virus

HLA : Human Leukocyte Antigen

H5N1 : hemagglutinin (5) – neuraminidase (1)

ICS : Intracellular Cycokine Staining

IFN (IFN-y) : Interferons

LGP2 : Laboratory of Genetics and Physiology 2

IL : Interleukin

MAVS : Mitochondrial Antiviral Signaling

MDA5 : Melanoma Differentiation-Associated Protein 5

MERS-CoV : Middle East Respiratory Syndrome Coronavirus

MHC : Major Histocompatibility Complex

NGS : Next-Generation Sequencing

NLR : Nucleotide-Binding Oligomerization (NOD)-Like Receptors

NLRP3 : NLR Family Pyrin Domain Containing 3

NOD1 : Nucleotide-Binding Oligomerization Domain-Containing Protein 1

NTD : N-Terminal Domain

ORF : Open-Reading Frame

PAMPs : Pathogen-Associated Molecular Patterns

PBMC : Peripheral Blood Mononuclear Cell

PLC : Peptide-Loading Complex

PRR : Pattern-Recognition Receptor

RBD : Receptor-Binding Domain

RLR : Retinoic Acid-Inducible Gene I (RIG-I)-Like Receptors

RNA : Ribonucleic Acid

SARS-CoV : Severe Acute Respiratory Syndrome Coronavirus

SARS-COV-2 : Severe Acute Respiratory Syndrome Coronavirus 2

scRNA-seq : Single-Cell Ribonucleic Acid Sequencing

scTCR-seq : Single-Cell T Cell Receptor Sequencing

ssRNA : Single-Stranded Ribonucleic Acid

TAP : Transporter associated with Antigen Processing

TCR: T Cell Receptor

$T_h$ : T Helper Cell

TLR : Toll-Like Receptors

TMD : Transmembrane Domain

TNF-*a* : Tumour Necrosis Factor (Alpha)

VOC : Variant of Concern

VOI : Variant of Interest

VUM : Variant Under Monitoring

WHO : World Health Organization

**Acknowledgements**

I would like to thank my co-supervisors Etienne Caron and Julie Hussin, who were both instrumental in helping me define and refine my ideas, who provided me with critical advice, and who helped me complete this project. This work would not have been possible without their help. I would also like to thank Jean-Christophe Grenier for being so helpful, and so efficient in providing assistance, as well as Isabelle Sirois for her great support and scientific advice, and for encouraging myself as well as other team members to always mind work-life balance as well as mental health. I would like to give a special thanks to Dominique Fournelle for her great work with SANTA SIM.

I would also like to thank all members of Julie and Etienne's groups for their support, advice, and for being great, fun team members. I want to especially thank Etienne for helping me refine my presentation skills and for helping me prepare for presenting this work at the MIM symposium, 35e Congrès de la recherche, and HUPO.

I would finally like to thank my family, my friends, and my partner for being so supportive throughout this journey.

## INTRODUCTION

The current pandemic caused by Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2), a zoonotic virus causing acute infections of the lower respiratory tract, is the latest in a series of pandemic events that have plagued humanity throughout history. Unlike epidemics, which consist of a sudden rise in cases related to an infectious disease, a pandemic is declared when the rate of infection rises exponentially across multiple countries. As was the case with previous pandemics, such as the Bubonic plague of the 14$^{th}$ century, the current pandemic has had tremendous impacts on our economies, social norms, and public health. Efforts to mitigate the situation have been marked by a high level of collaboration, resulting in part in the rapid development and approval of numerous vaccines within one year. This unprecedented collaborative ecosystem has also enabled the rapid characterization of SARS-CoV-2 biochemistry, epidemiology as well as its relationship with the human immune system. However, the rapid spread of the virus across the globe has also resulted in significant genomic and proteomic diversification. Advanced genomic surveillance initiatives as well as novel lineage nomenclature schemes have allowed the scientific community to closely monitor the emergence of new lineages in the context of epidemiologically relevant outbreaks (1–4). As a result of these efforts, a set of Variants of Concerns (VOCs) deemed to pose risks to public health were identified. Although many features of the viral infection cycle and virulence have been unveiled, many questions remain. In light of the recent emergence of various widespread VOCs, a current point of interest consists of understanding the capacity of SARS-CoV-2 variants to evade immune recognition. In this dissertation, I will discuss the current understanding of viral biology, its relationship with the adaptive immune system, as well as its ability to evade recognition by the adaptive immune system. I will then present a manuscript (Hamelin *et al*, 2021) in which collaborators and I investigated

the global mutational landscape of SARS-CoV-2 during the first year of the pandemic, and its ability to enable CD8+ T lymphocyte escape in an HLA supertype-dependent manner.

# 1 CHAPTER I: Literature review

## 1.1 SARS-CoV-2

### *1.1.1 Introduction of the virus*

The current Coronavirus Disease 2019 (COVID-19) pandemic, caused by the Severe Acute Respiratory Syndrome Coronavirus 2, has resulted in 5.2 million deaths and 260 million cases worldwide. Every large-scale pandemic experienced by humanity has been mitigated using approaches befitting of the era. The black plague of the 14$^{th}$ century, the first recorded pandemic, was mitigated by the first known instances of public health measures including limited travel and quarantines. Several centuries later, the Great Influenza of 1918 was mitigated by much more sophisticated, rapid health response. Recent analyses demonstrated that cities in which more stringent and rapid social restrictions were implemented experienced lower mortality rates (5,6). The Human Immunodeficiency Virus (HIV) pandemic of the late 20$^{th}$ century benefited from modern science, with the development of assay-based diagnostic tools and Antiretroviral Therapy (ART). However, the COVID-19 pandemic marks a turning point in the medical journey of our species. The rise of cutting-edge discoveries in conjunction with the advent of globalization has equipped the nations of the world with a unique set of tools to tackle this pandemic. Since the emergence of the virus in December 2019, astounding global research efforts have been aimed at contributing to our understanding of the inner workings of the virus, its pathogenicity, as well as the various defense mechanisms deployed by the human body to counteract infection. Our approach to this pandemic did not only make use of an unprecedented level of international collaboration but also of the many recent advancements of modern science. For example, early in the pandemic, research groups were able to rapidly characterize many features of human immune

response to SARS-CoV-2; the field of structural biology quickly solved the structure of the Spike glycoprotein, responsible for viral entry into host cells via an interaction with the ACE2 receptor; the early development of diagnostic tools was able to quickly inform public health response. These scientific achievements, along with many others, quickly established an informed platform upon which pharmaceutical companies and public health organizations could generate solutions. In this section, we will discuss the biology of SARS-CoV-2 as well as its relationship with the human adaptive immune system.

#### 1.1.1.1 Close relatives and previous outbreaks

In the wake of the COVID-19 pandemic, companies such as Pfizer, AstraZeneca and Moderna were able to rapidly put forth putative vaccines. Although this feat can be attributed to the impressive wave of global collaboration that ensued the declaration of a global pandemic, many of the biological, immunologic and pathogenic features of the virus were already partially understood due to outbreaks from its predecessors. The coronaviruses constitute a large family of zoonotic viruses which can be separated into the following genera: Alpha, Beta, Gamma, and Delta coronavirus (7). First identified in 1962, coronaviruses were originally associated with mild gastrointestinal and respiratory infections in mammals (8,9). The *Sarbecovirus* subgenus of Betacoronavirus has been of greatest medical and epidemiologic interest throughout the last few decades, as the species identified within this subgenus have been associated with fatal infections in the lower respiratory tract of humans (10,11). These include Severe Acute Respiratory Syndrome Coronavirus (SARS-CoV) which caused an outbreak resulting in~8000 cases in 2002/2003, as well as the Middle East Respiratory Syndrome Coronavirus (MERS-CoV), causing an outbreak in Saudi Arabia in 2012 (12,13). Genomic analyses have revealed that SARS-CoV and MERS-CoV possess 79% and 50% conservation, respectively, with SARS-CoV-2 (14,15).

The virus with highest similarity to SARS-CoV-2 was identified as the horseshoe bat coronavirus RatG13 (*Rhynolophus Affinis*) with 96% sequence similarity, suggesting horseshoe bats as the most likely origins for SARS-CoV-2 (14). However, accurately and confidently tracing the evolutionary path of SARS-CoV-2 proves to be a challenge due to the high recombination rate characteristic of coronaviruses (16). Regardless of the complete origin of SARS-CoV-2, the high mutation rate in the Receptor Binding Domain (RBD) of the Spike Glycoprotein has been identified as a key determinant of host repertoire (16,17) and has paved the way for SARS-like viruses to infect humans. Indeed, genomic modifications to the RBD enabled the Spike Glycoprotein to interact with the human ACE2 surface receptor (see section 1.3.1.2 for more details) (17). Studies attempting to compare SARS-CoV-2 to its relative, SARS-CoV, have identified key features within the Spike Glycoprotein differentiating the two viruses. These include an enhanced binding interface between Spike and ACE2 to facilitating viral entry into host cells, as well as the addition of a furin cleavage site within the Spike protein to improve the infectious cycle (17). Structural and biochemical investigations demonstrated these changes to play key roles in the increased infectivity of SARS-CoV-2.

The emergence of SARS-CoV and MERS-CoV have both sparked much research into the epidemiology, infectivity, and biology of SARS-like coronaviruses. These investigations have led findings pertaining to the identification and characterization of the proteins making up the proteome of SARS-CoV (and SARS-CoV-2), the pathologies and life cycle of SARS-like viruses, as well as the relationship between SARS-like viruses and the human immune system. These findings all contributed to the initial development of diagnostic tools, as well as the rapid development of highly effective vaccines within one year.

#### 1.1.1.2 SARS-CoV-2 Protein composition

SARS-CoV-2, like most viruses, is elegantly simple in its genomic and proteomic composition. Composed of a 29,903 base-pair single-stranded ribonucleic acid (ssRNA) genome, the SARS-CoV-2 proteome constitutes 14 Open-Reading Frames (ORFs) resulting in 27 proteins. These are made up of a combination of structural as well as non-structural, and in-frame as well as out-of-frame proteins (Figure 1.1).

The 5' end of the viral genome encodes for the genes ORF1a and ORF1ab, resulting in the polyproteins pp1a and pp1ab, respectively. These genes are the largest, taking up around two-third of the viral genome. Following translation, these polyprotein products are cleaved into 15 non-structural proteins (nsps). The polyprotein pp1a results in nsps 1-10, while pp1ab results in nsps 12-16. Together, these non-structural proteins play a variety of roles in involved in processes including viral replication, protein translation, and resistance to host innate immunity. The remainder of the SARS-CoV-2 genome is made up of four structural proteins, namely the Spike Glycoprotein (S), the Nucleocapsid protein (N), the Envelope protein (E), and the Membrane protein (M), as well as an assortment of additional open reading frames referred to as the accessory proteins. Briefly, the E protein interacts with the human PALS1, altering the tight junction formation and promoting the pathogenesis of SARS-CoV-2 (18). The M protein is involved in the assembly of virions, as well as their budding following viral replication (19). The N protein is a multifunctional protein with roles in viral replication. The N protein to possesses two RNA-binding domain, the N-terminal Domain (NTD) and the C-terminal Domain (CTD) linked by a disordered, serine/arginine-rich region. In part due to the positively charged linker region, this protein was shown form a complex with viral RNA, facilitating the transportation of RNA to the

replication transcription complex (RTC) (20,21). However, the N protein was also shown to be involved in RNA transportation in collaboration with the M protein following RNA replication, thus indicating the multivalent roles of this protein (20,22).

The S protein has unequivocally been subject to the greatest proportion of scientific investigations aimed at SARS-CoV-2 proteins. Being amongst the most abundant of the SARS-CoV-2 proteins, and being a key player in facilitating viral entry into host cells, the Spike protein has been of great interest in the development of both therapeutic as well as prophylactic treatments. The S protein, 1273 amino acids in length, is composed of two subunits: S1 and S2 (23). The former is composed of an N-terminal Domain as well as a Receptor-Binding Domain, and the latter is composed of five discrete domains: the Fusion Peptide (FP, or S2'), HeptaPeptide Domains 1 and 2 (HPD1/2), a Transmembrane Domain (TMD), as well as a Cytoplasm domain (CD) (23–26). The primary role of the Spike protein is to bind to the human cell surface receptor Angiotensin Converting Enzyme 2 (ACE2), thus permitting viral entry into the cell (24,27). The Spike S1 RBD was shown to be the primary point of contact within the confines of this interaction, making it a key mediator of viral entry into host cells (27,28). As such, disruption of the RBD-ACE2 binding interface has proven to be the most direct route to inhibit viral entry into the cell, and has been exploited by antibodies (infection- and vaccine-induced, inhibitors as well as therapeutic monoclonal antibodies) (25,27–31). Finally, the SARS-CoV-2 proteome is composed of a variety of accessory proteins, namely ORF3a/b/c/d, ORF6, ORF7a/b, ORF8, ORF9b/c, and ORF10. (32). These accessory proteins are much less well understood than those mentioned above, and the features that are understood were acquired during previous investigations of SARS-CoV and MERS-CoV (33). Inquiries thus far have demonstrated that this collection of accessory proteins are in fact not essential to the life cycle and replication of SARS-CoV-2, but rather are involved

in its pathogenicity and in host-virus interactions (34). In corroboration with these statements, mutations within a number of these accessory proteins were described in Variants of Concerns (VOCs) associated with increased infectivity.

Previous SARS-CoV and MERS-CoV outbreaks as well as the current SARS-CoV-2 pandemic have provided the scientific community with ample opportunity to investigate the various structural and non-structural proteins making up coronaviruses. However, only a minority of these proteins have been considered as targets of interest for prophylactic and therapeutic treatments.

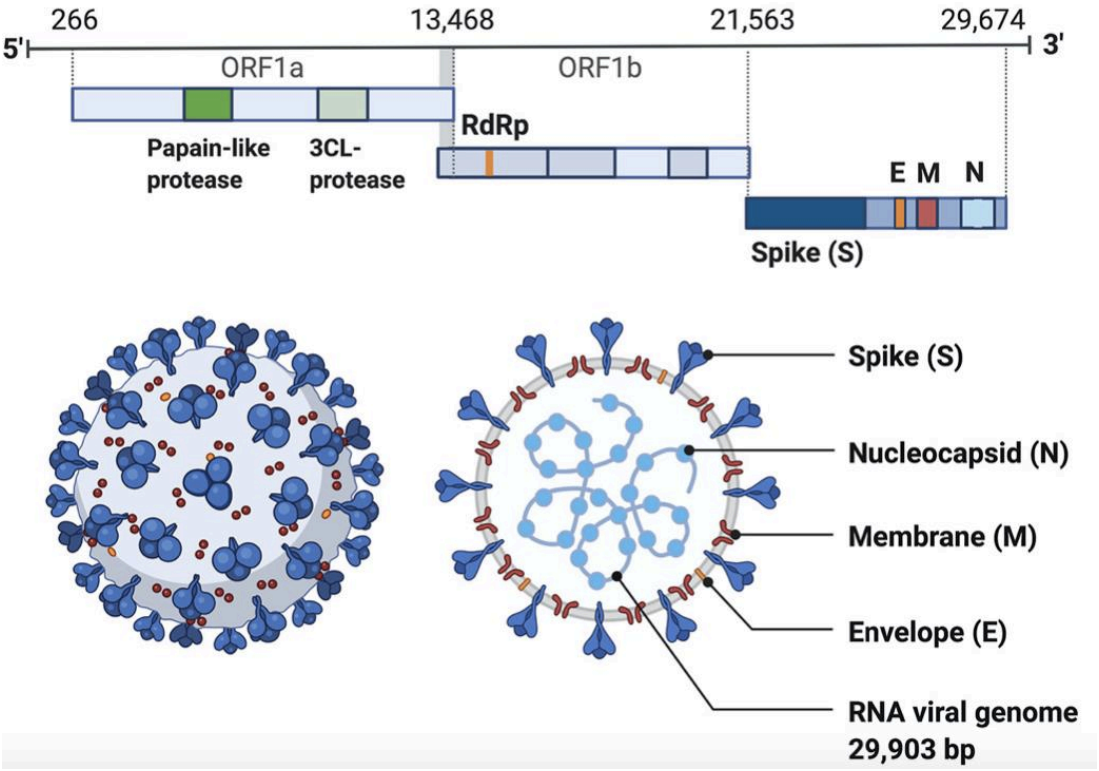



Figure 1.1. Organization of the SARS-CoV-2 genome. Adapted from Rando et al. 2021 (35)

### *1.1.2 Evolution of the virus throughout the pandemic*

The current SARS-CoV-2 pandemic represents the most recent addition to an extensive list of deadly infectious pandemics that have challenged humans. Other well-known components of the list include the Black Plague (*Yersinia pestis*, 14$^{th}$ century), the Great Influenza (1918), and HIV/AIDS (1980-present). However, although the current pandemic will join previous pandemics when judged on its medical, societal, and economic impacts, it will mark a turning point in history due to an unprecedented ability to mediate the disease with such rapid response time. One of the initial responses to the pandemic was the initiation of a never-before-seen level of international collaboration, resulting in massive sharing of data and a push toward open science. Downstream effects of this global collaborative effort included the rapid characterization of SARS-COV-2 biology, the identification of putative therapeutic targets/agents, as well as the design and approval of multiple highly effective vaccines within a year. Another key outcome was a global initiative to share virtually all sequenced SARS-CoV-2 genomic data, thus drastically facilitating the global monitoring of SARS-CoV-2 evolution. As such, the scientific community was able to track in near-real-time the evolution of the virus in a global pandemic setting. This effort allowed for the rapid characterization of emerging viral strains, and the eventual identification of Variants of Concern. The current pandemic provided a unique opportunity to track the evolution of the virus, and to identify adaptation events with the potential to jeopardize the effectiveness of public health interventions.

#### 1.1.2.1 SARS-CoV-2 sequencing technologies

Beyond the unprecedented collaborative efforts observed over the course of the pandemic, the ability of the scientific community to closely monitor the evolution of SARS-

COV-2 in near real-time can be attributed to the multiple rapid and accurate sequencing technologies that have been developed and applied in the context of virology. To this end, guidelines were established to standardize and optimize sample collection so as to ensure the generation of high-quality sequencing data (36,37). In accordance to said guidelines, high quality SARS-CoV-2 genomes were successfully sequenced from samples originating from the upper and lower respiratory tracts, although genomes were also acquired form feces and urine samples. Next-Generation Sequencing (NGS) has quickly become the method of choice for the sequencing of viral genomes, and SARS-CoV_2 is no exception. Multiple NGS-based approaches as well as protocols for the preparation of libraries were generated for the sequencing of SARS-CoV-2 (38–40).

*Shotgun Metatranscriptomics*. Shotgun Metatranscriptomics is a powerful method enabling the sequencing of all genomic material within a culture-independent sample. Metatranscriptome sequencing was involved in the initial discovery of SARS-CoV-2 (14,41) and has since yielded an array of complete and near-complete genomic assemblies of SARS-CoV-2. However, despite its ability to capture numerous microbial species, SARS-CoV-2 strains, as well as host genetics, metatranscriptomics is not without its drawbacks. In the context of the current pandemic consists of the limiting cost-effectiveness of the method when it comes to large-scale genomic surveillance, due to the high sequencing depth required (42).

*Amplicon-based sequencing*. While metagenomics/metatranscriptomics enables the analysis of all genomic material within a sample, amplicon-based approaches provide a much more targeted route of analysis. Nevertheless, the latter requires substantial prior knowledge regarding the organism of interest and can therefore not be utilized in species/strain discovery. Amplicon-based sequencing involves the initial enrichment of targeted genomic material via

first-strand cDNA synthesis. Genomes are then amplified using a multiplex PCR approach (Multiplex PCR Targeted Amplicon Sequencing, or MTA-seq). Due to its high specificity and robustness, MTA-seq approaches require less material than metagenomic approaches. Dohl *et al.* optimized a highly efficacious and cost-effective amplicon-based method, which was implemented by the ARTIC network as a SARS-CoV-2 genomic surveillance approach (42). This method was later further optimized using nanopore-based long read sequencing (43).

*Hybrid capture-enrichment Sequencing.* In this approach, shotgun libraries are denatured and subsequently hybridized to nucleic acid probes, which are generally biotinylated to enable streptavidin-based isolation of genomes (44,45). In the context of SARS-CoV-2, hybrid capture was found to have lower sensitivity than amplicon-based approach (46). Nevertheless, due to the longer probes used (~120bp), hybrid capture approaches have a higher tolerance for mutations within the targe sequences, thus minimizing the risk of probe failures caused by probe-target mismatches . As the majority of sequences deposited on the Global Initiative for Sharing Avian Influenza Data (GISAID) were generated by amplicon-based sequencing, we may experience the accumulation of a bias caused by SARS-CoV-2 mutation-driven amplicon failure.

Although diverse in utilities, these methods have contributed to the discovery and the on-going genomic monitoring of SARS-CoV-2.

#### 1.1.2.2 Initial evolution (First months/first year)

The mutation rate of coronaviruses is lower than other RNA viruses due to the presence of a proofreading mechanism carried out by nsp14 (3'-5' exoribonuclease), making the genomic diversification of SARS-CoV-2 a relatively slow process (47). Nevertheless, Mutations within the SARS-CoV-2 genome were quickly identified as a potential challenge in the public health response

against COVID-19. Although most mutations are neutral and do not modulate viral fitness, some may become fixed within the population due to positive selection. The accumulation of such genomic variations may confer the ability of zoonotic RNA viruses to overcome their host's various defense mechanisms. A mutation substituting an Aspartic acid (D) for a Glycine (G) at the 614$^{th}$ position of the spike protein (S:D614G) in SARS-CoV-2 was identified in January 2020 and consisted of the first Spike protein variation to become predominant (48). This mutation was characterized as a recurrent mutation, gaining prevalence in distinct geographical regions simultaneously. This mutation was not only associated with higher viral loads in the upper respiratory tracts of infected individuals but was also associated with enhanced growth in pseudoviral experiments, suggesting it capacity to confer a fitness advantage to the virus (48). The early identification of this viral fitness-enhancing variant evidenced the importance of monitoring the global evolution of the virus for subsequent mutations with possible impacts on viral fitness, infectivity, and pathogenicity.

#### 1.1.2.3 SARS-CoV-2 surveillance and data-sharing platforms

Early in the pandemic, several initiatives were put in place to facilitate the tracking of genomic variations on a global setting. These included GISAID as well as NextStrain (49,50) (see section 1.3.2.2 for more detail). Both platforms were put in place several years prior to the pandemic, in response to infectious disease outbreaks such as SARS-CoV, MERS-CoV, H5N1, Ebola and Acquired Immunodeficiency Syndrome (AIDS). Due to their prior establishment and proven track record, they represented ideal platforms to quickly initiate the genomic surveillance of SARS-CoV-2.

*GISAID*. Launched in 2008, GISAID was established to incentivize and facilitate the sharing of infectious disease-related genomic data. Although the open sharing of scientific data provides many advantages, the concept of open science has received received criticism. Some researchers fear that others will prevail in the 'Race-to-publication' using their data, while certain countries may want to closely monitor the sharing of data to encourage well-defined scientific collaborations. Additionally, countries may use caution when sharing data regarding the outbreak of novel pathogens to avoid being held accountable. The aim of GISAID was to promote international data-sharing while addressing concerns related to open science. This was achieved with the development of a data-sharing agreement providing protection to contributors while ensuring proper accreditation of all authors.

Prior to the current pandemic, GISAID was already established as a prominent contributor to open science. By 2016, it had accumulated over 650,000 viral sequences within its own database, EpiFlu$^{TM}$, and these were submitted by more than 850 institutions. In addition, GISAID and EpiFlu$^{TM}$ were already widely used by the World Health Organization (WHO). However, the paradigm-shifting feature of the data-sharing platform came to light over the course of the COVID-19 pandemic. Having accumulated more than 6,162,679 SARS-CoV-2 sequences by late 2021, GISAID provided a unique opportunity for research groups around the world to contribute to the rapid and in-depth genomic surveillance of SARS-CoV-2 (51). Additionally, GISAID developed a series of visualization tools to allow the public and scientific community to monitor the dispersion of the virus and its various strains (52). However, the platform is not without its caveats. The majority of SARS-CoV-2 sequences were contributed by a few wealthy countries such as the United States and the United Kingdom, while numerous countries lacking the proper viral surveillance infrastructure have been significantly underrepresented. Examples include El

Salvador and Lebanon, which have contributed very few sequences despite having significant exposure to COVID-19 (53). This disproportionate contribution of sequences was criticized for introducing bias when considering the global prevalence of genomic variants. Nevertheless, EpiFlu$^{TM}$ quickly became the largest database of SARS-CoV-2 genomic sequences, and has enabled the scientific community to contribute to our understanding of the virus while promoting scientifically informed public health initiatives.

*NextStrain*: Released in 2015, the primary role of NextStrain was to facilitate the analysis and visualization of phylodynamic trends pertaining to viral outbreaks, epidemics and pandemics (49). Using a python-based framework, NextStrain was developed to build and maintain a database composed of viral sequences from other publicly available repositories, including GISAID and NCBI (Figure 1.2). In addition, the NextStrain platform is equipped with a suite of phylogenetic tools enabling a series of analyses pertaining to the temporal, geographical, and phylogenetic features of viral epidemics, culminating in the implementation of a Maximum Likelihood phylodynamic-based framework (54,55). Importantly, the NextStrain platform was designed to be adaptable to any existing or novel viruses, making it an ideal tool to tackle the current SARS-CoV-2 pandemics.

*Other Platforms*: Additional tools were developed to further address the accumulation and diversification of SARS-CoV-2 sequences made available to researchers around the world. These tools, which include CoVariant, CoVizu and Outbreak.info amongst others, were designed to complement and enhance the analytical and visualization tools made available through NextStrain and GISAID (56–58). Importantly, the development and establishment of these tools was enabled by the large-scale, open sharing of viral sequences by GISAID. The outbkreak.info platform acts as a standardized repository for a wide range of SARS-CoV-2-related data types, including

COVID-19 epidemiology (cases/deaths), emerging SARS-CoV-2 variants, as well as the sharing of SARS-CoV-2-related publications, clinical trials, laboratory protocols, and datasets (58). The CoVariant platorm consists of a comprehensive web-based tool providing detailed information regarding existing and emerging lineages of interests. The information pertains to relevant VOC-specific mutations, their evolutionary context, their clinical relevance, and structural modelling of mutated proteins (56). Finally, CoVizu provides additional means of analyzing the diversification of SARS-CoV-2 over the course of the pandemic by constructing phylogenetic trees in a temporal context (57). CoVizu then utilizes beadplots to allow users to visualize near real-time analysis of epidemiologically relevantSARS-CoV-2 strains.

Together, the data-sharing infrastructure of GISAID combined with the data analytics and visualization tools provided by NextStrain and other tools established an ideal ecosystem to facilitate the global, collaborative, and in-depth genomic surveillance of SARS-CoV-2 outbreaks.

#### 1.1.2.4 Methods used to characterize mutations

The enormous amount of data made available to the scientific community by GISAID and NextStrain had for effect to engender the development and application of a flurry of established and emerging bioinformatic tools aimed at interrogating the evolution of SARS-CoV-2 (51,52). However, the wide array of phylogenetic and evolutionary investigations that ensued, in the context of an active pandemic, resulted in an urgent need for robust naming conventions. In efforts to standardize the naming of SARS-COV-2 strains, three prevailing nomenclature schemes were introduced: the clade nomenclatures developed separately by NextStrain and GISAID, as well as Pango lineages, pioneered by Rambaut *et al.* (2020) (Figure 1.3) (1,3,52). The clade nomenclature systems were designed to monitor the general, high-level diversification of SARS-CoV-2. In

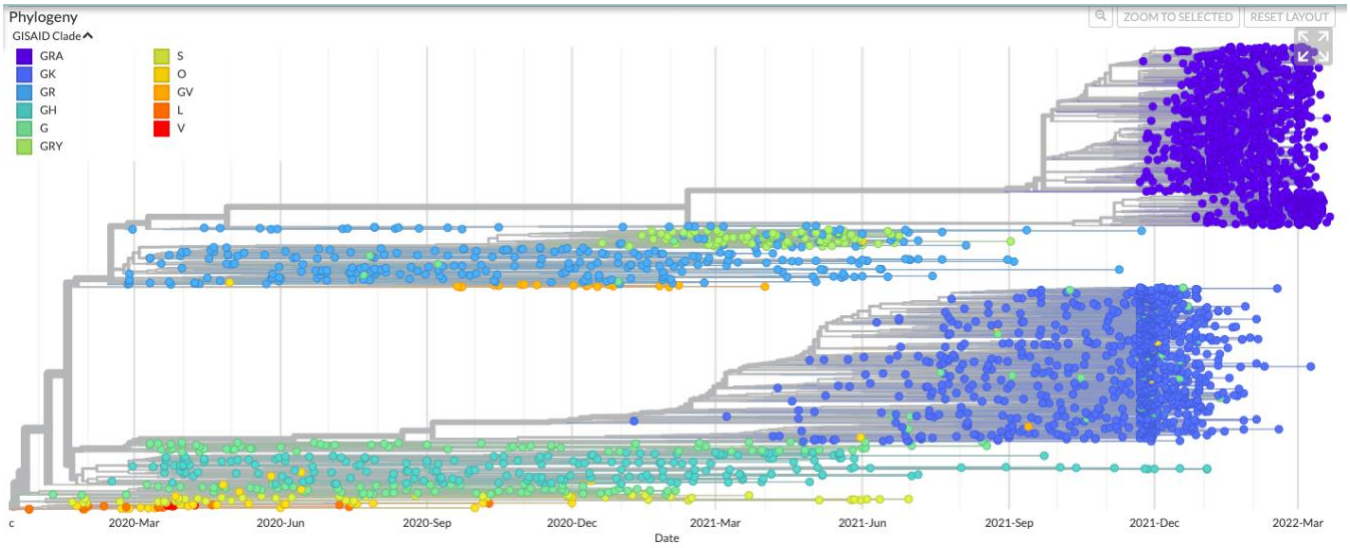



Figure 1.2. Phylogenetic representation of SARS-CoV-2 evolution powered by NextStrain, acquired from the GISAID platform. Here, lineages are colored according to the GISAID clade nomenclature (acqruied from GISAID website) (59).

(51,52)(1,3,52)contrast, Pango lineages were developed to conduct fine-grain surveillance while actively tracking outbreak events.

*Clade nomenclature (NextStrain)*. The NextStrain clade nomenclature consists of a naming system aimed at monitoring highly frequent and geographically widespread viral strains (60). In this nomenclature system, a new clade is defined when it attains a 20% global frequency and possesses a minimum of 2-mutations difference from the parent clade (60). New clades are then named based on the year of their definition, as well as a letter corresponding to the order in which clades occurred. Following these guidelines, NextStrain initially defined five clades. The first two clades, 19A and 19B, were defined in Asia and were separated by the mutations C8782T (silent) and T28144C (Orf8:L84S). Upon gaining prevalence in Europe, clade 19A gave way to clade 20A, defined by the addition of mutations C3037T (silent), C14408T (ORF1ab:P314L), and A23403G (Spike:G614G). Clade 20A then gave way to clades 20B and 20C. The former, located in Europe was defined by the addition of a triple mutation, G28881A (N:R203K), G28882A (N:R203K) and

G28883C (N:G204R). The latter, localized in North America, was defined by the addition of C1059T (ORF1a:T265I) and G25563T (ORF3a:Q57H).

*Clade nomenclature (GISAID).* Unlike the NextStrain approach, the GISAID clade nomenclature utilized a statistical approach to define phylogenetic clusters (2). Namely, the Phylogenetic Clustering by Linear Integer Programming (PhyCLIP) was used to determine genome distances and define lineages (61). The latter were then combined based on similarities in mutational definitions, resulting in major clades. Using this approach, GISAID defined 7 distinct clades (in order of occurrence): S, L, V, G, GK, GH, GR, GV, and GRY. The mutational definitions have been outlined by GISAID (2).

*Pango Lineages.* Pango lineages, as defined by Rambaut *et al* (1), has become the most widely used nomenclature within the scientific community as well as major news outlets. While the methods employed by GISAID and NextStrain to generate their respective clade nomenclatures produced general phylogenetic trends, the approach helmed by the Pango method provides a much more detailed account of viral evolution (3). This innovative system was designed to actively monitor lineages contributing to viral spread while marking lineages that have likely become inactive. This scheme resulted in a continually evolving nomenclature system with an emphasis on outbreak-associated events (1). As such, this dynamic nomenclature system was not developed to continually track every single change occurring within the viral genome, but rather to shed light on important evolutionary events with relevance to the context of public health interventions. In efforts to develop such a nomenclature, Rambaut *et al.* developed the following set of rules (1): Lineages should not only diverge from the parent lineage, but expand in a new geographical area. Divergence from parent lineages should be demonstrated using the following logic: i) A new lineage should diverge from its parent by at least two mutation events; be identified in at least 5

high quality genomes (>95% coverage); result in a bootstrap value >70%; genomes within the lineage should share at least one mutation event. Furthermore, the system designed by Rambaut *et al.* dictates that major lineages be designated by a letter, and that descending lineages be assigned numerical annotations (ex. A.1, A.2.1, B.1, B.2.1). To encourage a simple and dynamic system, each major lineage is only afforded three levels (after which a new major lineage is designated (ex: A.1.1.1.2 → C1).

The initial designations made using this nomenclature corroborated those made by the clade systems discussed above. Two major lineages, A and B, were identified based on mutation events at nucleotide positions C8782 and T28144. Subsequent designations included major outbreak events in the USA (A.1, A.3) and Europe (A.2, A.5), as well as some important initial outbreak events in Italy (B.1, B.2), the UK (B.3) as well as (probably) Iran (B.4). The Pango lineages have contributed to the designation of variants of concern, including the variants designated as Alpha (B.1.1.7), Beta (B.1.351), Gamma (P.1) and Delta (B.1.617).

*Haplotype networks*. The use of phylogenetics to assess the relationship between sequences is computationally taxing and suffers a reduction in efficiency with an increasing number of genomic sequences, which could prove problematic given the 10 million SARS-CoV-2 sequences currently located in the GISAID EpiFlu database. An alternative approach consists of assessing mutation frequencies in a temporal context, and to build a haplotype network based on the temporal frequencies of driver mutations (62–64). Phylogenetic trees expect sequences belonging to ancestor nodes to no longer occur and to only be represented in said ancestor nodes. In contrast, haplotype networks are compatible with scenarios where sequences belonging to ancestral nodes and those arising in current nodes may be observed in the same temporal timeframe, which is the case with the current sampling of SARS-CoV-2. Haplotype networks were in fact applied to

SARS-CoV-2 sequences early in the pandemic (64). To further take advantage of the compatibility between the haplotype network approach and the current sampling scheme of SARS-CoV-2, our group generated a haplotype network (62). The latter was applied to characterize the mutational diversity of SARS-CoV-2 in a temporal context throughout the first year of the pandemic. To this end, a set of mutations found to rapidly expand over the course of the pandemic were identified and used to define haplotypes. Derived Allele Frequencies (DAF) were considered for the selection of these mutations and resulted in 22 genomic positions. These were found, for the most part, to corroborate the genomic positions considered in the generation of the Pango- and clade- based lineages discussed above. Notable mutations include the early mutations C241U, C3037U, C4408U and D23403G, as well as the triple mutation G28881A, G28882A, G28883C. This haplotype analysis, conducted on sequences from the first year of the pandemic, resulted in 122 distinct haplotypes containing each at least 10 sequences. Of these, 17 haplotypes were found to be representative of the dominant lineages describing the evolution of the virus throughout the first 12 months of the pandemic. Comparison between the lineages identified using the haplotype network, Pangolin, and the NextStrain approaches resulted in a high level of similarity, although some discrepancies arose. Namely, the haplotype network was able to achieve a higher level of granularity than the NextStrain clade system (62). For example, the NextStrain clades 19A and B were separated into haplotypes I and IV, and haplotypes III and IX, respectively. In contrast, the haplotypes identified using the 22 genomic positions differed from the Pango lineage designations with regards to lineage B.1. The sequences composing this lineage were in fact assigned to three evolutionarily distant haplotypes, labeled as haplotypes II, III and VIII. The use of haplotype networks therefore offers a level of granularity unobserved in alternative methods. However, the evolutionary resolution achieved by the haplotype network approach could not differentiate all

VOCs. Using the set of chosen genomic positions to conduct haplotype definitions, the Alpha and Gamma variants were grouped within haplotype XV.

Overall, each of the four nomenclature schemes described here present specific advantages as well as caveats, offering a range of granularity. An optimal approach for the monitoring of SARS-CoV-2 evolution would likely involve a combination of these nomenclatures.

#### 1.1.2.5 Variants Of Concern

The diversification of SARS-CoV-2 has been a point of concern for public health responses around the world. Although the majority of mutations result in changes that are either neutral or detrimental to viral biology, some genomic variations may improve viral fitness, pathogenicity or infectivity (65–68). Such mutation events could eventually impede the ability of public health initiatives to detect, treat or prevent SARS-CoV-2 infections (69). Initial surveillance largely resulted in the identification of single mutations of interest, including the D614G mutation in the spike glycoprotein. Although other mutations became fixed early in the pandemic, this substitution was extensively investigated given it's demonstrated impact on infectivity and transmissibility (48,70). Additionally, the early genome cluster characterized by this mutation, referred to as the B.1 lineage, was also accompanied by other linked mutations including P323L within ORF1ab. These may have aided the Spike:D614G mutation by conferring viral fitness.

The accumulation of mutations within the SARS-CoV-2 genome, in combination with effective nomenclature systems, eventually led to the characterization of SARS-CoV-2 lineages composed of numerous co-occurring mutations. Epidemiologic and pathogenic inquiries into these

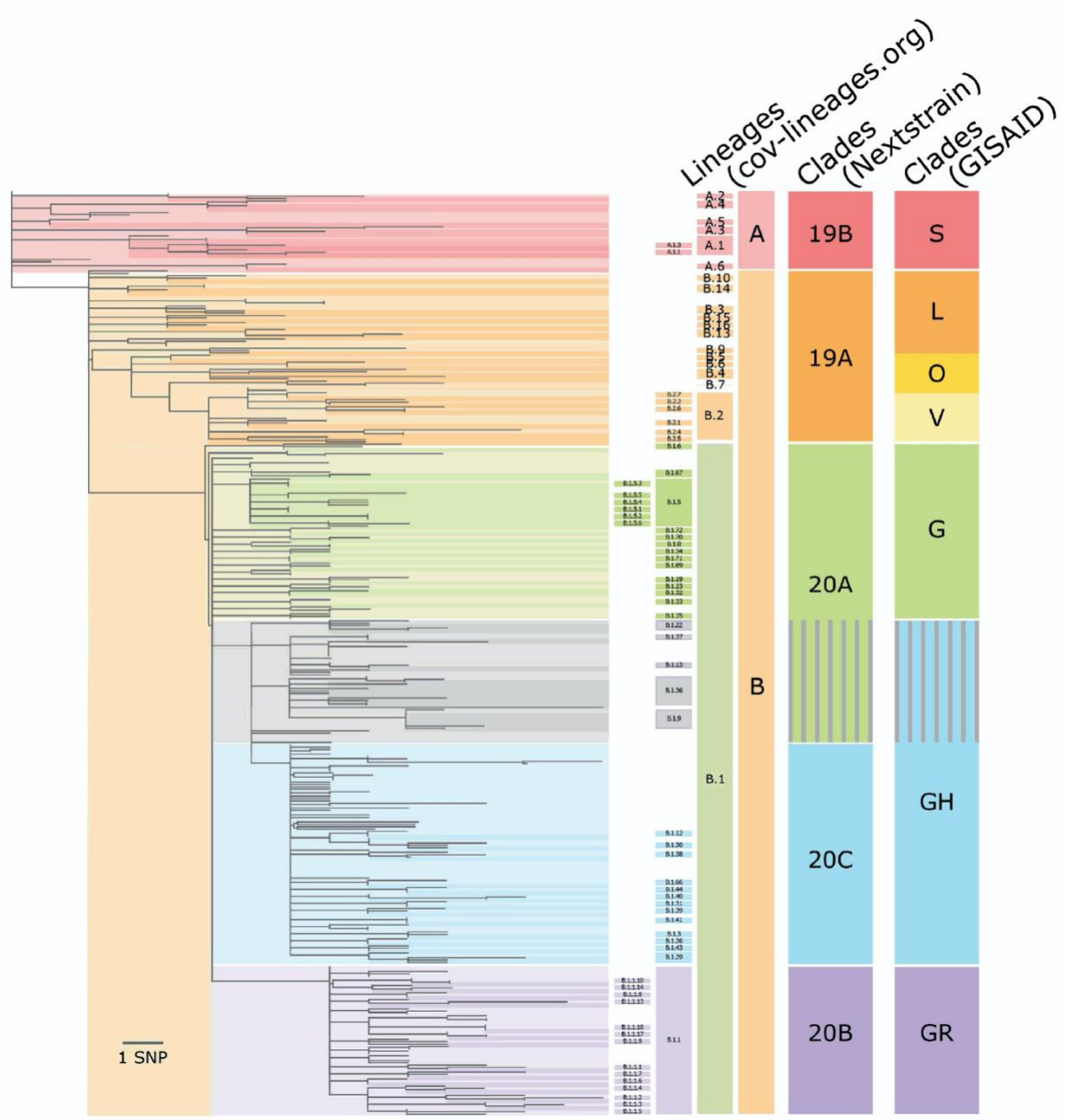



Figure 1.3. Graphical representation comparing the various widely accepted nomenclature systems, namely clade systems respective to GISAID and NextStrain, as well the Pango lineages. Figure adapted from Alm et al. 2020 (3).

lineages have resulted in terminology to characterize their relevance to public health. These include Variants Of Concern (VOCs), Variants Of Interest (VOIs), Variants Under Monitoring, and Variants Of High Consequence (VOHC) (69,71).

*VOCs*. SARS-CoV-2 lineages are flagged as VOCs when their impact on the infectivity, transmissibility, or pathogenicity of the virus has direct relevance to public health (67,68,72). In July 2021, the World Health Organization (WHO) published a naming scheme for referring to VOCs (4). The use of nomenclatures developed by NextStrain, GISAID and Pangolin resulted in complex names, and the use of countries of origin (ex. UK variant) produced stigmatization. In efforts to facilitate the dialogue between public health, the scientific community, and the public, the WHO established the use of Greek letters to denote VOCs (4). The first such variant, Alpha (B.1.1.7), was associated with 22 genomic variations including 7 non-synonymous substitutions and 3 deletions within the Spike glycoprotein. Original phylogenetic and epidemiologic analyses of the virus suggested its increased transmissibility (73,74). In addition, several mutations have been shown to impact the life cycle of the virus. The N501Y mutation was shown to facilitate the entry of the virus into the cell by enhancing the interaction between the Spike glycoprotein and ACE2 (75,76); the P681H mutation was shown to inhibit the cleaving of the Spike glycoprotein by furin, thus impeding on the infectious cycle of SARS-CoV-2 (77,78); deletions of residues 69 and 70 were found to interfere with recognition by antibodies and enhance viral entry into host cells (79). Subsequent VOCs, including Beta (B.1.135) and Gamma (P.1), were also associated with higher transmissibility, as well as mutations increasing viral entry into host cells and promoting immune evasion (80–86). The fourth VOC, Delta (B.1.617), has been of great interest to the public health sector and scientific community. This interest stemmed from its rapid spread across the globe, causing it to become the dominant lineage (87). Of the numerous substitutions making up this variant, several Spike protein mutations have been shown to contribute to public health-relevant phenotypes (88). These include L452R, E484K, T478K and P681R. The former two were found to enhance the Spike-ACE2 interface, thus improving viral entry into host cells;

P681R was found to facilitate Spike cleaving by furin; and although T478K has not been directly investigated, a similar T478 substitutions (T478I) was found to affect viral recognition by antibodies (89). The SARS-CoV-2 lineage most recently classified as a VOC was Omicron (B.1.1.529). This variant was first identified in South Africa in November 2021, and quickly acquired the status of VOC due to an unprecedented number of mutations as well as elevated transmissibility. This variant contains over 50 amino acid mutations, over 30 of which are found in the spike protein. Strikingly, 15 of these mutations were found within the Spike RBD. As previous studies demonstrated the immunodominance of the RBD and the ability of RBD-localized mutations to modulate the neutralizing activity of antibodies, the scientific community rapidly hypothesized as to the antibody-escaping potential of Omicron (90,91). Indeed, multiple studies confirmed the ability of Omicron-associated mutations to enable significant evasion from antibodies while facilitating transmission (92–94). Additionally, Omicron was experimentally shown to partially evade vaccine-induced neutralizing antibodies (95).

*VOIs*. The current list of VOCs has been the focus of recent research efforts and public health initiatives due to their ability to enhance transmissibility, pathogenicity, infectivity and/or immune evasion, as well as leading to drastic rises in infection rates. Beyond these VOCs, attention has been allocated to a growing list of variants under surveillance, known as Variants Of Interest (VOIs). The WHO has defined VOIs as lineages containing non-synonymous mutations suspected or known to modulate infection phenotypes, and to lead to increased transmission in at least one country. Former examples of VOIs included several lineages characterized by the E484K mutations found in multiple VOCs, such as VOIs Zeta (P.2), Theta (P.3) and Iota (B.1.526) (69,71). Eta, a VOI identified in both the United Kingdom and Africa, possesses the E484K substitution as well as several deletions found within the Alpha. These lineages have now been re-

designated as Variants Under Monitoring (VUMs). The current list of VOIs is dynamic and will continue to change with the on-going evolution of the virus, and with the expanding capacity of genomic surveillance initiatives.

*VUMs*. Variants Under Monitoring (VUMs) consist of SARS-CoV-2 lineages for which careful monitoring is warranted due to the possibility for future risks to public health (69,71). At the time of designation, however, the pathogenic and epidemiologic features of VUM lineages remain poorly understood and therefore require further investigation.

### *1.1.3 SARS-CoV-2 and the adaptive immune system*

Understanding humans' biological defense lines against SARS-CoV-2 is the first step in the development of effective prophylactic treatments. As it is the case with other coronaviruses, SARS-CoV-2 engages all three arms of the host's adaptive immune response upon infection: B cells (antibodies), CD8+ T cells, and CD4+ T cells. Humoral response is in fact the first adaptive immunity-based line of defense against the virus, with IgG and IgM being the main antibodies.

*Adaptive immune response in the context of SARS-CoV.* In the case of SARS-CoV, following the 2003 outbreak, antibodies against the Nucleocapsid (N) as well as Spike Glycoprotein (S) were found to be associated with better viral clearance and disease outcome(96–99). The humoral response was characterized as being composed primarily of IgG and IgM antibodies, with IgG titers being detectable in all patients within 14-210 days following infection, while IgM antibodies reached peak titers at 30 days and were undetectable after 120 days (97). Notably, viral neutralization was shown to correlate with convalescence (96,97,100). However, despite the undeniable value of humoral immunity with regards to SARS-CoV viral clearance, antibodies remain relatively short-lived. Anti-SARS-CoV antibodies were in fact found to be undetectable

after 2 to 3 years following SARS-COV infections (101–103). In the case of SARS-CoV-2, neutralizing antibodies have proven to be an effective diagnostic tool, an indicator of acuteness of infection as well as playing an important role in the clearance of the virus. However, as with SARS-CoV, antibody responses in SARS-CoV-2 have proven to wane relatively quickly, with antibody titers reaching baseline as quicky as 50 days after infection in some cases(104). In contrast, T cell memory appears to persist for much longer, with SARS-CoV-specific CD8+ T cells detected up to 11 years following infections in some cases (105). The role of CD4+ and CD8+ T cells in establishing protection against SARS-CoV as well as MERS-CoV has been thoroughly investigated and demonstrated over the last two decade (100,106–111). Studies consisted primarily of analyses of SARS-CoV and Middle East Respiratory Syndrome-related Coronavirus (MERS-CoV) convalescent peripheral blood mononuclear cells (PBMCs), or BALB/c mice. The consensus of these studies is that a large proportion of the cellular-based immune response is aimed at structural proteins, namely the N, S and Membrane (M) protein, although cytokine release has been detected in response to ORF3a. Airway CD4+ memory T cells were shown to effectively lead to cross-protection against SARS-CoV and MERS-CoV in BALB/c mice through induction of interferon-gamma (IFN-*y*) production as well as CD8+ T cell activation (112). A 2008 study of 128 convalescent SARS-CoV patients found that CD4+ T cell responses were correlated with disease severity (107). With regards to functionality, polychromatic flow cytometric analyses demonstrated that individuals with mild-to-moderate disease symptoms were associated with CD4+ T cells producing a single type of cytokine, whereas severe SARS-CoV cases were highly correlated with the production of multiple cytokines (INF-*y*, IL-2 and TNF-*a*). CD8+ T Lymphocyte-mediated immunity was shown to play a key role in fighting off SARS-CoV infections. Long-lasting CD8+ T cell immune responses were characterized against

immunodominant epitopes sourced from the Spike Glycoprotein, the Nucleocapsid, as well as ORF3a (109–111,113–118). As an example, multiple studies validated the efficacy of Spike protein epitopes in inducing strong CD8+ T-cell immune responses in transgenic mice models in the context of HLA-A02:01 (109,113). In agreement with these findings, the CD8+ T cell response to two immunodominant SARS-Spike protein peptides, S436 and S525 were investigated in mice challenged with a lethal dose of SARS-CoV (118). S436 and S525 were shown to not only lead to the significant expansion of S436/S525-specific CD8+ T cells in infected mice compared to control mice, but the CD8+ T cells were also associated with significant production of INF-*y*, TNF-*a*, and granzyme-B. The resulting CD8+ T cells shown to possess high cytotoxicity *in vivo*.

*Adaptive immune response in the context of SARS-CoV-2.* As a result of the vast array of coronavirus-related investigations sparked by the SARS-CoV (2003-2004) and MERS-CoV (2012) epidemics, the on-going SARS-CoV-2 global pandemic was preceded by over 15 years of research on closely related lethal viruses. Nevertheless, the initial 6 months of the SARS-COV-2 global pandemic saw the generation of a massive wealth of information concerning the adaptive immune response against SARS-CoV-2. The majority of early findings regarding T Lymphocyte-based immunity against SARS-CoV-2 originated from *in silico* investigations, principally due to the time and cost-effectiveness of such approaches. These were largely aimed at predicting HLA-dependent Cytotoxic T lymphocytes (CTL) epitopes in the context of vaccine development (119–122). The first such study, by Grifoni *et al*, investigated the three antigen-dependent arms of the adaptive immune system (119). Using sequence homology between SARS-CoV-2 and SARS-CoV as well as bioinformatic predictions, a vast array of putative B cell as well as T cell epitopes were identified in the hopes of facilitating vaccine design. In a separate study, Major Histocompatibility Complex (MHC)-binding predictions of 32,257 peptides from SARS-CoV-2's proteome were

computed against 145 HLA genotypes to characterize the propensity of all 145 HLA types to present SARS-CoV-2 epitopes (121). This extensive analysis led to the ranking of numerous HLA types based on their predicted ability to present viral epitopes, with HLA-B15:03 predicted to be amongst the HLA types able to present the highest diversity of SARS-CoV-2 epitopes, and HLA-B46:01 was predicted to present the lowest diversity of SARS-CoV-2 epitopes. While entirely predictive, this analysis proved informative with regards to establishing a link between HLA genotype and the extent of anti-SARS-CoV-2 CTL response. Beyond solely predictive analyses, experimental research efforts quickly led to the detailed characterization and validation of CD4+ and CD8+ T lymphocyte viral epitope landscapes along with the breadth of CTL responses against validated and predicted SARS-CoV-2 epitopes (122,123). In a study by Grifoni *et al* , using activation induced marker (AIM) assays and intracellular cytokine staining, CTL responses were quantitatively (magnitude) and qualitatively (cytokines) queried for a large array of CD4+ and CD8+ T-cell epitopes in the form of 'MegaPools' (MP). CD4+ response was determined to be predominantly aimed at the Spike glycoprotein, although strong responses were also observed for M, N, and numerous non-structural proteins. In the case of CD8+ T cell-mediated responses, co-dominant responses were observed for the Spike and M proteins, with ORF6, N and ORF3a also resulting in strong CTL responses. Although vaccines have been largely aimed at Spike protein, results presented here strongly suggest that vaccines should optimally consider a combination of SARS-CoV-2 antigens to maximize the magnitude of CTL responses induced upon immunization. In a subsequent study, the three antigen-dependent arms of the adaptive immune system (antibodies, CD4+ T Lymphocytes, and CD8+ T Lymphocytes) were measured in convalescent and acute SARS-CoV-2 patients in the context of disease severity in order to establish their role in infection control and/or disease resolution (122). Surprisingly, neutralizing antibody titers were

poorly correlated with disease severity. In contrast, SARS-CoV-2-specific CD4+ and CD8+ T cells strongly correlated (inversely) with disease severity, with stronger responses associated with lesser disease severity. Strikingly, a convalescent cohort patient was able to clear the infection with detectable SARS-CoV-2-specific CD4+ and CD8+ T cell responses, but with no detectable antibodies, suggesting the importance of cell-mediated immune responses in SARS-CoV-2 infection resolution (122). In addition, several recent longitudinal studies further shed light on the dynamics of adaptive immunity over up to 10 months following infection and/or vaccination (122,124–128). These studies corroborated earlier findings regarding the relatively rapid waning of antibody titers following immunization by SARS-CoV. Measurement of neutralizing antibodies targeting either the SARS-CoV-2 Spike protein or its RBD indicated a decline in titers over 8 months following infection (124,125). In contrast, memory CD4+ T cells were found to persist for as long as 10 months following infection. In study by Jennifer Dan *et al.*, CD4+ T cell memory was detectable in 93% of convalescent individuals one month following infection (53/57 individuals), and in 92% of individuals 6-8 months following infection (33/36 individuals) (124). However, in the same study, CD8+ T cell memory was found to undergo a steady decline in the months following infection, with the rate of detection going from 70% after one month to 50% after 6-8 months. The decline in memory CD8+ T cells observed was corroborated by several other studies. These findings contrast studies pertaining to SARS-CoV, which found detectable CD8+ T cells as long as 11 years following infection (105). At any rate, it is currently difficult to directly compare the dynamics of adaptive immunity specific to SARS-CoV and SARS-CoV-2, given the short time span elapsed since the start of the SARS-CoV-2 pandemic. Longitudinal studies conducted in years to come will continue to shed light on the long-term dynamics of the adaptive immune system following COVID-19 convalescence.

Overall, the findings presented here, regarding T lymphocyte-mediated immune responses in the context of SARS-CoV, MERS-CoV, and SARS-CoV-2, strongly point to the importance of CD4+/CD8+ T cell-based immunity in clearing CoV infection and producing long-lasting T-lymphocyte memory to viral antigens. However, these findings also beg the following question: given the critical role played by T cells in disease resolution, what would be the impact of T-cell epitope-associated mutations on the breadth of CTL responses and disease outcome?

#### 1.1.3.1 Antibody escape

Since the start of the pandemic, SARS-CoV-2 has acquired a significant amount of genomic diversification. As a result of widespread collaborations as well as close genomic monitoring the viral strains, the first mutations were characterized shortly after the identification and naming of the virus. Although the majority of mutations have little to no impact of viral viability and pathogenicity, some mutations play non-trivial roles in viral mutations. In the context of SARS-CoV-2, mutations have been associated with a variety of impacts, including mediating viral entry by enhancing the interaction between the spike glycoprotein and ACE2 receptor; affecting the release of virions following viral replication; and promoting immune evasion. The latter will be the subject of the following sections (65,87,88,129).

Over the course of the COVID-19 pandemic, SARS-CoV-2 has managed to develop several immune evasion mechanisms. In particular, SARS-CoV-2 variants leading to immune escape has been the object of extensive global scientific scrutiny. Antibodies, the primary components of the humoral immune response, were shown to be highly correlated with COVID-19 disease severity (122,130,131). Composed mainly of IgM, IgA and IgG, anti-SARS-CoV-2 antibodies were found to primarily target the Spike glycoprotein as well as the nucleocapsid. The

neutralizing effect of SARS-CoV-2 was shown to manifest itself by disrupting the viral interaction between he Spike protein and the human ACE2 receptor, effectively preventing its entry into host cells (65,83,132). As such, mutations able to convey antibody evasion were hypothesized to occur within the Spike glycoprotein and Nucleocapsid. Such mutations first came to light with the identification of the VOCs. The first VOC, B.1.1.7 (Alpha), initially identified in September 2020, was found to convey partial resistance to neutralizing antibodies. Of the many mutations making up this lineage, 7 missense variations and 3 deletions were located within the spike protein. Numerous studies demonstrated the ability of B.1.1.7 to partially evade recognition by neutralizing antibodies by both convalescent (~3-fold reduction) and by vaccinated (~2-fold) sera (133,134). Of the S-protein variations, N501Y on the RBD, and the Y144 deletion on the NTD were associated with the highest resistance to neutralizing antibodies (133,134). Interestingly, the N501Y mutation was not in fact found to directly disrupt the interaction with neutralizing antibodies, but to enhance the interaction between the RBD and ACE2. This enhanced interaction allows RBD to outcompete neutralizing antibodies in interacting with ACE2. As the NTD does not interact with ACE2, but rather with alternative receptors in cells lacking the ACE2 receptor, the Y144 likely disrupts such interactions (129). Unsurprisingly, the B.1.1.7 variations responsible for the partial reduction in neutralizing activity by convalescent/vaccinated sera were found to completely escape monoclonal antibodies, as a single mutation is sufficient to abrogate the mAb-epitope binding interface. In contrast, the polyclonal nature of sera antibodies allows to circumvent the impact of a few mutations. This feature is especially relevant to the efficacy of monoclonal antibody treatments against SARS-CoV-2.

The VOC B.1.351 (Beta) has been of particular interest for the evasion of neutralizing antibodies. Although this variant possesses 7 mutations and 3 deletions in the spike protein, only

three mutations located in the RBD (N501Y, K417N, E484K) were associated a decrease in neutralization by antibodies (84,86). As in the B.1.1.7 variant, N501Y was shown to enhance the interaction between the RBD and ACE2, therefore outcompeting neutralizing antibodies which would otherwise disrupt this binding interface. K417N and E484K were both found a synergistically enhance the RBD-ACE2 binding interface in combination with N501Y, therefore further competing against the corresponding neutralizing antibodies. Interestingly, the decrease in neutralization efficacy conveyed by the B.1.351 lineage was found to be particularly prominent in individuals with lower neutralizing antibody titers (82). As such, due to the much lower antibody titers observed in unvaccinated convalescent individuals in contrast with vaccinated individuals, this puts unvaccinated individuals at much higher risk of losing neutralizing activity following infection by a virus of the B.1.351 lineage. In fact, this VOC was found to result in a 11-33-fold decrease in neutralizing antibodies in unvaccinated, convalescent individuals, as opposed to a 3.4-8.5 fold decrease in vaccinated individuals.

The P.1 (Gamma) lineage was associated with a decrease in neutralizing activity, although not as severe as the B.1.351 lineage. This may in part be attributed with differences in mutations within the NTD (83,132). However, a disparity was still observed between the impact of the VOC on unvaccinated convalescent and vaccinated convalescent individuals. The B.1.427 and B.1.429 lineages (epsilon) were of interest as they introduced a new mutation, L452R, capable of modulating the activity of neutralizing antibodies. Although this mutation does not directly mediate the RBD-ACE2 interface, is was shown to allosterically enhance the binding interface.

The B.1.617.2 lineage (Delta) has become of great interest to the scientific and medical community due to its significantly superior infectivity and pathogenicity. These increased features can be attributed to the mutations L452R and E484Q, which enhance the ACE2-binding interface,

as well as the P681R mutation within the S furin-cleaving site, enhancing the cleaving of the spike glycoprotein by furin (77,135,136). The antibody escape properties of this VOC were associated with the RBD mutations L452R, E484Q and T478K. E484Q was found to have a lesser impact of neutralization activity than the E484K mutations observed in other VOCs, but was still shown to result in a 10-fold decrease in neutralization by antibodies. Although the mutation T478K it was not directly investigated, a similar mutation, T478I was interrogated in vitro and shown to reduce the neutralizing activity of convalescent sera (89).

Finally, the B.1.1.529 lineage (Omicron) was shown to extensively evade neutralization antibodies (92–94). The humoral immune evasion observed can be explained by 15 of the 50 mutations defining Omicron being found within the Spike RBD (92). As the latter was identified as the primary target of neutralizing antibodies, hyper-mutation of the Spike RBD may be expected to modulate the neutralization of the SARS-CoV-2 Spike protein (90,91). Importantly, Omicron-asociated mutations were found to partially evade neutralizing antibodies induced by Pfizer's BNT162b2 vaccine (95).

Overall, the current set of VOCs were shown to reduce the efficacy of neutralizing antibodies found in convalescent sera, largely due to their ability to enhance the RBD-ACE2 interface. However, it remains to be shown whether this demonstrated decrease in neutralizing activity directly impacts the increased infectivity of this set of VOCs, and to what extent they impact the severity and outcome of the disease.

#### 1.1.3.2 T-cell escape

The genomic diversification of SARS-CoV-2 over the course of the current COVID-19 pandemic has led scientists to closely monitor the potential evolutionary adaptations pertaining to immune evasion. While prevailing VOCs were rapidly associated with a decrease in neutralization activity, none were found to significantly evade recognition by T cells (137,138). Several studies demonstrated that the faction of the immune response attributed to the activation of CD8+ and CD4+ T cells remained largely unaffected by VOCs Alpha, Beta, Gamma and Delta in either convalescent or vaccinated individuals (137,138). However, the studies did not consider the high level of polymorphism associated with the HLA locus throughout the human population. Due to this significant polymorphism, different individuals may recognize different repertoires of T cell epitopes and therefore be affected differently by SARS-CoV-2 proteomic variations. Additionally, the studies presented above were highly focussed on the current set of VOCs, which together, represent a very small proportion of the mutations currently in circulation.

To partially address these concerns, several studies investigated the impact of individual mutations on the activation of T cell epitopes in the context of HLA types. In the earlier stages of the pandemic, Agerer *et al.* conducted an in-depth study of the impact of individual SARS-CoV-2 variations on the ability of SARS-CoV-2-convalescent PBMCs to amount an adequate cytotoxic immune response against the virus (139). Performing analyses of both wild type as well as mutated epitopes, they demonstrated that a subset of mutations found within viral CD8+ T cell epitopes were able to negatively impact the breadth and quality of immune responses. Using epitope presentation predictions (netMHCpan 4.0) complimented by HLA-epitope destabilization assays, Agerer *et al.* identified a subset of circulating mutations predicted to abrogate the binding of mutated epitopes by specific HLA alleles. To further investigate the immunological implications

of these findings, the ability of these mutations to impact the proliferation and activation of CD8+ T cell epitopes were confirmed by Enzyme-Linked Immunosorbent Spot (ELISpots) as well as Intracellular Cytokine Staining (ICS) for IFN-*y*. Finally, Agerer *et al*. used single-cell T Cell Receptor sequencing (scTCR-seq) as well as scRNA-seq on expanded T cells to assess the diversity of T cell populations activated by a single mutated T cell epitope compared to wild type. The results further corroborated a lower level of T cell expansion in the mutant. Expectedly, the TCRs associated with peptide-specific activation were identical for both the wild type and mutant epitope. While TCRs are known to recognize the center-portion of T cell epitopes (residues 3-6), mutations abrogating HLA-epitope binding are generally found at anchor residues (residue 2, 9) of the HLA binding groove, and would not be expected to directly interfere with TCR-recognition. However, analysis of cytotoxic genes resulting from scRNA-seq of peptide specific T cell clusters indicated a lower quality of cytotoxic activation in mutant peptide-specific T cells. Overall, these results establish the ability of a subset of SARS-CoV-2mutations occurring within CD8+ T cell epitopes to reduce the proliferation and activation of CD8+ T cells by abrogating the presentation of epitopes by HLA molecules in an HLA allele-dependant manner.

Although the study introduced above was performed early in the pandemic on a subset of variations that had not yet become fixed in the population, a subsequent study investigated the ability of a two extensively fixed mutation to mitigate the breadth and quality in T cell responses. Motozono *et al.* investigated the ability of the mutations L452R and Y453F to confer resistance to HLA-A24 (65). The former contributes to lineages B.1.427/429 and B.1.617, whereas the latter contributes to lineage B.1.1.298. These are therefore highly epidemiologically relevant. Analyses were focused on a single epitope, an RBD-specific 9-mer spanning residues 448-456 (NF9)

containing both mutation events, which was demonstrated to be an immunodominant HLA-A24 epitope by three independent studies. Stimulation of PBMCs by mutant or wild type epitopes followed by the isolation of CD8+ T cell subsets by FACS indicated that both mutations have the ability to dramatically reduce the activation of CD8+ T cells. Incidentally, the mutation L452R has been of great interest due to its ability to enhance the Spike RBD-ACE2 binding interface, therefore leading to both increased infectivity as well as antibody escape.

Recently, Silva *et al*. Utilized ELISpots as well as TCR repertoire sequencing to identify a series of mutations leading to the complete abrogation of CD8+ T cell response against immunodominant epitopes (140). These included the disruption of a CD8+ ORF3a epitope by Q213K, the disruption of a CD8+ Nucleocapsid epitope by P13X, as well as the disruption of a second Nucleocapsid epitope by T362I/P365S.

Yet another study by Zhang *et al*. investigated four prevalent mutations, K417T (Spike), K417N (Spike), Y144- (Spike) and L452R (Spike) found in variants P.1, B.1.351, B.1.1.7 and B.1.617.2 respectively (141). These mutations caused reductions in CD8+ T cell activations in the context of prevalent HLA types HLA-A*02:01, HLA-A*02:07, HLA-A*11:01 and HLA-A*24:02 respectively. These mutations effectively disrupted the activation of CD8+ T cells to epitopes that were previously experimentally shown to stimulate T cells in most tested samples. T cell activation was obtained by ICS and MHC-epitope interactions were investigated by means of x-ray crystallography. These findings suggest T cell evasion to be a common characteristic amongst variants and to be considered as a putative factor alongside neutralizing antibody escape and enhanced viral entry when interrogating their increased infectivity.

In the context of Omicron, two separate studies suggested its numerous mutations to convey a limited impact on the overall CD4+ and CD8+ T lymphocyte activation and proliferation,

as shown by the stimulation of vaccinated and/or convalescent PBMCs with Omicron or wild-type peptide pools (142,143). Nevertheless, in both studies, a subset of participants sustained a drastic reduction in CD8+ T cell activation (> 50% in some cases) when stimulated with the Omicron peptide pool compared to its wild-type counterpart. Although neither studies proceeded to experimentally investigate a potential link between HLA type and Omicron-mediated T-cell evasion, both groups hypothesized that the significant reduction in CD8+ T cell activation observed in a subset of participant could be explained by variations in HLA profiles.

Collectively, the findings yielded by these investigations indicate the ability of several SARS-CoV-2 mutations to disrupt T cell response in an HLA-dependant manner. Despite the limited adaptative advantage provided by a T cell escape mutation given the high diversity of HLA types throughout the population, the fixation of mutations leading to the abrogation of immunodominant epitopes in the context of prevalent HLA types could impact the quality of cellular immune response in a number of individuals. Given the on-going evolution of SARS-CoV-2, the accumulation of genomic variations and the emergence of new lineages, surveillance for variations (or combinations of variations) conferring T cell evasion continues to be relevant to our understanding of SARS-CoV-2 infectivity and virulence.

### *1.1.4 SARS-CoV-2 and the innate immune system*

The role of the innate immune system in resolving SARS-CoV-2-driven infections is beyond the scope of this dissertation. Nevertheless, innate immunity plays an important role in mediating viral infections and is therefore worth a brief mention. The innate immune system plays a crucial role in the host response to viral infections and does so by impeding key steps of

the infection cycle and by aiding the adaptive immune system. It is largely mediated by molecules known as pattern-recognition receptors (PRRs) which are released by a variety of leukocytes, and manifests itself through the induction of inflammatory pathways (144,145). Found in endosomes, cytoplasm, and on the membrane, PRRs recognize molecular patterns specific to pathogens, known as pathogen-associated molecular patterns (PAMPS). In addition, they can recognize damage-associated molecular patterns (DAMPs), molecular entities indicative of cellular damage (144). Recent decades have shed much light on the various members of PRR families, which include Toll-like receptors (TLRs), C-type lectin receptors, retinoic acid-inducible gene I (RIG-I)-like receptors (RLRs), absent in melanoma 2 (AIM-2)-like receptors, and nucleotide-binding oligomerization (NOD)-like receptors (NLRs) (144). Several studies have sought to elucidate the relationship between SARS-CoV-2 and PRR activity during infections. And are briefly discussed below.

*TLRs.* Multiple groups corroboratively demonstrated the role played by one particular TLR, TLR2, in sensing SARS-CoV-2 and subsequently facilitating pro-inflammatory mechanisms (146). TLR2 was shown to induce pro-inflammatory cytokines and IL-6 production through the recognition of the SARS-CoV-2 E protein. Stimulation of either human macrophages treated with TLR2 inhibitors or murine macrophages deficient in TLR2 with the E protein led to a decrease in inflammation. Other TLRs, including but not limited to TLR1, TLR3, TLR4 and TLR6 were hypothesized to mediate SARS-CoV-2-specific innate immunity, although further conclusive experimental evidence is needed (147,148).

*RLRs.* SARS-CoV-2-specific innate immunity was shown to be mediated by MDA5 and LGP2, two RLRs well-studied in the context of IFN regulation (149,150). Using small interfering RNAs, Yin *et al.* (2021) knocked down 16 viral RNA sensors as well as MAVS in

Calu-3 cells (an airway epithelial cell line) infected by SARS-CoV-2 (150). Knockdown of MDA5 and LGP2 was found to reduce type 1 IFN-expression and increase viral replication. Surprisingly, the same was observed for NOD1, an intracellular bacterial peptidoglycan sensor.

*NLRs.* NRLs were shown to be involved in mediating SARS-CoV-2 infections by modulating the generation of inflammasomes and type 1 IFN (151). NLRP3, a very well-characterized NRL shown to recognize both PAMPs and DAMPs as well as to induce caspase-1 and cell-death, was associated with SARS-CoV-2 sensing. SARS-CoV-2 proteins ORF3a and N and single-stranded viral RNA were shown to induce NLRP3 inflammasome, caspase-1 and cell death in studies using HEK293 cell lines (152–154).

Although the innate immune system was shown to play a crucial role in clearing SARS-CoV-2 infection through the PRR-mediated release of IFN and pro-inflammation cytokines (155), it was shown to have dichotomous implications. When dysregulated, the leukocyte-mediated release of cytokines, known as a cytokine storm, results in a potentially lethal condition. Karki *et al* (2021) demonstrated that the excessive production of cytokines TNF-*a* and IFN-*y* resulted in inflammatory cell death, also known as PANoptosis (156). By replicating the synergistic impact of these two cytokines in mice, Karki *et al*. induced a lethal shock akin to that observed in COVID-19 patients experiencing a cytokine storm. Such cytokine storms were associated with disease severity as well as multi-organ damage (156,157).

## 1.2 Epitope presentation

The adaptive immune system is a highly complex component of jawed-organisms that has evolved to become an exquisitely multivalent and adaptable defense mechanism. At its core, the adaptive immune system is made up of two arms: the cellular immune system as well as the

humoral immune system. Although these two arms best function complimentarily and are both required for an individual to achieve the full breadth of its immune response, they both manifest their roles in very different ways, and are made up of different components. In this section, we will review the mechanisms enabling the presentation of viral epitopes by HLA class I and II molecules and their subsequent recognition by CD8+ and CD4+ T lymphocytes, respectively.

### *1.2.1 MHC class I and II antigen processing and presentation pathway*

Cellular immunity has been the subject of significant interrogation and scrutiny over the last several decades. T cells, which are at the epicenter of the cellular arm of the immune system, have been shown to be key players in the mediation of a wide range of conditions, including infectious diseases, autoimmune diseases, and cancer. Although T cells are intimately involved in each of these classes of conditions, the manner in which they carry out their roles and the impact that a strong T cell-based immune response has on the condition varies enormously. However, at the very center of the ominous presence of T cells throughout our immune system and defense mechanism lies one key questions: how do T cells know what to attack, and when to attack it? Whether we are talking about infectious diseases, autoimmune diseases, or cancer, this question introduces a key concept that is indispensable in understanding how our immune system functions, and how we can batter assist it. Although it remains an open-ended question, many research groups and numerous important scientific breakthroughs over several decades have managed to provide a highly detailed set of answers, which can be boiled down to one word: **epitope**. Significant scientific efforts have demonstrated that T cells are equipped with surface receptor, called the T Cell Receptor (TCR). Briefly, the interaction between a TCR recognizes its corresponding target, known as an epitope, triggers the activation and mass expansion of the T cell in question, therefore

providing the immune system the proper toolkit to eliminate the recognized insult. Epitopes consist of a short fragment of protein, called a peptide, and is presented to T cells by Major Histocompatibility Complexes (MHC) class I and II. The Human Leukocyte Antigen consists of the human version of this complex. HLA class I and II both play different, yet highly complementary roles in T cell activation. HLA class I is present on virtually all nucleated cells of the human body and presents epitopes to a subset of T cells called CD8+ T cells. The primary outcome of this interaction is to convert CD8+ T cells into CTLs, which have for primary role to eliminate the cells responsible for the activation. Examples consist of infected or cancerous cells. In contrast, HLA Class II are only displayed on professional Antigen Presenting Cells (APCs), including dendritic cells and macrophages. The will take up harmful agents circulating in the vascular system, such as viruses, process them and present the resulting epitopes to CD4+ T cells, also known as T helper cells ($T_h$ Cells). This name is quite appropriate as once activated, these T cells will facilitate many other factions of the immune system. In part, this interaction will cascade the activation of CTLs as well as the activation of B cells, resulting in the production of antibodies.

Another complementary question which has baffled scientists since the initiation of modern immunology is the following one: with so many different types of viruses, bacteria, cancers and other diseases, how could the immune system possibly be able to generate a highly dedicated and specific immune response for every possible insult to the human body? The answer to this question not only directly involves epitope recognition, but also has led people to think of the immune system as one of the most elegant components of human biology. HLA, the molecules responsible for presenting epitopes to T cells and therefore at heart of the interface between disease and immunity, are amongst the most polymorphic regions of the human genome. The HLA locus,

located on chromosome 6, is composed of multiple genes, which can be spliced into a vast array of combinations. Due to a high rate of variation throughout human populations, roughly, 25,000 different HLA molecules have been identified to date (158).

In this section, the various components conducive to the generation, presentation, and recognition of epitopes will be discussed.

#### 1.2.1.1 Proteasomes, TAP, HLA molecules

Epitope processing and presentation is central to the activation of the adaptive immune system. This review will explore the relationship between SARS-COV-2 and the T-lymphocyte arm of the adaptive immune system. Specifically, the impact of SARS-CoV-2's mutation landscape on T-cell epitope processing and presentation will be discussed. The antigen presentation pathway has been substantially interrogated throughout the last four decades and has been extensively reviewed elsewhere (159). Briefly, viral epitopes can lead to T lymphocyte activation via two distinct pathways: MHC Class I epitope presentation pathway and MHC Class II epitope presentation pathway (Figure 1.4). MHC class I molecules are ubiquitously presented on all nucleated cells, and present endogenously generated peptides. Upon viral entry into a target cell and viral replication, a proportion of viral proteins are degraded into peptides by the proteasome, a 20S cylindrical complex associated with two 19S complexes at each end. Following proteolysis, peptides are transported to the ER by the Transporter associated with Antigen Processing (TAP), a heterodimeric complex composed of TAP1 and TAP2 (160,161). In the ER lumen, peptides will be further trimmed at the N-terminus by ER aminopeptidases ERAP1 and ERAP2 (162,163) before forming a complex with several other proteins, including an MHC class I molecule, resulting in the Peptide-Loading Complex (PLC). Finally, the resulting epitope-loaded

MHC class I molecules are presented on the cell surface, where they may interact with their complimentary TCR presented by CD8+ CTL, resulting in the CTL-mediated destruction of infected cells.

In contrast, MHC class II molecules are strictly presented on professional APCs, including dendritic cells, macrophages and B cells. Unlike class I epitopes, class II epitopes are generated through the degradation of viral antigens via the endosomal pathway. Upon loading onto MHC class II, the MHC-epitope complex is transported to the cell surface and presented to CD4+ T Lymphocytes, which in turn plays numerous key roles within the confines of the innate as well as adaptive immune systems, including induction of B-cells as well as CD8+ Cytotoxic T Lymphocytes.

#### 1.2.1.2 Class I classification, supertypes

One particular hallmark of the peptide presentation pathway is the extremely high polymorphism associated with MHC molecule-encoding genes throughout the human population. Briefly, HLA genotypes have been traditionally classified into a genotype-specific nomenclature, classified by allele group (gene locus), HLA protein (in order of discovery), synonymous mutations and mutations in non-coding regions. This detailed and specific nomenclature system results in a highly heterogenous population composed of thousands of different HLA genotypes with a wide array of binding motifs. A binding motif is the preferred amino acid sequence required to obtain a stable binding interface between an HLA molecule and epitope, and it is defined by anchor residues. These play key roles in the interaction and are generally found at position two (P2) and the C-terminus of class I epitopes and interact with the B and F pockets of HLA binding

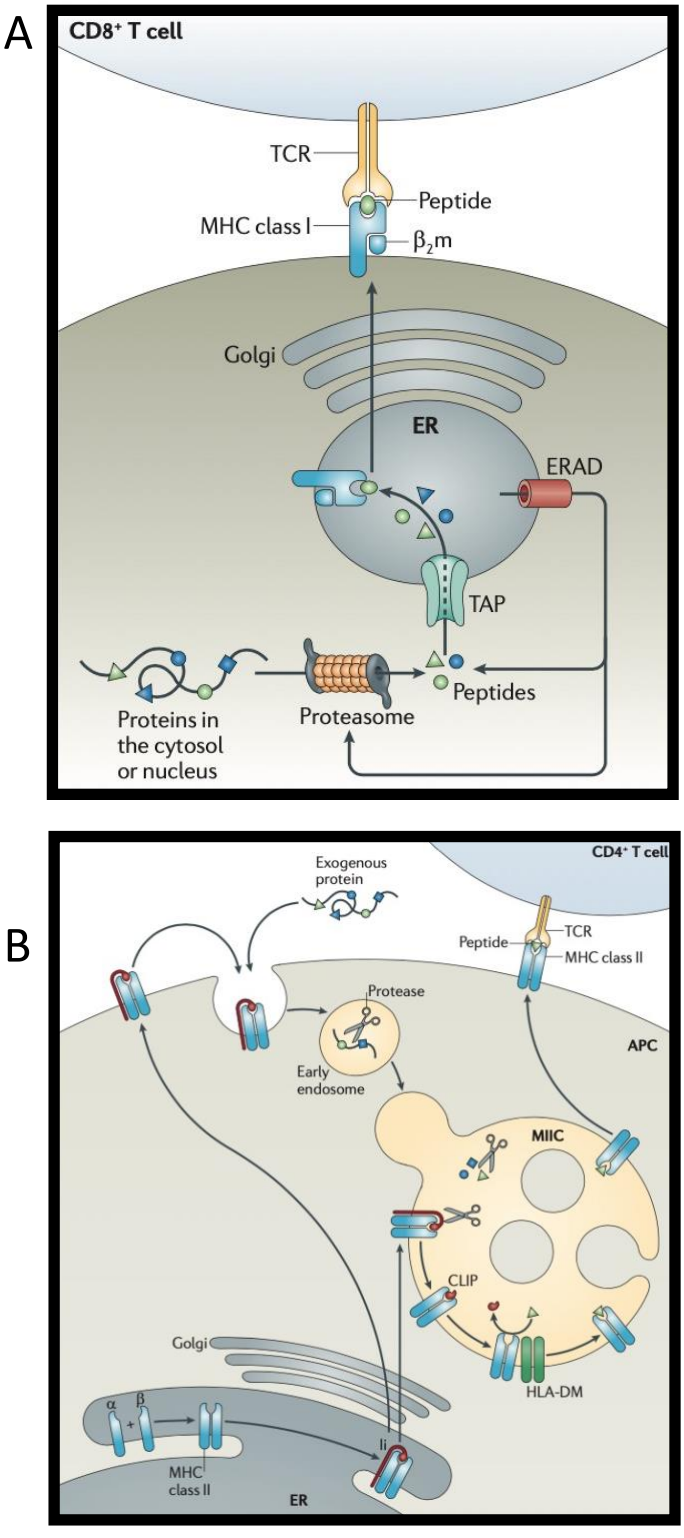



Figure 1.4. Presentation of cytosolic peptides by HLA molecules. Peptides may be presented by HLA class I (a) or HLA class II (b) molecules. Adapted from Neefjes et al. 2011 (159)

grooves (164,165). In an effort to overcome the high level of polymorphism and complexity within class I HLA molecules, these were clustered into groups called supertypes, based on similarities between their binding motifs, thus allowing analyses to consider hundreds of HLA types at once, as opposed to assessing HLA types individually. Supertype classifications have proven extremely useful and have allowed biding motif-based conclusions to reach much broader target groups (166).

### *1.2.2 Identification of HLA class I epitopes*

Although the identification of epitopes by T cells has been helmed as a central component of the cellular immune system, the comprehensive study of epitopes has proven to be a considerable challenge. Since its discovery, this knowledge gap has sparked a flurry of technological and computational innovations, and has led to the inception of numerous new fields of research and societies. The study of the collection of epitopes presented by an individual, now coined the 'immunopeptidome', dates from the early 1990s (167–171). The study of the immunopeptidome has brought to light many of the challenges making its investigation difficult. These include the high abundance and diversity of epitopes presented by HLA molecules, the differentiation of self and foreign epitopes, and the recognition of epitopes by T cells. These challenges have been addressed by innovations in mass spectrometry, immunology, proteogenomics, and computational tools. Together, these innovations have allowed scientists to paint a highly comprehensive and detailed picture of the immunopeptidome in the context of immunity and disease. Here, we will discuss the main techniques that have been developed to interrogate the immunopeptidome and its relationship with T cells.

#### 1.2.2.1 Epitope presentation predictions

T cells are a key component of the cellular arm of the adaptive immune system and are essential in the elimination of foreign agents as well as the elimination of disease. One of the central features allowing T cells to fulfill their function lies in the ability of TCRs to recognize epitopes processed by the peptide presentation pathway and presented by HLA molecules on the surface of cells. The identification of immunogenic epitopes modulating the ability of T cells to eliminate disease has been a central component within the field of immunology. In the context of oncology, epitope identification has enabled the identification of cancer-specific neoepitopes, facilitating the development of immunotherapies. In the context of virology, the identification of immunogenic viral epitopes has contributed to our understanding of antiviral immunity while aiding vaccine development.

Initial experimental approaches for the identification of T cell epitopes were low throughput and costly. Epitope-prediction tools presented a paradigm-shift in immunology and are now central to the characterization of epitopes. Although epitope predictions are not flawless and plagued by high rates of false positives, and must therefore be thoroughly experimentally validated, they are essential in shortlisting putatively clinically relevant epitopes in the context of immunotherapy, infectious diseases and vaccine design.

The majority of these tools rely on machine learning approaches trained on two possible types of data: eluted ligands (EL), consisting of epitopes eluted from MHC molecules by immunoprecipitation and identified by mass spectrometry, as well as binding affinities (BA), the experimentally determined binding affinity between epitopes and MHC molecules measured by binding assays. Inevitably, advancements in the field of machine learning as well as the expansion of EL and BA training datasets has enabled tremendous improvements in the precision and

accuracy of epitope prediction software. Early computational tools were developed in the late 1980s and 1990s (172–174). Since then, many different tools have been released, which will not all be discussed here. However, in 2020, Paul *et al*. compared 15 different epitope-prediction tools in an in-depth benchmarking analysis and concluded NetMHCpan 4.0 and MHCFlurry 2.0 to provide superior predictive power (175–177). The NetMHCpan algrithm series are robust and have been extensively used by the scientific community since their inception (initial version in 2007) (175,178–180). NetMHCpan predictions rely on a combination of EL and BA datasets, and are made using a neural network-based approach named NNAlign (181). The latter in fact is the central component of a series of very successful predictions, including NetMHC, NetMHCII, and NetMHCIIpan. MHCflurry 2.0 is a relatively recent tool (initial version in 2018). Like NetMHCpan, MHCFlurry was trained on both BA and EL datasets and relies on neural network.

#### 1.2.2.2 Mass Spectrometry

The immunopeptidome is composed of tens thousands of different epitopes presented by HLA class I and II, making it challenging to study (182–184). Although predictions from genomic sequences have proven incredibly useful in identifying putative T cell epitopes, they are still plagued with a high rate of false positives. To this day, the method of choice to accurately quantitate and characterize immunopeptidomes remain mass spectrometry (MS). Traditionally, the interrogation of immunopeptidomes by MS involves the immunoprecipitation of HLA-bound peptides. In this process, HLA-peptide complexes are initially captured by HLA-specific antibodies, after which the peptides are separated from HLA molecules using acid-elution. Fragment Ion Spectra are then generated for each peptide detectable by MS in a process known as Data-Dependant Acquisition (DDA) (185,186). DDA-based immunopeptidomics has proven to be

a robust analytical approach for a variety of tissue and cell types, with the ability to detect north of 10,000 unique epitopes depending on the instrument (182–184). The MS-based identification of immunopeptidomes has been successfully utilized in the context of a wide range of immune-driven diseases, including cancers, autoimmune diseases and infectious diseases (187–193). In the context of viral infections, which constitute the object of this dissertation, DDA-MS has been used to identify viral epitopes presented by both HLA class I and II, in efforts to identify putative vaccine targets (194).

However, DDA experiments are limited in their reproducibility, making it difficult to consistently characterize and quantify HLA peptides across different samples, conditions, or tissue types. A rapidly-emerging approach, known as Data-Independent Acquisition (DIA) has been suggested to address this caveat. Sequential Window Acquisition of all Theoretical Mass Spectra (SWATH-MS) is a concrete implementation of DIA-MS, and has shown to provide a strong alternative to DDA-MS, greatly improving the reproducibility and sensitivity across multiple samples (195).

#### 1.2.2.3 T cell activation

There are two folds to the investigation of T cell epitopes. The first consists of the identification of epitopes, an endeavor helmed by mass spectrometry. The second is the identification of clinically and immunologically relevant epitopes. In a clinically relevant context, an interest consists of identifying epitopes capable of activating T cells and leading to a robust immune response. Although the ability of HLA molecules to present epitopes does not guarantee its immunogenicity, experimentally-determined HLA-epitope IC50 thresholds have been highly conducive to the identification of cancer- and virus-specific T cell epitopes (176,196–198).

Nevertheless, experimentally assessing the ability of epitopes to activate T cells has been a crucial component of the identification of clinically relevant T cell epitopes. Many techniques have been developed to address this need, spanning a wide range of tools and biological features of T cell activation. Methods commonly used to assess the activation of CD8+ and CD4+ T cells include, but are not limited to Enzyme-linked immunosorbent spot (ELISpot) assays, Intracellular Cytokine Staining, Activation-Induced Markers (AIM) assays, and Target:Effector assays. As these have been described in detail in many reviews, they will not be discussed here. However, such techniques have played central roles in the identification SARS-CoV-2 epitopes (123,199).

#### 1.2.2.4 Immunosequencing

The adaptive immune system has proven to be a highly complex system, embellished by the high level of polymorphism in genes responsible for the specificity of MHC molecules as well as TCRs, and B cell receptors (BCR) (200,201). Immunosequencing is a method that was developed to address and interrogate this high level of complexity. Although immunosequencing may be applied to both BCR and TCR sequencing, this section will focus on the latter. Briefly, TCRs are receptors on the surface of T cells and are responsible for antigen specificity through the recognitions of MHC-peptide complexes. TCRs are composed of two chains, with the majority being made up of chains alpha and beta (TCR$\alpha\beta$), although unconventional TCR may be composed of chains gamma and delta (TCR$\chi\delta$) (200–202). These chains result from the combination of randomly-selected V (variable), D (diversity) and J (joining) gene segments. The random selection of these gene segments, known as V(D)J recombination, determines the conformational diversity of Complementarity Determining Regions (CDR) loops (three per chain, for a total of six CDR loops), which determine the antigen specificity of T cells. CDR3 accounts

for the most variable region of the TCR and therefore plays a significant role in the diversity of antigens targeted by T cells.

Interrogating the diversity and dynamics of T cell populations has allowed researchers to develop a deep and highly comprehensive understanding of the relationship between cellular immunity and disease. This has been achieved by conducting high throughput sequencing (HTS) of CDR3 loops from T cells isolated from patients (200). A flurry of computational tools has been developed to interrogate the resulting genomic data, inferring the diversity of T cells, the dynamics of immune responses, the identification of clonally expanded T cells, and the inference of antigens responsible for such clonal expansions. Briefly, immunosequencing approaches include bulk sequencing, as well as single-cell sequencing.

*Bulk sequencing*. Following TCR enrichment, RNA is sequenced using available HTS methods, with Illumina MiSeq and HiSeq being commonly used (200). As single-nucleotide differences are crucial in analyzing and clustering TCRs, a series of pipelines have been developed to extract TCR repertoires while accounting for sequencing errors. One of the limitations of bulk repertoire sequencing is the inability to pair alpha and beta chains, an important step in determining antigen specificity. However, many applications are still achievable. Several approaches allow the assessment of clonal diversity in an individual in the context of disease as well as the comparison of T cell clonal diversity between individuals.

*Single cell TCR sequencing (scTCR-seq)*. scTCR-seq has become popular in recent years, as it helps overcome some of the limitations of bulk sequencing. Namely, this approach permits the pairing of alpha and beta sequences from each individually analyzed T cell, therefore providing detailed information on clonal expansion as well as allowing the inference of antigen specificity (203–205). Previous methods involved complete transcriptome sequencing at the single cell level

using microfluidics, using methods such as Smart-Seq2 (206). Resulting RNA-seq data could subsequently be mined for TCRα and TCRβ pairs. However, such methods were costly and allowed the analysis of a limited number of cells (hundreds). Other recent advances have increased the number of cells to be analyzed, while reducing costs. Such methods employ cDNA Illumina short-read sequencing, which can be performed using 3' or 5' sequencing. More recently, a new method employing a combination of long-read (Oxford Nanopore) and short-reads alongside high-throughput droplet-based methodologies (202). This workflow, called RAGE-seq, allows for the accurate sequencing of full TCRs on thousands of cells.

The sequencing of TCR repertoires has provided access to a rich source of information regarding the dynamics of adaptive immune responses. As such, many computational tools have been developed to interpret the resulting genomic datasets. For example, the GLIPH2 software was developed to cluster T cell clonotypes based on antigen-recognition similarities (207). In addition, advances in structural modelling and well as the accumulation of crystal structures in the Protein Data Bank (PDB) have enabled the *in sillico* modelling of TCRs alone, TCR-antigen complexes, as well as TCR-antigen-HLA complexes (208,209). Such algorithms have not only allowed to deepen our understanding of the molecular dynamics and selection process behind TCR-antigen recognition, but have also facilitated the structure-based *in silico* identification of antigens from TCR repertoire analyzes. Finally, numerous tools were developed to infer antigen specificity form TCR sequences, relying on TCR-epitope datasets (210–213).

Overall, immunosequencing provides a unique opportunity to study not only the dynamics of adaptive immune responses, but also to acquire to acquire in-depth and comprehensive knowledge of key antigens responsible for T cell clonotype expansion in relation to disease.

## 1.3 Immune escape

### *1.3.1 Disruption of peptide presentation*

Since the start of the pandemic, SARS-CoV-2's mutational landscape has been extensively described and investigated, with mutations such as D614G in the spike glycoprotein and P214L in ORF1ab showing strong signs of selection early on in the pandemic (48,214–216). Beyond quantitatively and qualitatively characterizing the virus' mutational landscape through phylogenetic analyses, studies have been conducted to establish the role of mutations in pathogenicity, and disease severity (48,214). Such investigations included cohort studies aimed at correlating mutations to symptom severity and disease outcome, as well as structural biology studies aiming to understand the impact of certain SARS-CoV-2 mutations on viral protein functionality, and on virus-host interaction dynamics (217–219).

Despite the extent and importance of these studies, one aspect of virus-host interaction that has yet to be interrogated: The impact of SARS-CoV-2 mutations on peptide presentation and T-Cell activation. A thorough investigation of the relationship between SARS-CoV-2 and the peptide-presentation pathway is urgently needed to understand the impact that such a relationship would have on disease outcome and vaccine efficacy. Although not yet studied in SARS-CoV-2, CTL escape variants have been previously studied in numerous organisms.

Viral epitope-associated mutations were by Pircher *et. al* in 1990, wherein CTL responses to epitope variants were investigated in transgenic mice models (220). Viral epitope-associated mutations have since been extensively investigated in HIV (Human immunodeficiency virus) type 1 as a mode of immune escape, with the first evidence of CTL escape mutants in HIV-infected

cohort participants published in 1991 by Philips R.E. *et. al* (221). In this study, naturally occurring and accumulating mutations in 1 HLA-B*27 and 3 HLA-B*8 restricted epitopes within HIV-1's *GAG* protein were investigated in 6 patients throughout longitudinal studies, and multiple mutations were found to dramatically impact CTL responses specific to these epitopes. Over the last two decades, numerous studies proceeded to investigate CTL escape mutations at population and individual levels. Three mechanisms of escape have thus far been described with regards to CTL-escape variants (Figure 1.5).

#### 1.3.1.1 Disruption of epitope processing

In the first mechanism, CTL-escape mutations within or up(down)stream of the affected epitopes lead to the impairment of antigen processing (222–224). In a study by Yokomaku *et al*, certain mutations were shown to have little to no impact on in-vitro binding assays yet were able to successfully eliminate CTL responses. Draenert *et al* experimentally validated a mode of action for this type of CTL escape, consisting of impairing peptide trimming at the $NH_2$-terminus by the ER-associated aminopeptidase-I prior to HLA-loading.

#### 1.3.1.2 Disruption of HLA-epitope binding

In the second mechanism, CTL-escape mutations act by disrupting the HLA-epitope binding interface (225–232). HLA-peptide interactions have been well characterized for a wide range of HLA class I allotypes and are known to be highly dependent on crucial interactions involving anchor residues. In HLA class I, these anchor positions generally occur at position 2 and/or the C-terminus of epitopes. Disruption of these anchor positions by mutation events has been shown to be highly detrimental to the interaction between epitopes and the B/F binding

pockets of HLA molecules. In a study by Carlson *et al* in which HLA-associated polymorphism were analysed in a cohort of 1,888 infected (treatment-naïve) individuals, anchor residue-associated mutations occurred 1.8-fold more than at other positions amongst 9-mer epitopes, and were predicted to induce a 10-fold reduction in HLA-binding affinity (226). From a functional perspective, multiple studies have experimentally connected peptide binding and CTL-response impairment to epitope-associated mutations. Examples include the well-established I135R mutation in HIV-1's DNA Polymerase (Pol) protein, which was shown to decrease binding of peptide TAFTIPSI to HLA-B*51, and to subsequently reduce the breadth of B*51-TAFTIPSI-specific CD8+ T cell responses (229). R264X mutations at anchor position 2 of the immunodominant B*27-restricted epitope KRWIILGLNK (*gag* 263-272) resulted in significant reduction of binding leading to a reduction in B*27-KRWIILGLNK-specific CD8+ T cell response (230,233,234). R264T and R264Q were shown to lead to a 348 and 30-fold decrease in binding affinity, respectively, and to both dramatically impair CTL response.

#### 1.3.1.3 Disruption of epitope-TCR binding

In the third mechanism, epitope-associated mutations may specifically disrupt the interaction between a T-Cell Receptor (TCR) and a peptide-HLA complex. In such cases, the epitope-associated mutation will minimally disrupt peptide-HLA interactions but will result in a reduction in CTL response to the peptide-HLA complex(235–237). While escape mutants have been extensively described in HIV-1, their role in SARS-CoV-2 infection remains unknown. The shaping of anti-SARS-CoV-2 CTL responses by SARS-CoV-2's mutational landscape could not only provide a more comprehensive understanding of disease severity, but could also shape vaccine development.

### *1.3.2 SARS-CoV-2 T cell epitopes*

#### 1.3.2.1 SARS-CoV-2 T cell epitope databases

An important component for the characterization of the relationship between a virus and the cellular immune system consists of identifying the repertoire of clinically relevant T cell epitopes. Important features include HLA restrictions and immunodominance. Over the course of the current pandemic, numerous groups have investigated and identified CD8+ and CD4+ T cell SARS-CoV-2 epitopes (238–241). To facilitate the analysis of clinically relevant T cell epitopes, Quadeer e*t al.* developed a database along with a wed interface providing information to the scientific community regarding the most up-to-date set of experimentally validated CD8+ and CD4+ T cell SARS-CoV-2 epitopes (242). Thus far, Quadeer *et al.* reviewed 25 separate studies, from which 843 epitope-HLA pairs in the context of 60 distinct HLA types were acquired. In addition, the database information regarding the relevance of each epitope, in the form of standard frequency response (RF) as well as the number of distinct studies reporting each individual epitope-HLA pair. The web application can be found at https://www.mckayspcb.com/SARS2TcellEpitopes/.

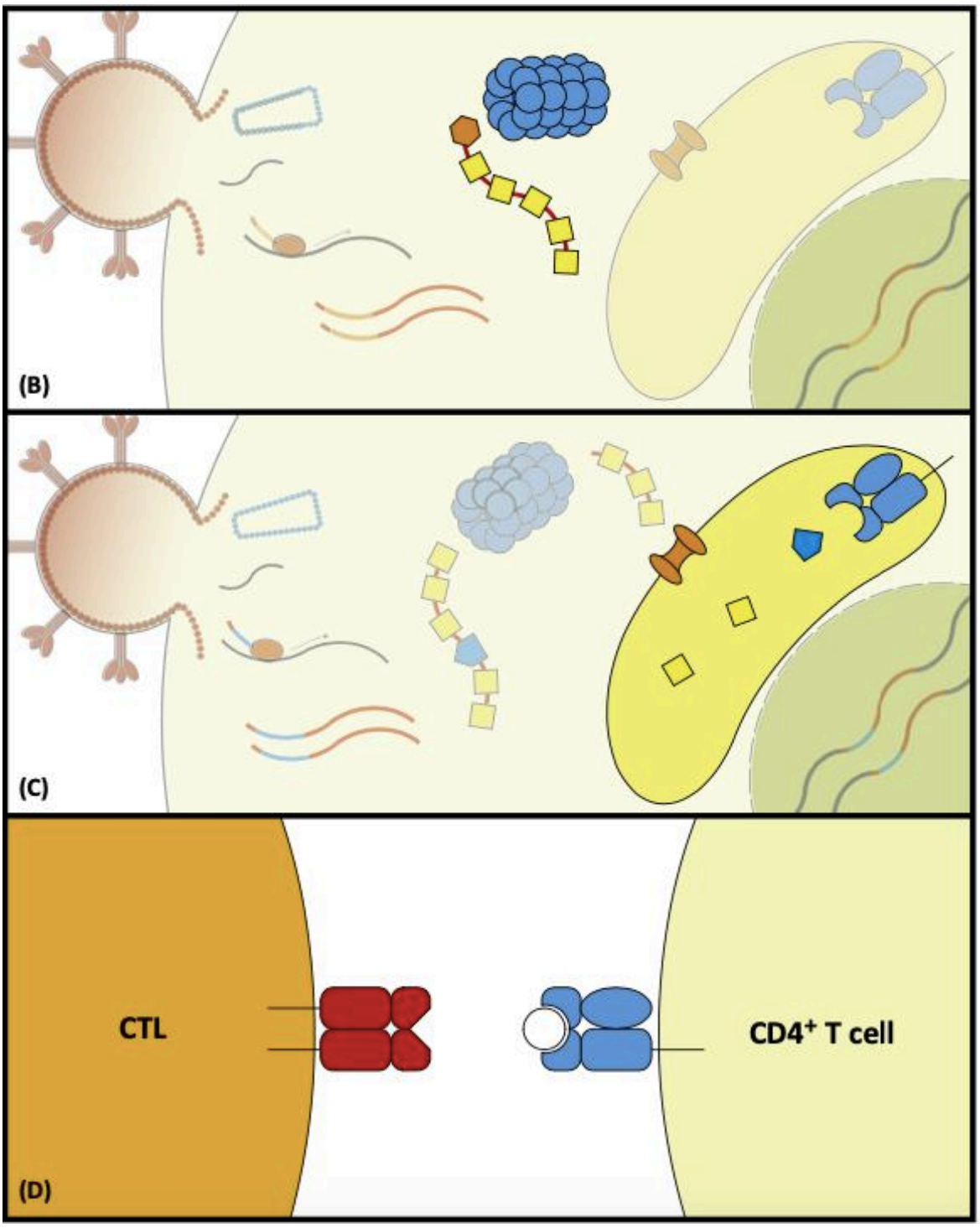



Figure 1.5. Mechanisms of T cell-based immune evasion by HIV-1. Viral mutations may evade T cell recognition by disrupting HLA-epitope binding (a), disrupting epitope processing (b), or disrupting TCR-epitope binding (c). Adapted from Carlson et al. 2014 (243).

## 1.4 Hypotheses and Objectives

Hypothesis: We hypothesize that the mutation types describing the global SARS-CoV-2 mutational landscape over the course of the first year of the pandemic will not randomly distributed, but rather governed by specific mutation types. Moreover, we hypothesize that the observed mutational patterns will disproportionately impact the presentation of CD8+ T cell epitopes in an HLA-dependant manner.

Objectives: *i)* Using a combination of in-house and standardized software, we will analyze the SARS-CoV-2 viral sequences from the first year of the pandemic available on GISAID in order to characterize the global mutational landscape. *ii)* Using a combination of HLA-peptide binding assays, peptide presentation predictors, as well as datasets of externally validated CD8+ epitopes, we will assess the impact of the global mutational landscape on the presentation of epitopes by HLA alleles. This analysis will be conducted in an HLA dependent as well as HLA supertype dependent manner.

# 2 CHAPTER II: ARTICLE

Titre: 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 A. 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)

## 2.1 Abstract

The rapid, global dispersion of SARS-CoV-2 has led to the emergence of a diverse range of variants. Here, we describe 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. We analyzed a total of 330,246 high quality SARS-CoV-2 genome assemblies, sampled across 143 countries and all major continents from December 2019 to December 2020 before mass vaccination or the rise of the Delta variant. 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; this 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.

## 2.2 Introduction

As of September 2021, the COVID-19 pandemic, caused by the novel Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2), has led to upwards 4.6 million deaths and 222 million confirmed cases worldwide (https://coronavirus.jhu.edu/map.html), making vaccine development and deployment an urgent necessity (244). 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 (245).

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 (246,247). 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 (248–257). 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 (258) and, in some cases, observed to be undetectable just a few months post-infection (259). In contrast, an increasing number of studies point CD4+ and CD8+ T cells as key regulators of disease severity (253,255,260–262). 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 (258,263). Similar observations have been made for the most closely related human coronavirus, SARS-CoV, for which T cells have been detected 11 years (264) and 17 years (251) after the initial infection, whereas antibodies were noted to be undetectable after 2-3 years (265–267). 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 (268). To this end, bioinformatics tools were developed to predict T cell-specific epitopes during the early phase of the pandemic (269). 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 (270). 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 (270). As of May 2021, ~700 HLA class I-restricted SARS-CoV-2-derived epitopes have been experimentally validated (https://www.mckayspcb.com/SARS2TcellEpitopes/) (271).

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 (269). 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 (272–277). This extensive diversification has resulted in widespread variants such as B.1.1.7 (Alpha), B.1.351 (Beta), B.1.617.2 (Delta) (66,67,88). Although the Delta lineage was not yet present in the human population during the first year of the pandemic, it is of the utmost importance to continually interrogate the relationship between emerging SARS-CoV-2 variants and the adaptive immune system (278). In addition, it is important to highlight here that the pool of mutations observed in SARS-CoV-2 sequences were shown to be 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 (273,279–286). Since shown for other viruses (287,288), 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.

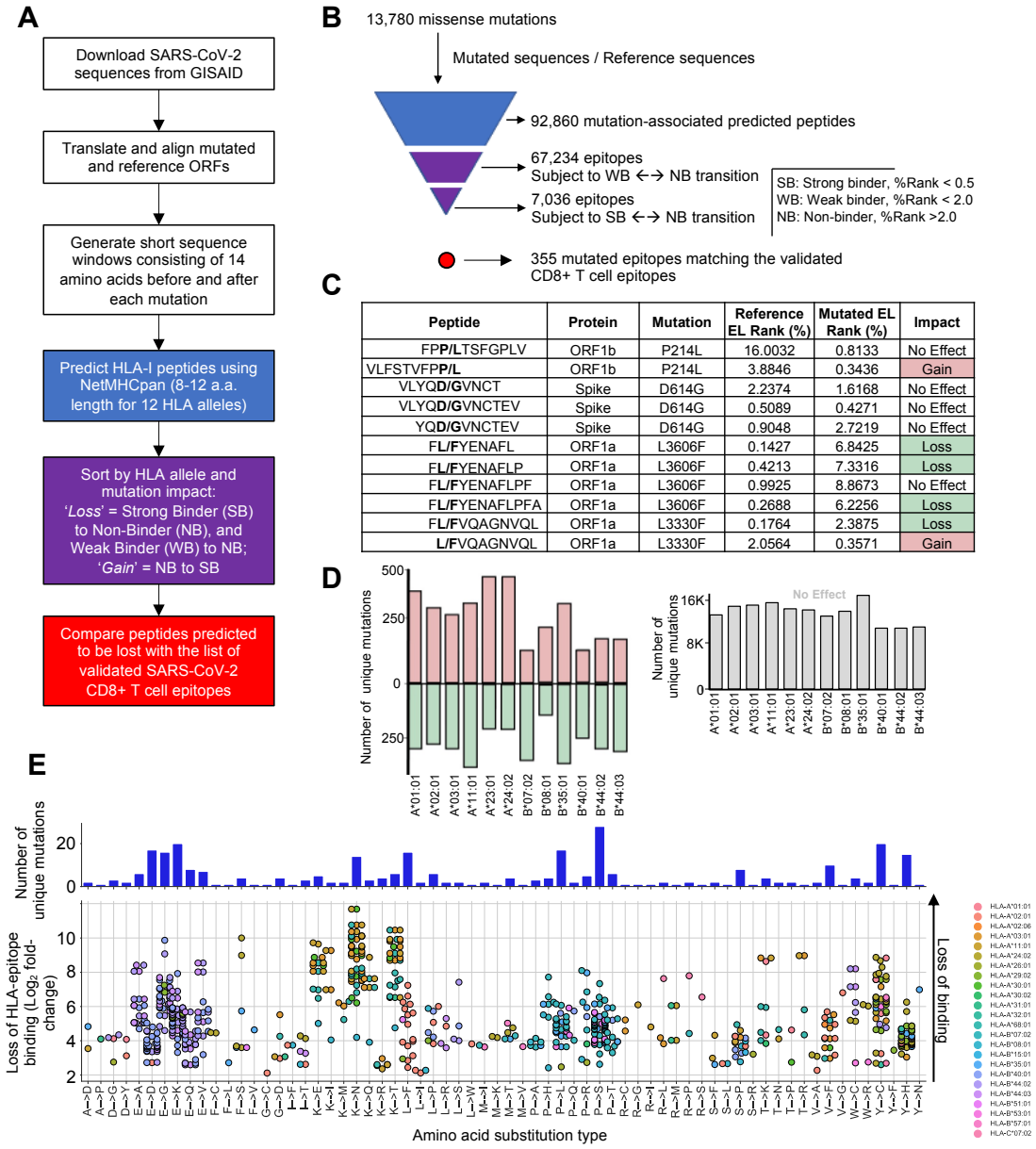
## 2.3 Results

### *2.3.1 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 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-CoV-2 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 2.1**). 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%) (274) (**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 (270,271) (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 2.1**). 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 2.2A,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 2.2B**). Epitopes located in the RBD vaccine locus were also impacted by mutations (**Figure 2.2C**).

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 2.S1**). 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 2.S1A,B**). Our analysis also identified a mutation in the HLA-B*07:02-restricted N105 epitope SPRWYFYYL, which is one of the most




| Peptide | Protein | Mutation | Reference EL Rank (%) | Mutated EL Rank (%) | Impact |
|---|---|---|---|---|---|
| FPP/LTSFGPLV | ORF1b | P214L | 16.0032 | 0.8133 | No Effect |
| VLFSTVFPP/L | ORF1b | P214L | 3.8846 | 0.3436 | Gain |
| VLYQD/GVNCT | Spike | D614G | 2.2374 | 1.6168 | No Effect |
| VLYQD/GVNCTEV | Spike | D614G | 0.5089 | 0.4271 | No Effect |
| YQD/GVNCTEV | Spike | D614G | 0.9048 | 2.7219 | No Effect |
| FL/FYENAFL | ORF1a | L3606F | 0.1427 | 6.8425 | Loss |
| FL/FYENAFLP | ORF1a | L3606F | 0.4213 | 7.3316 | Loss |
| FL/FYENAFLPF | ORF1a | L3606F | 0.9925 | 8.8673 | No Effect |
| FL/FYENAFLPFA | ORF1a | L3606F | 0.2688 | 6.2256 | Loss |
| FL/FVQAGNVQL | ORF1a | L3330F | 0.1764 | 2.3875 | Loss |
| L/FVQAGNVQL | ORF1a | L3330F | 2.0564 | 0.3571 | Gain |

**Figure 2.1. Impact of SARS-CoV-2 mutations on CD8+ T cell epitopes.** (**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**) Frequency of amino acid substitution types leading to loss of HLA binding for experimentally validated SARS-CoV-2 CD8+ T cell epitopes (from Quadeer et al. 2021). Mutations considered were those detected in more than 4 individuals (GISAID) and predicted to lead to a strong loss of HLA-epitope binding for common HLA-I alleles. Top: number of unique missense mutations for various amino acid substitution types. Bottom: $Log_2$ fold change (mutated / reference) of predicted loss of HLA-epitope binding (NetMHCpan4.1 %Rank) for the various amino acid substitution types. Each dot represents an epitope pair (mutated / reference). Color indicates HLA-I alleles affected by the mutations.

immunodominant SARS-CoV-2 epitope (254,270,289–292). 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 2.2B**) and is predicted to decrease HLA epitope binding by 47-fold (**Figure 2.4D**), 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 2.S2**) (293). 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.

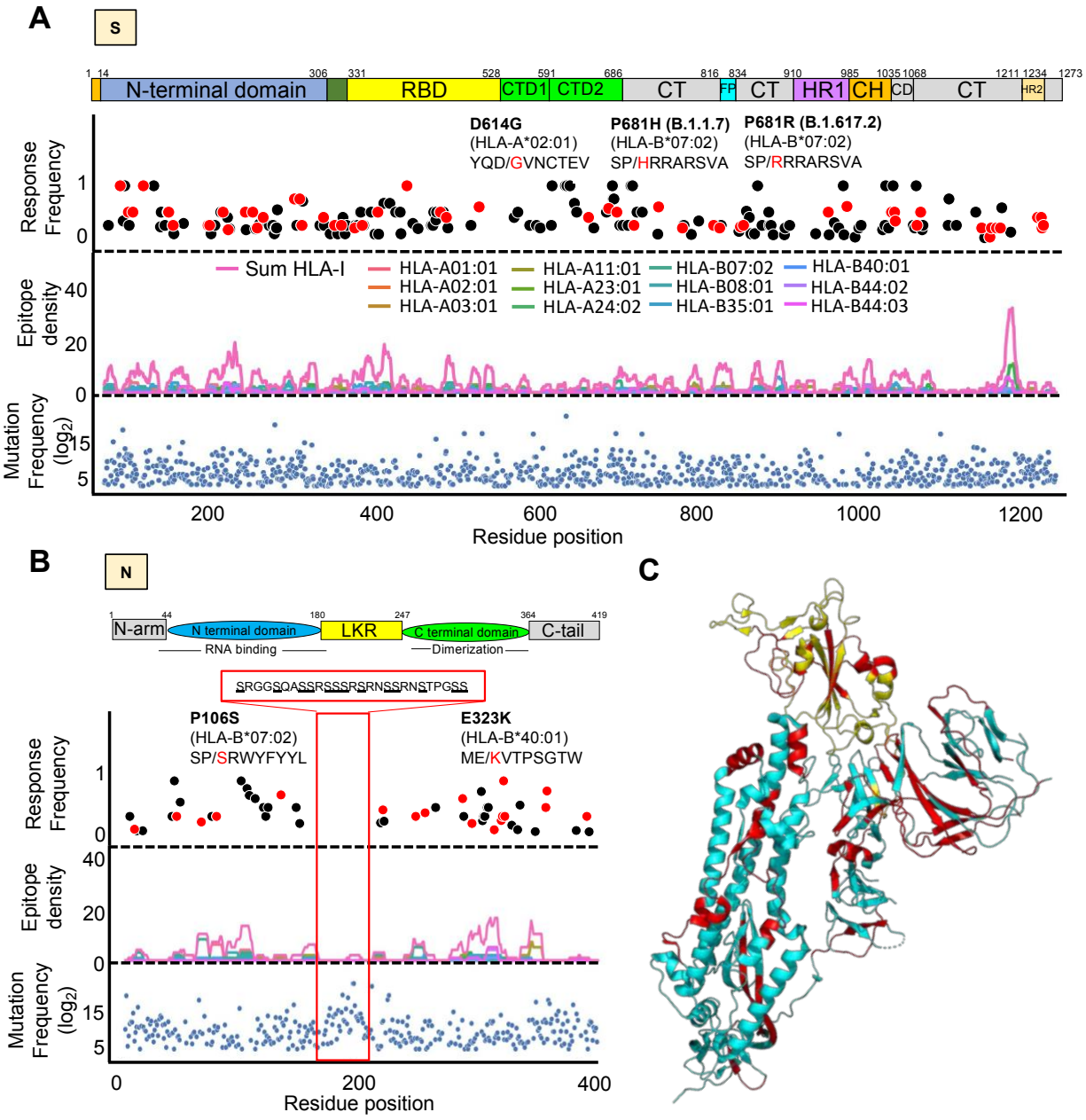



**Figure 2.2. Distribution of CD8+ T cell epitopes and their mutated variants across the immunodominant Spike (S) and Nucleocapsid (N) antigens.** (**A, B**) Lower panel: blue dots showing all mutations that occurred in at least 4 SARS-CoV-2 genomes (GISAID). Middle panel: epitope density showing the 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 Quadeer et al. 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 S 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.

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 2.1C,D** and **Figure 2.S3**). 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.

### *2.3.2 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 experimentally validated CD8+ T-cell epitopes, we observed that specific mutation types were over-represented while others were under-represented (**Figure 2.1E** and **Figure 2.S1C,D**). For instance, we found that 31% of the prevalent mutations (i.e. found in >100 genomes) predicted to abrogate the presentation of experimentally validated CD8+ T cell epitopes (Quadeer et al. 2021) led to the removal of proline residues (Pro→X) (**Figure 2.S1C**). These observations led to the hypothesis that the disproportionate presence of certain mutation types amongst mutations predicted to disrupt

peptide presentation 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 2.3**). Distinct mutational patterns at the amino acid level were observed amongst mutations detected in more than 100 genomes/individuals (**Figure 2.3),** referred to in this study as 'prevalent mutations' (see Methods and **Table 2.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 2.3**). 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 (294) (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 compositions 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 (identified in >100 genomes) listed in **Table 2.S2**.

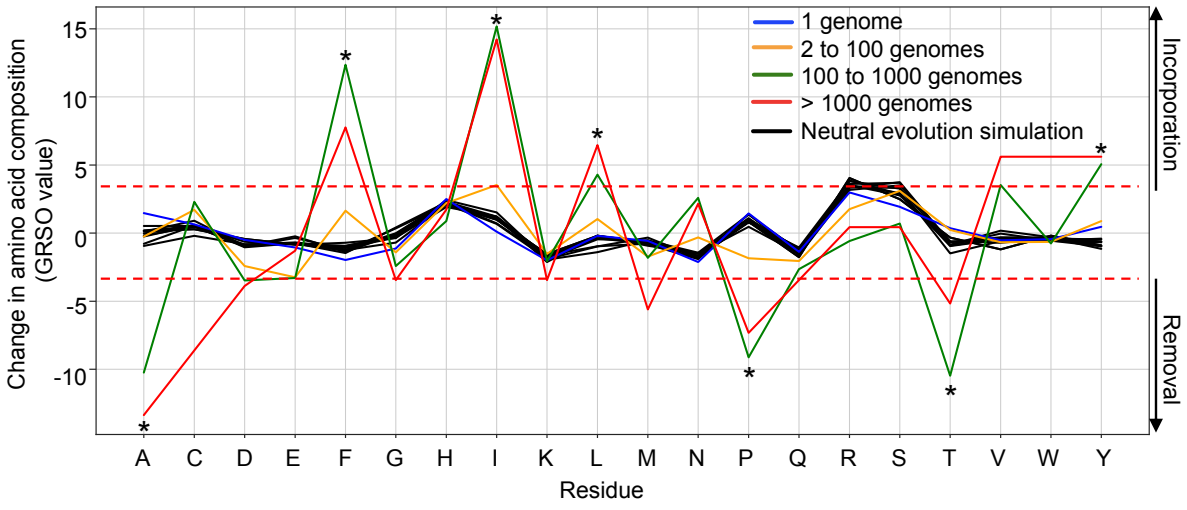



**Figure 2.3. Global amino acid mutational biases in SARS-CoV-2 proteomes.** A total of 330,246 SARS-CoV-2 genomes were translated into protein sequences and analyzed for the identification of any amino acid mutational bias. Amino acid residues (x-axis) that were removed and introduced in SARS-CoV-2 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 (yellow line), 100 to 1000 genomes (green line) and more than 1000 genomes (red line). Simulations of neutral evolution simulation (random mutations; black lines) were performed using the SANTA-SIM algorithm and serve 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).

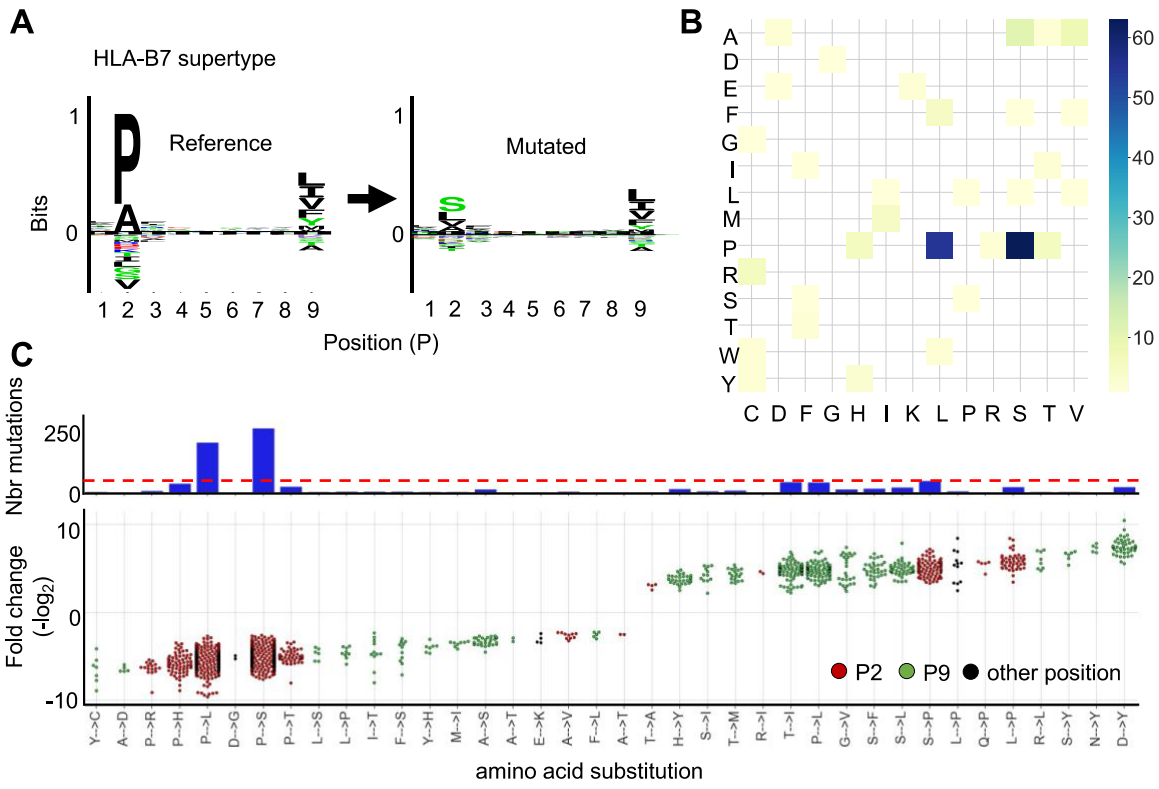
### *2.3.3 Prominent removal of proline residues leads to a predicted global loss of epitopes presented by HLA-B7 supertype molecules*

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 (295). Hundreds of different peptide binding motifs have been reported over the last decades (296). Overlapping binding motifs are qualified as "HLA supertypes" on the basis of their main anchor specificity (297,298). 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 (297). 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 2.3**) 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 (299) 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 2.S4B**). (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 à 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 PàL ($p = 8.6x10^{-35}$) and PàS ($p = 3.4x10^{-24}$) substitutions at anchor residue position P2 (**Figure 2.4A,B**). Reference/mutated peptide pairs from these categories were the most abundant, with > 250 mutated peptides per category (**Figure 2.4C**). 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 2.4D**).

In addition to the dominant PàS/L substitution type, other PàX substitutions were observed, including in variants of concern. For instance, our most recent analysis (August 2021) of mutations found in the Pangolin B.1.1.7 variant (Alpha) 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-B7 types tested. Notably, the more recent B.1.617.2 (Delta) variant was also found to disrupt the same epitope SPRRARSVA via a proline-to-arginine mutation in the Spike protein (Spike:P681R) (**Figure 2.2A**). 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 and B.1.617.2 variants. This finding lets 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.




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 2.4. Mutation of proline (P) 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**) Experimentally validated CD8+ T cell epitopes (from Quadeer et al. 2021) that are affected by the lost of a P residue. Mutated epitopes encoded by Spike (S), Nucleocapside (N), Open Reading Frame (ORF) 1a, 1b and 3a are shown as representative exemples. Effect of the P→X substitutions on predicted epitope binding affinities (NetMHCpan 4.1 %Rank) is shown. Data of magnitude of T cell response for reference epitopes were obtained from Quadeer et al. 2021.

### *2.3.4 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 2.3**). Most of these amino acids act as key epitope anchor residues for multiple HLA class I supertypes (**Figure 2.S4**). 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 2.5**). 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.

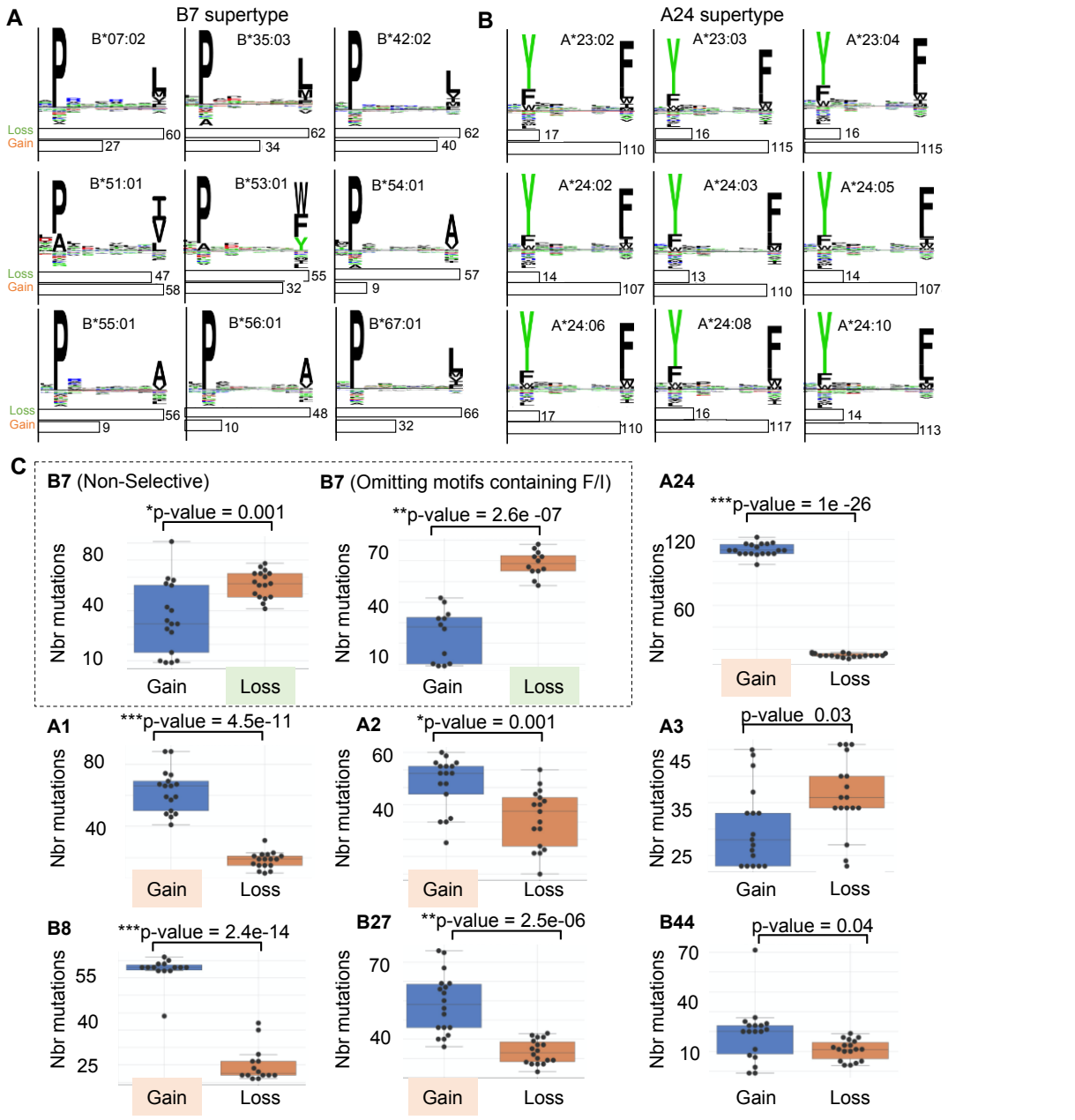



**Figure 2.5. Loss or gain of SARS-CoV-2 mutated epitopes for different HLA class I supertypes.** (**A, B**) Motif views showing established epitope binding motifs for different HLA-I alleles that belong to the HLA-B7 (A) and HLA-A24 (B) supertype family. Shaded squares highlight anchor residues that are preferentially removed (pale green) or introduced (pale orange) in SARS-CoV-2 proteomes (related to Figure 3), respectively. Histograms below the motif views 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 significant 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 S4) were used to generate the graphs and p-values (*$p \leq 0.001$, **$p < 1e\text{-}5$, ***$p < 1e\text{-}10$).

In order to compare supertypes to each other, we generated a 'Gain/Loss plot' for each supertype assessed (**Figure 2.5C**). 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. 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 ($p = 2.5x10^{-6}$). 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 3) at anchor residue P9 for certain HLA types (namely HLA-B*51:01 and HLA-B*53:01) (**Figure 2.5A**). In fact, omitting motifs containing isoleucine or phenylalanine increased the significance of epitope lost *versus* gained ($p = 2.6x10^{-7}$) (**Figure 2.5C**). Together, our results show that the amino acid mutational biases that feature the global diversity of SARS-CoV-2 proteomes can positively or negatively affect binding affinities of mutated epitopes for a wide range of HLA class I molecules in a supertype-dependent manner.

### *2.3.5 The C-to-U point mutation bias largely drives diversification of SARS-CoV-2 T cell epitopes*

Next, we sought to better understand the genetic determinants that drive the association between epitope presentation and the amino acid mutational biases found in the SARS-CoV-2 population. To this end, we analyzed the abundance of all the possible nucleotide mutation types (i.e. A-to-C, 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 common mutation type (43%), followed by G-to-U (28%), as well as A-to-G, G-to-A and U-to-C (from 9.7% to 11.6%) (**Figure 2.S5A**), in line with observations made by others (279–286).

Next, we aimed to determine the contribution of these different nucleic acid mutation types to the global mutational pattern observed at the amino acid level in Figure 3. To do so, we generated simulated population samples of 1000 SARS-CoV-2 genomes using SANTA-SIM (294), applying various extents of mutational biases corresponding to the two most common mutation types observed (i.e. C-to-U and G-to-U). The resulting simulated viral populations were then analyzed to elucidate the global amino acid mutational pattern engendered by these simulated nucleic acid point mutation biases, and whether they recapitulate the observed patterns. Indeed, our data show that the mutational pattern resulting from the simulated C-to-U bias very closely mimicked the mutational pattern observed in the real-life dataset (**Figure 2.6A**). 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 2.6A**). The G-to-U mutation bias also contributed to the introduction of isoleucine and phenylalanine (**Figure 2.S5B**). Together, these results show

that the predominant C-to-U point mutations largely contribute to shaping the global proteomic diversity of SARS-CoV-2.

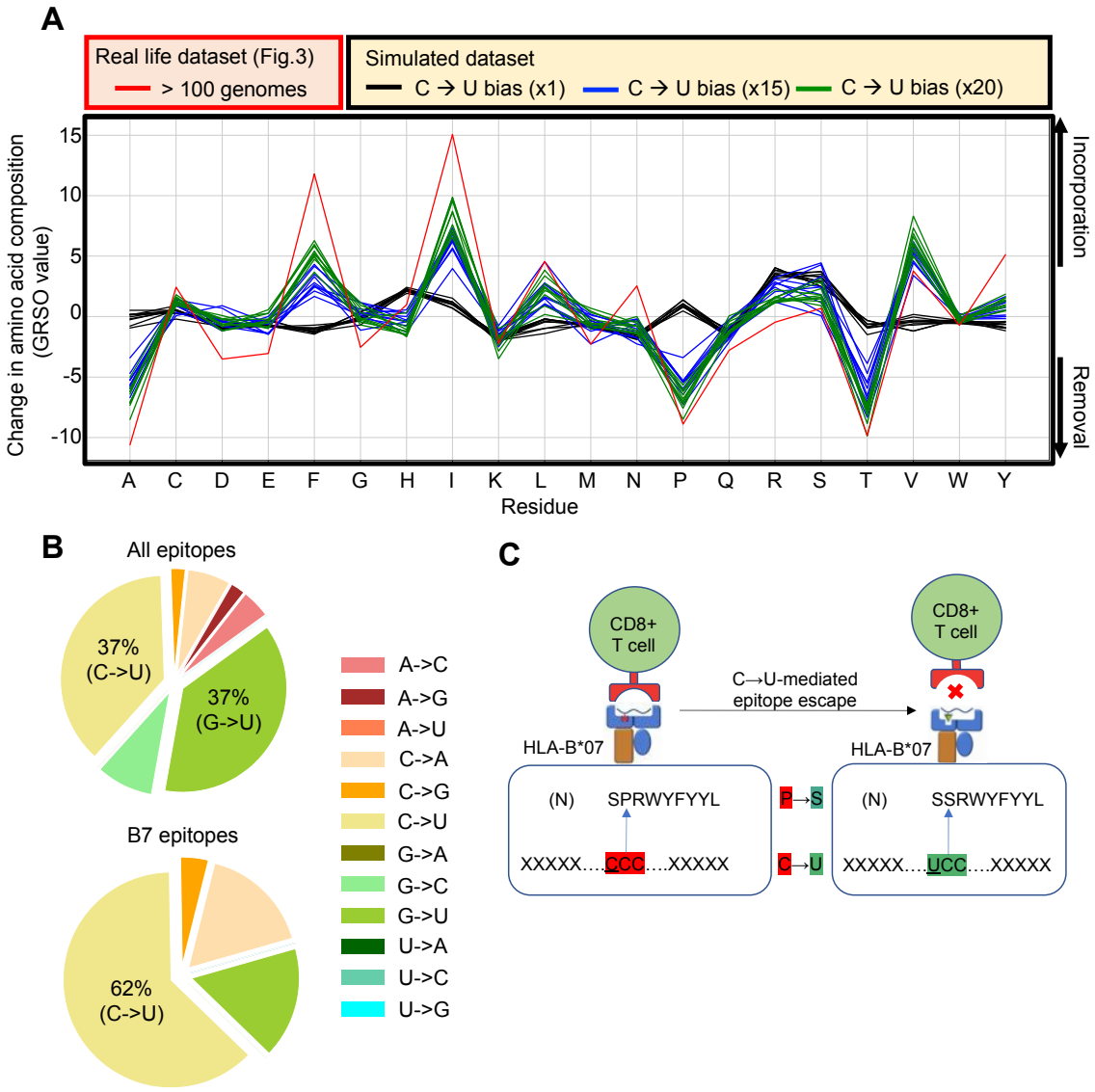



**Figure 2.6. The C-to-U point mutation bias largely drives the diversity of SARS-CoV-2 proteomes and CD8+ T cell epitopes.** (**A**) Comparison of global amino acid mutational patterns generated from real-life versus simulated SARS-CoV-2 genomes. Amino acid residues (x-axis) that were removed (y-axis; negative values) and introduced (y-axis; positive values) in real-life (red line) versus simulated (black, blue and green lines) SARS-CoV-2 are presented by %-difference in overall amino acid composition (y-axis; GRSO values), 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 3. (**B**) (Top) Pie chart showing the proportion of nucleotide substitution types from the list of validated CD8+ T cell epitopes in Quadeer et al. 2021. (Bottom) Pie chart showing the proportion of nucleotide substitution types from the list of validated CD8+ T cell epitopes that belong to the B7 supertype family in Quadeer et al. 2021. (**C**) Schematic illustrating the C-to-U-mediated epitope escape model. The observed P-to-S substitution in the immunodominant SPRWYLFYYL epitope from the Nucleocapsid (N) antigen is shown as an example.

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 PBMCs of acute and convalescent COVID-19 patients (270,271) 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 experimentally validated CD8+ T cell epitopes. 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 2.6B**). In addition, our data show that 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 2.6B**). These results strongly suggest that the dominant C-to-U point mutation bias found amongst prevalent SARS-CoV-2 mutants has the potential to 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 2.6C**), and hence, may contribute to molding T cell immunity against COVID-19 at the population level.

## 2.4 Discussion

Mutations contribute to the genetic diversity of SARS-CoV-2 and shape the progression of the COVID-19 pandemic (272,273,300). T cells are key players controlling COVID-19 disease severity. Therefore, determining whether and how the mutational landscape of SARS-CoV-2 shapes HLA-restricted T cell responses 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 (301,302) and influenza (303). In the case of a 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 downstream analysis. Our approach is universal and could be applied to other viruses in the future, given the development of distinct, prevalent mutational biases. Our global approach has led to

several conclusions to help understand how the increasing genomic diversity of SARS-CoV-2 may shape T cell immunity in human populations. Our findings have important implications that are discussed below in the context of disease severity, viral evolution and vaccine resistance.

In this study, we found that prevalent SARS-CoV-2 mutations are governed by defined mutational patterns, with C-to-U being a predominant mutation type, as previously shown by others (279–286). In fact, we show that the C-to-U mutation bias in SARS-CoV-2 genomes has a remarkably intimate relationship with the observed amino acid mutational biases, indicating that C-to-U mutations largely contribute to the global proteomic diversity of SARS-CoV-2. Moreover, we show that this mutational bias leads to the preferential substitution of proline residues with leucine or serine residues in the P2 anchor position of SARS-CoV-2 CD8+ T cell epitopes, and hence, drastically compromise epitope binding to B7 supertype molecules. These molecules, which represent ~35% of the human population, preferentially bind epitopes with proline at P2 (304)(Francisco et al., 2015). Therefore, the C-to-U mutational bias observed amongst prevalent mutants may partially disrupt SARS-CoV-2 T cell immunity in a very significant proportion of the human population. Noteworthy, this impact of C-to-U mutations on B7-dependent epitope escape was somehow predictable. In fact, proline residues originate from codons that are highly rich in C whereas serine and leucine residues originate from codons that are rich in U. One could therefore predict, at least to some extent, that a strong C-to-U bias would lead to proline-to-leucine 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. The impact of the loss of several B7 epitopes on the immune response of an individual, however, remains unclear.

In this study, we observed that proline→X mutations were more enriched amongst prevalent mutations (>100 genomes) predicted to abrogate the presentation of experimentally validated CD8+ T cell epitopes than across the global mutation landscape of SARS-CoV-2 proteomes (31% and 9.1%, respectively). These two percentages are in fact indicative of different phenomena. The former reflects the susceptibility of certain HLA alleles to specific mutational patterns (the removal of proline in this case), whereas the latter reflects the overall mutational biases observed across SARS-CoV-2 proteomes. This noticeable difference may suggest that certain mutation types play a particularly important role in HLA type-dependant cytotoxic T lymphocyte (CTL) escape. This concept becomes evident when considering the 13 common alleles investigated in this study. The detrimental impact of proline→X mutations on the presentation of peptides by B7 alleles is reflected in the higher proportion of proline→X mutations (31%) leading to the loss of epitopes. 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 drive 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

(273,279,280,305,306). In this regard, APOBEC activity has been shown to broadly drive viral evolution and diversity, including in human immunodeficiency virus (HIV) (307–313). In fact, APOBEC-induced mutations driving the evolution and diversification of HIV-1 were shown to have an intimate relationship with T cell immunity (312,314). 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 (287,288,312,314–316). 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 that is consistently 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-dependent 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 higher in B7 individuals compared to A24 individuals. In fact, mutated epitopes are predicted to be unfavorably and favorably presented by B7 and A24 supertypes, respectively (Figure 5). The supertype dependency is important here because it suggests that T cell responses are shaped differently across different human populations in response to infection by mutated forms of SARS-CoV-2. For instance, the predicted model lets us hypothesize that, within

the first year of the pandemic (from December 2019 to December 2020), human populations expressing the A24 supertype at higher frequency (e.g. >90% of people in specific geographical regions in Taiwan) may likely mount a T cell response upon infection by mutated forms of SARS-CoV-2 that will not be as readily disrupted by mutation events, in comparison to individuals expressing the B7 supertype (i.e. ~35% of the human population) (304). Interestingly, a recent computational study corroborated the propensity of HLA B*07:02 to lose epitopes due to SARS-CoV-2 variants (317). Our proposed model may therefore act as a contributing factor addressing the global diversity of immunological responses against SARS-CoV-2 variants as the pandemic progresses. Several studies have indeed interrogated associations between HLA alleles and COVID-19 disease severity (318–320) as well as mutations and T-cell evasion (293,321,322). However, to the best of our knowledge, this is the first study that proposes a connection between mutation biases, differential presentation of epitope variants (HLA supertype dependent), and variability in host responses to SARS-CoV-2 infection, all in the context of the continuously expanding genomic diversity of SARS-CoV-2 mutants. Additionally, the current study establishes a basis for investigating CTL-escape in the context of HLA (super)types strategically selected based on the diversification patterns of SARS-CoV-2.

With regard to the variants of concern, we noted that the B.1.1.7 (Alpha) variant was predicted to lose the B7 supertype-associated, experimentally validated epitope SP/HRRARSVA as a result of a proline-to-histidine substitution. The B.1.617.2 (Delta) variant was in fact also predicted to lead to the loss of the same epitope via a proline-to-arginine substitution (SP/RRRARSVA). As the B.1.617.2 variant has become the most widespread SARS-COV-2 lineage globally since July 2021, it would be of interest to experimentally interrogate the impact of this variant in the

activation of CTLs in $B7^+$ individuals. Although our study does not demonstrate that the disproportionate loss of proline across the SARS-CoV-2 mutation landscape is the cause for the increased infectivity of the discussed variants of concern, we propose that it may be a contributing factor in the context of certain populations. In this regard, while genomic surveillance is ongoing in different regions of the world, measuring the level of transmission of the B.1.1.7 and B.1.617.2 variants 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 (323), 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 (254,270,289–292). 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.

## 2.5 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 31$^{st}$ 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 dataset, which includes 1.7M SARS-CoV-2 genomes as of May 2021, may lead to additional insights in this regard.

## 2.6 Acknowledgements

We acknowledge and thank GISAID as well as all contributing laboratories for giving access to their SARS-CoV-2 genome sequences. We also 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.

## 2.7 Materials and Methods

### STAR METHODS

### RESOURCE AVAILABILITY

#### Materials Availability

This study did not generate new materials.

#### Data and Code Availability

- Source data statement. This paper analyzes existing, publicly available data. All sequence data used 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. A GISAID acknowledgment table containing a full list of the laboratories and authors who contributed to the extensive GISAID SARS-CoV-2 genome database queried in this study is available in supplementary materials.
- Code statement. All original code has been deposited at https://github.com/CaronLab/CoVescape and is publicly available as of the date of publication. DOIs are listed in the Key Resources Table.
- Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

### 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 4.1 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 4.1 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.

**In vitro HLA-peptide binding assays**

Peptide binding to class I HLA molecules was quantitatively measured using classical competition assays based on the inhibition of binding of a high affinity radiolabeled peptide to purified HLA molecules, as detailed elsewhere (324). Briefly, HLA molecules were purified from lysates of EBV transformed homozygous cell lines by affinity chromatography by repeated passage over Protein A Sepharose beads conjugated with the W6/32 (anti-HLA-A, -B, -C) antibody, following

separation from HLA-B and -C molecules by pre-passage over a B1.23.2 (antiHLA B, C) column. Protein purity, concentration, and the effectiveness of depletion steps was monitored by SDS-PAGE and BCA assay. Peptide affinity for respective class I molecules was determined by incubating 0.1-1 nM of radiolabeled peptide at room temperature with 1 μM to 1 nM of purified HLA in the presence of a cocktail of protease inhibitors and 1 μM B2microglobulin. Following a two-day incubation, HLA bound radioactivity was determined by capturing MHC/peptide complexes on W6/32 antibody coated Lumitrac 600 plates (Greiner Bioone, Frickenhausen, Germany). Bound cpm was measured using the TopCount (Packard Instrument Co., Meriden, CT) microscintillation counter. The concentration of peptide yielding 50% inhibition of the binding of the radiolabeled peptide was calculated. Under the conditions utilized, where [label]<[MHC] and IC50 ≥ [MHC], the measured IC50 values are reasonable approximations of the true Kd values. Each competitor peptide was tested at six different concentrations covering a 100,000-fold dose range, and in three or more independent experiments. As a positive control for inhibition, the unlabeled version of the radiolabeled probe was also tested in each experiment.

**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. (325) 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* à 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 S5A). 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* à *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) were compared to a neutral evolution as a negative control (N = 10 replicates) by way of twotailed independent T-Test.

**Statistical analysis**

A Two-tailed One-Sample T-Test was used to assess the statistical significance of the observed mutational biases against the neutral simulations (N = 10 replicates). A Two-tailed Independent T-Test assuming different variances was used to assess the statistical significances of 1) the simulated biased SARS-CoV-2 evolution, 2) the gain/loss plots in the context of supertypes, and

3) the statistical significance associated with the average fold change in %rank associated with each position-specific amino acid mutation type in the supertype analysis.

## 2.8 Author Contributions

The study presented in this chapter resulted from collaborative efforts involving multiple laboratories internationally. Nevertheless, I was the primary contributor to the body of this work. Firstly, with the advent of the COVID-19 pandemic, I conceived the project with the help of Dr. Etienne Caron and Dr. Julie Hussin. I then proceeded to generate all codes aimed at analyzing SARS-CoV-2 genomes acquired from GISAID and identify mutations; querying the global mutational landscape to determine mutation rates, prevalent mutations, and to characterize prevalent mutational biases; utilizing epitope presentation prediction software (netMHCpan 4.1) to identify all predicted mutation-containing epitopes; determining the impact of SARS-CoV-2 mutations on the presentation of predicted CD8+ epitopes as well as experimentally validated CD8+ epitopes; and at generating all figures presented in this chapter. I was also responsible for writing and submitting the manuscript in collaboration with Dr. Etienne Caron, and for addressing reviews throughout the peer-review process. Jean-Christophe Grenier, Fatima Mostefai, as well as Peter Kubiniok provided me with support in the design of bioinformatic pipelines. Dominique Fournelle was responsible for optimizing and conducting the evolution simulations of SARS-CoV-2 using SANTA-SIM. Finally, all co-authors provided helpful comments throughout the generation of analyses as well as in the reviewing of the manuscript.

## 2.9 Supplementary Figures

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 |

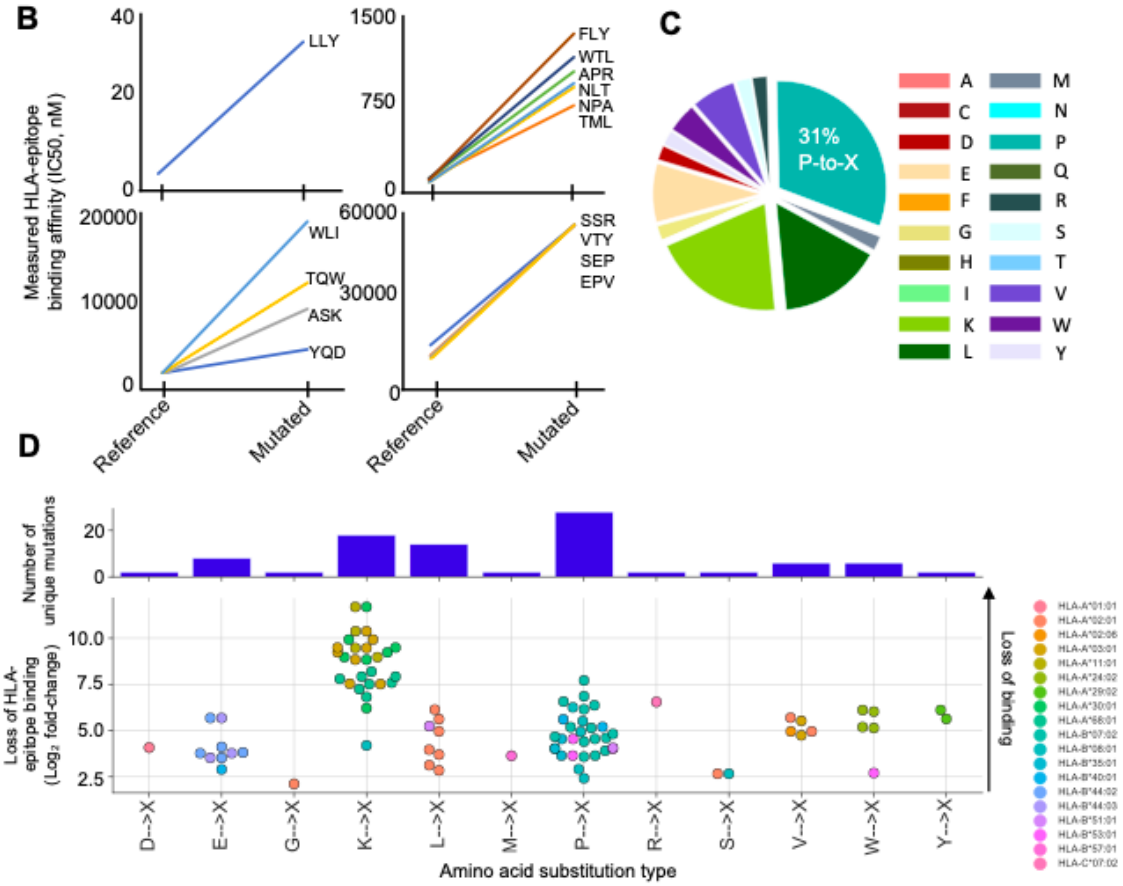



**Figure 2.S1. HLA peptide binding measurements and mutational biases in SARS-CoV-2 mutated epitopes, Related to Figure 2.1 and Figure 2.2.** (A) HLA binding assay was performed to determine the in vitro binding affinity (nM) of representative SARS-CoV-2 peptides for specific HLA class I alleles. Peptides 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 predicted to abrogate the presentation of experimentally validated CD8+ T cell epitopes (Quadeer et al. 2021).

Proline (P)-to-X is the most dominant substitution type. (D) Predicted loss of HLA-epitope binding clustered by substitution type from the list of experimentally validated CD8+ T cell epitopes in Quadeer et al. 2021. 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 | A*02:01 | M-L90F | 38 | 0.5 | 11.9 |
| ME/KVTPSGTWL | B*40:01 | N-E323K | 23 | 0.3 | 10.1 |

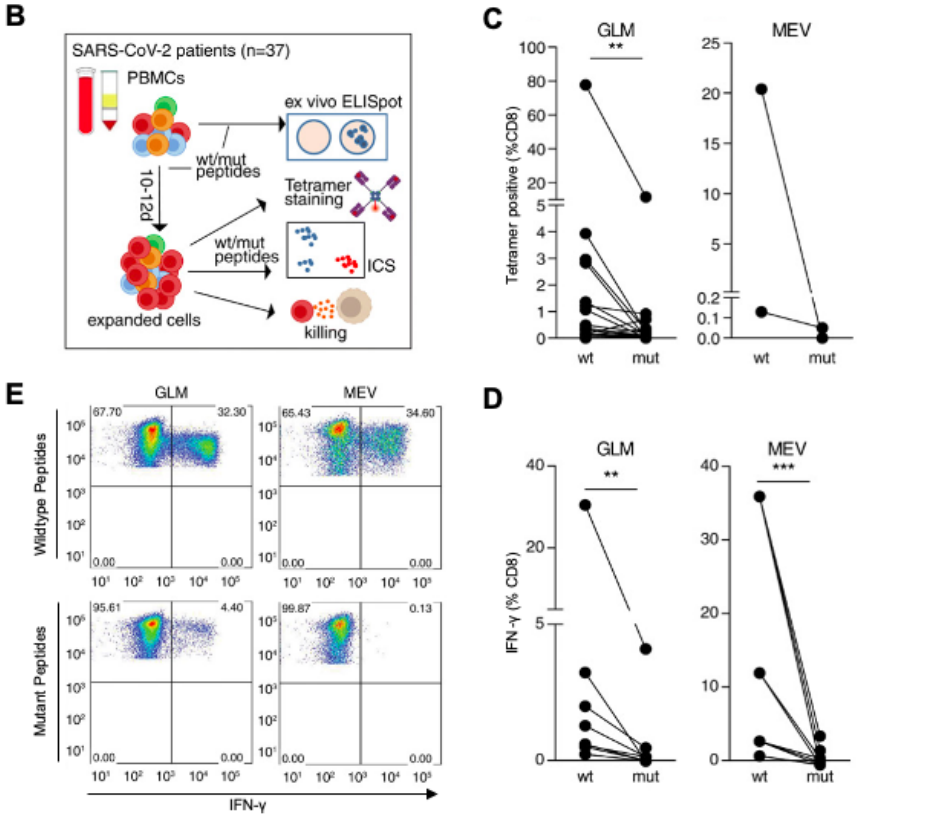



**Figure 2.S2. Identification of two SARS-CoV-2 mutated epitopes that were previously associated with decreased CD8+ T cell responses, Related to Figure 2.2.** (A) The mutated epitopes GFMWLSYFI (A*02) and MKVTPSGTWL (B*40) were detected in 38 and 23 genomes/individuals in this study (GISAID) and their T cell immunogenicity was thoroughly investigated in Agerer et al. 2021 (copywright 2021, with permission from AAAS). The epitopes derive from the Membrane (M) and the Nucleocapsid (N) antigen. The NetMHCpan %rank is indicated for the reference and mutated form of the epitopes. (B-E) T cell recognition of the reference and mutated epitopes (figure panels published in 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 (wt) 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 wt or mutant peptides and stained with tetramers containing the wt 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 wt or mutant peptide. (E) Representative FACS plots for (D).

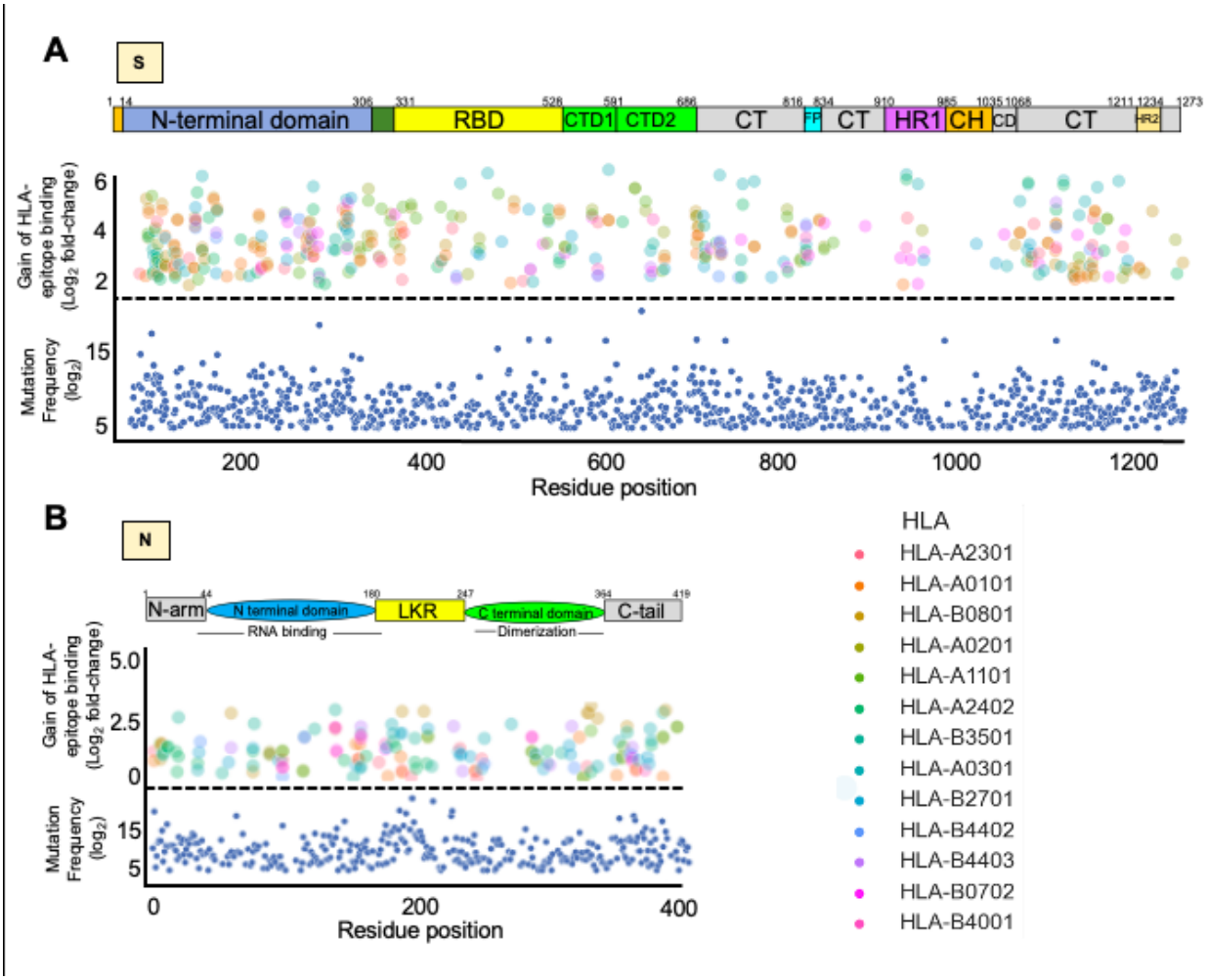



**Figure 2.S3. Impact of mutations on gain of peptide binding to various HLA class I molecules across the immunodominant Spike (S) and Nucleocapsid (N) antigens, Related to Figure 2.1.** (A, B) Lower panel: blue dots showing all mutations that occurred in at least 4 SARS-CoV-2 genomes (GISAID). Upper panel: dots showing predicted peptides subjected to a strong gain of binding (see also Figure 1C,D) to one of 12 highly common HLA-I alleles queried (color coded) due to a mutation.

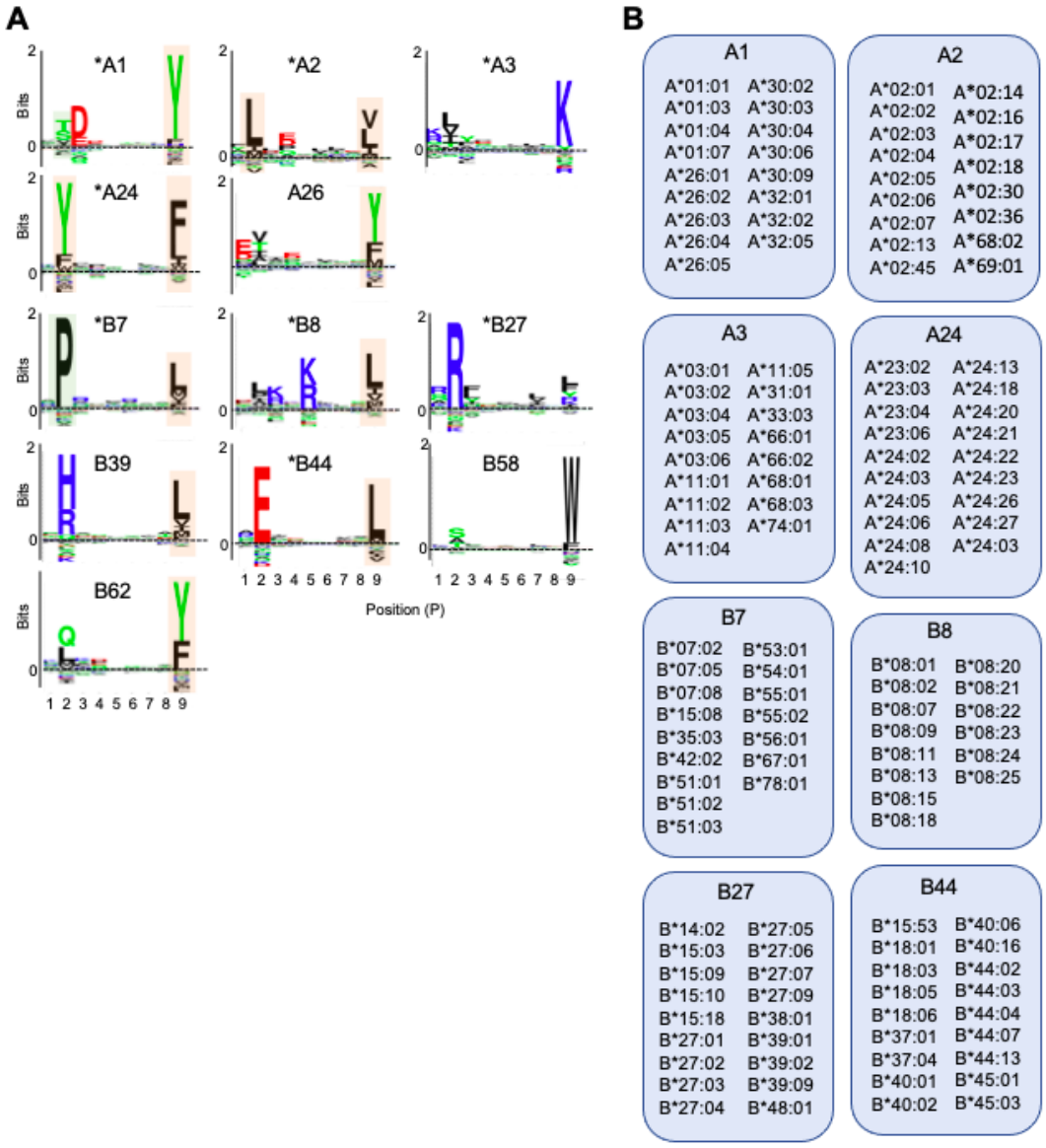



**Figure 2.S4. Analysis of HLA class I supertypes, Related to Figure 2.5.** (A) Epitope binding motifs for several HLA class I supertypes. Anchor residues are located at P2 and P9. Pale orange and green squares cover amino acid residues that are preferentially introduced (F, I, L, Y) and removed (A, 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' in Figure 5.

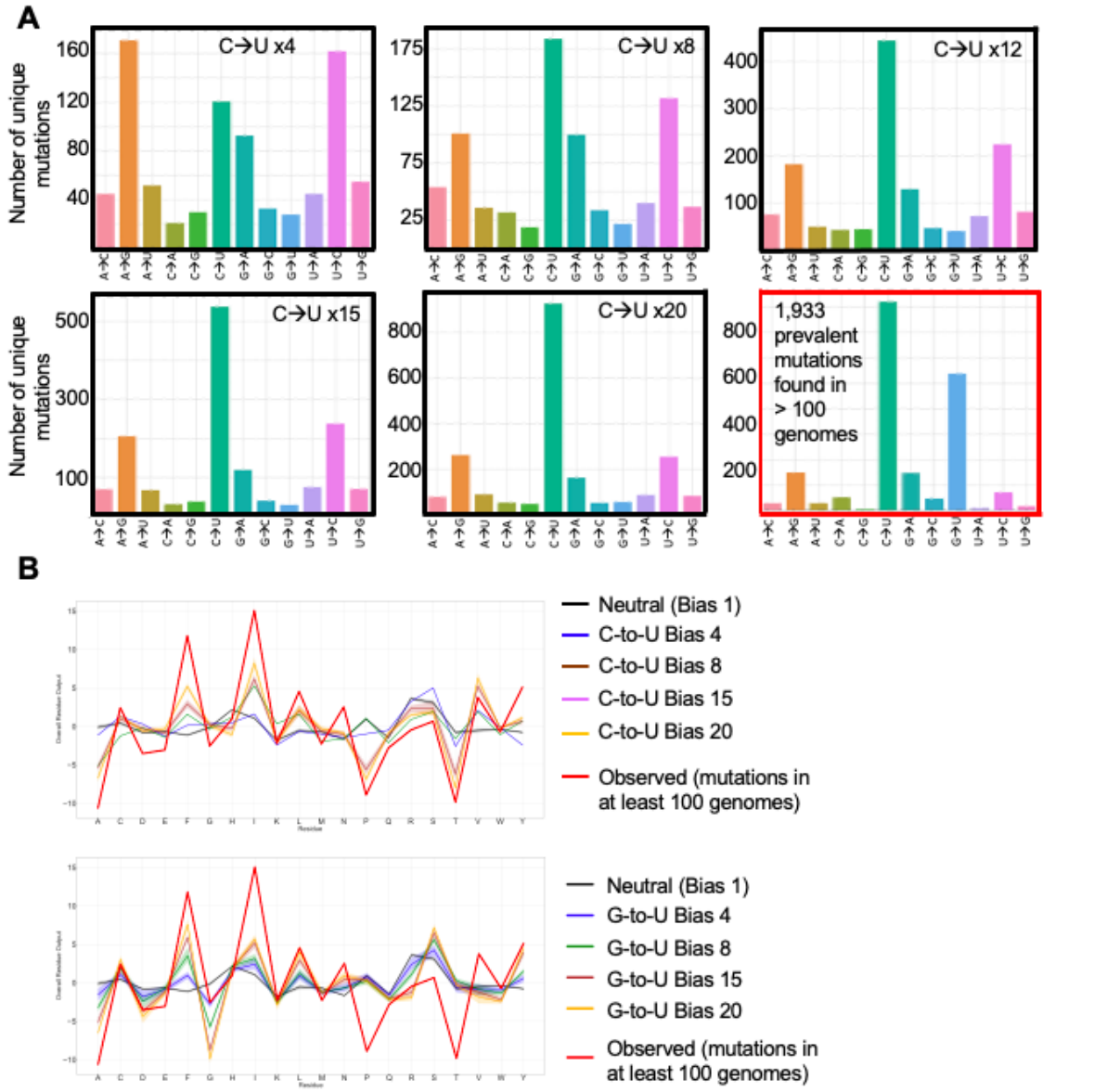



**Figure 2.S5. Comparison of mutation biases between real-life/observed and simulated data, Related to Figure 2.3 and Figure 6.** (A) Histograms showing the number of unique mutations identified for each mutation type (A-to-C, A-to-G, etc.) after simulating the evolution of SARS-CoV-2 genomes through 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.

# 3 CHAPTER III: DISCUSSION

## 3.1 Relevance of assessing the impact of SARS-CoV-2 mutations on specific populations (HLA-dependant)

Using a combination of genomic approaches, computational predictions, MHC binding assays, as well as databases of experimentally validated SARS-CoV-2 T cell epitopes, my work investigates the ability of SARS-CoV-2 variants to evade cellular immune recognition in an HLA-dependant manner. Importantly, we lay the groundwork to encourage the scientific community to pay close attention to HLA-types when experimentally investigating the T-cell evasion potential of SARS-CoV-2 variants. Indeed, our analyses demonstrated that the strong C→U mutational pattern observed in the first year of the pandemic, resulting in the removal of Pro, Thr and Ala from viral proteomes, was immunologically relevant to HLA alleles belonging to the proline-dependant B7 supertype.

The relatively recent emergence, identification, and characterization of several VOCs has encouraged the research community to investigate their immune-evasion potential. Multiple research initiatives taking place over the last months have interrogated the ability of VOCs Alpha, Beta, Gamma and Delta to evade cellular immunity (137,138). These investigations demonstrated their limited impact on cellular immunity. Nevertheless, the potential capacity of SARS-CoV-2 to evade cellular immunity cannot be ruled out. An important aspect of cellular immunity lies in the high amount of polymorphism in HLA alleles across human populations. An individual may possess a combination of 6 HLA class I alleles, allowing every individual to develop a propensity for a unique pool of SARS-CoV-2 class I epitopes. Given the diversity of binding motifs described across HLA class I types and supertypes, the ability of SARS-CoV-2 mutations to promote T-cell evasion may differ drastically from one individual to the next. This diverse range of impacts of mutations on T cell escape is contrasted by neutralizing antibody escape, which has been shown

to be much more universal across human populations, and a putative driver of viral evolution. Cellular immunity was in fact hypothesized to be an unlikely driver of the evolution of an acute virus such as SARS-COV-2 (140,326–328). Given the high diversity of HLA type combinations found throughout the human population, it is thought that adaptations to an individual with a particular assortment of HLA alleles would be futile in a subsequent infection faced with entirely different HLA alleles. This contrasts with chronic viruses such as HIV, which experiences significant intra-host evolution and may therefore adapt to a single individual's immune system. The unlikely relationship between SARS-CoV-2 adaptations and T cell evolution may explain the negligible impact of VOCs on T cell activation. Nevertheless, regardless of the evolutionary pressures driving the set of adaptations specific to each lineage, we cannot eliminate the putative impact of VOCs on subpopulations based on their HLA alleles.

Another important aspect remains that despite their epidemiological relevance, the set of mutations making up these VOCs constitutes a relatively small proportion of the total global mutational landscape. The identification of VOCs is largely based on the ability of mutations to modulate the transmissibility of the virus rather than its disease severity, as lineages are generally flagged as VOIs/VOCs with regards to their international prevalence. Studies aimed at characterizing the pathogenicity and virulence of VOCs have thus far been reactive to their initial labelling. However, we would like to entertain the existence of viral lineages conveying low to moderate epidemiologic success (lower transmissibility) while carrying out substantial impacts on disease severity. We hypothesize that although clinically relevant, such variants would likely fall under the radar of public health. As robust T cell responses were strongly associated with lesser disease severity in multiple studies, one may ask whether lineages composed of mutations capable of evading T cell escape could modulate disease severity in an HLA-dependant manner while

failing to acquire widespread epidemiologic prevalence. This hypothesis suggests the incorporation of additional considerations when interrogating the clinical relevance of lineages. Whilst lineage prevalence plays an important role in the identification of VOCs, we propose to add considerable weight to the ability of lineages to modulate disease severity, and to do so by considering the impact of SARS-CoV-2 non-synonymous mutations on T cell escape amongst HLA-defined subpopulations. The importance of even single mutation events on the clinical outcome of viral infections in the context of HLA and T cell escape was well demonstrated in HIV/AIDS. The acquisition of a mutation (Y135F) within an immunodominant epitope of the HIV *Nef* protein was found to result in significantly lower viral suppression and higher viral loads in HLA-B35:01$^{+}$ individuals (231). In the context of SARS-CoV-2, several mutations have been individually assessed for their ability to evade T cells in the context of certain HLA types. Notably, the L452R mutation within the Spike Glycoprotein RBD was found to enable T cell evasion in the context of HLA-A*24 (65). Several other T cell-evading mutations were identified by Agerer *et al* (139). However, in studies assessing the ability of VOCs to evade cellular immunity, the HLA types were not taken into consideration with the same weight. Therefore, it is yet unknown if VOCs, VOIs or lineages of moderate epidemiological success modulate disease severity in human subpopulations in an HLA-dependant manner.

Given these various arguments, close monitoring of the impact of emerging variants on HLA-dependant immune evasion will be critical in the months to come.

### 3.2 The future of T-cell evasion for SARS-CoV-2

As thoroughly shown in the literature, the cellular arm of the adaptive immune system is instrumental in clearing SARS-CoV-2 infection. The quality of both the CD8+ and CD4+

components of the cellular immune response were also shown to be highly correlated with disease severity (122). As such, the careful and comprehensive monitoring of the relationship between viral mutations and T cells will continue to be highly relevant to public health. Importantly, as the virus continues to accumulate neutral as well as positively selected genomic variations, the number of potentially clinically impactful mutations comprising individual lineages is expected to increase. This is showcased by the most recent VOC, Omicron (B.1.529), which carries as many as 50 mutations with 30 in the spike protein (329), considerably more than VOCs Alpha, Beta, Gamma and Delta. As lineages become increasingly divergent from the reference virus, it may become more likely that a single lineage comprising dozens of mutations could affect numerous T cell epitopes, therefore posing a greater clinical risk.

Moreover, as the pandemic presses on and the public health response progresses, it may be of interest to monitor potential evolutionary pressures that public health initiatives may impose on the virus. For example, interrogating the dynamics of the mutational landscape prior to, and following the introduction of large-scale vaccination worldwide may shed light on putative evolutionary pressures imposed onto the virus by prophylactic strategies. Commonly used vaccine treatments currently in circulation target the Spike glycoprotein or one of its domains, the ACE2-targeting Receptor-Binding Domain (RBD). The large-scale deployment of such strategies may confer adaptations within the targeted regions of the viral proteome. Given the essential role played by adequate cellular immune responses against SARS-CoV-2 in mediating symptom severity and resolving infections, it will continue to be of great importance to closely monitor existing and emerging SARS-CoV-2 lineages for mutations associated with T cell evasion (122,239,327).

### 3.3 The impact of T-cell escape on memory T-cells, and on the long-term success of vaccines

The success of any adaptive immune response is determined not only by its capacity to amount a strong defense against active infections, but also to generate long-lasting infection-specific lymphocytes. These lymphocytes, known as memory B cells and T cells, continue circulating the host's lymph nodes following convalescence to re-ignite the adaptive immune system upon a subsequent contact with the original infectious agent. For example, antibody titers are well known to wane over the months following infection. However, memory B cells, which are responsible for producing antibodies, will replenish antibody titers upon re-infection. Memory CD4+ and CD8+ T cells have also been found to play key roles in the long-lasting effectiveness of adaptive immune defenses. Importantly, memory lymphocytes are critical to the success of prophylactic strategies. The generation of memory lymphocytes following vaccine administration will ensure a long-lasting protection against the virus. The current set of mRNA vaccines used to immunize against SARS-CoV-2 were shown to not only activate all arms of the adaptive immune system, but also to induce the production of both memory CD4+ and CD8+ T cells as well as memory B cells (330–335). Whereas antibody titers have originally been considered as the main determinant of vaccine efficacy, recent studies demonstrated that strong activation CD8+ and CD4+ memory T cells were contributors to prophylactic protection (333–335).

Upon the emergence and identification of VOCs, a major concern has been the ability of mutations found in the current set of VOCs to reduce the efficacy of SARS-CoV-2 vaccines. These concerns partially originated from the demonstrated ability of certain mutations to evade recognitions by antibodies, allowing such mutations to reduce the protection conferred by vaccine-generated antibodies. Indeed, multiple studies indicated the ability of VOCs to modulate the efficacy of several vaccines and as well as their ability to generate neutralizing antibodies. Namely,

individuals vaccinated with the AstraZeneca ChAdOx1 and Novavax COVID-19 were shown to be differentially impacted by the ancestral, B.1.1.7 and B.1.351 lineages, with the variants leading to reductions in production of neutralizing antibodies (336–338). An alternative strategy by which mutations might negatively impact the prophylactic efficacy of vaccines consists of reducing antigen-recognition by memory T cells. Such mutations would result in the reduced ability of memory T cells induced by the vaccine (ancestral SARS-CoV-2 lineage) to recognize mutated epitopes. To this effect, several studies were conducted to assess the protection provided by vaccine-induced T cells against VOCs (326,330). Close monitoring of the impact of variants on the quantity and quality of vaccine-induced memory T cells will continue to be of great importance. As previously stated, T cells were found to play key roles in not only disease severity, but also in reducing SARS-CoV-2 viral loads (122,127,339). As lineages continue to accumulate mutations, the ongoing surveillance of emerging lineages may eventually lead to the identification of VOCs with the ability to disrupt the efficacy of ancestral lineage-induced memory T cells. The Omicron VOC constitutes one such example. Although its impact on the quantity and quality of memory B and T cells has not yet been queried, the sheer number of mutations (50 within the entire proteome, 30 within the Spike glycoprotein) is cause for concern.

### 3.4 Future work

As SARS-CoV-2 continues to adapt to the human immune system, the monitoring of SARS-CoV-2 lineages will continue to play an important role in our ability to mediate and mitigate the impact of the virus on human health. Although no VOC has yet been associated with the clinically significant disruption of cellular immunity, the evolution of SARS-CoV-2 is an on-going process resulting in the continued emergence of new lineages, and possibly new VOCs. Additionally, the

current identification of VOCs is highly dependent on epidemiologic features such as viral transmissibility as opposed to virulence factors. As such, it is possible that the current VOC-identification scheme fails to detect lineages associated with worse disease outcomes albeit with milder epidemiologic success. Future work will involve the continued surveillance of SARS-CoV-2 evolution while assessing the impact of circulating as well as emerging lineages on cellular immunity. In this scheme, although attention will be given to the prevalence of lineages, significant weight will be allocated to their ability to disrupt cellular immunity. This approach may result in VOC definitions diverging from accepted guidelines. Additionally, the current COVID-19 pandemic presents an ideal opportunity to investigate the rapid, global fixation of evolutionary patterns in a real-time fashion. Public health interventions are often known to drive the diversification of viruses responsible for acute infections. As such, it will be of interest to investigate the dynamics of viral evolution prior to and following mass vaccination. The sheer number of sequences found within the EpiFlu$^{TM}$ database of GISAID, now containing over 9 million sequences, is ideal to conduct such a large-scale analysis. In this section, we will briefly discuss the various features of viral evolution that will be investigated in our future work.

### *3.4.1 Evolving mutational biases*

The manuscript presented in this dissertation evidenced the emergence of well-defined mutation biases across the mutational landscape of SARS-CoV-2 throughout the first year of the pandemic. These biases were primarily dominated by C→U and G→U. These mutational biases not only dictated the genomic diversification of SARS-CoV-2, but also its proteomic diversification. As demonstrated in this study, these dominant mutation types were found to diversify the amino acid composition of CD8+ T cell epitopes, therefore shaping the repertoire of class I epitopes. To this effect, host-mediated RNA-editing enzymes such as APOBECs were suggested as a mechanism

of action responsible for the observed biases. The cytidine deamination activity certain members of the APOBEC family have been associated with the promotion of C→U mutations. Recently, Kim K *et al* (2021) experimentally demonstrated the propensity of two members of the APOBEC family, APOBEC3A and APOBEC1 to enable the APOBEC-mediated editing of SARS_CoV-2 RNA, resulting in a high C-to-U mutation frequency (340). Moreover, recent analyses performed by our group yielded early signs for the dilution of such mutation biases. As the pandemic progresses, it will be of interest to continue monitoring the dynamics of global mutation landscape. Such information may provide unique insights into the evolutionary behaviour of an acute zoonotic infectious agent when faced with the rapid globalization observed over the course of the current pandemic.

### *3.4.2 Tracking the long-term evolution of SARS-CoV-2*

#### 3.4.2.1 Relationship between emerging lineages and T cell escape

With the advent of mass vaccination, the impact of most viral lineages on the efficacy of approved vaccines is unknown. Although current efforts are highly directed at a small number of VOCs, there are actively hundreds of variants with unknown impacts on cellular immunity. In addition, on-going evolution of the virus is expected to lead to the emergence of new variants. The importance of active tracking of SARS-CoV-2 variants was evidenced by the recent emergence of the Omicron VOC. As the disruption of T cell-based immunity could affect the long-term immune protection observed in vaccinated and convalescent individuals, it will be of great importance to develop a methodology to identify lineages capable of compromising T cell-based protection. To this end, we propose to develop a scoring system allowing for the prioritization of such SARS-CoV-2 lineages by considering features specific to each circulating SARS-CoV-2 lineage. We will

then experimentally validate the impact of our prioritized variants on T cell recognition following vaccination by the Pfizer vaccine (targeting the Spike protein).

*Aim i) Prioritization of viral lineages***.** The ability of SARS-CoV-2 lineages to lead to T cell escape will be predicted by combining the following lineage-specific features into a scoring system: The number of missense mutations occurring within the spike protein (targeted by the Pfizer vaccine); the prevalence of lineages within the global population; the predicted ability of lineage-specific mutations to disrupt the presentation of validated epitopes by common HLA alleles, and to disrupt the predicted immunogenicity of validated epitopes. To these ends, an in-house pipeline has been developed to identify lineage-specific missense mutations, determine the protein of origin, and incorporate a variety of out-sourced bioinformatic tools used to predict the presentation by common HLA class I and II, and the immunogenicity of mutated and wild-type epitopes. These tools include netMHCpan-4.1, netMHCIIpan 4.0 and MHCflurry 2.0 for predictions of epitope presentation, and PRIME and MARIA for predictions of immunogenicity. The top five SARS-CoV-2 lineages will be selected for further analyses.

*Aim ii) Impact of SARS-CoV-2 variants on T cell recognition post-vaccination***.** Spike protein peptide pools will be generated for the top SARS-CoV-2 lineages selected in 'aim i' to assess their overall impact on vaccine effectiveness. Briefly, $OX40^{+}CD137^{+}CD4^{+}$ and $CD69^{+}CD137^{+}CD8^{+}$ T cells will be isolated from PMBCs of individuals vaccinated with the Pfizer vaccine (in collaboration with the RECOVER-2 project) using FACS. Isolated T cells will then be stimulated with mutated or wild type peptide pools and assessed by AIM assays. Alternatively, we will assess the impact of variants on T cell responses by Intracellular Staining assays (ICS).

*Aim iii) impact of SARS-CoV-2 variants on T cell expansion post-vaccination***.** To assess the impact of the selected SARS-CoV-2 lineages on vaccine effectiveness with a greater level of

granularity, we will employ Single-Cell V(D)J TCR sequencing using the RAGE-seq approach co-developed by Dr. Smith (202). This will allow us to determine the clonal expansion of T cell populations and to predict specific epitopes responsible for T cell clonotype expansion. To do so, we will simulate PBMCs from vaccinated individuals with mutated or wild-type peptide pools; sort the stimulated cells by FACS as in 'aim ii'; and conduct single-cell sequencing on the resulting cell populations using GridION (Oxford Nanopore technologies) as well as Illumina (NextSeq). Prevalent T cell clonotypes will be identified using 10X Genomics' Loupe Browser, and antigen-specific T-cell populations will be identified using GLIPH2. Finally, to identify mutations most responsible for variations in clonotype expansion, the sequence motifs of the paired *alpha* and *beta* TCR sequences will be extracted from each expanded clonotypes. Their corresponding epitopes will be identified using multiple outsourced algorithms (RACER, ERGO, TCRGP, TCellMatch, TCRpMHcmodels).

### *3.4.3 Are certain populations really more at risk?*

Over the course of the current pandemic, many questions have been asked regarding the epidemiology, transmissibility, and virulence of SARS-COV-2. Amongst the many questions that have plagued the scientific community, two have been asked time and time again: "*Who gets sicker, and why?*". Although many question marks remain, this question has been partially answered, with age and immunocompromising conditions being strong determinants of disease outcome. HLA diversity across populations has also been proposed as a putative determinant of disease outcome. Studies have shown that the abundance of T cell epitopes recognized by an individual will depend on their HLA types. The clinical relevance of these findings is reflected by

the role played by cellular immunity in resolving SARS-CoV-2 infections and in mediating disease severity.

With the emergence and identification of VOCs, the question stated above has been slightly modified to become the following: "*which variants are bad, and are they equally bad for everyone?*". Again, this question can be partially answered using age and immunocompromising conditions. As part of the manuscript described above, our group proposes the putative role played by HLA compositions in an individual-specific manner in determining the impact of variants. Indeed, due to the high binding specificity of HLA (super)types, each individual HLA allele can recognize a unique pool of epitopes. As such, each HLA molecule may have unique susceptibilities to the global pool of SARS-CoV-2 variants. Overall, each individual mutation may have limited impact on the overall quality of the cellular immune response. However, as lineages become increasingly divergent from the ancestral lineage, the growing set of genomic variations specific to each lineage may eventually disrupt a critical proportion of T cell epitopes. As such, taking HLA allele compositions into account may shed light on population-specific susceptibilities to SARS-CoV-2 variants.

#### 3.4.3.1 Laboratory validation of T-cell escape amongst B7+ individuals

A major finding introduced in the manuscript presented above consists of the increased susceptibility of individuals carrying the B7+ HLA allele to lose SARS-CoV-2 epitopes. This finding resulted from the observation that the global SARS-CoV-2 mutation landscape throughout the first year of the pandemic was dominated by mutation biases, with the most prevalent mutation type being C→U. This prevalent mutation type led to the preferential removal of proline from the proteome of SARS-CoV-2 variants. Predictions indicated that this mutation type led to the preferential loss of HLA-B7 epitopes. This can be explained by the second position of HLA B7

binding grooves which heavily rely on the presence of prolines for the binding and presentation of epitopes to occur. Future work will involve the experimental validation of this concept.

## 4 Conclusion

In this dissertation, I reviewed the biological, evolutionary, epidemiological and immunological implications of SARS-CoV-2 over the course of this pandemic. I then presented my scientific contributions to the field in the form of a peer-reviewed manuscript featuring an in-depth investigation of the mutational landscape of SARS-CoV-2 over the course of the first year of the pandemic. By looking at over 300,000 SARS-CoV-2 genomic sequences, I observed that the global mutational landscape of SARS-CoV-2 was dominated by mutational biases, with the most prevalent bias, C→U, resulting in a global loss of Proline. These mutational biases were predicted to diversify CD8+ T cell epitopes in an HLA-supertype manner, with the removal of proline leading to the preferential loss of CD8+ T cell epitopes in the context of the HLA-B7 supertype. Together, these findings establish a link between the global SARS-CoV-2 mutational landscape and immune evasion in a fashion that takes into account HLA diversity. Additionally, the model developed in this work introduces a strategy to identify human sub-populations (such as those carrying HLA alleles of the B7 supertype) at risk of having a reduced CTL-based immune response when faced with SARS-CoV-2 variants. This study lays the groundwork for further investigating the HLA-dependent impact of circulating and emerging SARS-COV-2 variants on immune evasion in a comprehensive manner.

# 5 References


1. Rambaut A, Holmes EC, O'Toole Á, Hill V, McCrone JT, Ruis C, et al. A dynamic nomenclature proposal for SARS-CoV-2 lineages to assist genomic epidemiology. Nat Microbiol. 2020;5(11):1403–7.

2. (GISAID) GI on SAID. Clade and lineage nomenclature aids in genomic epidemiology studies of active hCoV-19 viruses [Internet]. 2020. Available from: https://www.gisaid.org/references/statements-clarifications/clade-and-lineage-nomenclature-aids-in-genomic-epidemiology-of-active-hcov-19-viruses/

3. Alm E, Broberg EK, Connor T, Hodcroft EB, Komissarov AB, Maurer-Stroh S, et al. Geographical and temporal distribution of SARS-CoV-2 clades in the WHO European Region, January to June 2020. Eurosurveillance. 2020;25(32):2001410.

4. Konings F, Perkins MD, Kuhn JH, Pallen MJ, Alm EJ, Archer BN, et al. SARS-CoV-2 Variants of Interest and Concern naming scheme conducive for global discourse. Nat Microbiol. 2021;6(7):821–3.

5. Hatchett RJ, Mecher CE, Lipsitch and M. Public health interventions and epidemic intensity during the 1918 influenza pandemic. PNAS. 2007;

6. Boostma MCJ, Ferguson NM. The effect of public health measures on the 1918 influenza pandemic in U.S. cities. PNAS. 2007;

7. Woo PCY, Lau SKP, Lam CSF, Lau CCY, Tsang AKL, Lau JHN, et al. Discovery of Seven Novel Mammalian and Avian Coronaviruses in the Genus Deltacoronavirus Supports Bat Coronaviruses as the Gene Source of Alphacoronavirus and Betacoronavirus and Avian Coronaviruses as the Gene Source of Gammacoronavirus and Deltacoronavirus. J Virol. 2012;86(7):3995–4008.

8. Perlman S, Netland J. Coronaviruses post-SARS: update on replication and pathogenesis. Nat Rev Microbiol. 2009;7(6):439–50.

9. Kendall EJC, Bynoe ML, Tyrrell DAJ. Virus Isolations from Common Colds Occurring in a Residential School. Brit Med J. 1962;2(5297):82.

10. Hu B, Guo H, Zhou P, Shi Z-L. Characteristics of SARS-CoV-2 and COVID-19. Nat Rev Microbiol. 2021;19(3):141–54.

11. Gorbalenya AE, Baker SC, Baric RS, Groot RJ de, Drosten C, Gulyaeva AA, et al. The species Severe acute respiratory syndrome-related coronavirus: classifying 2019-nCoV and naming it SARS-CoV-2. Nat Microbiol. 2020;5(4):536–44.

12. Guan Y, Zheng BJ, He YQ, Liu XL, Zhuang ZX, Cheung CL, et al. Isolation and Characterization of Viruses Related to the SARS Coronavirus from Animals in Southern China. Science. 2003;302(5643):276–8.

13. Song H-D, Tu C-C, Zhang G-W, Wang S-Y, Zheng K, Lei L-C, et al. Cross-host evolution of severe acute respiratory syndrome coronavirus in palm civet and human. P Natl Acad Sci Usa. 2005;102(7):2430–5.

14. Zhou P, Yang X-L, Wang X-G, Hu B, Zhang L, Zhang W, et al. A pneumonia outbreak associated with a new coronavirus of probable bat origin. Nature. 2020;579(7798):270–3.

15. Lu R, Zhao X, Li J, Niu P, Yang B, Wu H, et al. Genomic characterisation and epidemiology of 2019 novel coronavirus: implications for virus origins and receptor binding. Lancet Lond Engl. 2020;395(10224):565–74.

16. Boni MF, Lemey P, Jiang X, Lam TT-Y, Perry BW, Castoe TA, et al. Evolutionary origins of the SARS-CoV-2 sarbecovirus lineage responsible for the COVID-19 pandemic. Nat Microbiol. 2020;5(11):1408–17.

17. Andersen KG, Rambaut A, Lipkin WI, Holmes EC, Garry RF. The proximal origin of SARS-CoV-2. Nat Med. 2020;26(4):1–3.

18. Teoh K-T, Siu Y-L, Chan W-L, Schluter MA, Liu Chia-Jen, Peiris JSM, et al. The SARS Coronavirus E Protein Interacts with PALS1 and Alters Tight Junction Formation and Epithelial Morphogenesis. Molecular Biology of the Cell.

19. Mousavizadeh L, Ghasemi S. Genotype and phenotype of COVID-19: Their roles in pathogenesis. J Microbiol Immunol Infect. 2021;54(2):159–63.

20. Verheije MH, Hagemeijer MC, Ulasli M, Reggiori F, Rottier PJM, Masters PS, et al. The Coronavirus Nucleocapsid Protein Is Dynamically Associated with the Replication-Transcription Complexes ▿ †. J Virol. 2010;84(21):11575–9.

21. Snijder EJ, Limpens RWAL, Wilde AH de, Jong AWM de, Zevenhoven-Dobbe JC, Maier HJ, et al. A unifying structural and functional model of the coronavirus replication organelle: Tracking down RNA synthesis. Plos Biol. 2020;18(6):e3000715.

22. Jack A, Ferro LS, Trnka MJ, Wehri E, Nadgir A, Nguyenla X, et al. SARS-CoV-2 nucleocapsid protein forms condensates with viral genomic RNA. Plos Biol. 2021;19(10):e3001425.

23. Huang Y, Yang C, Xu X, Xu W, Liu S. Structural and functional properties of SARS-CoV-2 spike protein: potential antivirus drug development for COVID-19. Acta Pharmacol Sin. 2020;41(9):1141–9.

24. Wrapp1 D, Wang N, Corbett KS, Goldsmith JA, Hsieh C-L, Abiona O, et al. Cryo-EM structure of the 2019-nCoV spike in the prefusion conformation. Science. 2020;

25. Walls AC, Park Y-J, Tortorici MA, Wall A, McGuire AT, Veesler D. Structure, Function, and Antigenicity of the SARS-CoV-2 Spike Glycoprotein. Cell. 2020;181(2):281-292.e6.

26. Xia S, Zhu Y, Liu M, Lan Q, Xu W, Wu Y, et al. Fusion mechanism of 2019-nCoV and fusion inhibitors targeting HR1 domain in spike protein. Cell Mol Immunol. 2020;17(7):765–7.

27. Lan J, Ge J, Yu J, Shan S, Zhou H, Fan S, et al. Structure of the SARS-CoV-2 spike receptor-binding domain bound to the ACE2 receptor. Nature. 2020;581(7807):215–20.

28. Wang Q, Zhang Y, Wu L, Niu S, Song C, Zhang Z, et al. Structural and Functional Basis of SARS-CoV-2 Entry by Using Human ACE2. Cell. 2020;181(4):894-904.e9.

29. Hastie KM, Li H, Bedinger D, Schendel SL, Dennison SM, Li K, et al. Defining variant-resistant epitopes targeted by SARS-CoV-2 antibodies: A global consortium study. Science. 2021;374(6566):472–8.

30. Taylor PC, Adams AC, Hufford MM, Torre I de la, Winthrop K, Gottlieb RL. Neutralizing monoclonal antibodies for treatment of COVID-19. Nat Rev Immunol. 2021;21(6):1–12.

31. Hoffmann M, Kleine-Weber H, Schroeder S, Krüger N, Herrler T, Erichsen S, et al. SARS-CoV-2 Cell Entry Depends on ACE2 and TMPRSS2 and Is Blocked by a Clinically Proven Protease Inhibitor. Cell. 2020;181(2):271-280.e8.

32. Shang J, Han N, Chen Z, Peng Y, Li L, Zhou H, et al. Compositional diversity and evolutionary pattern of coronavirus accessory proteins. Brief Bioinform. 2020;22(2):1267–78.

33. Liu DX, Fung TS, Chong KK-L, Shukla A, Hilgenfeld R. Accessory proteins of SARS-CoV and other coronaviruses. Antivir Res. 2014;109:97–109.

34. Rohaim MA, Naggar RFE, Clayton E, Munir M. Structural and functional insights into non-structural proteins of coronaviruses. Microb Pathogenesis. 2020;150:104641.

35. Rando HM, MacLean AL, Lee AJ, Lordan R, Ray S, Bansal V, et al. Pathogenesis, Symptomatology, and Transmission of SARS-CoV-2 through Analysis of Viral Genomics and Structure. Msystems. 2021;6(5):e00095-21.

36. Li C, Zhao C, Bao J, Tang B, Wang Y, Gu B. Laboratory diagnosis of coronavirus disease-2019 (COVID-19). Clin Chimica Acta Int J Clin Chem. 2020;510:35–46.

37. Organization WH. Laboratory testing for coronavirus disease (COVID-19) in suspected human cases: interim guidance, 19 March 2020. 2020.

38. Paden CR, Tao Y, Queen K, Zhang J, Li Y, Uehara A, et al. Rapid, Sensitive, Full-Genome Sequencing of Severe Acute Respiratory Syndrome Coronavirus 2. Emerg Infect Dis. 2020;26(10):2401–5.

39. Pillay S, Giandhari J, Tegally H, Wilkinson E, Chimukangara B, Lessells R, et al. Whole Genome Sequencing of SARS-CoV-2: Adapting Illumina Protocols for Quick and Accurate Outbreak Investigation during a Pandemic. Genes-basel. 2020;11(8):949.

40. Campos GS, Sardi SI, Falcao MB, Belitardo EMMA, Rocha DJPG, Rolo CA, et al. Ion torrent-based nasopharyngeal swab metatranscriptomics in COVID-19. J Virol Methods. 2020;282:113888–113888.

41. Wu F, Zhao S, Yu B, Chen Y-M, Wang W, Song Z-G, et al. A new coronavirus associated with human respiratory disease in China. Nature. 2020;579(7798):265–9.

42. Gohl DM, Garbe J, Grady P, Daniel J, Watson RHB, Auch B, et al. A rapid, cost-effective tailed amplicon method for sequencing SARS-CoV-2. Bmc Genomics. 2020;21(1):863.

43. Tyson JR, James P, Stoddart D, Sparks N, Wickenhagen A, Hall G, et al. Improvements to the ARTIC multiplex PCR method for SARS-CoV-2 genome sequencing using nanopore. Biorxiv. 2020;2020.09.04.283077.

44. Gaudin M, Desnues C. Hybrid Capture-Based Next Generation Sequencing and Its Application to Human Infectious Diseases. Front Microbiol. 2018;9:2924.

45. Kozarewa I, Armisen J, Gardner AF, Slatko BE, Hendrickson CL. Overview of Target Enrichment Strategies. Curr Protoc Mol Biology. 2015;112(1):7.21.1-7.21.23.

46. Xiao M, Liu X, Ji J, Li M, Li J, Yang L, et al. Multiple approaches for massively parallel sequencing of SARS-CoV-2 genomes directly from clinical samples. Genome Med. 2020;12(1):57.

47. Robson F, Khan KS, Le TK, Paris C, Demirbag S, Barfuss P, et al. Coronavirus RNA Proofreading: Molecular Basis and Therapeutic Targeting. Mol Cell. 2020;79(5):710–27.

48. Korber B, Fischer WM, Gnanakaran S, Yoon H, Theiler J, Abfalterer W, et al. Tracking changes in SARS-CoV-2 Spike: evidence that D614G increases infectivity of the COVID-19 virus. Cell. 2020;182(4):812-827.e19.

49. Hadfield J, Megill C, Bell SM, Huddleston J, Potter B, Callender C, et al. Nextstrain: real-time tracking of pathogen evolution. Bioinformatics. 2018;34(23):4121–3.

50. Elbe S, Buckland-Merrett G. Data, disease and diplomacy: GISAID's innovative contribution to global health. Global Challenges. 2017;1(1):33–46.

51. Hufsky F, Lamkiewicz K, Almeida A, Aouacheria A, Arighi C, Bateman A, et al. Computational strategies to combat COVID-19: useful tools to accelerate SARS-CoV-2 and coronavirus research. Brief Bioinform. 2020;22(2):bbaa232-.

52. Munnink BBO, Worp N, Nieuwenhuijse DF, Sikkema RS, Haagmans B, Fouchier RAM, et al. The next phase of SARS-CoV-2 surveillance: real-time molecular epidemiology. Nat Med. 2021;27(9):1518–24.

53. Maxmen A. One million coronavirus sequences: popular genome site hits mega milestone. Nature. 2021;593(7857):21–21.

54. Volz EM, Koelle K, Bedford T. Viral Phylodynamics. Plos Comput Biol. 2013;9(3):e1002947.

55. Sagulenko P, Puller V, Neher RA. TreeTime: Maximum-likelihood phylodynamic analysis. Virus Evol. 2018;4(1):vex042.

56. Hodcroft EB. CoVariants: SARS-CoV-2 Mutations and Variants of Interest [Internet]. 2021. Available from: https://covariants.org/

57. Ferreira R-C, Wong E, Gugan G, Wade K, Liu M, Baena LM, et al. CoVizu: Rapid analysis and visualization of the global diversity of SARS-CoV-2 genomes. Virus Evolution. 2021;

58. Mullen JL, Tsueng G, Latif AA, Alkuzweny M, Cano M, Haag E, et al. outbreak.info [Internet]. 2020. Available from: https://outbreak.info/

59. GISAID. GISAID - Genomic Epidemiology of hCoV-19 [Internet]. 2023. Available from: https://www.gisaid.org/

60. Year-letter Genetic Clade Naming for SARS-CoV-2 on Nextstrain.org [Internet]. 2020. Available from: https://nextstrain.org/blog/2020-06-02-SARSCoV2-clade-naming

61. Han AX, Parker E, Scholer F, Maurer-Stroh S, Russell CA. Phylogenetic Clustering by Linear Integer Programming (PhyCLIP). Mol Biol Evol. 2019;36(7):1580–95.

62. Mostefai F, Gamache I, Huang J, N'Guessan A, Pelletier J, Pesaranghader A, et al. Data-driven approaches for genetic characterization of SARS-CoV-2 lineages. Biorxiv. 2021;2021.09.28.462270.

63. Selle ML, Steinsland I, Lindgren F, Brajkovic V, Cubric-Curik V, Gorjanc G. Hierarchical Modelling of Haplotype Effects on a Phylogeny. Frontiers Genetics. 2021;11:531218.

64. Forster P, Forster L, Renfrew C, Forster M. Phylogenetic network analysis of SARS-CoV-2 genomes. P Natl Acad Sci Usa. 2020;117(17):9241–3.

65. Motozono C, Toyoda M, Zahradnik J, Saito A, Nasser H, Tan TS, et al. SARS-CoV-2 spike L452R variant evades cellular immunity and increases infectivity. Cell Host Microbe. 2021;29(7):1124-1136.e11.

66. Frampton D, Rampling T, Cross A, Bailey H, Heaney J, Byott M, et al. Genomic characteristics and clinical effect of the emergent SARS-CoV-2 B.1.1.7 lineage in London, UK: a whole-genome sequencing and hospital-based cohort study. Lancet Infect Dis. 2021;21(9):1246–56.

67. Tegally H, Wilkinson E, Giovanetti M, Iranzadeh A, Fonseca V, Giandhari J, et al. Detection of a SARS-CoV-2 variant of concern in South Africa. Nature. 2021;592(7854):438–43.

68. Faria NR, Mellan TA, Whittaker C, Claro IM, Candido D da S, Mishra S, et al. Genomics and epidemiology of the P.1 SARS-CoV-2 lineage in Manaus, Brazil. Science. 2021;

69. Prevention C for DC and. SARS-CoV-2 Variant Classifications and Definitions [Internet]. 2021. Available from: https://www.cdc.gov/coronavirus/2019-ncov/variants/variant-classifications.html?CDC_AA_refVal=https%3A%2F%2Fwww.cdc.gov%2Fcoronavirus%2F2019-ncov%2Fvariants%2Fvariant-info.html#anchor_1632158775384

70. Peacock TP, Penrice-Randal R, Hiscox JA, Barclay WS. SARS-CoV-2 one year on: evidence for ongoing viral adaptation. J Gen Virol. 2021;102(4).

71. Organization WH. Tracking SARS-CoV-2 variants [Internet]. 2021. Available from: https://www.who.int/en/activities/tracking-SARS-CoV-2-variants/

72. England PH. Investigation of novel SARS-COV-2 variant Variant of Concern 202012/01—technical briefing. [Internet]. 2021. Available from: https://www.gov.uk/government/publications/phe-investigation-of-novel-sars-cov-2-variant-of-concern-20201201-technical-briefing-3-6-january-2021

73. Davies NG, Abbott S, Barnard RC, Jarvis CI, Kucharski AJ, Munday JD, et al. Estimated transmissibility and impact of SARS-CoV-2 lineage B.1.1.7 in England. Science. 2021;372(6538).

74. Volz E, Mishra S, Chand M, Barrett JC, Johnson R, Geidelberg L, et al. Assessing transmissibility of SARS-CoV-2 lineage B.1.1.7 in England. Nature. 2021;593(7858):266–9.

75. Gu H, Chen Q, Yang G, He L, Fan H, Deng Y-Q, et al. Adaptation of SARS-CoV-2 in BALB/c mice for testing vaccine efficacy. Science. 2020;369(6511):1603–7.

76. Starr TN, Greaney AJ, Hilton SK, Ellis D, Crawford KHD, Dingens AS, et al. Deep Mutational Scanning of SARS-CoV-2 Receptor Binding Domain Reveals Constraints on Folding and ACE2 Binding. Cell. 2020;182(5):1295-1310.e20.

77. Hoffmann M, Kleine-Weber H, Pöhlmann S. A Multibasic Cleavage Site in the Spike Protein of SARS-CoV-2 Is Essential for Infection of Human Lung Cells. Mol Cell. 2020;78(4):779-784.e5.

78. Peacock TP, Goldhill DH, Zhou J, Baillon L, Frise R, Swann OC, et al. The furin cleavage site in the SARS-CoV-2 spike protein is required for transmission in ferrets. Nat Microbiol. 2021;6(7):899–909.

79. Meng B, Kemp SA, Papa G, Datir R, Ferreira IATM, Marelli S, et al. Recurrent emergence of SARS-CoV-2 spike deletion H69/V70 and its role in the Alpha variant B.1.1.7. Cell Reports. 2021;35(13):109292.

80. Coutinho RM, Marquitti FMD, Ferreira LS, Borges ME, Silva RLP da, Canton O, et al. Model-based estimation of transmissibility and reinfection of SARS-CoV-2 P.1 variant. Commun Medicine. 2021;1(1):48.

81. Wang P, Nair MS, Liu L, Iketani S, Luo Y, Guo Y, et al. Antibody resistance of SARS-CoV-2 variants B.1.351 and B.1.1.7. Nature. 2021;593(7857):130–5.

82. Planas D, Bruel T, Grzelak L, Guivel-Benhassine F, Staropoli I, Porrot F, et al. Sensitivity of infectious SARS-CoV-2 B.1.1.7 and B.1.351 variants to neutralizing antibodies. Nat Med. 2021;27(5):917–24.

83. Hoffmann M, Arora P, Groß R, Seidel A, Hörnich BF, Hahn AS, et al. SARS-CoV-2 variants B.1.351 and P.1 escape from neutralizing antibodies. Cell. 2021;184(9):2384-2393.e12.

84. Zhou D, Dejnirattisai W, Supasa P, Liu C, Mentzer AJ, Ginn HM, et al. Evidence of escape of SARS-CoV-2 variant B.1.351 from natural and vaccine-induced sera. Cell. 2021;184(9):2348-2361.e6.

85. Li R, Ma X, Deng J, Chen Q, Liu W, Peng Z, et al. Differential efficiencies to neutralize the novel mutants B.1.1.7 and 501Y.V2 by collected sera from convalescent COVID-19 patients and RBD nanoparticle-vaccinated rhesus macaques. Cell Mol Immunol. 2021;18(4):1–3.

86. Garcia-Beltran WF, Lam EC, Denis KSt, Nitido AD, Garcia ZH, Hauser BM, et al. Multiple SARS-CoV-2 variants escape neutralization by vaccine-induced humoral immunity. Cell. 2021;184(9):2372-2383.e9.

87. Campbell F, Archer B, Laurenson-Schafer H, Jinnai Y, Konings F, Batra N, et al. Increased transmissibility and global spread of SARS-CoV-2 variants of concern as at June 2021. Eurosurveillance. 2021;26(24):2100509.

88. Cherian S, Potdar V, Jadhav S, Yadav P, Gupta N, Das M, et al. SARS-CoV-2 Spike Mutations, L452R, T478K, E484Q and P681R, in the Second Wave of COVID-19 in Maharashtra, India. Microorg. 2021;9(7):1542.

89. Liu Z, VanBlargan LA, Bloyet L-M, Rothlauf PW, Chen RE, Stumpf S, et al. Identification of SARS-CoV-2 spike mutations that attenuate monoclonal and serum antibody neutralization. Cell Host Microbe. 2021;29(3):477-488.e4.

90. Greaney AJ, Loes AN, Crawford KHD, Starr TN, Malone KD, Chu HY, et al. Comprehensive mapping of mutations in the SARS-CoV-2 receptor-binding domain that affect recognition by polyclonal human plasma antibodies. Cell Host Microbe. 2021;29(3):463-476.e6.

91. Piccoli L, Park Y-J, Tortorici MA, Czudnochowski N, Walls AC, Beltramello M, et al. Mapping Neutralizing and Immunodominant Sites on the SARS-CoV-2 Spike Receptor-Binding Domain by Structure-Guided High-Resolution Serology. Cell. 2020;183(4):1024-1042.e21.

92. Cameroni E, Bowen JE, Rosen LE, Saliba C, Zepeda SK, Culap K, et al. Broadly neutralizing antibodies overcome SARS-CoV-2 Omicron antigenic shift. Nature. 2022;602(7898):664–70.

93. Planas D, Saunders N, Maes P, Guivel-Benhassine F, Planchais C, Buchrieser J, et al. Considerable escape of SARS-CoV-2 Omicron to antibody neutralization. Nature. 2022;602(7898):671–5.

94. Cao Y, Wang J, Jian F, Xiao T, Song W, Yisimayi A, et al. Omicron escapes the majority of existing SARS-CoV-2 neutralizing antibodies. Nature. 2022;602(7898):657–63.

95. Cele S, Jackson L, Khoury DS, Khan K, Moyo-Gwete T, Tegally H, et al. Omicron extensively but incompletely escapes Pfizer BNT162b2 neutralization. Nature. 2022;602(7898):654–6.

96. Zhang L, Zhang F, Yu W, He T, Yu J, Yi CE, et al. Antibody responses against SARS coronavirus are correlated with disease outcome of infected individuals. J Med Virol. 2006;78(1):1–8.

97. Shi Y, Wan Z, Li L, Li P, Li C, Ma Q, et al. Antibody responses against SARS-coronavirus and its nucleocaspid in SARS patients. J Clin Virol. 2004;31(1):66–8.

98. Huang L, Chiu C, Yeh S, Huang W, Hsueh P, Yang W, et al. Evaluation of antibody responses against SARS coronaviral nucleocapsid or spike proteins by immunoblotting or ELISA. J Med Virol. 2004;73(3):338–46.

99. Liu X, Shi Y, Li P, Li L, Yi Y, Ma Q, et al. Profile of Antibodies to the Nucleocapsid Protein of the Severe Acute Respiratory Syndrome (SARS)-Associated Coronavirus in Probable SARS Patients. Clin Diagn Lab Immun. 2004;11(1):227–8.

100. Leung DTM, Tam 1 Frankie Chi Hang, Ma 1 Chun Hung, Chan 1 Paul Kay Sheung, Cheung 2 Jo Lai Ken, Niu 2 Haitao, et al. Antibody Response of Patients with Severe Acute Respiratory Syndrome (SARS) Targets the Viral Nucleocapsid. Journal of Infectious Diseases. 2004;

101. Wu L-P, Wang N-C, Chang Y-H, Tian X-Y, Na D-Y, Zhang L-Y, et al. Duration of Antibody Responses after Severe Acute Respiratory Syndrome - Volume 13, Number 10—October 2007 - Emerging Infectious Diseases journal - CDC. Emerg Infect Dis. 2007;13(10):1562–4.

102. Tang F, Quan Y, Xin Z-T, Wrammert J, Ma M-J, Lv H, et al. Lack of Peripheral Memory B Cell Responses in Recovered Patients with Severe Acute Respiratory Syndrome: A Six-Year Follow-Up Study. J Immunol. 2011;186(12):7264–8.

103. Liu W, Fontanet 1 Arnaud, Zhang 3 Pan-He, Zhan 1 Lin, Xin 1 Zhong-Tao, Baril 2 Laurence, et al. Two-Year Prospective Study of the Humoral Immune Response of Patients with Severe Acute Respiratory Syndrome. The Journal of Infectious Diseases. 2006;

104. Seow J, Graham C, Merrick B, Acors S, Pickering S, Steel KJA, et al. Longitudinal observation and decline of neutralizing antibody responses in the three months following SARS-CoV-2 infection in humans. Nat Microbiol. 2020;1–10.

105. Ng O-W, Chia A, Tan AT, Jadi RS, Leong HN, Bertoletti A, et al. Memory T cell responses targeting the SARS coronavirus persist up to 11 years post-infection. Vaccine. 2016;34(17):2008–14.

106. Peng H, Yang L, Wang L, Li J, Huang J, Lu Z, et al. Long-lived memory T lymphocyte responses against SARS coronavirus nucleocapsid protein in SARS-recovered patients. Virology. 2006;351(2):466–75.

107. Li CK, Wu H, Yan H, Ma S, Wang L, Zhang M, et al. T Cell Responses to Whole SARS Coronavirus in Humans. J Immunol. 2008;181(8):5490–500.

108. Zhao J, Alshukairi AN, Baharoon SA, Ahmed WA, Bokhari AA, Nehdi AM, et al. Recovery from the Middle East respiratory syndrome is associated with antibody and T-cell responses. Sci Immunol. 2017;2(14):eaan5393.

109. Wang B, Chen H, Jiang X, Zhang M, Wan T, Li N, et al. Identification of an HLA-A*0201-restricted CD8+ T-cell epitope SSp-1 of SARS-CoV spike protein. BLOOD. 2004;

110. Chen H, Hou J, Jiang X, Ma S, Meng M, Wang B, et al. Response of Memory CD8+ T Cells to Severe Acute Respiratory Syndrome (SARS) Coronavirus in Recovered SARS Patients and Healthy Individuals. J Immunol. 2005;175(1):591–8.

111. T-Cell Epitopes in Severe Acute Respiratory Syndrome (SARS) Coronavirus Spike Protein Elicit a Specific T-Cell Immune Response in Patients Who Recover from SARS.

112. Zhao J, Zhao J, Mangalam AK, Channappanavar R, Fett C, Meyerholz DK, et al. Airway Memory CD4+ T Cells Mediate Protective Immunity against Emerging Respiratory Coronaviruses. Immunity. 2016;44(6):1379–91.

113. Zhao K, Yang B, Xu Y, Wu C. CD8+ T cell response in HLA-A*0201 transgenic mice is elicited by epitopes from SARS-CoV S protein. Vaccine. 2010;28(41):6666–74.

114. Oh H-LJ, Chia A, Chang CXL, Leong HN, Ling KL, Grotenbreg GM, et al. Engineering T Cells Specific for a Dominant Severe Acute Respiratory Syndrome Coronavirus CD8 T Cell Epitope. J Virol. 2011;85(20):10464–71.

115. Huang J, Ma R, Wu C. Immunization with SARS-CoV S DNA vaccine generates memory CD4+ and CD8+ T cell immune responses. Vaccine. 2006;24(23):4905–13.

116. Yang L, Peng H, Zhu Z, Li G, Huang Z, Zhao Z, et al. Persistent memory CD4+ and CD8+ T-cell responses in recovered severe acute respiratory syndrome (SARS) patients to SARS coronavirus M antigen. J Gen Virol. 2007;88(10):2740–8.

117. Huang J, Cao Y, Du J, Bu X, Ma R, Wu C. Priming with SARS CoV S DNA and boosting with SARS CoV S epitopes specific for CD4+ and CD8+ T cells promote cellular immune responses. Vaccine. 2007;25(39–40):6981–91.

118. Channappanavar R, Fett C, Zhao J, Meyerholz DK, Perlman S. Virus-Specific Memory CD8 T Cells Provide Substantial Protection from Lethal Severe Acute Respiratory Syndrome Coronavirus Infection. J Virol. 2014;88(19):11034–44.

119. Grifoni A, Sidney J, Zhang Y, Scheuermann RH, Peters B, Sette A. A Sequence Homology and Bioinformatic Approach Can Predict Candidate Targets for Immune Responses to SARS-CoV-2. Cell Host Microbe. 2020;27(4):671-680.e2.

120. Kiyotani K, Toyoshima Y, Nemoto K, Nakamura Y. Bioinformatic prediction of potential T cell epitopes for SARS-Cov-2. J Hum Genet. 2020;65(7):569–75.

121. Nguyen A, David JK, Maden SK, Wood MA, Nellore A, Thompsona RF, et al. Human Leukocyte Antigen Susceptibility Map for Severe Acute Respiratory Syndrome Coronavirus 2. Journal of Virology. 2020;

122. Moderbacher CR, Ramirez SI, Dan JM, Grifoni A, Hastie KM, Weiskopf D, et al. Antigen-specific adaptive immunity to SARS-CoV-2 in acute COVID-19 and associations with age and disease severity. Cell. 2020;

123. Mateus J, Grifoni A, Tarke A, Sidney J, Ramirez SI, Dan JM, et al. Selective and cross-reactive SARS-CoV-2 T cell epitopes in unexposed humans. Science. 2020;eabd3871.

124. Dan JM, Mateus J, Kato Y, Hastie KM, Yu ED, Faliti CE, et al. Immunological memory to SARS-CoV-2 assessed for up to 8 months after infection. Sci New York N Y. 2021;371(6529):eabf4063.

125. Jung JH, Rha M-S, Sa M, Choi HK, Jeon JH, Seok H, et al. SARS-CoV-2-specific T cell memory is sustained in COVID-19 convalescent patients for 10 months with successful development of stem cell-like memory T cells. Nat Commun. 2021;12(1):4043.

126. Bert NL, Tan AT, Kunasegaran K, Tham CYL, Hafezi M, Chia A, et al. SARS-CoV-2-specific T cell immunity in cases of COVID-19 and SARS, and uninfected controls. Nature. 2020;584(7821):457–62.

127. Tan AT, Linster M, Tan CW, Bert NL, Chia WN, Kunasegaran K, et al. Early induction of functional SARS-CoV-2-specific T cells associates with rapid viral clearance and mild disease in COVID-19 patients. Cell Reports. 2021;34(6):108728–108728.

128. Matyushenko V, Isakova-Sivak I, Kudryavtsev I, Goshina A, Chistyakova A, Stepanova E, et al. Detection of IFNγ-Secreting CD4+ and CD8+ Memory T Cells in COVID-19 Convalescents after Stimulation of Peripheral Blood Mononuclear Cells with Live SARS-CoV-2. Viruses. 2021;13(8):1490.

129. Lazarevic I, Pravica V, Miljanovic D, Cupic M. Immune Evasion of SARS-CoV-2 Emerging Variants: What Have We Learnt So Far? Viruses. 2021;13(7):1192.

130. Hansen CB, Jarlhelt I, Pérez-Alós L, Landsy LH, Loftager M, Rosbjerg A, et al. SARS-CoV-2 Antibody Responses Are Correlated to Disease Severity in COVID-19 Convalescent Individuals. J Immunol. 2020;206(1):ji2000898.

131. Garcia-Beltran WF, Lam EC, Astudillo MG, Yang D, Miller TE, Feldman J, et al. COVID-19-neutralizing antibodies predict disease severity and survival. Cell. 2021;184(2):476-488.e11.

132. Wang P, Casner RG, Nair MS, Wang M, Yu J, Cerutti G, et al. Increased resistance of SARS-CoV-2 variant P.1 to antibody neutralization. Cell Host Microbe. 2021;29(5):747-751.e4.

133. Supasa P, Zhou D, Dejnirattisai W, Liu C, Mentzer AJ, Ginn HM, et al. Reduced neutralization of SARS-CoV-2 B.1.1.7 variant by convalescent and vaccine sera. Cell. 2021;184(8):2201-2211.e7.

134. Collier DA, Marco AD, Ferreira IATM, Meng B, Datir RP, Walls AC, et al. Sensitivity of SARS-CoV-2 B.1.1.7 to mRNA vaccine-elicited antibodies. Nature. 2021;593(7857):136–41.

135. Peacock TP, Sheppard CM, Brown JC, Goonawardane N, Zhou J, Whiteley M, et al. The SARS-CoV-2 variants associated with infections in India, B.1.617, show enhanced spike cleavage by furin. Biorxiv. 2021;2021.05.28.446163.

136. Zhang L, Mann M, Syed ZA, Reynolds HM, Tian E, Samara NL, et al. Furin cleavage of the SARS-CoV-2 spike is modulated by O-glycosylation. PNAS. 2021;

137. Jordan SC, Shin B-H, Gadsden T-AM, Chu M, Petrosyan A, Le CN, et al. T cell immune responses to SARS-CoV-2 and variants of concern (Alpha and Delta) in infected and vaccinated individuals. Cell Mol Immunol. 2021;18(11):2554–6.

138. Geers D, Shamier MC, Bogers S, Hartog G den, Gommers L, Nieuwkoop NN, et al. SARS-CoV-2 variants of concern partially escape humoral but not T-cell responses in COVID-19 convalescent donors and vaccinees. Sci Immunol. 2021;6(59):eabj1750.

139. Agerer B, Koblischke M, Gudipati V, Montaño-Gutierrez LF, Smyth M, Popa A, et al. SARS-CoV-2 mutations in MHC-I-restricted epitopes evade CD8+ T cell responses. Sci Immunol. 2021;6(57):eabg6461.

140. Silva TI de, Liu G, Lindsey BB, Dong D, Moore SC, Hsu NS, et al. The impact of viral mutations on recognition by SARS-CoV-2 specific T cells. Iscience. 2021;24(11):103353.

141. Zhang H, Deng S, Ren L, Zheng P, Hu X, Jin T, et al. Profiling CD8+ T cell epitopes of COVID-19 convalescents reveals reduced cellular immune responses to SARS-CoV-2 variants. Cell Reports. 2021;36(11):109708.

142. Keeton R, Tincho MB, Ngomti A, Baguma R, Benede N, Suzuki A, et al. T cell responses to SARS-CoV-2 spike cross-recognize Omicron. Nature. 2022;603(7901):488–92.

143. Naranbhai V, Nathan A, Kaseke C, Berrios C, Khatri A, Choi S, et al. T cell reactivity to the SARS-CoV-2 Omicron variant is preserved in most but not all individuals. Cell. 2022;185(6):1041-1051.e6.

144. Diamond MS, Kanneganti T-D. Innate immunity: the first line of defense against SARS-CoV-2. Nat Immunol. 2022;23(2):165–76.

145. Tartey S, Takeuchi O. Pathogen recognition and Toll-like receptor targeted therapeutics in innate immune cells. Int Rev Immunol. 2017;36(2):1–17.

146. Zheng M, Karki R, Williams EP, Yang D, Fitzpatrick E, Vogel P, et al. TLR2 senses the SARS-CoV-2 envelope protein to produce inflammatory cytokines. Nat Immunol. 2021;22(7):829–38.

147. Choudhury A, Mukherjee S. In silico studies on the comparative characterization of the interactions of SARS-CoV-2 spike glycoprotein with ACE-2 receptor homologs and human TLRs. J Med Virol. 2020;92(10):10.1002/jmv.25987.

148. Totura AL, Whitmore A, Agnihothram S, Schäfer A, Katze MG, Heise MT, et al. Toll-Like Receptor 3 Signaling via TRIF Contributes to a Protective Innate Immune Response to Severe Acute Respiratory Syndrome Coronavirus Infection. Mbio. 2015;6(3):e00638-15.

149. Rebendenne A, Valadão ALC, Tauziet M, Maarifi G, Bonaventure B, McKellar J, et al. SARS-CoV-2 Triggers an MDA-5-Dependent Interferon Response Which Is Unable To Control Replication in Lung Epithelial Cells. J Virol. 2021;95(8):e02415-20.

150. Yin X, Riva L, Pu Y, Martin-Sancho L, Kanamune J, Yamamoto Y, et al. MDA5 Governs the Innate Immune Response to SARS-CoV-2 in Lung Epithelial Cells. Cell Reports. 2021;34(2):108628–108628.

151. Christgen S, Kanneganti T-D. Inflammasomes and the fine line between defense and disease. Curr Opin Immunol. 2020;62:39–44.

152. Pan P, Shen M, Yu Z, Ge W, Chen K, Tian M, et al. SARS-CoV-2 N protein promotes NLRP3 inflammasome activation to induce hyperinflammation. Nat Commun. 2021;12(1):4664.

153. Xu H, Akinyemi IA, Chitre SA, Loeb JC, Lednicky JA, McIntosh MT, et al. SARS-CoV-2 viroporin encoded by ORF3a triggers the NLRP3 inflammatory pathway. Virology. 2022;568:13–22.

154. Campbell GR, To RK, Hanna J, Spector SA. SARS-CoV-2, SARS-CoV-1, and HIV-1 derived ssRNA sequences activate the NLRP3 inflammasome in human macrophages through a non-classical pathway. Iscience. 2021;24(4):102295.

155. Schultze JL, Aschenbrenner AC. COVID-19 and the human innate immune system. Cell. 2021;184(7):1671–92.

156. Karki R, Sharma BR, Tuladhar S, Williams EP, Zalduondo L, Samir P, et al. Synergism of TNF-α and IFN-γ Triggers Inflammatory Cell Death, Tissue Damage, and Mortality in SARS-CoV-2 Infection and Cytokine Shock Syndromes. Cell. 2021;184(1):149-168.e17.

157. Karki R, Kanneganti T-D. The 'cytokine storm': molecular mechanisms and therapeutic prospects. Trends Immunol. 2021;42(8):681–705.

158. Robinson J, Barker DJ, Georgiou X, Cooper MA, Flicek P, Marsh SGE. IPD-IMGT/HLA Database. Nucleic Acids Res. 2020;48(D1):D948–55.

159. Neefjes J, Jongsma MLM, Paul P, Bakke O. Towards a systems understanding of MHC class I and MHC class II antigen presentation. Nat Rev Immunol. 2011;11(12):823–36.

160. Endert PM van, Riganelli * Daniela, Greco $ Giulia, Fleischhauer § Katharina, Sidney ll J, Sette ' Alessandro, et al. The Peptide-binding Motif for the Human Transporter. J Exp Med. 1995;

161. Corradi V, Singh G, Tieleman DP. The Human Transporter Associated with Antigen Processing MOLECULAR MODELS TO DESCRIBE PEPTIDE BINDING COMPETENT STATES. J Biol Chem. 2012;287(33):28099–111.

162. Saric T, Chang S-C, Hattori A, York IA, Markant S, Rock KL, et al. An IFN-γ–induced aminopeptidase in the ER, ERAP1, trims precursors to MHC class I–presented peptides. Nat Immunol. 2002;3(12):1169–76.

163. Saveanu L, Carroll O, Lindo V, Val MD, Lopez D, Lepelletier Y, et al. Concerted peptide trimming by human ERAP1 and ERAP2 aminopeptidase complexes in the endoplasmic reticulum. Nat Immunol. 2005;6(7):689–97.

164. Madden DR, Gorga JC, Strominger JL, Wiley DC. The three-dimensional structure of HLA-B27 at 2.1 Å resolution suggests a general mechanism for tight peptide binding to MHC. Cell. 1992;70(6):1035–48.

165. Silva ML, Guo H-C, Strominger JL, Wiley DC. Atomic Structure of a human MHC molecule presenting an influenza virus peptide. Nature. 1992;

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

167. HUNT DF, HENDERSON RA, SHABANOWITZ J, SAKAGUCHI K, MICHEL H, SEVILIR N, et al. American Association for the Advancement of Science is collaborating with JSTOR to digitize,. Science. 1992;

168. Falk irsten, Rötzschke O, Deres K, Metzger J, Jung G, Rammensee H-G. Nonapeptides Allows Their Quantification in Infected Forecast. J Exp Med. 1991;

169. Rötzschke O, Falk K, Deres K, Schild H, Norda M, Metzger J, et al. Isolation and analysis of naturally processed viral peptides as recognized by cytotoxic T cells. Nature. 1990;348(6298):252–4.

170. Henderdson RA, Michel H, Sakaguchi K, Shabanowitz J, Appella E, Hunt DF, et al. HLA-A2.1-Associated Peptides from a Mutant Cell C. Science. 1992;

171. Hunt DF, Michel H, Dickinson TA, Shabanowitz J, Cox AL, Sakaguchi K, et al. Complex Molecule lAd. Science. 1992;

172. Sette A, Buus S, Appella E, Smith JA, Chesnut R, Miles C, et al. Prediction of major histocompatibility complex binding regions of protein antigens by sequence pattern analysis. Proc National Acad Sci. 1989;86(9):3296–300.

173. Parker KC, Bednarek MA, Coligan JE. Scheme for ranking potential HLA-A2 binding peptides based on independent binding of individual peptide side-chains. Journal of Immunology. 1993;

174. Rammensee H-G, Bachmann J, Emmerich NPN, Bachor OA, Stevanović S. SYFPEITHI: database for MHC ligands and peptide motifs. Immunogenetics. 1999;50(3–4):213–9.

175. Jurtz V, Paul S, Andreatta M, Marcatili P, Peters B, Nielsen M. NetMHCpan-4.0: Improved Peptide–MHC Class I Interaction Predictions Integrating Eluted Ligand and Peptide Binding Affinity Data. J Immunol. 2017;199(9):3360–8.

176. Paul S, Croft NP, Purcell AW, Tscharke DC, Sette A, Nielsen M, et al. Benchmarking predictions of MHC class I restricted T cell epitopes in a comprehensively studied model system. Plos Comput Biol. 2020;16(5):e1007757.

177. O'Donnell TJ, Rubinsteyn A, Laserson U. MHCflurry 2.0: Improved Pan-Allele Prediction of MHC Class I-Presented Peptides by Incorporating Antigen Processing. Cell Syst. 2020;11(1):42-48.e7.

178. Nielsen M, Lundegaard C, Blicher T, Lamberth K, Harndahl M, Justesen S, et al. NetMHCpan, a Method for Quantitative Predictions of Peptide Binding to Any HLA-A and -B Locus Protein of Known Sequence. Plos One. 2007;2(8):e796.

179. Reynisson B, Alvarez B, Paul S, Peters B, Nielsen M. NetMHCpan-4.1 and NetMHCIIpan-4.0: improved predictions of MHC antigen presentation by concurrent motif deconvolution and integration of MS MHC eluted ligand data. Nucleic Acids Res. 2020;48(W1):gkaa379-.

180. Nielsen M, Andreatta M. NetMHCpan-3.0; improved prediction of binding to MHC class I molecules integrating information from multiple receptor and peptide length datasets. Genome Med. 2016;8(1):33.

181. Nielsen M, Andreatta M. NNAlign: a platform to construct and evaluate artificial neural network models of receptor–ligand interactions. Nucleic Acids Res. 2017;45(Web Server issue):W344–9.

182. Bergseng E, Dørum S, Arntzen MØ, Nielsen M, Nygård S, Buus S, et al. Different binding motifs of the celiac disease-associated HLA molecules DQ2.5, DQ2.2, and DQ7.5 revealed by relative quantitative proteomics of endogenous peptide repertoires. Immunogenetics. 2015;67(2):73–84.

183. Bassani-Sternberg M, Pletscher-Frankild S, Jensen LJ, Mann M. Mass Spectrometry of Human Leukocyte Antigen Class I Peptidomes Reveals Strong Effects of Protein Abundance and Turnover on Antigen Presentation*. Mol Cell Proteom Mcp. 2015;14(3):658–73.

184. Hassan C, Kester MGD, Ru AH de, Hombrink P, Drijfhout JW, Nijveen H, et al. The Human Leukocyte Antigen–presented Ligandome of B Lymphocytes*. Mol Cell Proteomics. 2013;12(7):1829–43.

185. Caron E, Kowalewski DanielJ, Koh CC, Sturm T, Schuster H, Aebersold R. Analysis of Major Histocompatibility Complex (MHC) Immunopeptidomes Using Mass Spectrometry*. Mol Cell Proteomics. 2015;14(12):3105–17.

186. Pak H, Michaux J, Huber F, Chong C, Stevenson BJ, Müller M, et al. Sensitive Immunopeptidomics by Leveraging Available Large-Scale Multi-HLA Spectral Libraries, Data-Independent Acquisition, and MS/MS Prediction. Mol Cell Proteom Mcp. 2021;20:100080.

187. Gonzalez-Duque S, Azoury ME, Colli ML, Afonso G, Turatsinze J-V, Nigi L, et al. Conventional and Neo-antigenic Peptides Presented by β Cells Are Targeted by Circulating Naïve CD8+ T Cells in Type 1 Diabetic and Healthy Donors. Cell Metab. 2018;28(6):946-960.e6.

188. Schumacher TN, Scheper W, Kvistborg P. Cancer Neoantigens. Annu Rev Immunol. 2019;

189. Bianchi V, Harari A, Coukos G. Neoantigen-Specific Adoptive Cell Therapies for Cancer: Making T-Cell Products More Personal. Front Immunol. 2020;11:1215.

190. Ebrahimi-Nik H, Michaux J, Corwin WL, Keller GLJ, Shcheglova T, Pak H, et al. Mass spectrometry driven exploration reveals nuances of neoepitope-driven tumor rejection. Jci Insight. 2019;4(14).

191. Chikata T, Paes W, Akahoshi T, Partridge T, Murakoshi H, Gatanaga H, et al. Identification of Immunodominant HIV-1 Epitopes Presented by HLA-C*12:02, a Protective Allele, Using an Immunopeptidomics Approach. J Virol. 2019;93(17):e00634-19.

192. Nagler A, Kalaora S, Barbolin C, Gangaev A, Ketelaars SLC, Alon M, et al. Identification of presented SARS-CoV-2 HLA class I and HLA class II peptides using HLA peptidomics. Cell Reports. 2021;35(13):109305.

193. Weingarten-Gabbay S, Klaeger S, Sarkizova S, Pearlman LR, Chen D-Y, Gallagher KME, et al. Profiling SARS-CoV-2 HLA-I peptidome reveals T cell epitopes from out-of-frame ORFs. Cell. 2021;184(15):3962-3980.e17.

194. Bettencourt P, Müller J, Nicastri A, Cantillon D, Madhavan M, Charles PD, et al. Identification of antigens presented by MHC for vaccines against tuberculosis. Npj Vaccines. 2020;5(1):2.

195. Caron E, Espona L, Kowalewski DJ, Schuster H, Ternette N, Alpízar A, et al. An open-source computational and data resource to analyze digital maps of immunopeptidomes. Elife. 2015;4:e07661.

196. Bjerregaard A-M, Nielsen M, Jurtz V, Barra CM, Hadrup SR, Szallasi Z, et al. An Analysis of Natural T Cell Responses to Predicted Tumor Neoepitopes. Front Immunol. 2017;8:1566.

197. Koşaloğlu-Yalçın Z, Lanka M, Frentzen A, Premlal ALR, Sidney J, Vaughan K, et al. Predicting T cell recognition of MHC class I restricted neoepitopes. Oncoimmunology. 2018;7(11):e1492508.

198. Peters B, Nielsen M, Sette A. T Cell Epitope Predictions. Annu Rev Immunol. 2020;

199. Grifoni A, Weiskopf D, Ramirez SI, Mateus J, Dan JM, Moderbacher CR, et al. Targets of T Cell Responses to SARS-CoV-2 Coronavirus in Humans with COVID-19 Disease and Unexposed Individuals. Cell. 2020;181(7):1489-1501.e15.

200. Teraguchi S, Saputri DS, Llamas-Covarrubias MA, Davila A, Diez D, Nazlica SA, et al. Methods for sequence and structural analysis of B and T cell receptor repertoires. Comput Struct Biotechnology J. 2020;18:2000–11.

201. Bassing CH, Swat W, Alt FW. The Mechanism and Regulation of Chromosomal V(D)J Recombination. Cell. 2002;109(2):S45–55.

202. Singh M, Al-Eryani G, Carswell S, Ferguson JM, Blackburn J, Barton K, et al. High-throughput targeted long-read single cell sequencing reveals the clonal and transcriptional landscape of lymphocytes. Nat Commun. 2019;10(1):3120.

203. Wang P, Jin X, Zhou W, Luo M, Xu Z, Xu C, et al. Comprehensive analysis of TCR repertoire in COVID-19 using single cell sequencing. Genomics. 2021;113(2):456–62.

204. Minervina AA, Komech EA, Titov A, Koraichi MB, Rosati E, Mamedov IZ, et al. Longitudinal high-throughput TCR repertoire profiling reveals the dynamics of T cell memory formation after mild COVID-19 infection. Elife. 2021;10:e63502.

205. Fernandez DM, Rahman AH, Fernandez N, Chudnovskiy A, Amir ED, Amadori L, et al. Single-cell immune landscape of human atherosclerotic plaques. Nat Med. 2019;25(10):1576–88.

206. Picelli S, Björklund ÅK, Faridani OR, Sagasser S, Winberg G, Sandberg R. Smart-seq2 for sensitive full-length transcriptome profiling in single cells. Nat Methods. 2013;10(11):1096–8.

207. Huang H, Wang C, Rubelt F, Scriba TJ, Davis MM. Analyzing the M. tuberculosis immune response by T cell receptor clustering with GLIPH2 and genome-wide antigen screening. Nat Biotechnol. 2020;38(10):1194–202.

208. Jensen KK, Rantos V, Jappe EC, Olsen TH, Jespersen MC, Jurtz V, et al. TCRpMHCmodels: Structural modelling of TCR-pMHC class I complexes. Sci Rep-uk. 2019;9(1):14530.

209. Liu I-H, Lo Y-S, Yang J-M. Genome-wide structural modelling of TCR-pMHC interactions. Bmc Genomics. 2013;14(Suppl 5):S5–S5.

210. Montemurro A, Schuster V, Povlsen HR, Bentzen AK, Jurtz V, Chronister WD, et al. NetTCR-2.0 enables accurate prediction of TCR-peptide binding by using paired TCRα and β sequence data. Commun Biology. 2021;4(1):1060.

211. Lin X, George JT, Schafer NP, Chau KN, Birnbaum ME, Clementi C, et al. Rapid assessment of T-cell receptor specificity of the immune repertoire. Nat Comput Sci. 2021;1(5):362–73.

212. Tong Y, Wang J, Zheng T, Zhang X, Xiao X, Zhu X, et al. SETE: Sequence-based Ensemble learning approach for TCR Epitope binding prediction. Comput Biol Chem. 2020;87:107281.

213. Fischer DS, Wu Y, Schubert B, Theis FJ. Predicting antigen specificity of single T cells based on TCR CDR3 regions. Mol Syst Biol. 2020;16(8):e9416.

214. Plante JA, Liu Y, Liu J, Xia H, Johnson BA, Lokugamage KG, et al. Spike mutation D614G alters SARS-CoV-2 fitness and neutralization susceptibility. Biorxiv. 2020;2020.09.01.278689.

215. Groves DC, Rowland-Jones SL, Angyal A. The D614G mutations in the SARS-CoV-2 spike protein: Implications for viral infectivity, disease severity and vaccine design. Biochem Bioph Res Co. 2020;

216. Mercatelli D, Giorgi FM. Geographic and Genomic Distribution of SARS-CoV-2 Mutations. Front Microbiol. 2020;11:1800.

217. Isabel S, Graña-Miraglia L, Gutierrez JM, Bundalovic-Torma C, Groves HE, Isabel MR, et al. Evolutionary and structural analyses of SARS-CoV-2 D614G spike protein mutation now documented worldwide. Sci Rep-uk. 2020;10(1):14031.

218. Yurkovetskiy L, Wang X, Pascal KE, Tomkins-Tinch C, Nyalile T, Wang Y, et al. Structural and Functional Analysis of the D614G SARS-CoV-2 Spike Protein Variant. Cell. 2020;183(3):739-751.e8.

219. Fernández A. Structural Impact of Mutation D614G in SARS-CoV-2 Spike Protein: Enhanced Infectivity and Therapeutic Opportunity. Acs Med Chem Lett. 2020;11(9):1667–70.

220. Pircher H, Moskophidis D, Rohrer U, Bürki K, Zinkernagel HH& RM. Viral escape by selection of cytotoxic T cell-resistant virus variants in vivo. Nature. 1990;

221. Phillips RE, Rowland-Jones S, Nixon DF, Gotch FM, Edwards JP, Ogunlesi AO, et al. Human immunodeficiency virus genetic variation that can escape cytotoxic T cell recognition. Nature. 1991;

222. Draenert R, Gall SL, Pfafferott KJ, Leslie AJ, Chetty P, Brander C, et al. Immune Selection for Altered Antigen Processing Leads to Cytotoxic T Lymphocyte Escape in Chronic HIV-1 Infection. J Exp Medicine. 2004;199(7):905–15.

223. Yokomaku Y, Miura H, Tomiyama H, Kawana-Tachikawa A, Takiguchi M, Kojima A, et al. Impaired Processing and Presentation of Cytotoxic-T-Lymphocyte (CTL) Epitopes Are Major

Escape Mechanisms from CTL Immune Pressure in Human Immunodeficiency Virus Type 1 Infection. J Virol. 2004;78(3):1324–32.

224. Kim V, Green WR. The Role of Proximal and Distal Sequence Variations in the Presentation of an Immunodominant CTL Epitope Encoded by the Ecotropic AK7 MuLV. Virology. 1997;236(2):221–33.

225. A Mutation in the HLA-B 2705-Restricted NP Epitope Affects the Human Influenza A Virus-Specific Cytotoxic T-Lymphocyte Response In Vitro.

226. Carlson JM, Brumme CJ, Martin E, Listgarten J, Brockman MA, Le AQ, et al. Correlates of Protective Cellular Immunity Revealed by Analysis of Population-Level Immune Escape Pathways in HIV-1. J Virol. 2012;86(24):13202–16.

227. Bronke C, Almeida C-AM, McKinnon E, Roberts SG, Keane NM, Chopra A, et al. HIV escape mutations occur preferentially at HLA-binding sites of CD8 T-cell epitopes. Aids. 2013;27(6):899–905.

228. Sun X, Shi Y, Akahoshi T, Fujiwara M, Gatanaga H, Schönbach C, et al. Effects of a Single Escape Mutation on T Cell and HIV-1 Co-adaptation. Cell Reports. 2016;15(10):2279–91.

229. Kawashima Y, Pfafferott K, Frater J, Matthews P, Payne R, Addo M, et al. Adaptation of HIV-1 to human leukocyte antigen class I. Nature. 2009;458(7238):641–5.

230. Ammaranond P, Zaunders J, Satchell C, Bockel DV, Cooper DA, Kelleher AD. A New Variant Cytotoxic T Lymphocyte Escape Mutation in HLA-B27-Positive Individuals Infected with HIV Type 1. 2005;21(5):395–7.

231. Murakoshi H, Koyanagi M, Akahoshi T, Chikata T, Kuse N, Gatanaga H, et al. Impact of a single HLA-A*24:02-associated escape mutation on the detrimental effect of HLA-B*35:01 in HIV-1 control. Ebiomedicine. 2018;36:103–12.

232. 1 PJG, Phillips RE, Colbert RA, McAdam S, Ogg G, Nowak MA, et al. Late escape from an immunodominant cytotoxic T-lymphocyte response associated with progression to AIDS. Nature Medicine. 1997;

233. Kelleher AD, Long * Chad, Holmes ‡ Edward C., Allen § Rachel L., Wilson * Jamie, Conlon * Christopher, et al. Clustered Mutations in HIV-1 gag Are Consistently Required for Escape from HLA-B27–restricted Cytotoxic T Lymphocyte Responses. J Exp Med. 2001;

234. Schneidewind A, Brockman MA, Sidney J, Wang YE, Chen H, Suscovich TJ, et al. Structural and Functional Constraints Limit Options for Cytotoxic T-Lymphocyte Escape in the Immunodominant HLA-B27-Restricted Epitope in Human Immunodeficiency Virus Type 1 Capsid ▿ . J Virol. 2008;82(11):5594–605.

235. Cale EM, Bazick HS, Rianprakaisang TA, Alam SM, Letvin NL. Mutations in a Dominant Nef Epitope of Simian Immunodeficiency Virus Diminish TCR:Epitope Peptide Affinity but not Epitope Peptide:MHC Class I Binding. J Immunol. 2011;187(6):3300–13.

236. Theodossis A, Guillonneau C, Welland A, Ely LK, Clements CS, Williamson NA, et al. Constraints within major histocompatibility complex class I restricted peptides: Presentation and consequences for T-cell recognition. Proc National Acad Sci. 2010;107(12):5534–9.

237. Iglesias MC, Almeida JR, Fastenackels S, Bockel DJ van, Hashimoto M, Venturi V, et al. Escape from highly effective public CD8+ T-cell clonotypes by HIV. Blood. 2011;118(8):2138–49.

238. Schulien I, Kemming J, Oberhardt V, Wild K, Seidel LM, Killmer S, et al. Characterization of pre-existing and induced SARS-CoV-2-specific CD8+ T cells. Nat Med. 2020;1–8.

239. Tarke A, Sidney J, Kidd CK, Dan JM, Ramirez SI, Yu ED, et al. Comprehensive analysis of T cell immunodominance and immunoprevalence of SARS-CoV-2 epitopes in COVID-19 cases. Cell Reports Medicine. 2021;2(2):100204.

240. Nelde A, Bilich T, Heitmann JS, Maringer Y, Salih HR, Roerden M, et al. SARS-CoV-2-derived peptides define heterologous and COVID-19-induced T cell recognition. Nat Immunol. 2020;1–12.

241. Kared H, Redd AD, Bloch EM, Bonny TS, Sumatoh HR, Kairi F, et al. SARS-CoV-2-specific CD8+ T cell responses in convalescent COVID-19 individuals. J Clin Invest. 2021;131(5).

242. Quadeer AA, Ahmed SF, McKay MR. Landscape of epitopes targeted by T cells in 852 individuals recovered from COVID-19: Meta- analysis, immunoprevalence, and web platform. Cell Reports Medicine. 2021;

243. Carlson JM, Le AQ, Shahid A, Brumme ZL. HIV-1 adaptation to HLA: a window into virus–host immune interactions. Trends Microbiol. 2015;23(4):212–24.

244. Callaway E. The race for coronavirus vaccines: a graphical guide. Nature. 2020;580(7805):576–7.

245. Krammer F. SARS-CoV-2 vaccines in development. Nature. 2020;586(7830):516–27.

246. Stephens DS, McElrath MJ. COVID-19 and the Path to Immunity. Jama. 2020;324(13):1279–81.

247. Sette A, Crotty S. Adaptive immunity to SARS-CoV-2 and COVID-19. Cell. 2021;

248. Altmann DM, Boyton RJ. SARS-CoV-2 T cell immunity: Specificity, function, durability, and role in protection. Sci Immunol. 2020;5(49):eabd6160.

249. Braun J, Loyal L, Frentsch M, Wendisch D, Georg P, Kurth F, et al. SARS-CoV-2-reactive T cells in healthy donors and patients with COVID-19. Nature. 2020;1–8.

250. Grifoni A, Weiskopf D, Ramirez SI, Mateus J, Dan JM, Moderbacher CR, et al. Targets of T cell responses to SARS-CoV-2 coronavirus in humans with COVID-19 disease and unexposed individuals. Cell. 2020;

251. Bert NL, Tan AT, Kunasegaran K, Tham CYL, Hafezi M, Chia A, et al. SARS-CoV-2-specific T cell immunity in cases of COVID-19 and SARS, and uninfected controls. Nature. 2020;584(7821):457–62.

252. Meckiff BJ, Ramírez-Suástegui C, Fajardo V, Chee SJ, Kusnadi A, Simon H, et al. Imbalance of regulatory and cytotoxic SARS-CoV-2-reactive CD4+ T cells in COVID-19. Cell. 2020;

253. Moderbacher CR, Ramirez SI, Dan JM, Grifoni A, Hastie KM, Weiskopf D, et al. Antigen-specific adaptive immunity to SARS-CoV-2 in acute COVID-19 and associations with age and disease severity. Cell. 2020;

254. Sekine T, Perez-Potti A, Rivera-Ballesteros O, Strålin K, Gorin J-B, Olsson A, et al. Robust T cell immunity in convalescent individuals with asymptomatic or mild COVID-19. Cell. 2020;

255. Weiskopf D, Schmitz KS, Raadsen MP, Grifoni A, Okba NMA, Endeman H, et al. Phenotype and kinetics of SARS-CoV-2-specific T cells in COVID-19 patients with acute respiratory distress syndrome. Sci Immunol. 2020;5(48):eabd2071.

256. Long Q-X, Liu B-Z, Deng H-J, Wu G-C, Deng K, Chen Y-K, et al. Antibody responses to SARS-CoV-2 in patients with COVID-19. Nat Med. 2020;26(6):845–8.

257. Long Q-X, Tang X-J, Shi Q-L, Li Q, Deng H-J, Yuan J, et al. Clinical and immunological assessment of asymptomatic SARS-CoV-2 infections. Nat Med. 2020;26(8):1200–4.

258. Dan JM, Mateus J, Kato Y, Hastie KM, Yu ED, Faliti CE, et al. Immunological memory to SARS-CoV-2 assessed for up to 8 months after infection. Science. 2021;eabf4063.

259. Seow J, Graham C, Merrick B, Acors S, Pickering S, Steel KJA, et al. Longitudinal observation and decline of neutralizing antibody responses in the three months following SARS-CoV-2 infection in humans. Nat Microbiol. 2020;5(12):1598–607.

260. Schub D, Klemis V, Schneitler S, Mihm J, Lepper PM, Wilkens H, et al. High levels of SARS-CoV-2 specific T-cells with restricted functionality in severe course of COVID-19. Jci Insight. 2020;5(20).

261. Zhou R, To KK-W, Wong Y-C, Liu L, Zhou B, Li X, et al. Acute SARS-CoV-2 infection impairs dendritic cell and T cell responses. Immunity. 2020;

262. Liao M, Liu Y, Yuan J, Wen Y, Xu G, Zhao J, et al. Single-cell landscape of bronchoalveolar immune cells in patients with COVID-19. Nat Med. 2020;26(6):842–4.

263. Peng Y, Mentzer AJ, Liu G, Yao X, Yin Z, Dong D, et al. Broad and strong memory CD4+ and CD8+ T cells induced by SARS-CoV-2 in UK convalescent individuals following COVID-19. Nat Immunol. 2020;1–10.

264. Ng O-W, Chia A, Tan AT, Jadi RS, Leong HN, Bertoletti A, et al. Memory T cell responses targeting the SARS coronavirus persist up to 11 years post-infection. Vaccine. 2016;34(17):2008–14.

265. Wu L-P, Wang N-C, Chang Y-H, Tian X-Y, Na D-Y, Zhang L-Y, et al. Duration of Antibody Responses after Severe Acute Respiratory Syndrome - Volume 13, Number 10—October 2007 - Emerging Infectious Diseases journal - CDC. Emerg Infect Dis. 2007;13(10):1562–4.

266. Tang F, Quan Y, Xin Z-T, Wrammert J, Ma M-J, Lv H, et al. Lack of Peripheral Memory B Cell Responses in Recovered Patients with Severe Acute Respiratory Syndrome: A Six-Year Follow-Up Study. J Immunol. 2011;186(12):7264–8.

267. Liu W, Fontanet A, Zhang P-H, Zhan L, Xin Z-T, Baril L, et al. Two-Year Prospective Study of the Humoral Immune Response of Patients with Severe Acute Respiratory Syndrome. J Infect Dis. 2006;193(6):792–5.

268. Liu G, Carter B, Bricken T, Jain S, Viard M, Carrington M, et al. Computationally Optimized SARS-CoV-2 MHC Class I and II Vaccine Formulations Predicted to Target Human Haplotype Distributions. Cell Syst. 2020;

269. Grifoni A, Sidney J, Zhang Y, Scheuermann RH, Peters B, Sette A. A Sequence Homology and Bioinformatic Approach Can Predict Candidate Targets for Immune Responses to SARS-CoV-2. Cell Host Microbe. 2020;

270. Tarke A, Sidney J, Kidd CK, Dan JM, Ramirez SI, Yu ED, et al. Comprehensive analysis of T cell immunodominance and immunoprevalence of SARS-CoV-2 epitopes in COVID-19 cases. Cell Reports Medicine. 2021;100204.

271. Quadeer AA, Ahmed SF, McKay MR. Landscape of epitopes targeted by T cells in 852 convalescent COVID-19 patients: Meta-analysis, immunoprevalence and web platform. Cell Reports Medicine. 2021;2(6):100312.

272. Popa A, Genger J-W, Nicholson MD, Penz T, Schmid D, Aberle SW, et al. Genomic epidemiology of superspreading events in Austria reveals mutational dynamics and transmission properties of SARS-CoV-2. Sci Transl Med. 2020;eabe2555.

273. Dorp L van, Richard D, Tan CCS, Shaw LP, Acman M, Balloux F. No evidence for increased transmissibility from recurrent mutations in SARS-CoV-2. Nat Commun. 2020;11(1):5986.

274. Korber B, Fischer WM, Gnanakaran S, Yoon H, Theiler J, Abfalterer W, et al. Tracking changes in SARS-CoV-2 Spike: evidence that D614G increases infectivity of the COVID-19 virus. Cell. 2020;

275. Laamarti M, Alouane T, Kartti S, Chemao-Elfihri MW, Hakmi M, Essabbar A, et al. Large scale genomic analysis of 3067 SARS-CoV-2 genomes reveals a clonal geo-distribution and a rich genetic variations of hotspots mutations. Plos One. 2020;15(11):e0240345.

276. Mercatelli D, Triboli L, Fornasari E, Ray F, Giorgi FM. coronapp: A Web Application to Annotate and Monitor SARS-CoV-2 Mutations. J Med Virol. 2020;

277. Mercatelli D, Giorgi FM. Geographic and Genomic Distribution of SARS-CoV-2 Mutations. Front Microbiol. 2020;11:1800.

278. Tarke A, Sidney J, Methot N, Yu ED, Zhang Y, Dan JM, et al. Impact of SARS-CoV-2 variants on the total CD4+ and CD8+ T cell reactivity in infected or vaccinated individuals. Cell Reports Medicine. 2021;100355.

279. Kosuge M, Furusawa-Nishii E, Ito K, Saito Y, Ogasawara K. Point mutation bias in SARS-CoV-2 variants results in increased ability to stimulate inflammatory responses. Sci Rep-uk. 2020;10(1):17766.

280. Giorgio SD, Martignano F, Torcia MG, Mattiuz G, Conticello SG. Evidence for host-dependent RNA editing in the transcriptome of SARS-CoV-2. Sci Adv. 2020;6(25):eabb5813.

281. Rice AM, Morales AC, Ho AT, Mordstein C, Mühlhausen S, Watson S, et al. Evidence for strong mutation bias towards, and selection against, U content in SARS-CoV-2: implications for vaccine design. Mol Biol Evol. 2020;38(1):msaa188-.

282. Simmonds P. Rampant C→U Hypermutation in the Genomes of SARS-CoV-2 and Other Coronaviruses: Causes and Consequences for Their Short- and Long-Term Evolutionary Trajectories. Msphere. 2020;5(3):e00408-20.

283. Klimczak LJ, Randall TA, Saini N, Li J-L, Gordenin DA. Similarity between mutation spectra in hypermutated genomes of rubella virus and in SARS-CoV-2 genomes accumulated during the COVID-19 pandemic. Plos One. 2020;15(10):e0237689.

284. Li Y, Yang X, Wang N, Wang H, Yin B, Yang X, et al. Mutation profile of over 4500 SARS-CoV-2 isolations reveals prevalent cytosine-to-uridine deamination on viral RNAs. Future Microbiol. 2020;15(14):1343–52.

285. Wang R, Hozumi Y, Zheng Y-H, Yin C, Wei G-W. Host Immune Response Driving SARS-CoV-2 Evolution. Viruses. 2020;12(10):1095.

286. Matyášek R, Kovařík A. Mutation Patterns of Human SARS-CoV-2 and Bat RaTG13 Coronavirus Genomes Are Strongly Biased Towards C>U Transitions, Indicating Rapid Evolution in Their Hosts. Genes-basel. 2020;11(7):761.

287. Monajemi M, Woodworth CF, Zipperlen K, Gallant M, Grant MD, Larijani M. Positioning of APOBEC3G/F Mutational Hotspots in the Human Immunodeficiency Virus Genome Favors Reduced Recognition by CD8+ T Cells. Plos One. 2014;9(4):e93428.

288. Grant M, Larijani M. Evasion of adaptive immunity by HIV through the action of host APOBEC3G/F enzymes. Aids Res Ther. 2017;14(1):44.

289. Ferretti AP, Kula T, Wang Y, Nguyen DMV, Weinheimer A, Dunlap GS, et al. Unbiased Screens Show CD8+ T Cells of COVID-19 Patients Recognize Shared Epitopes in SARS-CoV-2 that Largely Reside outside the Spike Protein. Immunity. 2020;53(5):1095-1107.e3.

290. Schulien I, Kemming J, Oberhardt V, Wild K, Seidel LM, Killmer S, et al. Characterization of pre-existing and induced SARS-CoV-2-specific CD8+ T cells. Nat Med. 2021;27(1):78–85.

291. Kared H, Redd AD, Bloch EM, Bonny TS, Sumatoh HR, Kairi F, et al. SARS-CoV-2-specific CD8+ T cell responses in convalescent COVID-19 individuals. J Clin Invest. 2021;131(5).

292. Saini SK, Hersby DS, Tamhane T, Povlsen HR, Hernandez SPA, Nielsen M, et al. SARS-CoV-2 genome-wide T cell epitope mapping reveals immunodominance and substantial CD8+ T cell activation in COVID-19 patients. Sci Immunol. 2021;6(58):eabf7550.

293. Agerer B, Koblischke M, Gudipati V, Montaño-Gutierrez LF, Smyth M, Popa A, et al. SARS-CoV-2 mutations in MHC-I-restricted epitopes evade CD8+ T cell responses. Sci Immunol. 2021;6(57):eabg6461.

294. Jariani A, Warth C, Deforche K, Libin P, Drummond AJ, Rambaut A, et al. SANTA-SIM: simulating viral sequence evolution dynamics under selection and recombination. Virus Evol. 2019;5(1):vez003.

295. Falk K, Rötzschke O, Stevanovic S, Jung G, Rammensee H-G. Allele-specific motifs revealed by sequencing of self-peptides eluted from MHC molecules. Nature. 1991 Mar 23;351(6324):290 296.

296. Gfeller D, Bassani-Sternberg M. Predicting Antigen Presentation-What Could We Learn From a Million Peptides? Front Immunol. 2018;9:1716.

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

298. Greenbaum J, Sidney J, Chung J, Brander C, Peters B, Sette A. Functional classification of class II human leukocyte antigen (HLA) molecules reveals seven different supertypes and a surprising degree of repertoire sharing across supertypes. Immunogenetics. 2011 Jun;63(6):325 335.

299. Reynisson B, Alvarez B, Paul S, Peters B, Nielsen M. NetMHCpan-4.1 and NetMHCIIpan-4.0: improved predictions of MHC antigen presentation by concurrent motif deconvolution and integration of MS MHC eluted ligand data. Nucleic Acids Res. 2020;gkaa379-.

300. Dorp L van, Acman M, Richard D, Shaw LP, Ford CE, Ormond L, et al. Emergence of genomic diversity and recurrent mutations in SARS-CoV-2. Infect Genetics Evol. 2020;83:104351.

301. Brumme ZL, Brumme CJ, Heckerman D, Korber BT, Daniels M, Carlson J, et al. Evidence of Differential HLA Class I-Mediated Viral Evolution in Functional and Accessory/Regulatory Genes of HIV-1. Plos Pathog. 2007;3(7):e94.

302. Kawashima Y, Pfafferott K, Frater J, Matthews P, Payne R, Addo M, et al. Adaptation of HIV-1 to human leukocyte antigen class I. Nature. 2009;458(7238):641–5.

303. Woolthuis RG, Dorp CH van, Keşmir C, Boer RJ de, Boven M van. Long-term adaptation of the influenza A virus by escaping cytotoxic T-cell recognition. Sci Rep-uk. 2016;6(1):33334.

304. Francisco R dos S, Buhler S, Nunes JM, Bitarello BD, França GS, Meyer D, et al. HLA supertype variation across populations: new insights into the role of natural selection in the evolution of HLA-A and HLA-B polymorphisms. Immunogenetics. 2015;67(11–12):651–63.

305. Salter JD, Bennett RP, Smith HC. The APOBEC Protein Family: United by Structure, Divergent in Function. Trends Biochem Sci. 2016;41(7):578–94.

306. Olson ME, Harris RS, Harki DA. APOBEC Enzymes as Targets for Virus and Cancer Therapy. Cell Chem Biol. 2018;25(1):36–49.

307. Peretti A, Geoghegan EM, Pastrana DV, Smola S, Feld P, Sauter M, et al. Characterization of BK Polyomaviruses from Kidney Transplant Recipients Suggests a Role for APOBEC3 in Driving In-Host Virus Evolution. Cell Host Microbe. 2018;23(5):628-635.e7.

308. Cuevas JM, Geller R, Garijo R, López-Aldeguer J, Sanjuán R. Extremely High Mutation Rate of HIV-1 In Vivo. Plos Biol. 2015;13(9):e1002251.

309. Albin JS, Haché G, Hultquist JF, Brown WL, Harris RS. Long-Term Restriction by APOBEC3F Selects Human Immunodeficiency Virus Type 1 Variants with Restored Vif Function ▿ . J Virol. 2010;84(19):10209–19.

310. Sadler HA, Stenglein MD, Harris RS, Mansky LM. APOBEC3G Contributes to HIV-1 Variation through Sublethal Mutagenesis ▿ . J Virol. 2010;84(14):7396–404.

311. Haché G, Shindo K, Albin JS, Harris RS. Evolution of HIV-1 Isolates that Use a Novel Vif-Independent Mechanism to Resist Restriction by Human APOBEC3G. Curr Biol. 2008;18(11):819–24.

312. Wood N, Bhattacharya T, Keele BF, Giorgi E, Liu M, Gaschen B, et al. HIV Evolution in Early Infection: Selection Pressures, Patterns of Insertion and Deletion, and the Impact of APOBEC. Plos Pathog. 2009;5(5):e1000414.

313. Jern P, Russell RA, Pathak VK, Coffin JM. Likely Role of APOBEC3G-Mediated G-to-A Mutations in HIV-1 Evolution and Drug Resistance. Plos Pathog. 2009;5(4):e1000367.

314. Kim E-Y, Lorenzo-Redondo R, Little SJ, Chung Y-S, Phalora PK, Berry IM, et al. Human APOBEC3 Induced Mutation of Human Immunodeficiency Virus Type-1 Contributes to Adaptation and Evolution in Natural Infection. Plos Pathog. 2014;10(7):e1004281.

315. Squires KD, Monajemi M, Woodworth CF, Grant MD, Larijani M. Impact of APOBEC Mutations on CD8+ T Cell Recognition of HIV Epitopes Varies Depending on the Restricting HLA. Jaids J Acquir Immune Defic Syndromes. 2015;70(2):172–8.

316. Casartelli N, Guivel-Benhassine F, Bouziat R, Brandler S, Schwartz O, Moris A. The antiviral factor APOBEC3G improves CTL recognition of cultured HIV-infected T cells. J Exp Medicine. 2010;207(1):39–49.

317. Nersisyan S, Zhiyanov A, Shkurnikov M, Tonevitsky A. T-CoV: a comprehensive portal of HLA-peptide interactions affected by SARS-CoV-2 mutations. Nucleic Acids Res. 2021;gkab701-.

318. Pisanti S, Deelen J, Gallina AM, Caputo M, Citro M, Abate M, et al. Correlation of the two most frequent HLA haplotypes in the Italian population to the differential regional incidence of Covid-19. J Transl Med. 2020;18(1):352.

319. Naemi FMA, Al-adwani S, Al-khatabi H, Al-nazawi A. Association between the HLA genotype and the severity of COVID-19 infection among South Asians. J Med Virol. 2021;93(7):4430–7.

320. Tomita Y, Ikeda T, Sato R, Sakagami T. Association between HLA gene polymorphisms and mortality of COVID-19: An in silico analysis. Immun Inflamm Dis. 2020;8(4):684–94.

321. Geers D, Shamier MC, Bogers S, Hartog G den, Gommers L, Nieuwkoop NN, et al. SARS-CoV-2 variants of concern partially escape humoral but not T-cell responses in COVID-19 convalescent donors and vaccinees. Sci Immunol. 2021;6(59):eabj1750.

322. Motozono C, Toyoda M, Zahradnik J, Saito A, Nasser H, Tan TS, et al. SARS-CoV-2 spike L452R variant evades cellular immunity and increases infectivity. Cell Host Microbe. 2021;

323. Grifoni A, Sidney J, Vita R, Peters B, Crotty S, Weiskopf D, et al. SARS-CoV-2 Human T cell Epitopes: adaptive immune response against COVID-19. Cell Host Microbe. 2021;

324. Sidney J, Southwood S, Moore C, Oseroff C, Pinilla C, Grey HM, et al. Measurement of MHC/Peptide Interactions by Gel Filtration or Monoclonal Antibody Capture. Vol. 100. Curr Protoc Immunol; 2013.

325. Huddleston J, Barnes JR, Rowe T, Xu X, Kondor R, Wentworth DE, et al. Integrating genotypes and phenotypes improves long-term forecasts of seasonal influenza A/H3N2 evolution. Elife. 2020;9:e60067.

326. Tarke A, Sidney J, Methot N, Yu ED, Zhang Y, Dan JM, et al. Impact of SARS-CoV-2 variants on the total CD4+ and CD8+ T cell reactivity in infected or vaccinated individuals. Cell Reports Medicine. 2021;2(7):100355.

327. Ledford H. How 'killer' T cells could boost COVID immunity in face of new variants. Nature. 2021;590(7846):374–5.

328. Li Z-RT, Zarnitsyna VI, Lowen AC, Weissman D, Koelle K, Kohlmeier JE, et al. Why Are CD8 T Cell Epitopes of Human Influenza A Virus Conserved? J Virol. 2019;93(6):e01534-18.

329. Prevention C for DC and. Science Brief: Omicron (B.1.1.529) Variant [Internet]. 2021. Available from: https://www.cdc.gov/coronavirus/2019-ncov/science/science-briefs/scientific-brief-omicron-variant.html

330. Neidleman J, Luo X, McGregor M, Xie G, Murray V, Greene WC, et al. mRNA vaccine-induced T cells respond identically to SARS- CoV- 2 variants of concern but differ in longevity and homing properties depending on prior infection status. elife. 2021;

331. Goel RR, Painter MM, Apostolidis SA, Mathew D, Meng W, Rosenfeld AM, et al. mRNA vaccines induce durable immune memory to SARS-CoV-2 and variants of concern. Science. 2021;eabm0829.

332. Goel RR, Apostolidis SA, Painter MM, Mathew D, Pattekar A, Kuthuru O, et al. Distinct antibody and memory B cell responses in SARS-CoV-2 naïve and recovered individuals after mRNA vaccination. science immunology. 2021;

333. Mateus J, Dan JM, Zhang Z, Moderbacher CR, Lammers M, Goodwin B, et al. Low-dose mRNA-1273 COVID-19 vaccine generates durable memory enhanced by cross-reactive T cells. Science. 2021;374(6566):eabj9853.

334. Oberhardt V, Luxenburger H, Kemming J, Schulien I, Ciminski K, Giese S, et al. Rapid and stable mobilization of CD8+ T cells by SARS-CoV-2 mRNA vaccine. Nature. 2021;597(7875):268–73.

335. Painter MM, Mathew D, Goel RR, Apostolidis SA, Pattekar A, Kuthuru O, et al. Rapid induction of antigen-specific CD4+ T cells is associated with coordinated humoral and cellular immunity to SARS-CoV-2 mRNA vaccination. Immunity. 2021;54(9):2133-2142.e3.

336. Skelly DT, Harding AC, Gilbert-Jaramillo J, Knight ML, Longet S, Brown A, et al. Two doses of SARS-CoV-2 vaccination induce robust immune responses to emerging SARS-CoV-2 variants of concern. Nat Commun. 2021;12(1):5061.

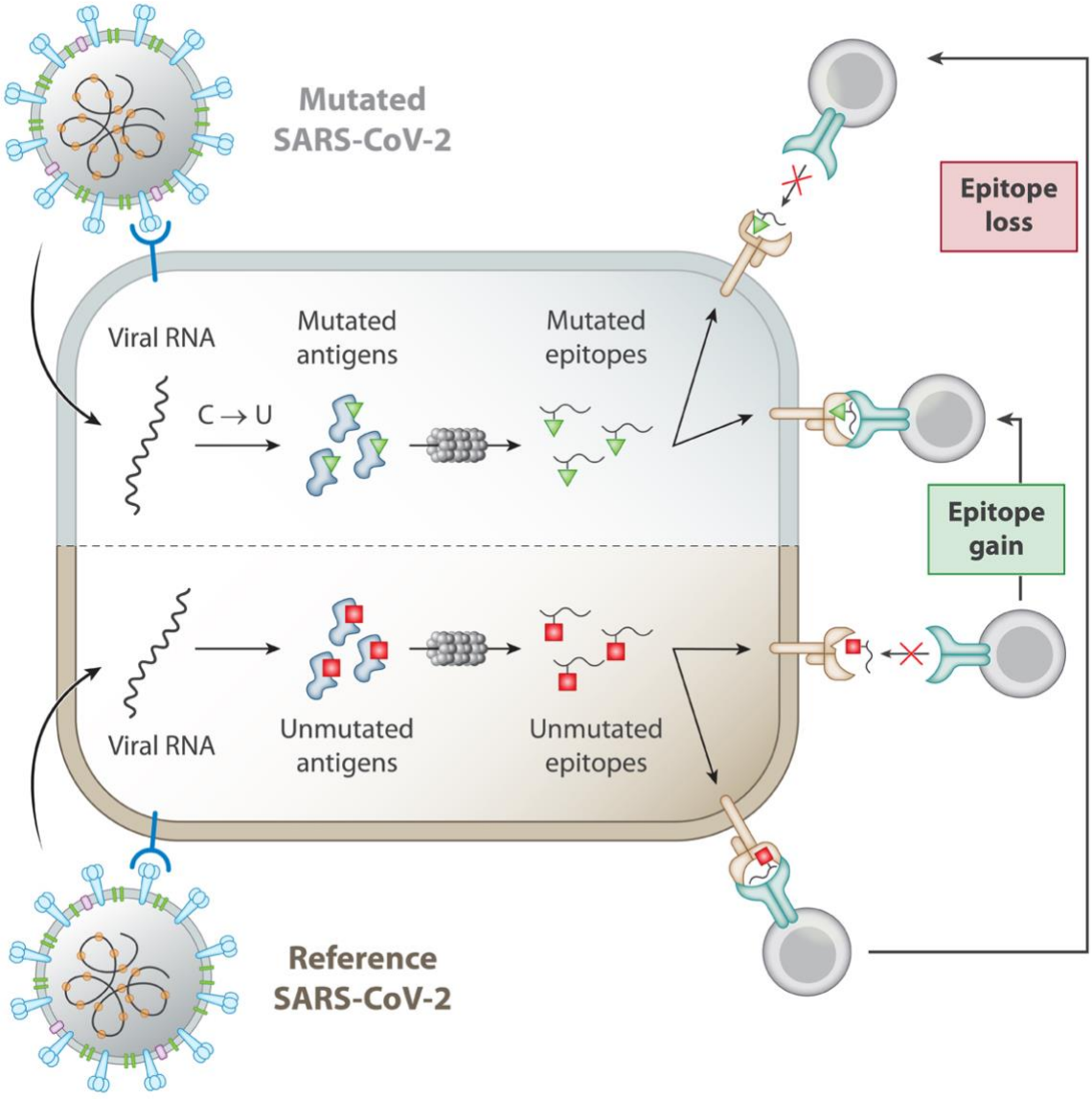
337. Voysey M, Clemens SAC, Madhi SA, Weckx LY, Folegatti PM, Aley PK, et al. Safety and efficacy of the ChAdOx1 nCoV-19 vaccine (AZD1222) against SARS-CoV-2: an interim analysis of four randomised controlled trials in Brazil, South Africa, and the UK. Lancet Lond Engl. 2021;397(10269):99–111.

338. Inc. N. Novavax COVID-19 Vaccine Demonstrates 89.3% Efficacy in UK Phase 3 Trial [Internet]. 2021. Available from: https://ir.novavax.com/2021-01-28-Novavax-COVID-19-Vaccine-Demonstrates-89-3-Efficacy-in-UK-Phase-3-Trial

339. Muñoz-Fontela C, Dowling WE, Funnell SGP, Gsell P-S, Riveros-Balta AX, Albrecht RA, et al. Animal models for COVID-19. Nature. 2020;586(7830):509–15.

340. Kim K, Calabrese P, Wang S, Qin C, Rao Y, Feng P, et al. APOBEC-mediated Editing of SARS-CoV-2 Genomic RNA Impacts Viral Replication and Fitness. Biorxiv. 2021;2021.12.18.473309.

# 6 Annexe




Graphical abstract illustrating the **loss** and **gain** of SARS-CoV-2 epitopes resulting from mutation events.