Université de Montréal

L’impact des mutations récurrentes du SARS-CoV-2 sur l’évasion immunitaire

Par
Dominique Fournelle

Département de Biochimie et Médecine Moléculaire, Faculté de Médecine

Mémoire présenté en vue de l’obtention du grade de M. Sc

en Bio-informatique

Août 2022

Université de Montréal
Département de Biochimie et Médecine Moléculaire, Faculté de Médecine

Ce mémoire intitulé

L'impact des mutations récurrentes du SARS-CoV-2 sur l'évasion immunitaire

Présenté par

Dominique Fournelle

A été évalué(e) par un jury composé des personnes suivantes

Vincent-Philipe Lavallée
Président-rapporteur

Julie Hussin
Directrice de recherche

Morgan Craig
Codirectrice

Simon Gravel
Membre du jury

Résumé

Nous sommes toujours aux prises avec la pandémie de SARS-CoV-2 plus de deux ans après son début. Le virus a depuis accumulé de nombreuses mutations qui ont mené à différentes souches virales au long de la pandémie. Plusieurs de ces mutations sont récurrentes: il y a un excès de mutations C > U dans les génomes viraux, certains codons sont fréquemment mutés vers différents acides aminés et certaines mutations sont convergentes, c'est-à-dire que la même substitution est apparue de manière indépendante sur différentes lignées. Dans ce mémoire, nous avons identifiés différentes manières par lesquelles le SARS-CoV-2 évolue à travers l'étude de ces mutations récurrentes et évaluons leur impact sur l'évasion immunitaire. Premièrement, nous avons déterminés que les mutations C > U sont responsable de l'introduction et du retrait préférentiel d'acides aminés spécifiques dans les épitopes viraux. Nous avons déterminé la significativité statistique de ces patrons de mutation à l'aide de simulations génomiques virales. Deuxièmement, nous avons participé à la surveillance des variants au Québec durant la deuxième vague de la pandémie, qui s'est déroulée d'août 2020 à mars 2021. C'était une période intéressante pour la diversité virale, puisque les restrictions de déplacement ont créé de multiples poches de variants locaux en compétition les uns avec les autres qui partagent des mutations convergentes. Notamment, nous reportons que les lignées B.1.160 et B.1.1.176 comptaient pour 50% des échantillons séquencés au sommet de la deuxième vague dans la province. Finalement, nous avons analysé les patrons mutationnels intra-hôte qui sont apparus *de novo* dans le contexte d'infections au SARS-CoV-2 de longue durée chez des patients atteints de cancers hématologiques. Une de ces patientes est une patiente québécoise infectée par B.1.160 et dans laquelle nous avons identifié la présence d'un réservoir viral. Nous avons également trouvé des éléments probants montrant différentes quasiespèces virales avec des propriétés d'évasion immunitaire. Nos résultats permettent de mieux comprendre les différentes manières dont les pressions sélectives façonnent l'évolution virale.

Mots-clés : SARS-CoV-2, COVID-19, évasion immunitaire, évolution virale, mutations convergentes.

Abstract

We are still living in the SARS-CoV-2 pandemic over two years after its start. The virus has since accumulated many mutations that have led to different viral strains throughout the pandemic. Several of these mutations are recurrent: there is an excess of C > U substitutions in viral genomes, some codons are frequently mutated to different amino acids, and some mutations are convergent, meaning that the same substitution has occurred independently on different lineages. In this thesis, we identified different ways in which SARS-CoV-2 evolves through these recurrent mutations and assess their impact on immune escape. First, we determined that C > U mutations drive the preferential introduction and removal of specific amino acids in viral epitopes. Using genetic simulations, we determined the statistical significance of these patterns. Second, we participated in the surveillance of variants in Quebec during the second wave of the pandemic that went from the end of August 2020 to the end of March 2021. This was an interesting period of viral diversity owing to imposed travel restrictions that created competition between multiple pockets of local strains that share convergent mutations. Notably, we found that lineages B.1.160 and B.1.1.176 account for 50% of samples sequenced at the height of the second wave in the province. Finally, we analyzed intra-host mutational patterns that arose de novo in the context of long-term infections of patients with hematological cancers, one of which was from Québec and infected by B.1.160. We have identified a pattern consistent with the presence of a viral reservoir in this patient. We have also found evidence of different viral quasispecies with immune escape properties. These results shed light on different ways in which selective pressures shape the evolution of SARS-CoV-2.

Keywords : SARS-CoV-2, COVID-19, immune escape, viral evolution, convergent mutations.

Table des matières

Liste des tableaux

Liste des figures

Liste des sigles et abréviations

SARS-CoV-1: *Severe Acute Respiratory Syndrome Coronavirus 1*

SARS-CoV-2: *Severe Acute Respiratory Syndrome Coronavirus 2*

COVID-19: Maladie à Coronavirus 2019

OMS: Organisation Mondiale de la Santé

CoV: Coronavirus

S: Protéine spicule d'un coronavirus

M: Protéine membranaire d'un coronavirus

E: Protéine enveloppe d'un coronavirus

N: Protéine nucléoplasmide d'un coronavirus

NSP: Protéine non structurale d'un coronavirus

β-CoV: Betacoronavirus

MERS: *Middle East Respiratory Syndrome*

WIV: Institut de Virologie de Wuhan

ORF: *Open Reading Frame*

IDR: *Intrinsically disordered region*

RBD: *Receptor Binding Domain*

ACE2: *Angiotensin converting enzyme 2*

TMPRSS2: Protéase transmembranaire à sérine 2

NPR1: Neuropiline 1

MIS-C: Syndrome inflammatoire multisystémique de l'enfant

AVC: Accident vasculaire cérébrale

GWAS: *Genome Wide Association Studies*

OAS: enzymes oligoadénylate synthétases

VOC: *Variant of Concern*

VOI: *Variant of Interest*

GRSO: Global residue substitution output

S-S: SANTA-SIM

SNP: *Single nucleotide polymorphism*

SNV: *Single nucleotide variant*

iSNV: *Intra-host single nucleotide variant*

TLR: Toll-like receptor

INFs: Interférons

CMH: Complexe majeur d'histocompatibilité

ADAR: *Adenosine deaminase acting on RNA*

APOBEC: *Apolipoprotein B editing complex*

ARN: Acide ribonucléique

ARNm: ARN messager

ADN: Acide désoxyribonucléique

IgE: Immunoglobulines E

NGS: *Next-generation sequencing*

TGS: *Third-generation sequencing*

PCR: *Polymerase chain reaction*

GISAID: Global Initiative on Sharing All Influenza Data

LSPQ: Laboratoire de santé publique du Québec

EDO: Équation différentielle ordinaire

IG: Intergénique

Ts/Tv: Ratio de transitions versus transversions

DPP9 : Dipeptidyl peptidase 9

GTR: *General Time Reversible*

mAbs: Anticorps monoclonaux

Remerciements

Je voudrais commencer par remercier Julie Hussin de m'avoir pris sous son aile dès ma première année de Baccalauréat et de m'avoir encouragé tout au long de mon parcours. Je suis reconnaissante pour la confiance qu'elle m'a accordé dans les divers projets auxquels j'ai pris part au sein de son laboratoire. Je remercie également Morgan Craig pour son support et ses précieux conseils.

Un gros merci à toute l'équipe du MHI-OMICs qui m'a accueilli à bras ouverts. Faire ma maîtrise en télétravail n'aurait pas été aussi agréable sans tous les fous rires sur Slack que j'ai partagé avec vous. Ce fut vraiment un honneur de faire avancer la science à vos côtés. Un merci tout particulier à Jean-Christophe Grenier, l'homme qui murmure à l'oreille des serveurs pour toute son aide. Merci également à Raphael Poujol à qui j'ai probablement fait générer au moins 150 RaphGraphs. Merci à Isabel Gamache, Cantin Baron, Camille Rochefort-Boulanger et Fatima Mostefai pour leur aide à la correction de ce mémoire.

Finalement, merci à ma famille et à mes amis pour leur soutien tout au long de cette aventure.

Chapitre 1 – Introduction

1.1 Mise en contexte

Le virus *Severe Acute Respiratory Syndrome coronavirus 2* (SARS-CoV-2), responsable de la maladie à coronavirus 2019 (COVID-19), a été rapporté pour la première fois en décembre 2019 à Wuhan, une ville de la province de Hubei en Chine (1). Le virus s'est rapidement propagé autour du globe au début de l'année 2020 et l'Organisation Mondiale de la Santé (OMS) a déclaré que cette épidémie avait atteint le stade de pandémie le 11 mars 2020 (2). En date du 31 août 2022, plus de 602 millions de cas ont été rapportés dans le monde, ainsi que plus de 6.4 millions de décès (3).

Presque trois ans après la découverte du premier cas, plusieurs questions d'importance clinique se posent encore sur le SARS-CoV-2, notamment au sujet de son évolution dans le contexte d'infections prolongées chez des patients immunosupprimés ou au sujet du syndrome post-COVID-19, aussi appelé COVID longue (4).

Ce mémoire contient quatre chapitres. Ce premier chapitre contient une revue extensive de la littérature sur le SARS-CoV-2, sur la COVID-19, ainsi que sur les différentes techniques d'analyses utilisées pour étudier l'évolution virale. Il présente également mon hypothèse et mes objectifs de recherche. Le deuxième chapitre présente mes analyses réalisées dans le cadre de deux projets distincts, soient l'utilisation de simulations génomiques virales pour étudier des patrons de mutations trouvées dans les données réelles et la caractérisation des variants présents au Québec lors de la deuxième vague. Le troisième chapitre présente un manuscrit qui contient une étude de cas réalisée sur des patients atteint de cancers hématologiques qui ont eu une infection au SARS-CoV-2 de longue durée. Finalement, au chapitre 4, j'interprète les résultats obtenus et conclus ce mémoire.

1.2 Revue de littérature

1.2.1 SARS-CoV-2

1.2.1.1 Les coronavirus

Les coronavirus (CoV) sont des virus à acide ribonucléique (ARN) à simple brins positifs qui infectent les mammifères et les oiseaux (5). Cette famille de virus utilise, entre autres, les chauves-

souris comme hôte réservoir, c'est-à-dire que l'espèce peut être atteinte du virus sans développer de symptômes, favorisant ainsi l'évolution virale (6). Ces virus peuvent ensuite être transmis aux humains de manière zoonotique, c'est-à-dire à partir d'un animal. Le génome de ce virus encode typiquement quatre protéines structurales, la protéine de Spicule (S), la protéine Membranaire (M), la protéine d'Enveloppe (E) et le Nucléoplasmide (N), ainsi qu'un nombre variable de protéines accessoires et non structurales (*nonstructural proteins*, NSP) (7). Il y a présentement sept types de CoV pouvant infecter l'humain, quatre d'entre eux causent des symptômes bénins de rhume, et trois peuvent causer une maladie grave, voire fatale : le SARS-CoV-1, le MERS et le SARS-CoV-2 (8).

1.2.1.1.1 SARS-CoV-1

Le premier virus documenté causant un syndrome respiratoire aigu sévère chez l'humain (*Severe Acute Respiratory Syndrome,* SARS-CoV-1) est un autre CoV détecté pour la première fois en 2002 (9), le SARS-CoV-1. Tout comme SARS-CoV-2, il appartient à la classe des Betacoronavirus (β-CoV) (7). Ce virus cause une maladie sévère dont les symptômes incluent de la fièvre, des symptômes grippaux ainsi que des symptômes gastro-intestinaux (10) et le taux de mortalité était de 9.6% (11). Selon l'OMS, 8096 cas ont été rapportés dans 26 pays et le dernier cas a été rapporté aux États-Unis en juillet 2003 (12).

1.2.1.1.2 MERS-CoV

Le *Middle East Respiratory Syndrome related coronavirus* (MERS-CoV) est un autre β-CoV infectant l'humain et dont le premier cas a été détecté en 2012 (13). L'épidémie de MERS-CoV n'est pas terminée puisqu'il existe encore un risque de transmission de nos jours. Bien que peu de cas ont été répertoriés dans le monde, soit 2442 cas, ce virus a un taux de mortalité de 35.5% (14).Le nombre de patients infectés qui doivent être hospitalisés est de 57% (15), et 29% doivent être placés sous respirateur (16). Contrairement aux deux SARS, MERS-CoV est rarement transmis d'humain à humain. La majorité des cas répertoriés sont survenus suite à une transmission zoonotique du dromadaire à l'humain (17). Des mesures sanitaires visant spécifiquement les communautés en contact avec ces animaux, mais également les voyageurs qui entrent et sortent des zones touchées permettent de contenir l'éclosion et ainsi prévenir une pandémie potentielle (14).

1.2.1.2 Les origines du SARS-CoV-2

Malgré de nombreux efforts de recherche, les origines exactes du SARS-CoV-2 sont encore

inconnues. Les deux hypothèses les plus courantes au début de la pandémie étaient une erreur de manipulation à l'Institut de Virologie de Wuhan (WIV), qui étudie entre autres les coronavirus, ainsi qu'une éclosion au marché de Huanan dans lequel des animaux vivants sont vendus. Suite à une enquête de l'OMS, l'hypothèse de la transmission zoonotique au marché est la plus plausible (18).

La première infection est estimée aux alentours du 18 novembre 2019, le marché d'Huanan est géographiquement dans l'épicentre de la première éclosion et deux des trois premiers patients répertoriés s'y sont retrouvés (19,20). L'espèce de l'animal responsable du premier événement de transmission n'a pas encore été identifié.

1.2.1.3 L'organisation génomique du SARS-CoV-2

Le génome du SARS-CoV-2 est composé de 29,903 nucléotides, ce qui est dans la norme pour un CoV (19). Il est constitué d'Adénine (A) et d'Uracile (U) à 62.02% (21). Le premier gène, *ORF1ab,* fait plus des deux tiers de son génome avec 21,555 nucléotides. Ce gène est composé de deux cadres de lecture ouverts (*open reading frame*, ORF) qui se chevauchent et encodent 16 protéines non structurales. Les protéines provenant de l'*ORF1a* ont des fonctions d'expression virales et celles de l'*ORF1b* ont des fonctions de réplication (22). Le prochain gène est le *S*, composé de la sous-unité S1 qui contient le domaine de fixation au récepteur (*Receptor Binding Domain*, RBD) responsable de la liaison au récepteur sur la cellule de l'hôte et de la sous-unité S2 qui est responsable de la fusion membranaire et de l'entrée virale dans la cellule de l'hôte (23). La protéineE est une protéine transmembranaire dont la totalité des fonctions est méconnue. Elle peut notamment servir de canal à ions et a un rôle à jouer dans l'encapsidation et la réplication virale (24). La protéine M est responsable de l'intégrité structurale du virus (25). La protéine N joue un rôle dans l'encapsidation virale, dans la transcription et dans l'assemblage de virions, une forme extracellulaire du matériel génétique du virus (26). Elle est composée de régions intrinsèquement désordonnées ce qui lui confère une structure flexible lui permettant de former des liens avec plusieurs motifs ARN. Finalement, les protéines structurelles décrites plus haut sont entrecoupées de plusieurs autres ORFs qui se chevauchent et produisent des protéines non structurelles et accessoires aux fonctions multiples (27). Un schéma de l'organisation physiologique et génomiquedu virus est disponible à la Figure 1.1.

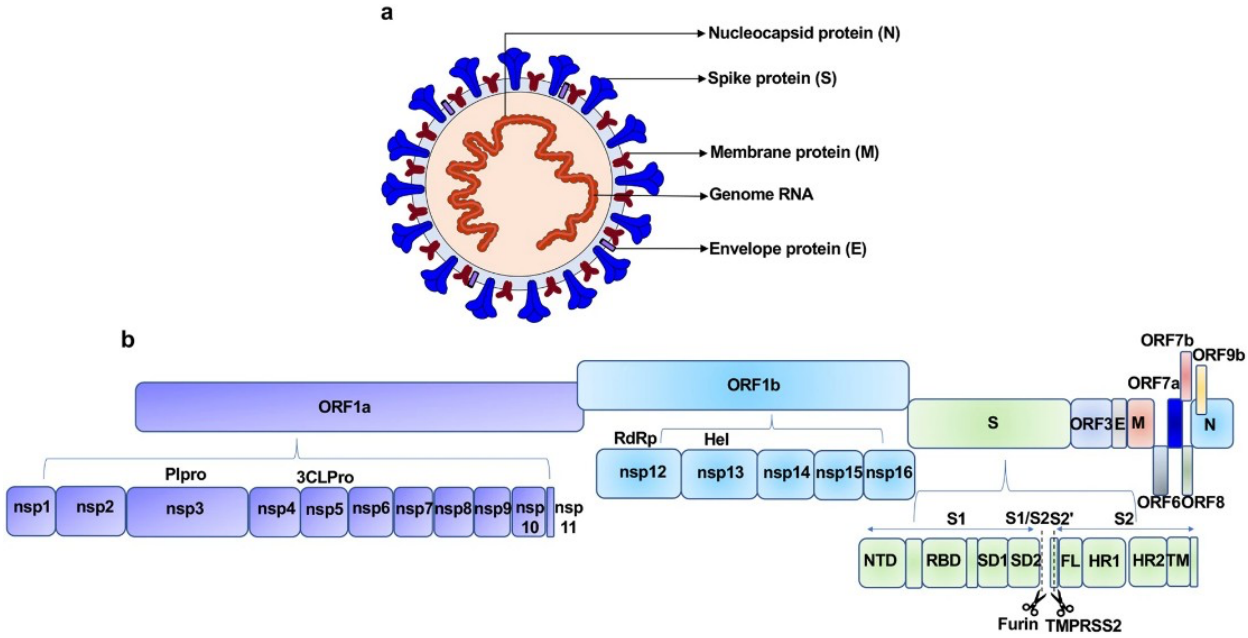


Figure 1.1 - Organisation génomique du SARS-CoV-2.
Image tirée de Zhang et al. (28).

1.2.1.4 Comparaison entre les génomes du SARS-CoV-2 et d'autres coronavirus

Une étude comparative de plusieurs génomes de coronavirus a démontré que les virus les plus proches du SARS-CoV-2 sont deux β-CoVs affectant les chauves-souris, nommés bat-SL-CoVZC45 et bat-SL-CoVZXC21, qui ont une homologie de séquence de 87.99% et 87.23% respectivement (29). Les auteurs de cette étude ont rapporté que les séquences codantes pour les protéines M, E et N sont identiques à plus de 90%, mais que l'homologie est plus basse pour la séquence codant pour la protéine S, qui est de 80%. Une analyse phylogénique de la protéine S a déterminé que celle-ci était plus apparentée à celle d'un autre β-CoV affectant les chauves-souris, soit RaTG13 avec laquelle elle partage 93.1% de sa séquence (30). En étudiant la séquence de la protéine S plus en détail, on remarque que la région de la sous-unité S1, chargée de se lier au récepteur de la cellule hôte, est identique à l'exception d'un nucléotide près à la région équivalente chez un coronavirus affectant le pangolin (pangolin-CoV). Ceci suggère un ou plusieurs événement(s) de recombinaison avec pangolin-CoV (6). Cependant, puisque que cette similitude est limitée à une petite fraction du génome et qu'il est possible que cette similitude soit une coïncidence due à l'adaptation du virus à un nouvel hôte, il est peu probable que le pangolin soit directement lié au premier évènement de transmission zoonotique (31).

1.2.2 Évolution génomique du SARS-CoV-2

1.2.2.1 Génétique des populations virales

La génétique des populations est la science qui étudie le changement dans la composition génétique d'une population. Ce changement, ou l'absence de changement, est communément appelé l'évolution et se mesure par la variation des fréquences alléliques au cours du temps. Puisque les virus sont des entités haploïdes, les théorèmes classiques basés sur la fréquence allélique d'échantillons diploïdes, comme le principe d'Hardy-Weinberg ne s'appliquent pas directement aux populations virales (32). Cela dit, d'autres concepts tels que la sélection, la dérive génétique, les variations de tailles de population (réductions et expansions) et le flux génétique sont pertinents pour l'étude des virus.

La sélection est l'ensemble des processus par lesquels un individu a plus de chance de transmettre ses gènes à une génération subséquente (33). Cette valeur adaptative est communément appelée *fitness*. Il existe plusieurs types de sélection. La sélection positive agit en augmentant la fréquence de mutation qui augmente les chances de transmission de matériel génétique (34). La sélection négative est l'inverse de la sélection positive: les parents porteurs de mutation délétères auront moins de chance de transmettre leur matériel génétique (33). Le dernier type de sélection est la sélection balancée qui maintient plusieurs allèles dans la population à des fréquences plus importantes qu'attendues. Dans le contexte de génomes haploïdes, ceci peut donner lieu à des sites avec trois allèles à haute fréquences (tri-alléliques) (35). Un exemple de sélection dans un contexte d'évolution virale serait une souche ayant développé une mutation résistante à un traitement anti- viral, et qui pourrait donc y survivre et continuer de proliférer (36). Il est important de noter que toute mutation ne confère pas nécessairement d'avantage ou de désavantage sélectif, il est attendu que la majorité des mutations soient neutres, c'est-à-dire qu'elles n'aient aucun effet sur l'aptitudede réplication ou de transmission d'un virus.

La dérive génétique est le processus par lequel les fréquences alléliques fluctuent de façon aléatoire d'une génération à l'autre. La dérive est intimement liée à la taille de la population et est forte lorsque sa taille de population est petite. Du coup, lors d'un goulot d'étranglement, caractérisé par une réduction drastique de la taille de la population (37,38), la dérive peut créer d'importants changements dans la composition génétique d'une population. Dans le contexte de population virales, les fluctuations de taille de populations lors de la transmission du virus d'un hôte à l'autre

et lors de l'introduction du virus dans une population naïve sont des facteurs importants de l'évolution.

Les goulots d'étranglement sont observés lors de la transmission des infections, parce que seulement une petite fraction des virus infectant un individu sera transmise à l'hôte suivant lors d'une nouvelle infection (39). La population virale dans le nouvel hôte connaîtra par la suite une croissance exponentielle, ce qui peut entraîner un enrichissement en allèles rares (40). Un autre exemple de dérive entraînée par un facteur démographique est l'effet fondateur. Dans le cas d'évolution virale, cela se produit entre autres lorsqu'un individu infecté se rend sur un nouveau territoire et y débute une nouvelle chaîne de transmission. Dans ce cas, l'ensemble de la diversité virale infectant les individus de la population originale n'est pas transmise à la nouvelle population. Analyser la diversité virale dans une nouvelle population permet donc d'estimer la provenance de la souche virale ainsi qu'une fourche de dates dans laquelle le premier événement de transmission a eu lieu (41).

Finalement, le flux génétique est l'échange de matériel génétique entre deux populations. Dans le cas des organismes qui se reproduisent de manière asexuée comme les virus, ceci prend souvent la forme d'un transfert de gène horizontal. Un exemple serait la méthode de réplication des rétrovirus qui leur permet d'injecter leur matériel génétique dans les cellules de son hôte (43). Un autre exemple serait la recombinaison, qui peut survenir lorsqu'au moins deux génomes viraux infectent une cellule. À travers divers mécanismes, ces génomes peuvent créer un nouveau génome hybride qui contient des fragments des génomes parents (42). Bien que disputé en début de pandémie, il y a aujourd'hui des preuves de recombinaison dans des génomes du SARS-CoV-2 (44,45).

Une substitution d'exactement un nucléotide au même emplacement dans le génome chez plusieurs individus d'une population est appelé un SNP (single nucleotide polymorphism) (46). Certains SNPs sont toujours retrouvés ensemble, car ils sont apparus sur la même souche virale, et créent un ensemble que l'on appelle haplotype. Les haplotypes peuvent ainsi servir à définir les différentes souches virales, ou variants (47). Lorsqu'on étudie l'effet potentiel d'un SNP sur le phénotype viral, il est important de le considérer dans le contexte de son haplotype parce que différentes combinaisons de polymorphismes peuvent avoir différents effets (48). On parle d'interaction génique lorsque le phénotype d'un mutant simple est différent du phénotype de ce mutant en combinaison à d'autres mutants (49).

1.2.2.1.1 Mutations récurrentes

Dans le cadre de l'évolution du SARS-CoV-2, on remarque que certains codons sur des lignées indépendantes mutent de manière récurrente. Un type de mutation récurrente est la mutation convergente, dans laquelle le même acide nucléique est muté, donnant lieu à une mutation synonyme ou à une mutation non-synonyme vers le même acide aminé (50). Plusieurs mutations convergentes ont été identifiées pour SARS-CoV-2 et certaines coévoluent ensembles dans ce que l'on pourrait appeler des haplotypes convergents (51). Il arrive également que différents variants aient le même codon muté vers plus d'un acide aminé différent (52,53). Ces mutations surviennent souvent dans le RBD de diverses protéines du SARS-CoV-2, suggérant une évolution soussélection.

1.2.2.1.2 Biais de transition

Les cinq acides nucléiques sont divisés en deux classes selon leur structure chimique: les purines et les pyrimidines. Les Adénines et les Guanines (G) sont des purines et les Cytosines (C), les Thymines (T) et les Uraciles sont des pyrimidines. Une substitution entre deux acides nucléiques d'une même famille est appelée une transition et une substitution entre deux acides aminés de différentes familles est appelée une transversion. Parce que la structure chimique est plus semblable entre les membres d'une même famille, il est attendu que les transitions soient plus fréquentes que les transversions dans un contexte d'évolution sans pression sélective causée par l'environnement (54). Ce phénomène s'appelle biais de transition (55) et est quantifié par le ratio de transition *versus* transversion (Ts/Tv) pour un organisme donné.

Ce biais peut changer selon l'organisme, puisqu'il ne dépend pas uniquement de la composition chimique des acides nucléiques. Il a été démontré que les transversions avaient un effet négatif sur la *fitness* de certains virus à ARN comme l'influenza et le virus d'immunodéficience humaine (VIH) (56). De plus, il semblerait que les transitions ne soient pas toutes équivalentes chez les virus à ARN. En effet, le taux de transition Cvers U (C > U) est plus élevé que les autres types de transitions chez plusieurs familles de virus (57). SARS-CoV-2 en particulier a un biais de transition C > U huit fois plus élevé que le biais de transition de U > C, et ce, malgré le fait que son génome contient plus de U que de C (58). De plus, il semble aussi y avoir un biais de substitution G > U plus élevé qu'attendu (59).

1.2.2.2 Phylogénie

La phylogénie est la science qui tente de reconstituer des événements évolutifs à l'aide de données de séquençage et dont les résultats sont représentés sous forme d'arbre, c'est-à-dire des branches qui se terminent par des feuilles qui représentent les diverses séquences. Un arbre phylogénétique peut être enraciné, avec l'ancêtre commun le plus récent à la racine, ou non enraciné. La phylogénie a joué un rôle clé dans l'étude de l'évolution virale, notamment pour déterminer la taxonomie de virus émergents, dont le SARS-CoV-2 (7,29). Diverses méthodes en phylogénies seront discutées à la section 1.2.5.3.1. On parle de phylodynamique lorsqu'on étudie l'ensemble des procédés épidémiologiques, immunologiques et évolutifs par lesquels la phylogénie virale est influencée (60).

1.2.2.3 Nomenclature des variants du SARS-CoV-2

Il est devenu évident depuis la première apparition de variants du SARS-CoV-2 qu'il était nécessaire d'avoir un système de nomenclature pour définir les nouveaux variants. Plusieurs systèmes ont été proposés, dont les clades de l'initiative *Nextstrain*, un projet *open-source* pour l'analyse phylogénétique des données sur le génome d'agents pathogènes (61) et les annotations Pango développées par Rambaut et collègues basés sur diverses méthodes au fil des versions (62) (voir section 1.2.2.3.2). La méthode qu'utilise Nextstrain pour réaliser ses analyses phylogénétiques sera discutée en détail à la section 1.2.5.3.2.

1.2.2.3.1 Les clades

Les clades sont un système de classification utilisé pour séparer différents variants en groupes distincts génétiquement. Typiquement, la distance génétique entre ces groupes est relativement égale (62). Les clades peuvent être divisés en sous-groupes lorsqu'un variant n'est pas assez distinct pour avoir son propre clade. Les séquences qui ne sont pas définies dans un groupe établi sont mises dans un groupe "non-classé" jusqu'à ce que suffisamment de séquences similaires soient récoltées pour créer un clade ou un sous-groupe. C'est le cas notamment pour le VIH qui a les clades *Main*, *Outlier* et *Non-main Non-outlier* (63). Chez les virus dont les nouveaux variants remplacent rapidement les variants parentaux, comme c'est le cas de l'Influenza A H5N1 par exemple, une mise à jour fréquente du système de clade est nécessaire (64).

Dans le cas de SARS-CoV-2, les clades définis par *Nextstrain* sont créés lorsque qu'un variant a au moins deux mutations le séparant de son parent et qu'il atteint une fréquence de 20% à travers

le monde (65). Ce système a l'avantage d'être robuste puisque basé sur plusieurs milliers d'échantillons, en revanche, il manque de granularité et doit être mis à jour fréquemment.

1.2.2.3.2 Les annotations Pango

Les annotations Pango sont probablement la méthode de nomenclature la plus utilisée pour la surveillance de variants locaux. Les séquences peuvent être assignées manuellement ou automatiquement à l'aide des programmes pangoLEARN ou UShER (66). La première version de pangoLEARN, utilisée pour Pango 2.0, utilisait un modèle de régression logistique multinomiale qui est une méthode de classification pour une tâche ayant plus de deux catégories (67) et un arbre de décision pour situer une séquence sur un arbre phylogénétique. La deuxième version, utilisée pour Pango 3.0, n'utilisait que l'arbre de décision. Cette version avait quelques défauts. Par exemple, plusieurs séquences étaient classées dans les catégories fourre-tout B.1 et B.1.1. Mostefai et collègues ont cependant réussi à reclasser ces séquences à l'aide d'un algorithme d'imputation et de réseau d'haplotype et ont montré que les annotations Pango ne reflétaient pas toujours la structure haplotypique (68). La version actuelle, utilisée pour Pango 4.0 utilise un algorithme de forêt d'arbres décisionnels dans lequel plusieurs arbres décisionnels sont entraînés sur différents sous-ensembles d'un jeu de données et arrivent à un consensus par rapport à la classification d'une séquence.

Ce système de nomenclature a l'avantage d'être dynamique et applicable à petite échelle pour la surveillance de variants émergents. Cela dit, des erreurs de classification ont été reportées (68). De plus, les annotations changent d'une version à l'autre, ce qui peut confondre les analyses. Un autre point important à souligner est que les différents algorithmes introduits lors des mises à jour de l'outil n'ont pas été rapporté dans des publications révisées par des pairs.

1.2.2.4 Classification selon le risque posé à la santé publique

Certains variants du SARS-CoV-2 sont plus contagieux et/ou donnent lieu à des infections plus sévères que d'autres. Pour pouvoir mieux suivre l'évolution de ces variants, l'OMS a mis en place leur propre système de classification.

1.2.2.4.1 Variants à suivre

Les variants à suivre (ou *Variant of Interest,* VOI) sont des variants pour lesquels on observe ou on prévoit des changements au niveau de la transmissibilité, de l'évasion immunitaire, de la sévérité

de la maladie, de la résistance au traitement ou pour lesquels le diagnostic est plus difficile avec les technologies de dépistage. De plus, ces variants doivent circuler dans la population de plusieurs pays. Il n'y a plus de VOI en circulation à la fin de l'été 2022 (69).

1.2.2.4.2 Variants préoccupants

Les variants préoccupants (ou *Variant of Concern,* VOC) sont des variants pour lesquels on note une augmentation de la transmissibilité, de la sévérité et/ou une modification de la présentation clinique ainsi qu'une diminution de l'efficacité des mesures sanitaires, des moyens de préventions tels les vaccins, des traitements et/ou des outils de dépistage.

Le premier VOC à émerger fut Alpha en Grande-Bretagne en décembre 2020. Ce variant avait une capacité de réplication de 43 à 90% plus élevée que les autres variants qui circulaient à cette époque-là (70). Deux autres variants notoires sont Beta et Gamma, originaires d'Afrique du Sud et du Brésil, respectivement (71,72). Ces variants possèdent entre autres la mutation de E vers K au codon 484 de la protéine S (S:E484K) qui est associée à la capacité du virus à réinfecter un hôte (73).

Le seul VOC en circulation à la fin de l'été 2022 est Omicron, qui est responsable de plus de 98% des cas de COVID-19 dans le monde. Omicron a lui-même plusieurs sous-lignées comportant des mutations qui affectent les caractéristiques du virus de manière distinctes. On parle donc d'une lignée de variants préoccupants pour désigner l'ensemble de la diversité génétique de cette souche virale (74). Omicron et ses multiples sous-lignées résultent en une infection moins sévère que d'autres VOC (75), et ce, même chez les personnes non-vaccinées (76). En revanche, ils sont plus transmissibles et sont associés à un haut taux de réinfection (77). De plus, ces variants ont une grande capacité d'évasion immunitaire ce qui réduit l'efficacité des traitements tels que les mAbs, les vaccins et le sérum de patient convalescent (78,79).

1.2.2.5 Mutations intra-hôte

Lorsque l'on parle de mutations, on parle généralement de mutations inter-hôte qui sont partagées et transmises dans la communauté. Cependant, une infection virale est causée par une population de plusieurs millions de virus individuels dans son hôte. À chaque réplication virale, il y a un risque d'apparition de mutations qui ne seraient présentes que chez cet individu, appelées SNV intra-hôte (ou *intra-host single nucleotide variant,* iSNV). Chaque hôte possède donc une diversité virale

intra-hôte qui diffère de la diversité inter-hôte dont il est question lorsque l'on parle de variants populationnels.

Une partie de la diversité intra-hôte est acquise lors de migrations de populations virales vers divers tissus du corps depuis le site d'inoculation (80). Chaque colonisation d'un nouvel organe résulte en un goulot d'étranglement qui peut altérer les fréquences alléliques. De plus, les populations virales de différents organes peuvent évoluer de manière indépendante les unes des autres. En étudiant la diversité intra-hôte dans les voies respiratoires et gastro-intestinales, Wang et collègues ont trouvé qu'il n'y avait pas de iSNV partagés entre les deux tissus chez un seul patient (81).

1.2.2.5.1 Le principe de quasiespèce

Le principe de quasiespèce est une théorie selon laquelle un organisme évolue pour former diverses populations de mutants distincts qui améliorent la *fitness* de la population plutôt que celle d'un virus individuel (82). On pourrait penser, par exemple, au VIH qui développe *de novo* des populations virales capables de résister à divers antirétroviraux avant même le début des traitements (83). Les mutations conférant cette résistance peuvent impacter négativement la capacité du virus à se reproduire ou à infecter de nouveaux patients. Dans ce cas, elles auront donc peu de chances d'être présentes à une fréquence élevée intra-hôte (84). La présence de différentes populations ayant des mutations qui leur sont propres est la raison pour laquelle il faut donner une combinaison de plusieurs antirétroviraux en même temps pour s'assurer que le traitement soit efficace (83). De plus, différentes classes d'antirétroviraux ciblent des mécanismes de reproduction virales différents, ce qui favorise l'élimination de diverses populations (85).

1.2.2.6 La théorie de l'évolution du SARS-CoV-2 à travers les patients immunosupprimés

Les patients immunosupprimés mettent en général plus de temps que les patients immunocompétents à combattre une infection au SARS-CoV-2. Des infections persistantes sur plusieurs mois ont été rapportées chez ce groupe de patients (45,86,87). Comme décrit plus haut, chaque réplication virale a le potentiel de mener à une mutation intra-hôte *de novo*, on peut donc se demander quel impact ces mutations peuvent avoir sur l'évolution génomique du virus.

Des études montrent que ces patients développent au fil de leurs infections des délétions sur la protéine S. Notamment, les délétions S:Δ144/145 (Alpha) et S:Δ243/244 (Beta) ont été détectées lors d'infection à long terme de patients immunosupprimés avant l'apparition des deux VOCs (87). Ces délétions tombent dans des épitopes de la protéine S, affectant négativement la capacité des

lymphocytes T de s'y lier. La question des épitopes et du système immunitaire en général sera abordée en détails à la section 1.2.3.3.

Les patients immunosupprimés sont également plus susceptibles d'acquérir des mutations réduisant la liaison des épitopes de la Spike, ce qui favorise l'évasion immunitaire. Par exemple, diverses mutations au résidu S:484 affectent cette liaison: S:E484K (Beta, Gamma) (73,88–90), S:E484Q (91,92), S:E484A (Omicron) (93). En plus d'affecter la capacité de l'hôte à se défendre contre l'infection, les diverses mutations répertoriées à ce site confèrent une résistance à un vaste éventail de mAbs et de plasma convalescent (94), ce qui permet la réinfection d'un hôte au SARS-CoV-2 (73). Plusieurs études de cas décrivant une infection persistante de SARS-CoV-2 chez des patients immunosupprimés ont noté une ou plusieurs mutations *de novo* à ce résidu, de manière séquentielle ou concomitante (86,95). Plusieurs autres mutations survenant dans des sites clés de la reconnaissance immunitaire ont également été répertoriées chez les patients immunosupprimés, c'est pourquoi de nombreux chercheurs adhèrent à la théorie selon laquelle ces infections persistantes sont un catalyseur de l'apparition de variants (90,95,96).

1.2.2.7 Les réservoirs viraux

Dans le contexte d'une infection virale, un réservoir est un site à l'intérieur de l'hôte dans lequel le virus s'accumule et où l'infection persiste (97). Les virus à l'intérieur de ces réservoirs échappent au système immunitaire, souvent parce que le virus entre en stade de latence durant lequel la prolifération de particules virales cesse ou que le réservoir est à un endroit où les lymphocytes T ont difficilement accès, ce qui est notamment le cas du VIH et des virus herpès (98,99).

Bien qu'il n'y ait pas de phase de latence documentée pour le SARS-CoV-2, la théorie du réservoir viral a été avancée pour expliquer certains cas de COVID longue (100). La présence de molécules intermédiaires de réplication du SARS-CoV-2 a été détectée jusqu'à 462 jours après le diagnostic positif, ce qui suggère la présence de virus viable dans certains patients atteints de COVID longue (101). Parmi les sites proposés pour un réservoir, il y a entre autres l'intestin, qui est un site connu pour être un réservoir de VIH (102), et le cerveau (103).

1.2.3 Caractéristiques de la COVID-19

1.2.3.1 Mécanisme d'entrée dans la cellule

1.2.3.1.1 Récepteur ACE2

Le récepteur de la cellule hôte, mentionné précédemment et auquel le RBD de la protéine S se fixe est l'enzyme de conversion de l'angiotensine 2 (*angiotensin converting enzyme 2*, ACE2) et est le même qu'utilisait SARS-CoV-1 (30). Des ARN messagers (ARNm) pour ce récepteur ont été trouvés chez l'humain dans 72 tissus, notamment dans les systèmes respiratoires, cardiovasculaires, rénaux et gastro-intestinaux (104). ACE2 a été détecté par immunohistochimie, entre autres, dans les cellules épithéliales des alvéoles pulmonaires, dans les entérocytes de l'intestin grêle, dans l'épithélium des voix nasopharyngiques, dans l'endothélium des veines et des artères, ainsi que les cellules des muscles lisses artériels (105).

1.2.3.1.2 TMPRSS2 et NRP1

Les protéines transmembranaires protéase transmembranaire à sérine 2 (TMPRSS2) et la neuropiline 1 (NRP1) contribuent à l'entrée du virus dans la cellule de l'hôte. Une fois le RBD de la sous-unité S1 attaché au récepteur ACE2, TMPRSS2 clive la protéine S en ses deux sous-unités, permettant ainsi à la sous-unité S2 d'activer le processus de fusion de la membrane virale avec celle de la cellule hôte (106). NRP1 facilite l'infection en s'attachant à la sous-unité S1 une fois la protéine S clivée. Le mécanisme précis de facilitation n'est toujours pas bien compris, mais il a été démontré que la capacité du virus à infecter les cellules hôtes est moins efficace lorsque l'action de NRP1 est bloquée (107,108).

1.2.3.2 Présentation clinique

Les premiers articles publiés sur la présentation clinique de la COVID-19 décrivent de la fièvre et une pneumonie sévère d'origine virale. De plus, les patients dont la sévérité de l'infection nécessite une admission à l'unité de soins intensifs développent souvent un choc cytokinique qui peut mener à une défaillance multiviscérale (109). Les cytokines sont des protéines du système immunitaire qui coordonnent la production et la fonction des cellules immunitaires et sanguines, en plus de moduler la réponse inflammatoire à une infection. Le terme choc cytokinique s'applique lorsque ladite réponse devient démesurée et devient néfaste pour l'hôte (110). La COVID-19 peut aussi causer plusieurs autres types de symptômes, notamment la fatigue, des troubles gastro-intestinaux, des accidents vasculaires cérébraux (AVC), des thromboses veineuses, des embolies

pulmonaires, des troubles d'ordre neurologiques, l'anosmie (la perte de l'odorat) et l'agueusie (la perte du goût) (81,111–113).

1.2.3.3 Réponse immunitaire de l'hôte

Le système immunitaire est composé d'un ensemble de cellules qui ont pour fonction de détecter les éléments étrangers présents dans le corps, appelés antigènes, et à les éliminer. Il existe deux types de réponse immunitaire, soient la réponse immunitaire innée qui est non spécifique et la réponse immunitaire adaptative qui est spécifique à un antigène.

1.2.3.3.1 Réponse immunitaire innée

Comme son nom l'indique, la réponse innée survient de façon immédiate en réaction à un antigène, même si celui-ci n'a jamais été rencontré par l'organisme auparavant. Les premiers en ligne de défense sont les phagocytes qui possèdent des récepteurs de type Toll (*Toll-like receptor,* TLR) qui reconnaissent diverses composantes qu'on ne retrouve pas normalement dans le corps, comme par exemple l'ARN double brin ou la flagelle des bactéries (114). Le phagocyte englobe ensuite l'antigène, puis fusionne avec un lysosome qui le détruit. Dans le cadre d'une infection virale, les cellules infectées produisent des peptides appelés interférons (INFs) qui signalent aux cellules environnantes la présence d'un pathogène (115). Ces cellules sécrètent alors d'autres peptides qui inhibent la réplication virale et contribuent à limiter l'infection (116). Un autre type de cellule impliquée dans la réponse immunitaire innée est la cellule tueuse naturelle (*natural killer*, NK). Elles ciblent les cellules tumorales et détectent les cellules infectées par un virus au début de l'infection (117). Les NK libèrent diverses protéines capables de tuer les cellules infectées ou tumorales, ou de coordonner la réponse immunitaire, telles les cytokines, les interleukines et les chimiokines (118).

Un autre mécanisme de défense inné dont le rôle principal est de combattre les infections virales est la modification des génomes viraux par les enzymes de la famille des adénosine désaminases (ou *adenosine deaminase acting on RNA*, ADAR) et des apolipoprotéines B (ou *apolipoprotein B editing complex,* APOBEC). Comme leur nom l'indique, les ADAR désaminasent les adenosines, c'est-à-dire qu'ils les transforment en une inosine (I), ce qui résultent en une substitution A > G (119). Ces enzymes agissent sur l'ARN double brin. Les CoVs sont des virus à ARN simple brin mais ils adoptent une conformation à double brin lors de la réplication virale, ce qui les rend vulnérables aux ADARs (120). Les enzymes de la famille des APOBEC désaminasent les cytosines

en uracil et affectent l'ARN ainsi que l'acide désoxyribonucléique (ADN) (121). L'introduction de ces mutations C > U peut avoir un effet délétère pour les virus, comme c'est le cas notamment pour VIH et l'hépatite B (122,123).

1.2.3.3.2 Réponse immunitaire adaptative

La réponse immunitaire adaptative s'acquiert après avoir été en contact avec un antigène spécifique. Elle repose sur deux types de lymphocytes, soient les lymphocytes T qui viennent du thymus et les lymphocytes B qui viennent de la moelle osseuse (124). Ces cellules possèdent des protéines transmembranaires, appelées récepteurs d'antigènes, qui se lient à une séquence spécifique sur un antigène, appelée épitope. Un agent antigénique a plusieurs épitopes différents. Chaque lymphocyte T ou B possède des milliers de récepteurs d'antigènes qui peuvent se lier à un seul épitope. Ces cellules se lient donc de manière spécifique à une seule partie d'un agent antigénique (116).

Lorsqu'un individu acquiert une protection contre un pathogène suite à une exposition à ce dernier, on parle de mémoire immunologique. Cette mémoire accélère la réponse immunitaire à un pathogène déjà rencontré, et améliore son efficacité. Lors d'une première exposition, il faut de 10 à 17 jours pour atteindre le maximum de production des cellules effectrice de la réponse immunitaire adaptative (116). Lors d'une réinfection au même pathogène, cette production de cellule effectrice ne prend que de deux à sept jours.

1.2.3.3.3 La présentation des épitopes aux lymphocytes T

Les cellules hôtes possèdent des protéines appelées molécules du complexe majeur d'histocompatibilité (CMH) qui sont responsables de présenter des fragments d'antigènes, appelés épitopes aux lymphocytes T à la surface de la cellule (125). Lorsqu'un lymphocyte T avec le récepteur d'antigène permettant de se lier à l'antigène croise la cellule en question, celui-ci se lie à la molécule du CMH. Il existe deux types de CMH, soient les types I et II. Toutes les cellules nucléées possèdent le type I, mais seulement les cellules dendritiques, les macrophages et les lymphocytes B possèdent le type II. Cette liaison entre un CMH de type II et un lymphocyte donne lieu à un échange de cytokines qui activent la réponse immunitaire (116).

Il existe de nombreux polymorphismes dans les CMH à travers les populations humaines, si bien qu'ils sont souvent regroupés en supertypes (126). Plusieurs études ont démontré que ces polymorphismes peuvent avoir un impact sur la capacité d'un CMH à se lier à un épitope (127–129). La réponse immunitaire à un pathogène peut donc différer d'une population à l'autre selon

le supertype de CMH.

1.2.3.3.4 Les lymphocytes B et la production d'anticorps

Contrairement aux cellules dendritiques et aux macrophages qui peuvent se lier à divers pathogènes par leur CMH de type II, les lymphocytes B ne peuvent se lier qu'à l'antigène pour lequel il a un récepteur (130). Lorsqu'un lymphocyte T CD4+ se lie à un lymphocyte B après que ce dernier lui ait présenté un épitope, le lymphocyte B s'active et des milliers d'anticorps sont produits. La fonctiondes anticorps est de se lier à l'épitope pour lequel ils ont un récepteur afin de neutraliser le pathogène et stimuler la phagocytose de ce dernier par les macrophages (116).

1.2.3.4 Interventions pour prévenir ou traiter la COVID-19

Le monde médical a été pris au dépourvu au début de la pandémie, entre autres parce qu'il n'y avait pas de traitements efficaces contre les coronaviridae. Le développement de traitements spécifiques pour prévenir ou lutter contre l'infection a contribué à réduire le taux de mortalité lié à la COVID-19 dès la première année de la pandémie. À titre d'exemple, le taux de mortalité dans une cohorte américaine de patients hospitalisés pour la COVID-19 est passé de 19.7% en avril 2020 à 9.3% en novembre 2020 et cette différence n'est pas expliquée par un changement de l'âge moyen des gens infectés (131).

1.2.3.4.1 Le sérum de patient convalescent

Le premier traitement qui a été développé contre la COVID-19 l'administration de sérum de patient convalescent aux patients malades. Le sérum contient les anticorps que le patient convalescent a développé contre le SARS-CoV-2. Cependant, ce genre de traitement est souvent spécifique au variant du premier patient, puisque les anticorps développés sont en fonction des résidus des épitopes auxquels il a été exposé lors de son infection (132).

1.2.3.4.2 Les anticorps monoclonaux

Les anticorps monoclonaux (mAbs) sont des molécules clonées en laboratoires à partir des anticorps produits par une culture d'un seul type de lymphocyte B. Ces lymphocytes ont préalablement été extraits d'un animal qui a été exposé au pathogène. Ils sont appelés monoclonaux parce qu'ils sont spécifiques à un épitope d'un antigène en particulier (116). Dans le cas du développement de traitements pour combattre les infections symptomatiques au SARS-CoV-2, plusieurs mAbs différents ont été développés et utilisés dans diverses combinaisons pour traiter les cas de COVID-19 modérés et sévères (133–135).

Certains variants, notamment Omicron qui circule majoritairement en 2022, ont développé des mutations capables d'échapper à ces mAbs (136). C'est pourquoi il est plus judicieux de se tourner vers des mAbs à large spectre, c'est-à-dire des anticorps qui ont la capacité de se lier aux épitopes de divers variants (137) ou vers des thérapies combinant différents types de traitements.

1.2.3.4.3 Les vaccins

Plusieurs types de vaccins contre le SARS-CoV-2 sont maintenant disponibles, soient les vaccins à ADN, les vaccins à ARNm, les vaccins utilisant un vecteur viral non-réplicatifs, les vaccins contenant le virus de SARS-CoV-2 inactivé et les vaccins contenant des sous-unités virales (138). Les vaccins à ADN utilisent des plasmides bactériens qui contiennent les gènes pour produire la protéine S ainsi que le peptide de signalement IgE (immunoglobulines E) (139). Ces plasmides intègrent les gènes d'antigènes dans les noyaux des cellules hôtes qui expriment ensuite les protéines virales afin d'entraîner une réponse immunitaire. Les vaccins à ARNm encapsulent des ARNm prêts à être traduits en protéines virales dans des nanoparticules lipidiques (140). Le fonctionnement est similaire à celui des vaccins à ADN dans le sens où l'hôte produit lui-même les protéines virales. Les vaccins utilisant un vecteur viral non-réplicatif contiennent un adénovirus qui n'a pas la capacité de se répliquer mais qui peut produire des protéines du SARS-CoV-2 (141). Les virus de SARS-CoV-2 inactivés utilisés pour les vaccins sont cultivés en laboratoire (142). Ils sont inactivés par β-propiolactone (un désinfectant aux propriétés antivirales) avant d'être injectés (143). Finalement, les vaccins contenant des sous-unités virales contiennent des fragments antigéniques qui permettent le déclenchement d'une réponse immunitaire (144).

Les vaccins ont contribué à prévenir les infections et le développement d'une forme sévère de la COVID-19 (145). Cependant, plusieurs campagnes de désinformation à leur sujet ont incité une partie de la population à adopter une attitude réfractaire par rapport aux campagnes de vaccination (146,147). De plus, certains pays ont rencontré des difficultés à se procurer des vaccins à distribuer à leur population, à savoir les pays du sud global (148). Le faible taux de personnes adéquatement vaccinées dans ces populations fait en sorte que les cas de COVID-19 nécessitant une hospitalisation sont plus élevés, ce qui met beaucoup de pression sur leurs systèmes de santé qui ont d'emblée moins de ressources que ceux du nord global (149).

1.2.3.4.4 Les antiviraux

Les antiviraux sont des médicaments qui sont utilisés pour combattre les infections virales. Ils

peuvent agir de diverses façons, entre autres en inhibant la réplication virale ou en affectant le mécanisme d'entrée du virus dans la cellule hôte (150). Un exemple de médicament antiviral utilisé contre le SARS-CoV-2 est Remdesivir, un inhibiteur de l'ARN polymérase virale (151). Ce médicament permet de réduire la durée d'hospitalisation des patients atteints d'une forme sévère de la COVID-19 en plus de diminuer le risque d'hospitalisation chez les personnes à risque lorsqu'il est donné au début de l'infection (152).

1.2.3.4.5 Les glucocorticoïdes

Les glucocorticoïdes sont un groupe d'hormones qui agissent entre autres sur le métabolisme du glucose en plus d'avoir une fonction anti-inflammatoire (116). Administrer des glucocorticoïdes synthétiques aux patients atteints des formes sévères et modérées de la COVID-19 permet de réduire le taux de mortalité, le nombre de jours sous assistance respiratoire ainsi que la durée d'hospitalisation (153). Ce traitement est toutefois à proscrire chez les personnes souffrant d'insuffisance respiratoire hypoxémique aiguë puisqu'une haute dose de glucocorticoïdes est associée à un taux de mortalité plus élevé chez cette population.

1.2.3.5 Comorbidités et résultats cliniques variables

Il a été établi rapidement au début de la pandémie que les personnes infectées par le SARS-CoV-2 peuvent avoir un large éventail de sévérité de leurs symptômes, et que certains ont une infection asymptomatique, c'est-à-dire qu'ils n'en développent pas du tout (154).

1.2.3.5.1 Complications associées à l'âge

Comme pour SARS-CoV-1, les enfants sont moins à risque de développer des symptômes sévères (155). Cependant, une complication affecte uniquement les enfants, il s'agit du syndrome inflammatoire multisystémique de l'enfant (MIS-C). Une étude sur une cohorte de patients de moins de 21 ans a déterminé que ce syndrome affecte 5.1 patients sur 1,000,000 personnes-mois (unité de mesure désignant un patient par mois de participation à la cohorte) (156). Chez les adultes, le risque de complications associées à la COVID-19 augmente après 40 ans (157) et le taux de mortalité augmente drastiquement pour les patients de plus de 70 ans (158). Un aspect surprenant de la courbe de sévérité selon l'âge est que les nonagénaires, les centenaires et les supercentenaires semblent avoir en moyenne des symptômes moins graves que les septuagénaires et les octogénaires (159). Il est courant de regrouper les patients de plus de 80 ou 85 ans ensemble lorsqu'on parle de taux de mortalité (160). Cependant, des études sur les patients très âgés démontrent que les patients

de 90 ans et plus ont des taux de mortalité différents que ceux des patients plus jeunes (161). La courbe de mortalité pour la COVID-19 est un autre exemple de pourquoi c'est une bonne pratique de séparer les décennies lors d'études sur les patients très âgés.

1.2.3.5.2 Comorbidités

Les patients souffrant de maladies chroniques telles que l'hypertension, le diabète, les maladies cardiovasculaires et l'obésité sont plus à risque de développer des symptômes sévères lors d'une infection à la COVID-19 (162,163). Il est cependant difficile de départager le rôle de l'âge et celui des comorbidités, puisque l'âge et le nombre de morbidités sont positivement corrélés (164).

En général, les personnes immunodéprimées ne semblent pas être autant à risque de complications sévères que la population générale (165). Ceci est dû au fait qu'ils ont moins de chance de développer une réponse immunitaire démesurée entraînant un choc cytokinique. Cela dit, lorsqu'ils sont infectés, ces individus peuvent mettre plusieurs mois à combattre l'infection, contrairement à une moyenne d'environ 14 jours pour la population générale (166,167). Ces infections de longue durée posent non seulement un risque pour leur santé mais aussi pour l'apparition de nouveaux variants viraux (166).

1.2.3.5.3 Différences génétiques entre patients

Plusieurs études d'association pangénomique (*Genome Wide Association Studies*, GWAS) ont trouvé des loci ayant un impact sur la sévérité des symptômes de la COVID-19 (168–170). Par exemple, un premier est situé sur le chromosome 19, locus 19p13.3, et code pour la dipeptidyl peptidase 9 (DPP9). Un SNP détecté à ce locus est lié à une inflammation dérégulée causant des lésions pulmonaires à un stade avancé de l'infection. Un deuxième est situé sur le chromosome 12, locus 12q24.13, comprenant un ensemble de gènes codant pour les enzymes oligoadénylate synthétases (OAS) induites par les interférons. Ce SNP dans le gène *OAS1* confère un effet protecteur contre l'infection au SARS-CoV-2 dans la population chinoise Han (171). Chez la population de descendance européenne, un haplotype d'origine néandertalien sur ce même gène confère également un effet protecteur contre l'infection (172). Finalement, un autre SNP est situé sur le chromosome 21, locus 21q22.1, et code pour le gène du récepteur d'interféron *IFNAR2*. Une faible expression de *IFNAR2* est associée à des symptômes plus sévères de la COVID-19 (173). Le nombre de loci associés à des effets protecteurs ou aggravants ne cesse d'augmenter (174,175). Il a été démontré qu'une approche de médecine personnalisée basée sur la présentation clinique réduit le taux de mortalité chez les patients atteints de la COVID-19 (176). L'information sur le génotype

du patient pourrait éventuellement être ajoutée aux paramètres cliniques afin de mieux orienter le traitement.

1.2.3.5.4 Syndrome post-COVID-19 ou COVID-longue

À ce jour, il n'y a pas de critères cliniques bien établis pour le syndrome post-COVID-19. L'éventail de symptômes pouvant affecter les patients atteints est vaste mais leur cause peut être distincte. Par exemple, plusieurs patients souffrent d'anosmie, probablement due à un dommage à l'épithélium olfactif (177). Un déficit cognitif a également été remarqué chez certains patients et une hypothèse par rapport à la cause de ce symptôme est l'hypoxie causée par des symptômes respiratoires sévères durant la phase aiguë de l'infection (178). D'autres symptômes sont connus pour émerger suite à d'autres types d'infections virales également, notamment l'apparition de nouvelles allergies ou intolérances alimentaires (179–181).

Comme le récepteur ACE2 est présent dans plusieurs tissus du corps, plusieurs systèmes ont le potentiel d'être affectés par la COVID-19 et ce, sans nécessairement causer de symptômes durant l'infection aiguë. Il n'y a pas de distinction entre une infection persistante, qui peut être entrecoupée de phases dans lesquelles une personne infectée ne teste plus positif, et entre un ensemble de symptômes qui perdurent une fois l'infection résorbée (4). Il est difficile d'établir un critère clair pour le syndrome post-COVID-19 parce qu'on ne comprend pas encore l'étendue de l'impact de l'infection dans les différents systèmes, ni la capacité de l'infection à causer des séquelles à moyen et à long terme.

1.2.4 Techniques d'analyses de l'évolution virale

1.2.4.1 Séquençage

Le séquençage de matériel génétique est la méthode par laquelle on reconstitue la séquence de nucléotides d'ADN ou d'ARN d'un organisme. La première méthode de séquençage fut développée par Sanger et al. en 1977 (182). Cette méthode consiste à synthétiser une séquence d'ADN in vitro à l'aide des mécanismes biochimiques de réplication d'ADN. Le processus débute par la fragmentation de l'ADN et sa dénaturation en ADN simple-brin qui servira de matrice. Ensuite, une courte séquence appelée amorce est ajoutée à l'extrémité 3'. Une amorce complémentaire se lie à celle de la matrice, permettant ainsi à l'ADN polymérase de débuter la synthèse du nouveau brin d'ADN en y incorporant les acides nucléiques complémentaires. Cette méthode a l'avantage d'avoir un taux d'erreur très bas (0.01%) et de produire des fragments de

séquences relativement longs (700-900 bp) (183,184). En revanche, il s'agit d'une méthode coûteuse et lente.

Le développement de technologies de séquençage de nouvelle génération (*next-generation sequencing*, NGS), telles que commercialisée par la compagnie Illumina a permis de réduire drastiquement les coûts liés au séquençage (185). De plus, les technologies de séquençage de troisième génération (*Third-generation sequencing*, TGS) dont Oxford Nanopore Technology nous permettent maintenant de séquencer des fragments de plusieurs milliers de paires de bases et de séquencer sur le terrain sans équipement de laboratoire (186).

1.2.4.1.1 Le séquençage de nouvelle génération

Le séquençage de nouvelle génération est une des techniques les plus répandues à travers le monde. Il existe plusieurs protocoles de NGS dont celui utilisé par la compagnie Illumina qui sera décrit ci-dessous. La première étape pour cette méthode consiste à générer de nombreuses copies des fragments d'ARN viral dans une technique appelée réaction de polymérisation en chaîne (*polymerase chain reaction*, PCR). Pour ce faire, l'ARN est fragmenté en segments d'un maximum de 300 paires de bases et deux adaptateurs sont ajoutés aux deux extrémités des fragments. L'adaptateur va permettre d'hybrider le fragment à un adaptateur complémentaire attaché à la puce de génotypage sur la machine de séquençage. Il y a par la suite une étape d'amplification en pont double-brin durant laquelle l'adaptateur à l'extrémité libre du fragment se lie à un autre adaptateur de la puce. Des amorces ainsi que les molécules nécessaires à la synthèse du brin complémentaire sont alors ajoutées, ce qui permet la synthèse d'un nouveau fragment. Le pont double-brin est ensuite dénaturé, les extrémités libres se lient à une autre amorce sur la puce, et cette étape est répétée jusqu'à avoir plusieurs agrégats distincts de fragments identiques sur la puce (187). Le séquençage débute lorsqu'un seuil d'amplification déterminé par l'équipe de recherche est atteint. Moins il y a de matériel génétique au départ et plus il faut de cycles pour atteindre le seuil. Ces cycles peuvent entraîner des erreurs qui seront possiblement vues comme des mutations dans le séquençage.

Lors du séquençage, une amorce est hybridée à l'adaptateur libre. Des acides nucléiques sur lesquels une molécule fluorescente associée est attachée sont mis à disposition de l'ARN polymérase afin de synthétiser le brin complémentaire. À chaque ajout de nucléotide, la molécule fluorescente est clivée et l'appareil Illumina prend une photo de la puce. Il est ensuite possible de reconstituer la séquence à l'aide du signal fluorescent capté au long de la synthétisation des fragments. Le taux d'erreur moyen de cette technique se situe entre 0.1 et 1%, dépendamment de

l'appareil et du protocole utilisé. Néanmoins, certains motifs spécifiques tels que GGC semblent augmenter le taux d'erreur.

1.2.4.1.2 Le séquençage de troisième génération

Les TGS sont les avancées les plus récentes dans le domaine de la technologie du séquençage. Un exemple de TSG est Oxford Nanopore, qui utilise une protéine appelée nanopore dans lequel on fait passer un courant constant d'ions. Chaque molécule qui passe dans le pore altère le courant de façon spécifique et il est possible d'identifier la molécule selon l'altération. Les molécules d'ARN sont donc séquencées en temps réel au fur et à mesure que ses nucléotides traversent le pore (188).

L'avantage de cette méthode dans un contexte pandémique est qu'il existe un protocole de séquençage d'ARN qui ne demande pas de préparer le matériel génétique en le fragmentant et en l'amplifiant (189). Il est donc possible de séquencer sur le terrain, ce qui est impossible avec la technologie Illumina. Cela dit, le taux d'erreur est de l'ordre de 5 à 15% (190), ce qui ne pose pas de problèmes pour générer un génome consensus mais rend l'étude de iSNV complexe.

1.2.4.2 Initiatives de partage de données

La surveillance des variants du SARS-CoV-2 et l'analyse de son évolution nécessite un grand nombre de séquences virales. Le *Global Initiative on Sharing All Influenza Data* (GISAID) est une initiative de recherche internationale qui permet le partage de séquences et de métadonnées virales (https://gisaid.org/). Le site a été créé en 2008 durant l'épidémie de grippe aviaire H5N1 afin d'encourager les équipes de recherche à partager leurs données le plus rapidement possible. Avant la création de cette plateforme, il était courant d'attendre d'avoir publié avant de partager ses séquences. Dans un contexte pandémique, cette pratique freine les avancées qui peuvent être cruciales pour mieux orienter les efforts de santé publique et les traitements. Un système comme GISAID qui permet de reconnaître la contribution des équipes qui ont soumis les séquences est un moyen idéal de contourner ce problème (191).

Il existe également une initiative similaire conçue spécifiquement pour le partage des génomes du SARS-CoV-2 au Canada, CanCOGeN (https://virusseq-dataportal.ca/). Ce portail permet de centraliser les données des différents organismes de santé publique provinciaux qui sont indépendants.

1.2.4.3 Analyses phylogénétiques

Tel que mentionné précédemment, la phylogénie reconstitue des évènements évolutifs et les représente sous forme d'arbre. Lorsque le taux de substitution nucléotidique est constant à travers l'arbre, le temps moléculaire peut être utilisé comme l'unité de longueur pour la taille des branches et il est donc possible de dater les séquences à l'aide de l'algorithme (192). Une autre façon de représenter l'évolution génétique à l'aide d'un arbre est d'utiliser la distance génétique entre deux séquences pour déterminer la longueur des branches. Dans ce cas, divers modèles permettant de spécifier différents ts/tv et de fréquence des nucléotides sont disponibles (193).

1.2.4.3.1 Algorithmes phylogénétiques

La méthode d'optimisation du maximum de vraisemblance (*maximum likelihood*) est un algorithme commun pour générer un arbre phylogénétique. Cette méthode génère une multitude de topologie d'arbres et desméthodes statistiques sont ensuite utilisées pour déterminer l'arbre qui reconstruit l'évolution des séquences étudiées de la façon la plus précise possible selon des modèles prédéterminés (194). Laméthode bayésienne a un fonctionnement similaire à la méthode de plausibilité maximale, mais elle utilise des informations qui proviennent d'autres analyses similaires, appelées *prior*. Un exemple de *prior* est une topologie d'arbre déjà générée (195). L'algorithme évalue la vraisemblance d'une topologie en fonction de sa divergence avec le *prior*. Finalement, la méthodede parcimonie est un autre algorithme commun qui génère une multitude de topologies d'arbres. La topologie qui demande le moins grand nombre d'évènements évolutifs est ensuite choisie (196).

1.2.4.3.2 Problèmes liés à l'utilisation de méthodes phylogénétiques pour étudier SARS-CoV-2

Depuis le début de la pandémie, plusieurs équipes ont essayé de démystifier l'origine du virus à l'aide d'outils phylogénétiques et de dater le premier cas de transmission zoonique. Malheureusement, cette tâche s'est avérée ardue: une étude comparant les résultats de différents articles qui ont obtenu une date de la première transmission à l'aide de méthodes phylogénétique a trouvé que les dates varient du 15 octobre au 8 décembre 2019, avec une variation de neuf mois dans l'intervalle de crédibilité de Bayes (197). Les 11 articles mentionnés dans l'étude sont parus en 2020 et utilisaient les mêmes séquences provenant de GISAID en plus du même logiciel pour 10 d'entre eux. La différence constatée s'explique donc par les différents modèles et paramètres utilisés ainsi que les différentes manières de l'échantillonnage des séquences.

Peu importe la méthode utilisée, lorsque les séquences utilisées sont très similaires les unes aux

autres, la topologie de l'arbre devient difficile à résoudre et plusieurs arbres sont générés avec la même plausibilité. Morel et al. (198) se sont penchés sur ce problème et ont remarqué que certains programmes n'avertissaient pas les utilisateurs lorsqu'une telle situation se produisait, une topologie choisie au hasard leur était simplement retournée (198). Les auteurs recommandent donc de faire attention lors de la génération et de l'interprétation de ces arbres car la quantité de séquences semblable rend la sélection de paramètres délicate.

1.2.4.3.3 Nextstrain

La plateforme *Nextstrain* est une base de données de génomes viraux, un pipeline d'analyse phylodynamique ainsi qu'un outil de visualisation des données interactif (60,61). Nextstrain utilise deux pipelines, Augur pour la génération de l'arbre et Auspice pour sa visualisation. Les séquences sont tout d'abord alignées avec MAFFT (199), un aligneur de séquences multiples. L'arbre est par la suite généré avec IQ-TREE (200) qui fait appel à la méthode de plausibilité maximale avec le modèle GTR (*General Time Reversible*) dans lequel les taux d'acide aminés et de substitution peuvent être différents (201). L'arbre obtenu est ensuite affiné avec l'outil TreeTime (60) qui utilise une méthode d'optimisation itérative pour peaufiner la longueur des branches.

1.2.4.4 Simulations de l'évolution génomique virale

Il existe deux types de simulations génomiques: les simulations basées sur la coalescence qui remonte le fil des événements évolutifs survenus dans le passé afin de trouver l'ancêtre commun le plus récent (202) ou les simulations de l'évolution future. Dans le cadre de l'analyse du SARS-CoV-2, les simulations coalescentes pourraient être utilisées afin d'en apprendre plus sur les origines du virus (203). Les simulations vers le futur permettent de réaliser des expériences *in silico* et d'observer l'évolution virale selon divers scénarios évolutifs. Par exemple, on pourrait vouloir faire varier la taille de la population virale afin d'observer les effets des divers principes de génétique des populations décris à la section 1.2.2.1. Les simulations vers le futur deviennent de plus en plus exigeantes selon le degré de spécificité du scénario évolutif à reproduire (204) et les programmes performants qui permettent de réaliser ce type de simulations sont relativement récents puisqu'ils peuvent devenir très intensifs au niveau computationnel.

SANTA-SIM (S-S) est un logiciel de simulation vers le futur spécifique aux virus qui a été développé récemment (205). S-S offre beaucoup de flexibilité en permettant aux usagers d'annoter diverses régions du génome et d'y appliquer divers paramètres évolutifs. Plusieurs fonctions de *fitness* sont disponibles, incluant les simulations neutres, les simulations dépendantes de la taille

de la population, l'assignation manuelle d'une valeur de *fitness* pour chaque acide aminé, ainsi qu'une valeur de *fitness* dépendante de l'âge (c'est-à-dire que les virus comportant les allèles les plus vieilles auront moins de chance de transmettre leur matériel génétique). S-S a également un outil de reconstruction phylogénétique intégré.

Les auteurs ne spécifient pas si le logiciel a été conçu pour simuler l'évolution intra- ou inter-hôte. Tous les exemples de l'article de l'article original (205) sont intra-hôtes, mais il serait possible derecréer manuellement l'effet du goulot d'étranglement de transmission, qui se produit lors de l'infection d'un hôte à l'autre en sous-échantillonnant les génomes obtenus au cours d'une simulation afin d'en commencer une nouvelle. S-S a quelques points faibles dans sa conception, notamment le fait qu'il ne permet de simuler que des génomes à populations constantes. Or, certains mécanismes de l'évolution virale dépendent de la taille de la population. De plus, les identifiants des séquences durant les simulations ne sont pas les mêmes que ceux utilisés par l'outil de phylogénie. Il n'y a donc aucun moyen de lier une séquence génétique à une feuille de l'arbre généré par S-S.

1.2.4.5 Modèles mathématiques de l'infection virale

Depuis la fin des années 1970, les modèles mathématiques jouent un rôle clé dans notre compréhension des infections et de la transmission virale, ainsi que des mécanismes de résistance aux médicaments (206). Il est possible de modéliser le modèle d'évolution quasiespèce décrit plus tôt à l'aide du système d'équation différentiel ordinaire (EDO) donné dans les Équations i-iii qui décrit le changement dans les cellules non infectées x en (i), dans les cellules infectées y_i en (ii) et dans virions v_i en (iii) (82).

(i) $$\frac{dx}{dt} = \lambda - \delta x - x \sum_i \beta_i v_i$$

(ii) $$\frac{dy_i}{dt} = x \sum_j \mathbf{Q}_{ji} \beta_j v_j - a_i y_i$$

(iii) $$\frac{dv_i}{dt} = k_i y_i - u_i v_i$$

Dans les équations ci-haut, λ est le taux de production de cellules non infectées (qui peut être ignoré lors d'infections aiguës), δ est le taux de mortalité des cellules non infectées et β_i est le taux d'infection par cellule. Le changement dans les cellules non infectées à travers le temps (Eq. i) peut donc être définie par la différence entre nombre de cellules non infectées produites et le nombre de

cellules éliminées soit par la mort naturelle, soit par l'infection par des virus libres ($x \sum_{\$} \beta_{\$} v_{\$}$). Q_{ij} représente la probabilité qu'une souche virale j mute en la souche virale i et a_i le taux de mortalité des cellules infectées. Le changement aux cellules infectées par la souche i dans le temps (Eq. ii) se mesure par la somme des nouvelles cellules infectées et de celles qui mutent vers la souche i moins celles qui meurent. k_i est le taux de production de nouveaux virus libres et u_i est le taux de mortalité des virus libres. Le changement des virions dans le temps (Eq. iii) est donc le taux de nouveaux virus libres moins ceux qui meurent (207). On peut ajouter diverses composantes à cet EDO pour modéliser des phénomènes plus complexes tels que la phase d'éclipse durant laquelle une cellule hôte qui vient tout juste d'être infectée ne produit pas encore de nouveaux virus, un taux de mortalité causé par un traitement antiviral ou spécifier différents taux d'infections pour différents tissus (208). Les modèles mathématiques comme celui-ci sont un excellent moyen de tester des hypothèses complexes qu'il serait difficile de valider expérimentalement *in vitro* ou *in vivo* (209).

1.2.4.6 L'apprentissage automatique

Bien que les modèles mathématiques nous aident à prédire des événements futurs (approche prospective), l'apprentissage automatique est utile pour le problème inverse, c'est-à-dire la reconstruction d'interactions en utilisant un grand nombre de données (approche rétrospective). L'apprentissage automatique est défini comme étant une branche de l'intelligence artificielle qui se base sur des principes mathématiques et statistiques pour imiter la façon dont le cerveau humain apprend (210). Spécifiquement dans le cadre de la génomique, la réduction de coûts de séquençage qu'a emmenée l'arrivée des technologies de NGS et TGS a augmenté de façon fulgurante le nombre de séquences d'ADN et d'ARN disponibles pour faire de la recherche (185). Des algorithmes d'apprentissage automatique sont utilisés pour traiter ces données dont le nombre est trop important pour que ce soit fait manuellement.

1.2.4.6.1 ImputeCoVNet

L'outil ImputeCoVNet fait appel à l'apprentissage profond pour imputer les positions manquantes dans les génomes du SARS-CoV-2 (211). Il arrive souvent qu'une séquence contienne des nucléotides manquants, ce qui peut poser problème lorsque cela arrive à une position d'intérêt, comme par exemple à un SNP clé pour identifier un variant. ImputeCoVNet utilise une technique d'imputation dite sans référence, c'est dans dire qu'elle se base sur la reconnaissance de motifs dans les données. Dans le cadre de SARS-CoV-2, ces patrons sont les haplotypes.

Cet outil contient un encodeur qui apprend une représentation, consistant en un vecteur à faible dimension, à partir des séquences d'entrée et un décodeur qui reconstruit la séquence à partir du vecteur. ImputeCoVNet a été utilisé pour imputer des nucléotides à 199 positions pour lesquelles la fréquence allélique dérivée était à plus de 1% durant la première vague de la pandémie (janvier à juillet 2020). Ceci a permis une classification de séquences de la première vague par haplotype qui se base sur 22 SNPs dans le génome du SARS-CoV-2 (68).

1.2.4.6.2 netMHCpan

netMHCpan est un outil utilisant un algorithme d'apprentissage automatique afin de prédire l'affinité d'un épitope à un CHM (212). Ces outils de prédiction sont généralement entraînés avec des données sur l'affinité de liaison ou avec des données sur les ligands élués en masse spectrométrie. Le désavantage du premier type de données est qu'il repose uniquement sur l'affinité de liaison et ignore les autres facteurs biologiques impliqués dans le processus. Le désavantage du second type de données est qu'il nécessite une annotation des ligands-CHM qui peut être biaisée vers les ligands les plus fréquents (213). netMHCpan résout les limitations citées ci-haut en combinant les deux types de données avec un algorithme d'apprentissage automatique qui annote les ligands en fonction des deux types de données.

1.2.5 La pandémie au Québec durant les deux premières vagues

1.2.5.1 La première vague

Le Québec est la première province canadienne à avoir fait face à la pandémie de SARS-CoV-2. Durant la première vague, définie comme allant jusqu'en juin 2020 dans la province, il y a eu 46,047 cas confirmés et 5,359 décès liés à la COVID-19 (214). Une étude sur les premières introductions du virus dans la province a déterminé qu'il y a eu approximativement 600 introductions du SARS-CoV-2 au Québec (215). Un élément d'introduction majeur a été la semaine de relâche (216), qui s'est déroulée du 2 au 6 mars 2020 et durant laquelle de nombreux québécois ont voyagé. Moins de 100 introductions ont été recensées des suites de cette période mais on peut y retracer les origines de 52 à 75% des infections de la première vague (215).

Durant cette période, le projet CoVSeQ a vu le jour, sous l'égide du consortium CanCOGeN mentionné à la section 1.2.5.2. Cette biobanque de séquence québécoise est partenaire du Laboratoire de Santé Publique du Québec qui est en charge de séquencer les données virales

(https://covseq.ca).

1.2.5.2 La deuxième vague

La deuxième vague au Québec a débuté en août 2020, après un été avec relativement peu de cas. Les mesures de la santé publique du Québec ont donc été relâchées, permettant, entre autres, une reprise de l'enseignement primaire et secondaire en personne à partir de septembre. Le nombre de cas rapportés a augmenté dans les semaines suivantes. Malgré de nouvelles restrictions sanitaires mises en place pour réduire le nombre de transmission, le nombre de cas a continué d'augmenter tranquillement jusqu'à la fin novembre, puis plus rapidement jusqu'à atteindre un sommet de 17,778 cas actifs à la fin du mois de décembre 2020 (214).

1.2.5.3 Séquençage de données virales durant la deuxième vague

La deuxième vague au Québec est définie comme allant de la fin du mois d'août 2020 à la fin du mois d'avril 2021. Parmi les échantillons récoltés lors des tests de dépistage, 15 à 30% ont été séquencés de manière aléatoire. Certains échantillons ont toutefois été séquencés de manière prioritaire, notamment ceux des voyageurs, ceux des cas graves chez les jeunes patients (moins de 50 ans), les éclosions ainsi que les cas de réinfection ou d'infection post-vaccination. De plus, à partir du mois de février 2021, tous les échantillons ont été envoyés au criblage afin de détecter des mutations signatures des variants préoccupants, notamment sur la S la délétion 69/70, S:N501Y, S:E484K et S:L452R. Les échantillons contenant une des mutations ciblées par le criblage ont par la suite été séquencés de manière prioritaire afin de confirmer le variant. L'échantillon de séquences dans la base de données est donc biaisé vers les variants préoccupants à partir de février 2021 (217).

1.3 Problématique de recherche

La deuxième vague au Québec était un moment intéressant dans l'évolution du SARS-CoV-2. Le virus était en circulation depuis suffisamment longtemps pour avoir acquis une diversité virale, mais les VOCs et VOIs ne dominaient pas encore le paysage épidémiologique international. Cela dit, on remarquait à l'époque que plusieurs variants acquéraient des mutations récurrentes et on se demandait quel impact ces mutations pouvaient avoir sur la transmissibilité du virus et la sévérité de l'infection. Bien que cela semble évident en 2022, la fonction de ces mutations, ainsi que l'importance de certaines combinaisons de mutations n'étaient pas bien comprises en 2020. Notre

hypothèse de recherche est que ces mutations récurrentes procurent un avantage évolutif au virus parce qu'elles lui confèrent une meilleure capacité d'évasion immunitaire.

Afin de vérifier cette hypothèse, nous étudierons l'impact des substitutions d'acides aminés les plus fréquentes sur la reconnaissance des épitopes du SARS-CoV-2. De plus, nous caractériserons la diversité virale au Québec durant la deuxième vague et comparerons les mutations propres à la province aux mutations d'autres variants dans le monde. Finalement, nous étudierons les patrons de mutation intra-hôte d'individus immunosupprimés dans le cadre d'une infection de longue durée au SARS-CoV-2 afin de déterminer si leurs mutations *de novo* correspondent à celles présentes dans les VOCs et VOIs.

Chapitre 2 – Analyses de la diversité virale de SARS-CoV-2

Dans ce chapitre, je présente les résultats d'analyses qui ont été réalisées dans le contexte de deux projets distincts auxquels j'ai contribué durant mes études de maîtrise. D'une part, il reporte des résultats inclus dans un article que nous avons publié dans la revue *Cell Systems* en février 2022 (218) sur l'effet des patrons mutationnels du SARS-CoV-2 sur ses épitopes, où je suis deuxième auteure (voir Annexe A). D'autre part, il reporte des analyses qui seront incluses dans un manuscrit en préparation, duquel je serai co-auteure, sur la diversité virale au Québec durant la deuxième vague de la pandémie. La liste des auteurs et le journal envisagé n'est pas encore déterminé pour ce deuxième manuscrit.

Dans l'article sur les épitopes, nous avons déterminé *in silico* quels acides aminés étaient le plus susceptibles d'être introduits ou retirés des génomes du SARS-CoV-2, ainsi que leurs effets sur la présentation d'épitopes. L'un des objectifs était de mesurer l'impact des différents types de substitutions d'acide nucléiques sur le patron mutationnel des résidus parmi les génomes soumis à GISAID lors de la première vague. Nous avions besoin de séquences virales dont les mutations étaient le fruit d'une évolution neutre afin d'établir à quel point nos résultats étaient inattendus sous neutralité. Au cours d'un stage réalisé à l'été 2020, avant mon entrée à la maîtrise, j'ai développé une expertise sur les simulations génomiques du SARS-CoV-2 à l'aide du logiciel SANTA-SIM (S-S). Les résultats présentés 2.2.1.1 proviennent du rapport de ce stage et mettent en contexte les analyses contenant des génomes simulés sous neutralité présentés en 2.2.1.2, réalisées durant ma maîtrise. Ces données simulées ont permis de déterminer que les acides aminés préférentiellement introduits ou retirés au-delà d'une certaine fréquence ne s'expliquaient pas par évolution neutre. De plus, j'ai réalisé d'autres simulations sous divers scénarios évolutifs afin de déterminer expérimentalement quelles substitutions d'acide nucléiques causent les mutations qui ont un effet sur la présentation d'épitopes. Les autres analyses de cet article ont été générés par David Hamelin dans le cadre de sa maîtrise.

Les résultats provenant du manuscrit en préparation sur la diversité virale au Québec décrivent deux variants qui circulaient dans la province durant la deuxième vague. Plusieurs personnes de mon groupe de recherche ont été impliqué dans le contrôle de qualité et l'analyse des séquences consensus, alors que je me suis focalisée sur le sous-groupe de séquences appartenant à ces deux

variants. De plus, j'ai identifié des mutations convergentes qui sont apparues dans un scénario évolutif semblable dans les deux variants Québécois ainsi qu'un variant international.

Ce chapitre est séparé en deux sections qui combinent les deux projets mentionnés ci-haut : la section "Méthodes" reprend toutes les méthodologies pour la préparation des données, analysées à la section "Résultats".

2.1 Méthodes

2.1.1 Simulations génomiques

L'outil SANTA-SIM (205) a été utilisé pour générer les génomes simulés avec les paramètres notés au Tableau 2.1.

Nom du paramètre	Valeur	Explications
Population Size	10 000	Taille de la population générée à chaque nouvelle génération – paramètre par défaut
Mutation Rate	2.04E-6	Obtenu à l'aide de la formule suivante: Taux de mutation par an/nombre de générations par an * taille du génome 24.5/400 * 29,903
Transition Bias	3.0	Estimation du biais de transition versus transversion à partir des données réelles puis confirmé *in silico* (voir figure 2.3)
Generation Count	400	Équivalent d'un an d'évolution du virus

Tableau 2.1 - Paramètres de SANTA-SIM

Les taux de mutation rapportés pour SARS-CoV-2 dans la littérature sont calculés par an à partir d'analyses phylogénétiques, mais S-S demande un taux de mutation par génération par paires de base, il a donc fallu déterminer combien de générations simulées équivaudrait à un an d'évolution. Basé sur une étude précédente utilisant S-S pour simuler des génomes d'influenza A/H3N2, ayant un taux de substitution par an à peu près deux fois plus élevé que celui du SARS-CoV-2 (A/H3N2: ~48.824, SARS-CoV-2: ~24.5) (61,219), nous avons choisi 400 pour le nombre de générations pour représenter une année d'évolution. À noter que le taux de substitution par an pour SARS-CoV-2 a fluctué entre 21 et 33 mutations par an depuis le début de la pandémie, la valeur choisie reflète le taux observé lors de la 2e vague. Nous avons généré 10 réplicats pour les scénarios évolutifs sous neutralité sans biais de substitution et avec les biais de substitution G > U, et 100 réplicats pour les scénarios avec un biais de substitution C > U afin d'avoir une meilleure puissance statistique. Les biais de substitution C > U et G > U sont bien caractérisés dans la littérature et

reflètent donc plus fidèlement l'évolution du virus (58,59). Les tests statistiques ont été réalisés en comparant les résultats obtenus aux simulations neutres qui servaient de contrôle à l'aide d'un test bilatéral indépendant.

2.1.2 Traitement des données de séquençage et génération de séquence consensus

Les librairies de séquençage ont été préparées avec les *primers* ARTIC (v3). Les échantillons séquencés par Illumina ont été préparés avec le kit Nextera Flex.

2.1.2.1 Illumina

Pour les échantillons séquencés par illumina, les *reads* bruts ont été élagués avec cutadapt (v2.10) (220), puis alignés à la référence Severe Acute Respiratory Syndrome Coronavirus-2 isolate Wuhan-Hu-1 (GenBank MN908947.3) (19) à l'aide du logiciel bwa-mem (v0.7.17) (221). Les *reads* alignés ont par la suite été filtrés à l'aide de sambamba (v0.7.0) (222) afin d'enlever les alignements de mauvaise qualité, c'est-à-dire les *reads* non-alignés, les alignements secondaires et *reads* appariés dont la longueur de la séquence entre les deux amorces ne se situe pas entre 60 et 300 nucléotides. Les amorces ARTIC restantes ont été élaguées avec iVar (v.1.3.4) (223). Des fichiers pileup, à savoir un fichier dans lequel le nombre des quatre nucléotides possibles est indiqué pour chaque position dans le génome, ont été produits avec Samtools (v1.9) (224). Ces fichiers ont par la suite été utilisés en entrée pour le logiciel FreeBayes (v1.2.2) (225) pour générer les séquences consensus ainsi que l'annotation de variants.

2.1.2.2 Oxford Nanopore Technologies

Dans le cas des échantillons séquencés par Nanopore, les bases du signal brut ont été annotées avec guppy (v3.4.4) (226) en utilisant l'algorithme High-Accuracy Model (dna_r9.4.1_450bps_hac). Les *reads* dont la taille ne tombait pas dans la fourchette attendue de 400 à 700 nucléotides ont été exclus de l'analyse. Les *reads* ont par la suite été alignés à la référence (Wuhan-Hu-1 (19)) à l'aide de minimap2 (v.2.17) (227) puis les *reads* alignés à la mauvaise amorce ont été enlevés. De plus, pour les régions dans lesquelles la profondeur de la couverture était supérieure à 800X, un sous-échantillonnage aléatoire a été pris pour ne conserver que 800 *reads* au maximum. Finalement, Nanopolish (v0.13.1) (228) a été utilisé pour annoter les allèles alternatifs dans les régions avec au moins 8X de profondeur par brin et flanquées d'au moins 10 nucléotides. Afin de générer les

séquences consensus, les allèles alternatives de Nanopolish ont par la suite été intégrées à la référence à l'aide de bcftools (v1.9) (224) si elles provenaient de régions avec une profondeur d'au moins 20X.

2.1.3 Bases de données virales

Les séquences virales annotées comme lignée Pango B.1.160, B.1.1.176 et B.1.1.317 (n=32,640, n=2,616 et n=2,244, respectivement) ont été téléchargées de GISAID en date du 01/11/2021 et de la base de données du Laboratoire de Santé Publique du Québec (LSPQ) en date du 16/11/2021. Les données du LSPQ sont fournies par nos collaborateurs dans le cadre du partenariat CoVSeQ.

2.1.4 Reconstruction phylogénétique

Les arbres phylogénétiques ont été générés par l'outil Nextstrain Viewer en utilisant les paramètres par défaut (61). Les séquences désignées comme étant des lignées Pango B.1.1.176 et B.1.160 (PangoLearn version 2021-11-09, Pangolin version 1.2.93) ont été extraites de la base de données du LSPQ pour leur arbre respectif. Afin de s'assurer de la qualité des séquences et d'éviter les doublons, seules les séquences ayant été soumises à GISAID ont été sélectionnées.

2.2 Résultats

2.2.1 Simulation de génome du SARS-CoV-2 sous neutralité

2.2.1.1 Patrons de substitution

2.2.1.1.1 Distance entre les positions pluri-alléliques

S-S étant un logiciel relativement nouveau et peu cité dans la littérature, nous avons tout d'abord voulu effectuer quelques tests pour nous assurer du bon fonctionnement du programme. En premier lieu, nous avons vérifié que nos simulations effectuées sous neutralité généraient réellement des mutations à des sites aléatoires dans le génome. À ces fins, nous avons étudié la distance entre les évènements de substitution dans les génomes simulés.

Après avoir effectué 10 réplicats de simulations sur 400 générations, la distance génomique entre les positions bi-alléliques, tri-alléliques et quadri-allélique a été mesurée afin de déterminer si certaines positions étaient plus mutées que les autres ou si certains segments du génome étaient conservés. Parce que la probabilité qu'un acide aminé donné soit mutée est constante à travers le génome, on prévoit que le nombre de positions mutées une fois soit plus élevé que le nombre de positions mutées deux fois, qui devrait être à son tour plus élevé que le nombre de positions

mutées trois fois. Tel qu'attendu, les positions bi-alléliques sont les plus fréquentes (Figure 2.1 a). La majorité surviennent à une distance de 1 à 10 bases, avec une distance moyenne de 5,40 bases parréplicat, et elles sont toutes à moins de 50 bases les unes des autres. Ceci nous indique que les mutations surviennent de manière stochastique dans les génomes simulés sous neutralité. Un total de 55 346 positions bi-alléliques ont été recensées au total. Les positions tri-alléliques se situent majoritairement à unedistance de moins de 200 bases, avec une pointe dans la classe des 50 à 75 bases (Figure 2.1 b). Oncompte 3 529 positions tri-alléliques et la distance moyenne entre ces positions est de 84,25 bases(Figure 2.1 c). Tel qu'attendu, les positions quadri-alléliques sont les moins fréquentes, avec 66 positions au total dans notre jeu de données simulé, et une distance moyenne de 3 575,5.

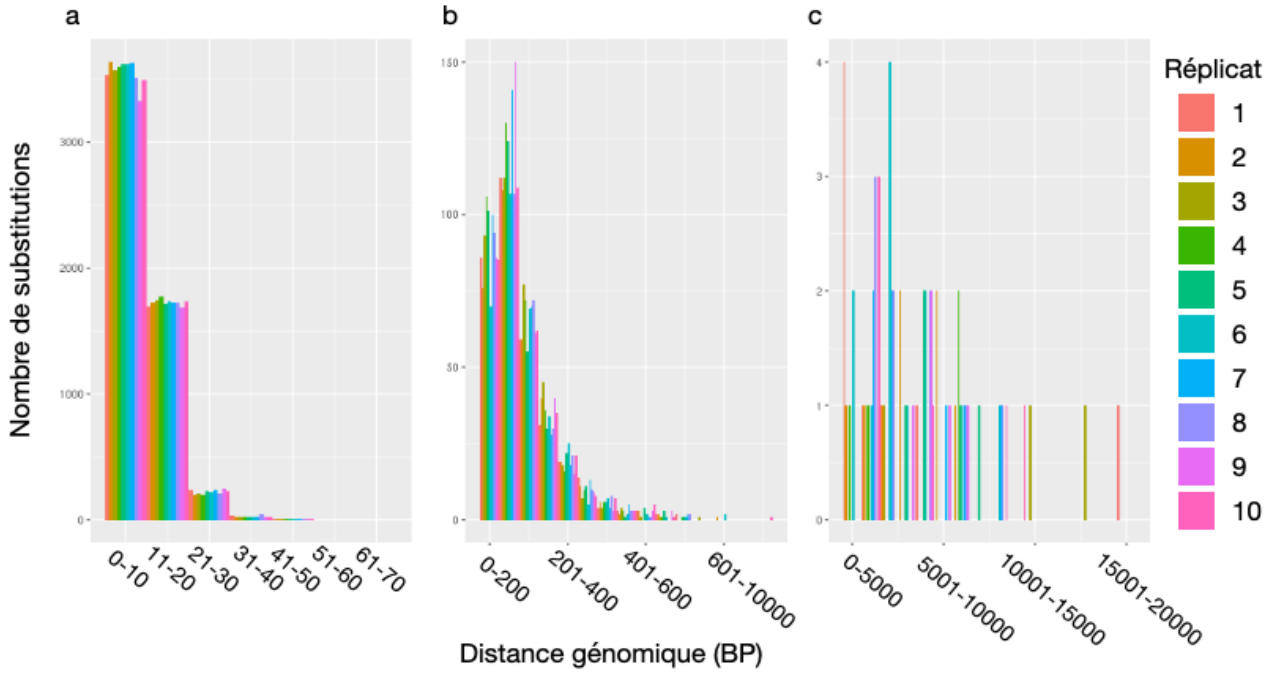


Figure 2.1 - Distance entre les positions pluri-alléliques de génomes simulés sous neutralité. Biais de transition = 3.0. **a**: positions bi-allèlique, amplitude de classe: 10; **b**: positions tri-alléliques, amplitude de classe: 50; **c**: positions quadri-alléliques, amplitude de classe: 1000.

2.2.1.1.2 Nombre de transitions versus transversions

Ensuite, nous voulions vérifier que la fonction qui gère le biais de transition fonctionne tel qu'attendu. Le nombre moyen des différentes substitutions possibles, selon le nucléotide initial et celui muté, a été calculé au cours des générations dans le cadre d'une évolution sous neutralité (Figure 2.2 a). On remarque que les types de substitution se regroupent en quatre grappes

distinctes. Les plus fréquentes sont U > C et A > G, suivies de G > A et C > U. Le troisième groupe comprend U > G, U > A, A > C et A > U. Finalement, le groupe avec les substitutions les moins fréquentes est constitué de G > U, G > C, C > G et C > A. Ces résultats correspondent aux patrons de substitution attendu sous neutralité avec biais de transition, c'est à dire que les transitions sont plus fréquentes et que les transversions, et les substitutions à partir des nucléotides enrichis dans un génome enrichi en A et U, soit que les mutations à partir de A et U, soient plus fréquentes que celles à partir de C et G. Nous voulions ensuite faire la même expérience mais en augmentant le taux de substitution C > U comparativement aux autres substitutions afin de vérifier le patron de substitution généré par S-S dans ce scénario qui se rapproche de celui observé dans les données réelles (Figure 2.2 b).

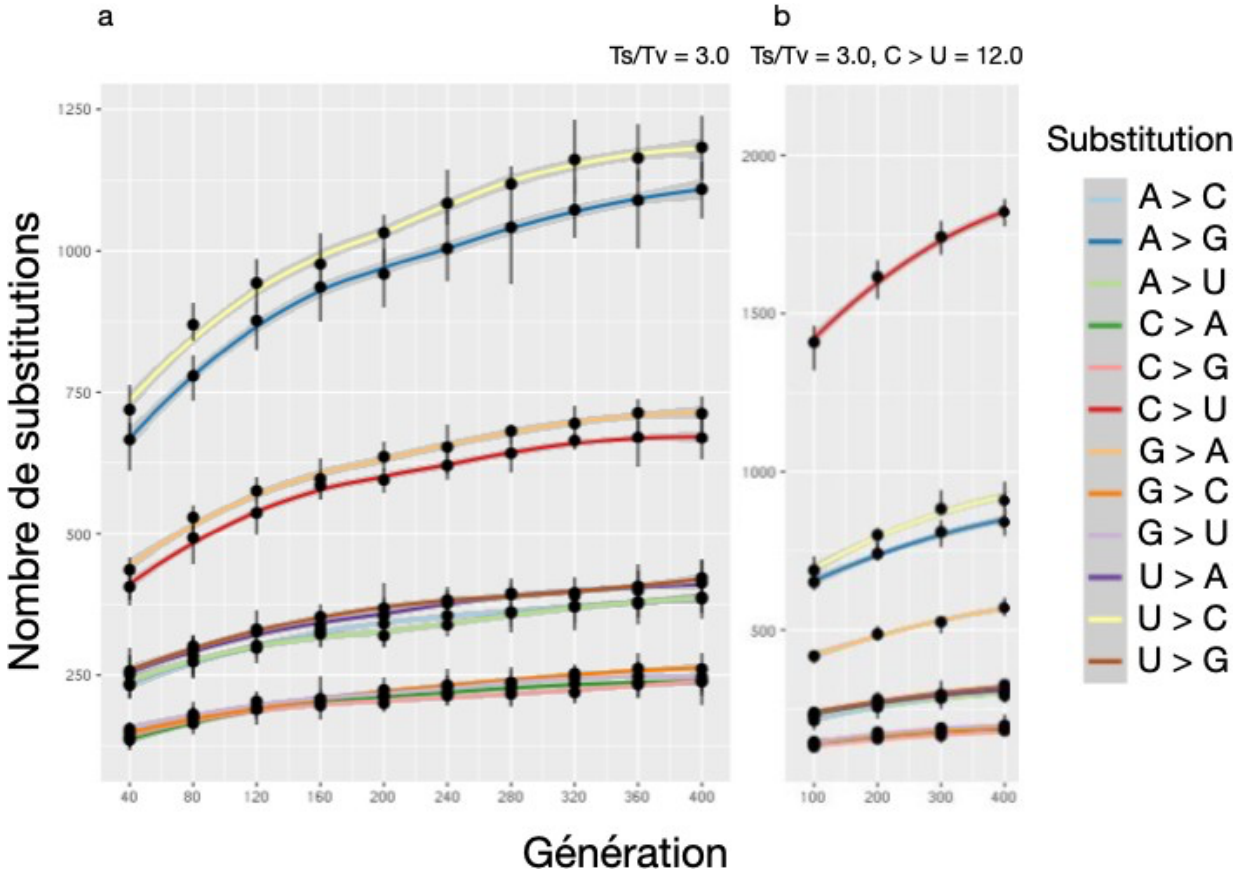


Figure 2.2 - Substitutions dans des génomes du SARS-CoV-2 simulés sous neutralité. Nombre de substitutions moyen pour n = 10 réplicats, biais de transition = 3.0. **a**: simulation neutre. **b**: simulation neutre avec taux de substitution C > U à 12.0.

2.2.1.2 Utilisation de génomes simulés comme contrôle lors d'analyse sur des donnéesréelles

Une fois les fonctions de S-S comprises et testées, il était possible de générer des génomes

simulés sous diverses conditions et de les utiliser comme contrôle dans des analyses *in silico*. L'étude décrite ci-dessous décrit l'impact des patrons de mutations du SARS-CoV-2 sur la liaison de ses épitopes aux lymphocytes T de l'hôte. Il a été démontré que la protection attribuable aux CD4+ et CD8+ mémoires dure plus longtemps que celle apportée par les anticorps (229). Comprendre la manière dont les mutations du SARS-CoV-2 affectent la liaison des épitopes permet de mieux prédire quels variants seront en mesure d'échapper aux réponses immunitaires développées dans le cadre d'une infection au virus ou de la vaccination.

2.2.1.2.1 Introduction ou retrait préférentiel de certains acides aminés dans les épitopesdu SARS-CoV-2

L'impact de diverses mutations dans les épitopes présentés par les CMH de type I identifiés par Quadeer et al. (230), a été mesuré par Hamelin et al. (218) (voir article en Annexe A). Les mutations non-synonymes réduisent l'affinité de liaison pour 31% et 10% des d'épitopes CD8+T des protéines S et N, respectivement. Au vu de ce résultat, on peut se demander si cette baisse d'affinité est causée par un patron mutationnel en particulier.

Taux global de substitution de résidus

Le *global residue substitution output* (GRSO) est un score qui rapporte la différence en du nombre de mutation introduisant un acide aminé X et du nombre de mutations qui enlève cet acide aminé X, sur l'ensemble des mutations de GISAID trouvées dans au moins quatre séquences. Ce score est calculé pour chaque acide aminé. Dans notre étude, le GRSO du protéome du SARS-CoV-2 a été calculé pour les mutations rapportées à diverses fréquences dans la base de données de GISAID durant la première année de la pandémie. En regardant cet ensemble de mutations, on remarque que certains acides aminés sont plus fréquemment mutés que d'autres; ils sont donc préférentiellement retirés du protéome (Figure 3). C'est le cas des Alanines (A), des Prolines (P) et des Thréonines (T). Au contraire, certains acides aminés sont introduits dans le protéome de manière préférentielle, notamment les Isoleucines (I), les Leucines (L), les Tyrosine (Y) et les Phenylalanines (F). Afin de vérifier si ce patron de mutation peut s'expliquer par les substitutions découlant uniquement d'une évolution sous neutralité avec biais de transition telle que décrite plus tôt, le GRSO observé a été comparé à celui obtenu à partir des génomes simulés sous neutralité (n=1000, 10 réplicats). On remarque que, pour les mutations trouvées dans plus de 100 génomes réels, les fréquences des acides aminés préférentiellement introduits ou retirés ont été multipliées ou divisées par un facteur (ou *fold change*) de plus de 4, correspondant à une valeur P de $1x10^{-11}$, calculée à partir des données simulées (Figure 2.3). De plus, le patron de substitution pour les

mutations trouvées dans un seul génome (aussi sommé singletons) et dans 2 à 100 génomes est presqu'identique à celui des simulations neutres. Ce résultat suggère que ces mutations peu fréquentes sont le fruit de processus stochastiques associés à l'évolution sous neutralité, mais les singletons peuvent aussi être enrichis pour des erreurs dans les séquences consensus de GISAID.

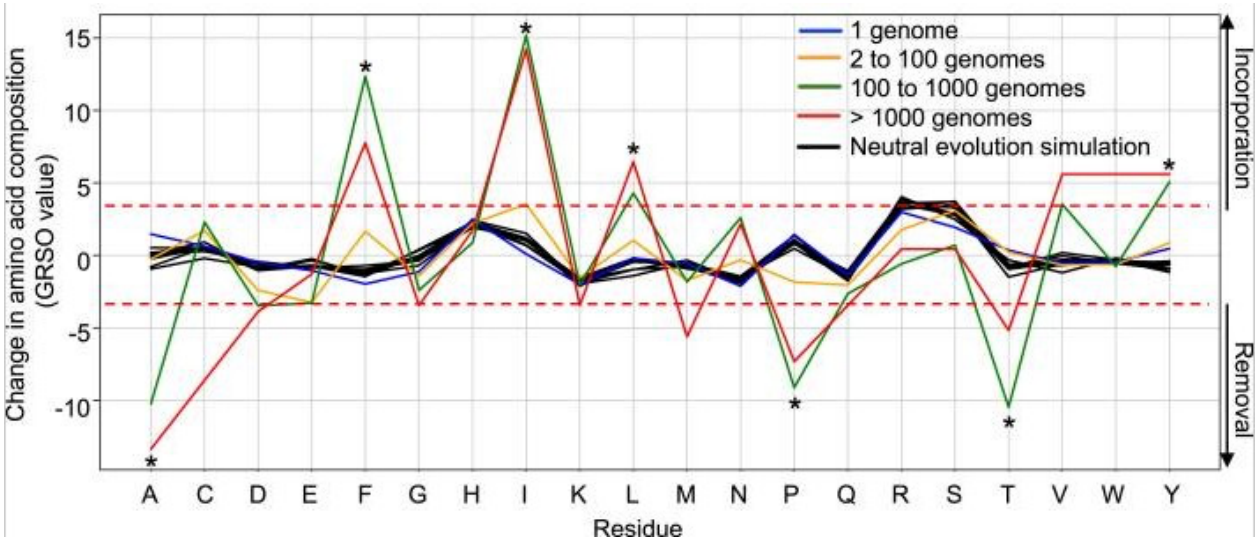


Figure 2.3 - Biais d'acides aminés dans le protéome du SARS-CoV-2.
Tirée de Hamelin et al. Protéomes de 330,246 génomes du SARS-CoV-2. Les acides aminés (axe des X) sont présentés selon la différence en pourcentage de leur introduction ou retrait après mutation (axe des Y). Le biais mutationnel est calculé pour les mutations survenant à diverses fréquences ainsi que dans les simulations génomiques neutres (ts/tv = 3.0), qui servent de contrôle pour déterminer la significativité statistique des patrons observés dans les données réelles. Les lignes rouges hachées montrent les valeurs limites (fold-change >4; valeur p < 1×10−11) utilisées pour définir les résidus préférentiellement introduits ou retirés (astérisques) (218).

Le rôle du biais de substitution C > U dans le patron mutationnel des épitopes des cellules T

Nous avons mesuré l'effet des mutations des acides aminés préférablement retirés ou introduits dans les épitopes des cellules T à l'aide de netMHCpan (212,218) (voir Annexe A). L'introduction préférentielle de F, I, L et Y, et le retrait favorisé de A, P et T affecte la liaison au CMH de manière positive ou négative pour divers supertypes d'antigènes des leucocytes humains (*human leukocyte antigen*, HLA). En effet, bien que les retraits de P représentent 9.1% des mutations à travers le génome du SARS-CoV-2, 31% des mutations qui abrogent lesépitopes sont des mutations de P vers un autre acide aminé. Spécifiquement, la capacité de liaison des CMH de classe HLA B7 aux épitopesqu'ils présentent est réduite par le retrait des P, et nos analyses prédisent que 62% des mutations affectant cette liaison est attribuable aux substitutions C > U (218). Ce résultat suggère que ce patron desubstitution contribue à l'évasion immunitaire.

Considérant les biais mutationnels de C > U et de G > U bien documentés dans la littérature (58,59), on peut se demander si ce biais a un effet sur les mutations causant une différence d'affinité entre CMH et épitopes. Des simulations génétiques ont été réalisées en testant différents paramètres pour les taux de substitution de C > U et G > U, en gardant constants les taux de transitions et transversions pour les autres combinaisons de nucléotides. Augmenter le taux de substitution de C > U de 15 et 20 fois par rapport au taux de base résulte en une valeur GRSO similaire à celle observée dans les données réelles (Figure 2.4 a). Notamment, ce biais a mené au retrait préférentiel des A, P et T de 6,7% ($p = 5.1x10^{-11}$), de 6,9% ($p=1.2x10^{-11}$) et de 8% ($p=4.8x10^{-12}$), respectivement, ainsi qu'à l'introduction préférentielle des I et des F de 8,2% ($p=1.3x10^{-8}$) et de 5,2% ($p=4.3x10^{-11}$) dans les génomes simulés (Figure 4a). Augmenter le biais de substitution G > U favorise l'introduction des F et L (Figure 2.4 b). On peut donc conclure que le taux élevé de substitution C > U contribue de façon importante au patron de mutation des acides aminés du SARS-CoV-2, et celui du G > U dans une moindre mesure.

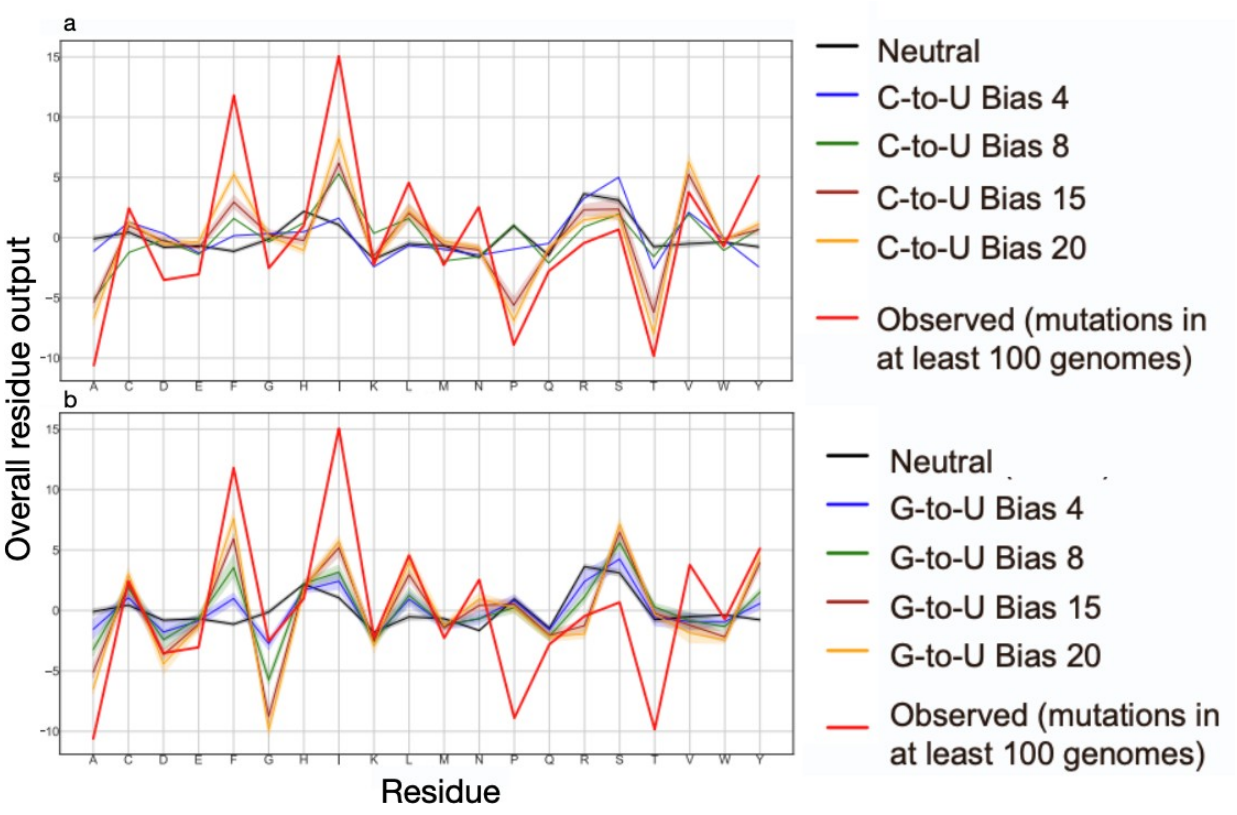


Figure 2.4 - Comparaison du biais de mutation observé dans les données réelles et les génomes simulés avec divers biais de substitution
Modifiée de Hamelin et al. (218) Le biais de transition général est de 3.0. Ceux de C > U et de G > Uvont de 4 à 20 pour les panneaux a et b, respectivement.

2.2.2 Portait de la deuxième vague au Québec

En août 2020, la majorité des séquences de SARS-CoV-2 avec une assignation Pango significative, c'est-à-dire qui n'est pas B.1 ou B.1.1, sont annotées comme B.1.157 et B.1.1.176 (Figure 2.5). Ces deux variants étaient présents en Ontario et au Québec durant la première vague, respectivement (215,231). B.1.160, aussi appelé Marseille-4 (232), a été introduit via la France en septembre (231). À partir de la mi-novembre 2020 jusqu'à la fin du mois de février 2021, autour de 50% des échantillons de SARS-CoV-2 séquencés au Québec sont annotés comme étant B.1.160 ou B.1.1.176.

La deuxième vague s'est terminée à la fin du mois de mars 2021, tout juste après l'introduction du VOC Alpha. Le dernier mois de cette vague a été caractérisé par une augmentation drastique du nombre de cas, la majorité de la lignée Alpha (Figure 2.5), accompagnée de la réduction presque totale des autres variants circulant au Québec.

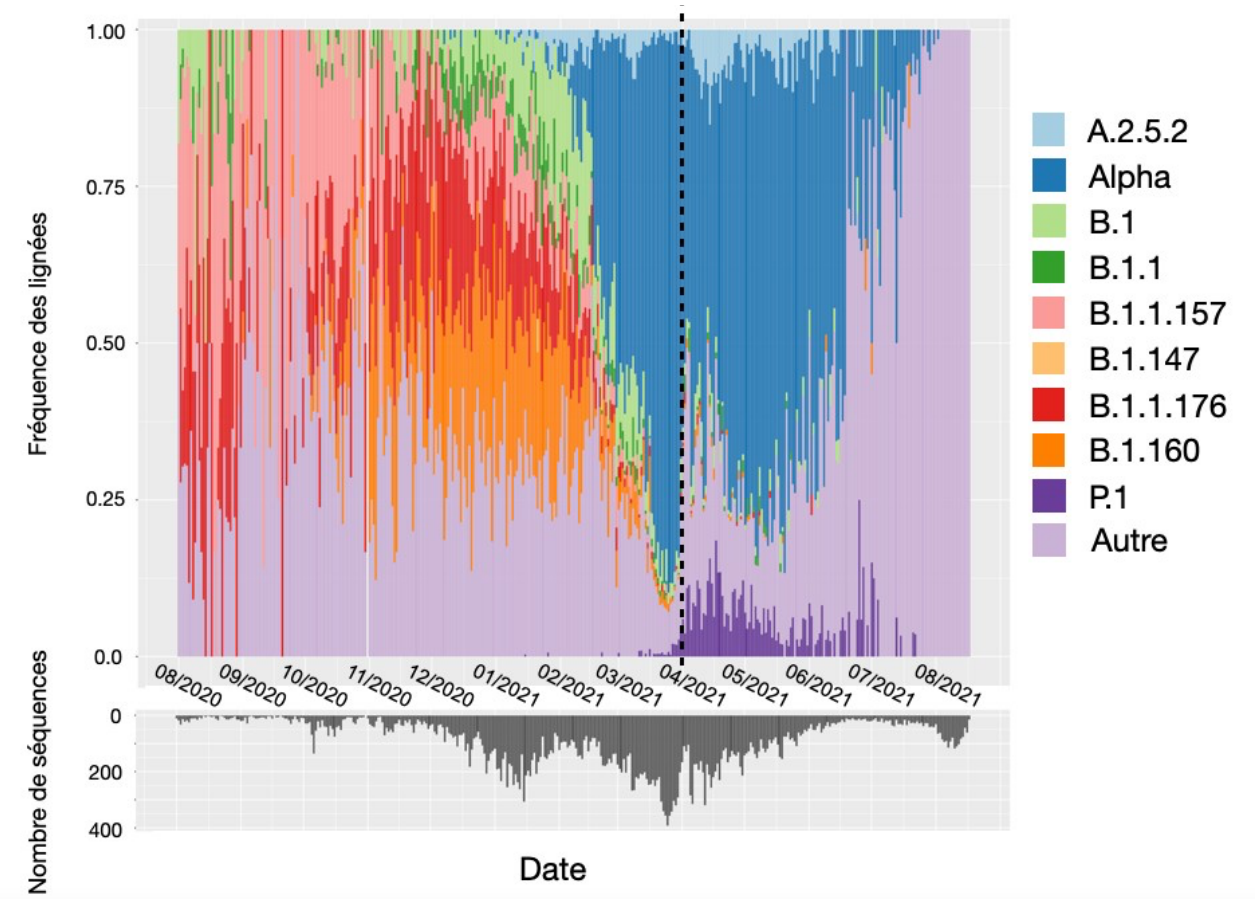


Figure 2.5 - Fréquences des lignées au Québec de août 2020 à août 2021.
Inclut la deuxième vague de la pandémie dans cette région (à gauche de la ligne pointillée). Le panneau du haut montre la fréquence quotidienne des neuf lignées les plus courantes et celui du bas le nombre de séquences échantillonnées, utilisées pour obtenir la fréquence.

2.2.2.1 B.1.1.176, un variant unique au Québec

B.1.1.176 a été détecté pour la première fois en mars 2020 et la dernière entrée dans GISAID pour ce variant date de juillet 2021. Au total, 94% des séquences B.1.1.176 proviennent du Québec et la majorité des cas de B.1.1.176 à l'extérieur du Québec proviennent de l'Ontario et du Nouveau Brunswick (Table 2), les deux provinces voisines. Le faible nombre de séquences de ce variant à l'extérieur du Québec suggère que les mesures visant à restreindre le déplacement entre les provinces ont réussi à limiter la propagation de B.1.1.176 au Canada.

Région	Nombre de séquences	Pourcentage
Québec	2659	94.4%
Ontario	90	3.2%
Nouveau Brunswick	47	1.7%
Colombie-Britannique	7	0.2%

Nouvelle Écosse	5	0.2%
Europe	2	0.07%
USA	1	0.03%
Asie	1	0.03%
Total	2816	

Tableau 2.2 - Répartition des cas de B.1.1.176 répertoriés dans GISAID.

La liste complète des mutations de ce variant se trouve à la Figure 2.6a. Mis à part la mutation S:D614G qui provient d'une lignée parentale, ce variant comporte deux mutations sur la protéine S, S:T20I et S:R357K, cette dernière est située dans le RBD. Ces deux mutations, détectées pour la première fois en mars 2020, sont apparues soit en même temps, soit dans un laps de temps très court l'une de l'autre puisqu'il n'y a aucune séquence qui contient une de ces mutations sans l'autre. On peut voir ces mutations sur les feuilles se terminant par des cercles dans l'arbre phylogénique de la Figure 2.6b. Le prochain événement mutationnel majeur est l'acquisition de N:A211V en juin 2020. Tel que mentionné à la section 1.2.6.2, il y a eu peu de cas rapportés de la COVID-19 au Québec durant l'été 2020 et par conséquent, il y a eu encore moins d'échantillons séquencés. La sous-lignée du variant B.1.1.176 contenant les mutations S:T20I, S:R357K et N:A211V est la seule sous-lignée qui circulait durant l'été et celle-ci a continué de se propager durant la deuxième vague. Il est difficile de déterminer si l'acquisition de ces trois mutations confère un avantage sélectif au variant ou si le succès de cette sous-lignée fût le fruit d'une éclosion non contrôlée durant l'été. Cela dit, le fait qu'elle ait été capable de survivre un été avec peu de transmission communautaire contrairement à d'autres variants de la première vague au Canada (231) suggère que le premier scénario est plus probable.

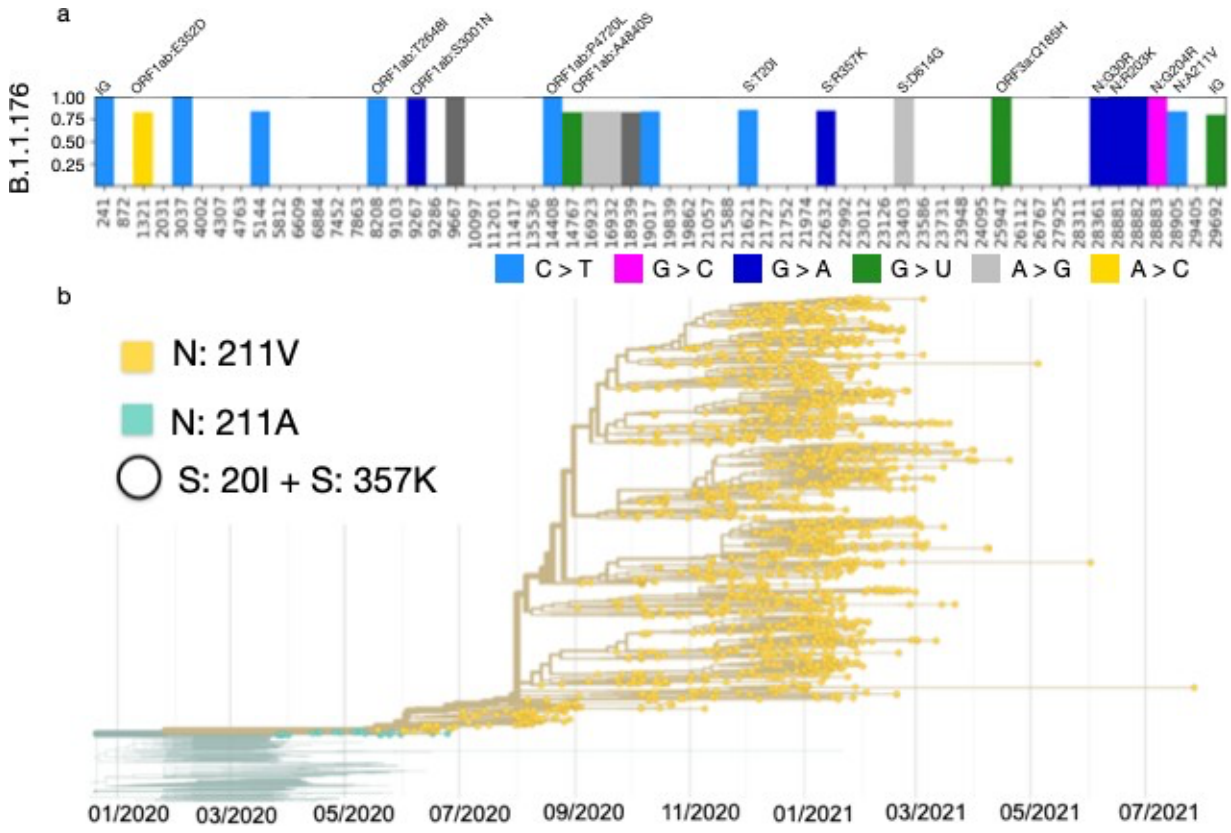


Figure 2.6 - Variant Québécois B.1.1.176.
a: fréquences des substitutions caractéristiques de B.1.1.176 pour 2,224 séquences annotées tel quel dans GISAID. Les mutations IG sont intergéniques. **b**: arbre de temps généré par Nextstrain viewer à partir de 2,979 séquences B.1.1.176 provenant du LSPQ.

2.2.2.2 B.1.160

Le premier échantillon B.1.160 a été séquencé au Québec en septembre 2020. Au Québec, 98.7% des séquences B.1.160 ont la mutation synonyme C222T, comparativement à 11% des séquences qui se trouvaient dans GISAID en date du téléchargement du jeu de données. Des 130 séquences B.1.160 + C222T dans GISAID datant du mois de septembre 2020 ou plus tôt, 123 proviennent de France et 7 proviennent d'autres pays Européens (Royaume-Uni, Italie, Danemark et Suisse). Ces données corroborent les résultats de McLaughlin et collègues, qui ont déterminé que B.1.160 a été introduit au Québec via la France (231). Comme C222T est une mutation synonyme, son omniprésence au Québec est probablement dû à un effet fondateur plutôt qu'à un avantage sélectif. B.1.160 acquiert également la mutation S:T20I au Québec uniquement. Les fréquences des mutations de ce variant au Québec comparées à l'ensemble des séquences B.1.160 sur GISAID se trouve à la Figure 2.7 a. On remarque la présence de la mutation S:S477N, qui est une mutation

caractéristique de ce variant. S:S477N est une mutation convergente apparue dans plusieurs variants à l'été 2020 (233). Cette mutation promeut l'évasion immunitaire et confère une résistance à certains mAbs (234). Un arbre de distance utilisant les 1,565 séquences B.1.160 du LSPQ a été généré (Figure 2.7 b). On remarque qu'une autre sous-lignée de ce variant sans C222T a été introduite dans la province plus tard. Cette sous-lignée comporte 23 mutations additionnelles et a donné lieu à une courte éclosion en février et en mars 2021.

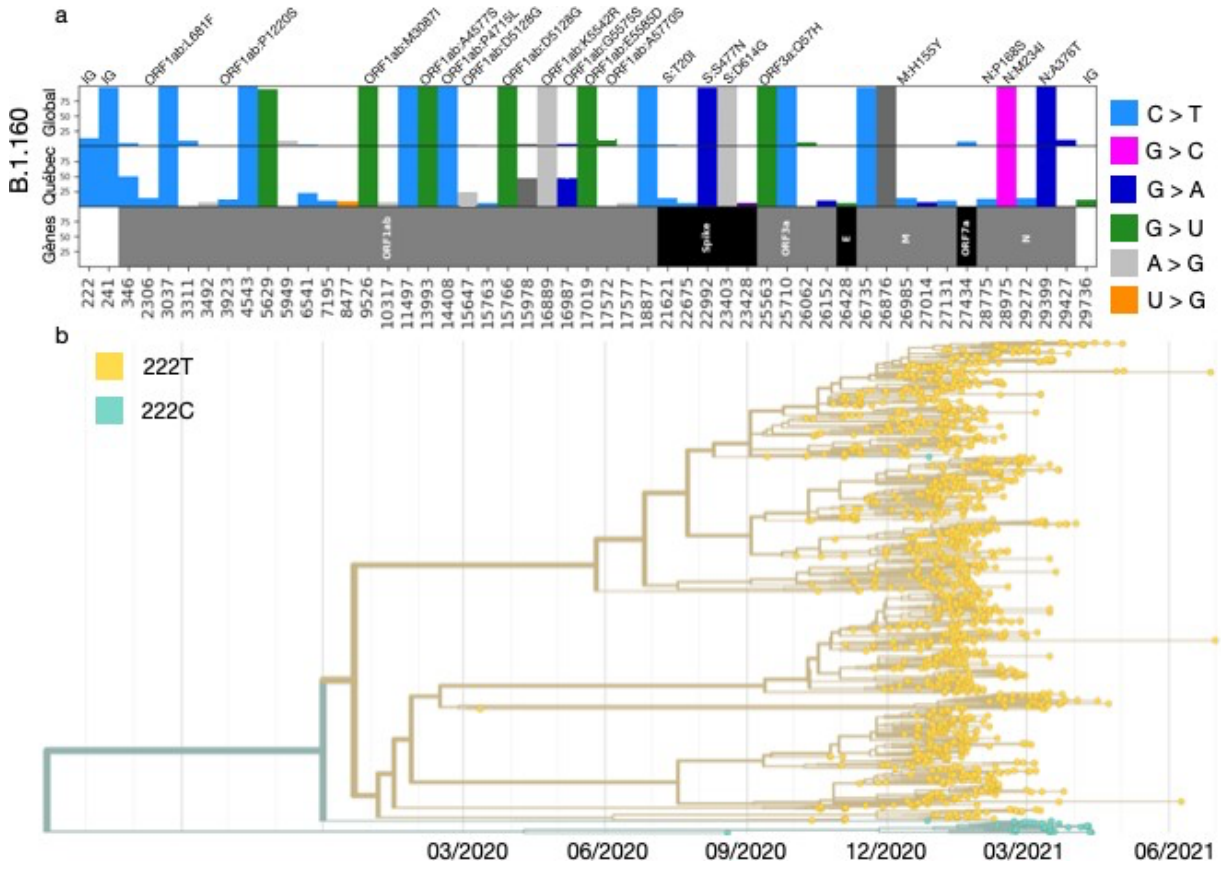


Figure 2.7 - Variant B.1.160 au Québec.
a: fréquences des substitutions pour toutes les séquences B.1.160 dans GISAID (n=32,640) versus celles du Québec (n=2,627). **b**: arbre de temps pour 1,564 séquences B.1.160 du LSPQ.

2.2.2.3 Mutations convergentes

2.2.2.3.1 Similarité entre B.1.1.176 et B.1.1.317

La sous-lignée B.1.1.176 est similaire au variant Russe B.1.1.317, un autre variant des première et seconds vagues qui circulait beaucoup en automne 2020 et en hiver 2021. B.1.1.317 a acquis la mutation N:A211V au printemps 2020, suivi de quatre mutations sur la protéine S qui ont été

détectées pour la première fois en été 2020 (S:D138Y, S:S477N , S:Q675R et S:A845S) (235) (Figure 2.8). La Russie a également eu moins de cas reportés à l'été, suivie d'une expansion des cas à l'automne. La sous-lignée de B.1.1.317 contenant N:A211V et les quatres mutations sur la S a pris de l'ampleur jusqu'à atteindre ~25% des échantillons séquencés au début de l'hiver 2021. Tout comme B.1.1.176, la fréquence de ce variant diminue drastiquement avec l'arrivée d'Alpha en Russie en avril 2021.

En plus d'une mutation convergente sur la N, B.1.1.317 et B.1.1.176 ont en commun l'apparition de mutations sur la S, incluant une dans son RBD, ainsi qu'une chronologie similaire. Bien que les mutations acquises sur la S ne soit pas les mêmes, elles coïncident avec une expansion de ses deux sous-lignées suite à un été avec peu de transmission.

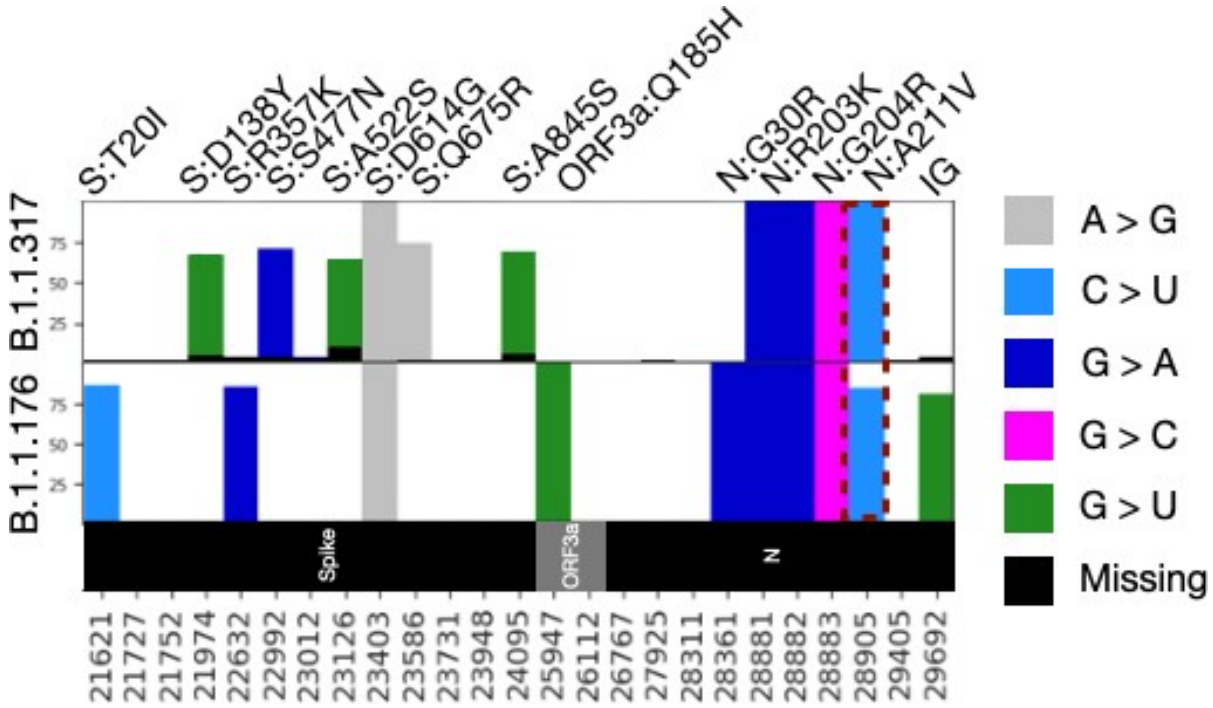


Figure 2.8 - Comparaison des variants B.1.1.176 (Québec) et B.1.1.317 (Russie).
La mutation convergente N:A211V est encadrée en rouge.

2.2.2.3.2 L'acquisition de S :T20I par B.1.1.176 et B.1.160

La sous-lignée B.1.160 +C222T au Québec a acquis la mutation S:T20I vers la fin octobre 2020 (date inférée par vraisemblance maximale (236): 23/10/2020, première séquence détectée 27/10/2020). Il s'agit du deuxième variant québécois qui développe cette mutation, le premier étant

B.1.1.176 en mars 2020. La présence de cette mutation convergente est une observation intéressante puisque cette mutation n'était pas fréquente à l'époque. Des 1.7 millions d'échantillons collectés avant avril 2021 sur GISAID, seulement 6550 ont S:T20I, dont 2,632 au Québec. Ce site est toutefois connu pour muter vers d'autres acides aminés que I. Parmi les échantillons de GISAID récoltés avant avril 2021, la mutation la plus fréquente à cette position est S:T20N (n=15,457, VOI Gamma). On retrouve aussi S:T20R (n=1), S:T20S (n=19), S:T20A (n=20), S:T20P (n=70) et S:T20F (n=5). Bien qu'on ne peut exclure la possibilité que cette mutation convergente soit due au hasard, on peut se demander si S:T20I confère un avantage sélectif pour SARS-CoV-2 dans le contexte québécois par exemple par rapport à l'environnement (température, humidité) (237,238).

2.2.3 Conclusion

En conclusion, j'ai utilisé, au fil des analyses présentées dans ce chapitre, diverses méthodes pour étudier l'évolution virale. En particulier, les simulations génomiques sont un outil intéressant pour tester des hypothèses qu'il serait difficile de tester avec des génomes réels. Également, l'étude de mutations récurrentes en début de pandémie permet d'identifier des positions clés dansle génome qui peuvent orienter les efforts de surveillance des variants futurs.

Chapitre 3 – Intra-host viral populations of SARS-CoV-2 in patients with hematological cancers

by

Dominique Fournelle M.Sc, Fatima Mostefai, Elsa Brunet-Ratnasingham, Raphael Poujol, Jean-Christophe Grenier, José Héctor Gálvez, Amélie Pagliuzza, Inès Levade, Sandrine Moreira, Simon Grandjean Lapierre, Nicolas Chomont, Daniel E. Kaufmann, Morgan Craig Ph.D., Julie G. Hussin

Author contributions

DF, MC and JGH conceived the study, interpreted the results and wrote the manuscript. DF, RP, JCG, JHG performed bioinformatics analyses and drafted methods sections. FM, RP and JCG created and maintained the intra-host database. EBR and DK recruited Q1 and sampled SARS-CoV-2 material. EBR, SGL, NC and AP performed experiments on Q1 SARS-CoV-2 temporal samples. IL and SM supervised the CoVSeQ initiative which sequenced Q1's SARS-CoV-2 samples. All authors revised and approved the final version of the manuscript.

3.1 Abstract

Throughout the SARS-CoV-2 pandemic, several variants of concern (VOC) have been identified, many of which share recurrent mutations in the spike protein's receptor binding domain (RBD). This region coincides with known epitopes and can therefore have an impact on immune escape. Protracted infections in immunosuppressed patients have been hypothesized to lead to an enrichment of such mutations and therefore drive evolution towards VOCs. Here, we show that immunosuppressed patients with hematologic cancers develop distinct populations with immune escape mutations throughout the course of their infection. Notably, by investigating the co-occurrence of substitutions on individual sequencing reads in the RBD, we found quasispecies harboring mutations that confer resistance to known viral monoclonal antibodies (mAbs) such as S:E484K and S:E484A. Neither patient was treated with antiviral monoclonal antibodies, yet they developed resistance mutations. Furthermore, we provide the first evidence for a viral reservoir based on intra-host phylogenetics. Our results suggest that protracted infections should be treated with combination therapies rather than by a single mAbs to clear pre-existing resistant mutations. Additionally, our findings on viral reservoirs can shed light on protracted infections interspersed with periods where the virus is undetectable as well as an alternative explanation for some long-COVID cases.

3.2 Introduction

Several SARS-CoV-2 VOCs have convergent mutations in the spike protein's RBD that coincide with known epitopes.(94) Mutations in this genomic region affect the ability of the spike (S) to enter the cell via the ACE2 receptor and have been linked with higher transmission rates and/or immune escape (53,239).

While in most cases, SARS-CoV-2 infections are cleared within a few days, key mutations develop *de novo* in long lasting infections in patients with immunosuppressive conditions. These infections can last for several months, and their viral mutation rate is higher than in shorter infections in immunocompetent patients (166). For this reason, it is suspected that protracted infections are one of the drivers of SARS-CoV-2's genomic evolution and a source of immune escape variants (95). One such example is S:E484K that was found in former VOCs Beta and

Gamma (90). This mutation has been shown to give the virus immune escape properties such as resistance to anti-viral monoclonal antibodies (mAbs) and convalescent sera as well as reinfection (73). Resistance to these treatments has become a growing concern during the past year as an increasing number of Omicron sub-lineages were found to be resistant to a variety of mAbs (240,241).

Despite the potential importance of these cases, few longitudinal datasets of sequences collected from immunosuppressed individuals at different time points during their infections are available. These datasets can give us insight into evolutionary events that are not observed in acute infections, such as an instance of recombination between two viral strains (45) or the presence of distinct viral populations with immune escaping mutations in a single sample (96).

Here, we describe the genetic events that arose in two patients with hematologic cancers that were infected by SARS-CoV-2 for several months. Samples from the first patient (Q1) were collected by the Public Health Laboratory of Québec (LSPQ), in Canada. Viral sequences from the second patient (K1) were generated by Lee et al. (86) from samples collected in Korea. Through phylogenetic and intra-host single nucleotide variant (iSNV) analysis, we show evidence for a mutational pattern suggestive of a viral reservoir as well as for several viral populations containing immune escape mutations in the spike's RBD.

3.3 Methods

3.3.1 Viral databases

SARS-CoV-2 consensus sequences data were obtained from the LSPQ database through the CoVSeQ consortium (https://covseq.ca/data-info?lang=en) on 16/11/2021. Only sequences that were covered at more than 90% and a mean depth of 50X for Illumina and 16X for Oxford Nanopore technologies (ONT) with no previously documented frameshift, less than 5% N at most 5 ambiguous bases were used. Serial sequences from two patients described by Lee et al. (86) (P1 and P2) were obtained from NCBI's Sequence Read Archive (study SRP357108). One of them (P1), did not have iSNV in the spike's RBD and was excluded from this study. A total of 147,537 representative SARS-CoV-2 Illumina libraries from 2020 and 2021 were downloaded from NCBI

and served as a reference dataset to compare patient data (see *Intra-host analysis* below). Metadata for Q1 were obtained as part of BQC-19, PMID: 34010280.

3.3.2 Whole-genome sequencing and consensus sequence generation

All LSPQ sequencing data were analyzed using the GenPipes (242) Covseq pipelines to produce variant calls and consensus sequences. Samples were sequenced on Illumina or ONT. Regardless of the sequencing technology, data was initially processed to remove any host sequences by aligning to a hybrid reference with both human (GRCh38) and SARS-CoV-2 (MN908947.3) (19). Any sequences that aligned to the human portion of the hybrid reference were removed from downstream analysis. For Illumina sequencing data, raw reads were first trimmed using cutadapt (v2.10) (220), then aligned to the reference using bwa-mem (v0.7.17) (221). Aligned reads were filtered using sambamba (v0.7.0) (222) to remove paired reads with an insert size outside the 60-300bp range, unmapped reads, and all secondary alignments. Then, any remaining ARTIC primers (v3) were trimmed with iVar (v.1.3.4) (243). To create a consensus representative of the most abundant species in the sample, a pileup was produced using Samtools (v1.9) (224) which was used as an input for FreeBayes (v1.2.2) (225). For ONT sequencing data, raw signals were basecalled using guppy (v3.4.4) (226) with the High-Accuracy Model (dna_r9.4.1_450bps_hac). Reads outside the expected size range (400-700bp) were removed from the analysis. Reads were then aligned to the reference using minimap2 (v.2.17) (227) and filtered to remove incorrect primer pairs and randomly downsampled to keep 800X depth per strand in high coverage regions. Finally, Nanopolish (v0.13.1) (228) was used to call mutations in regions with a minimum depth of 16X (8X per strand) and a flank of 10bp. After masking regions with coverage below 20X, mutations called by nanopolish were integrated into the reference using bcftools (v1.9) (224) to create a consensus sequence. In all cases, MN908947.3 was used as a reference genome. A full description of both pipelines can be found in the following URLs:

https://genpipes.readthedocs.io/en/genpipes-v4.1.2/user_guide/pipelines/gp_covseq.html

and

https://genpipes.readthedocs.io/en/genpipes-v4.1.2/user_guide/pipelines/gp_nanopore_covseq.html.

3.3.3 Phylogenetic analysis and mutational spectrum

The ‘Pangolin’ network was used to identify the sequence lineage for consensus sequences from Quebec (PangoLearn version 2021-11-09, Pangolin version 1.2.93) (244), and all sequences characterized as the B.1.160 lineage were used to generate a distance tree. The phylogenetic trees were generated with Nextstrain viewer (61) using the default settings.

3.3.4 Intra-host analysis

The dataset for the intra-host analysis consists of sequences from one patient from the LSPQ and one patient described by Lee et al. (86) The intra-host mutational patterns were compared to our in-house intra-host mutation database based on 147,537 representative samples. Each library was trimmed using TrimGalore! v0.6.0 and then mapped to the reference genome NC_045512.2 using bwa-mem v.0.7.17. The remaining amplicon sequences were trimmed using iVar with a hybrid amplicon definition file combining ARTIC v3, v4 and v4.1 designs. Primary reads were kept using Samtools v.1.15.1. iSNVs below 5% for Illumina and 10% for Nanopore that are not found at a higher frequency in at least one time point per patient are likely to be sequencing errors and were filtered out. Reads containing reference and alternative alleles for positions in the spike’s RBD were extracted from the BAM files using ctDNAtool (245). The number of reads containing different combinations of alternative and reference alleles was then compiled to determine the frequencies of the possible haplotypes.

3.4 Results

3.4.1 Description of patients

An immunosuppressed 73-year-old woman (Q1) with non-Hodgkin lymphoma first tested positive for SARS-CoV-2 (PANGO lineage B.1.160) on 08/01/2021 (Day 1, D1). She had undergone several courses of anti-CD20 (rituximab) and chemotherapy in the months preceding her COVID diagnosis. She was vaccinated with the Pfizer vaccine on 25/02/2021. The patient tested positive again on 28/04/2021 (D111). The full timeline of her infection is shown in Figure 3.1a. Because the sample sequenced on D111 had S:E484K, it was first assumed that this sample and all subsequent timepoints were from a reinfection. However, phylogenetic analysis of all time points shows that all samples came from the same infection that lasted at least 173 days, from 08/01/2021 to 29/06/2021 (Figure 3.1b). She passed away on 14/08/2021 from a non-COVID related complication.

One immunosuppressed South Korean 25 years old male patient (K1, described as P2 in Lee et al.) was infected with PANGO lineage B.1.497 in late 2020 and early 2021 (86). He had acute myelogenous leukemia and had received an allogeneic hematopoietic stem cell transplant one year prior. His infection lasted 73 days and 16 samples were collected over the span of the first 67 days (20/11/2020 to 26/01/2021). Neither of the two patients mentioned here were treated with mAbs or convalescent sera (86). The complete list of samples, dates, and tissues can be found in Table S 3.1.

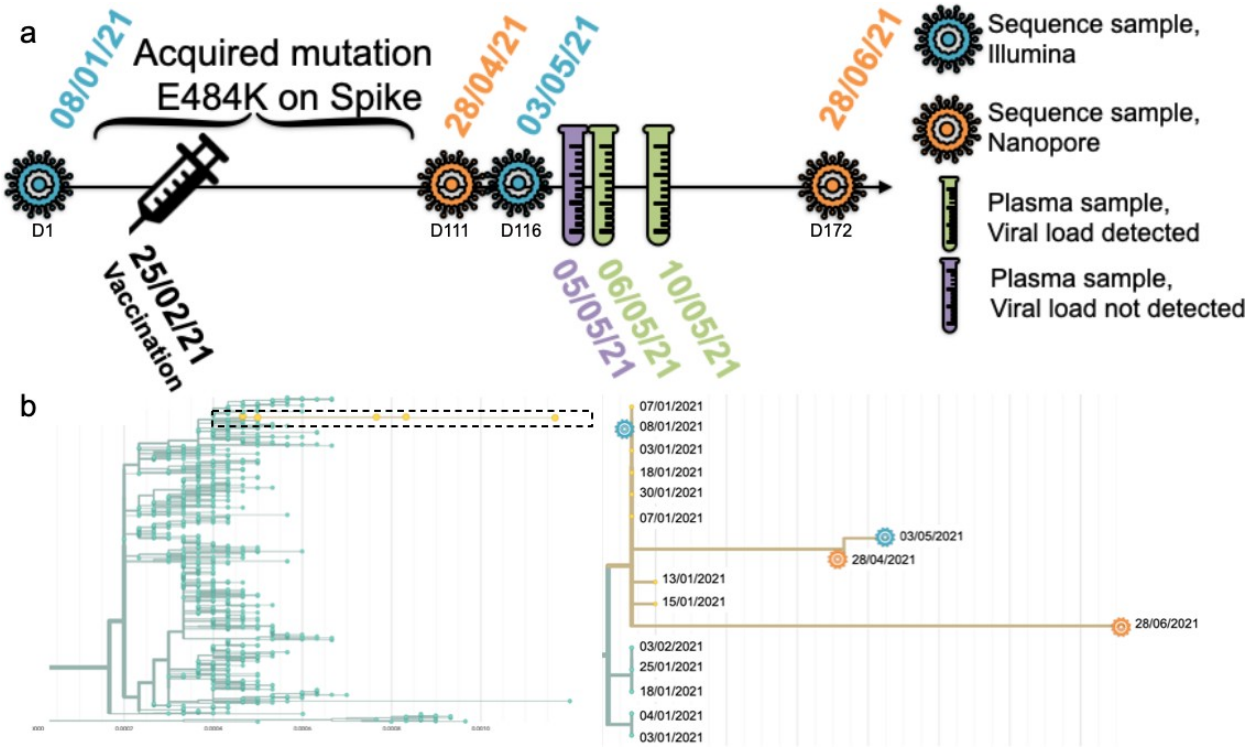


Figure 3.1 Description of Q1's infection.
a: Timeline of the infection and type of samples per date. **b**: Distance tree of lineage B.1.160 in Quebec on the left, close up of the box containing Q1's four consensus sequences on the right. The X axis is measured in substitutions per site per year.

3.4.2 Intra-host analysis of Q1's samples

At D1, Q1 had all the characteristic mutations of B.1.160 in Quebec, as well as eight additional mutations (Figure 3.2) shared with five other sequences in the LSPQ database (Figure 3.1b).

Sequences at D111 and D116 share ten new mutations that are not seen at D172. The additional mutations seen on D111 on Figure 1b but not in Figure 2 are low quality base calls that were filtered out from the Nextstrain analysis. The reversal of all consensus mutations acquired at D111 and D116 makes it unlikely that the substrain at D172 has evolved from the ones at D111/D116.

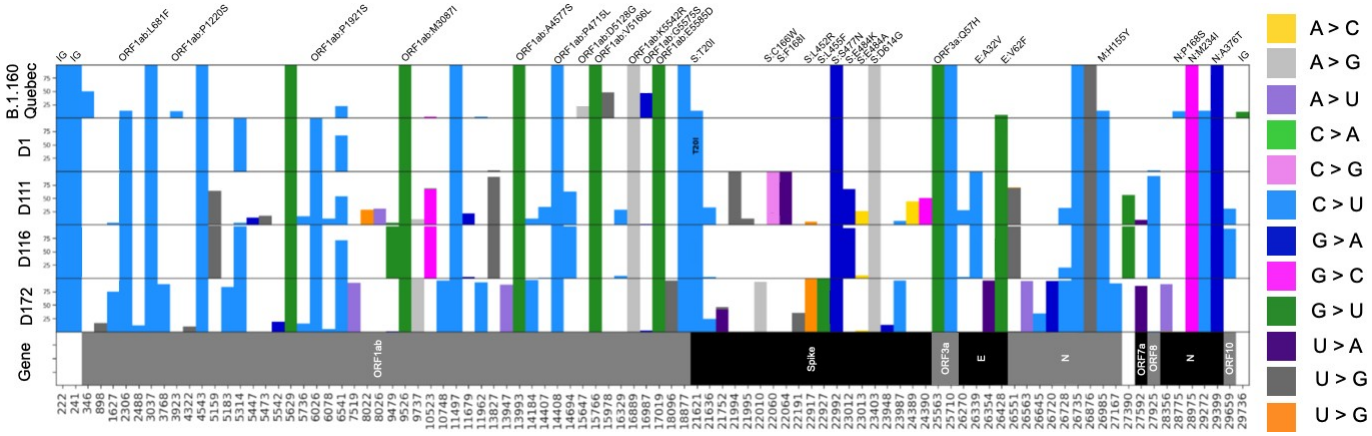


Figure 3.2 - Allelic frequencies in B.1.160 in Quebec and in Q1's infection.
The top row displays the frequencies for n = 2,627 B.1.160 consensus sequences from the LSPQ database. The four following rows show intra-host frequencies for Q1's mutations for each time point. Only mutations with intra-host frequencies above 5% for Illumina sequences (D1 and D116) and 10% for Nanopore sequences (D111 and D172) for at least one time point are presented. Because of the respective error rates of both sequencing technologies, discrepancies up to 5% for Illumina sequences (D1 and D116) and 10% for Nanopore sequences (D111 and D172) are likely to be sequencing artifacts. Non-synonymous mutations are written on top, and the color represents the nucleotide change.

3.4.3 Intra-host evidence of multiple viral populations with distinct immune escape mutations

Q1 and K1 had a combined total of nine substitutions resulting in a change of five residues in the spike's RBD (Figure 3.3a). Mutations at residues S:346, S:484, S:490, and S:494 confer resistance to an array of mAbs (246), and mutations at residue S:346 and S:348 have been linked with higher transmissibility (247). Both patients had substitutions G23012A and A23013C on S:484.

To determine the full extent of intra-host genetic diversity at a given time point, we analyzed individual reads to determine if the substitutions in the RBD belonged to different viral populations or if they co-occurred. For Q1, the substitutions on S:484 at D111 and D116 were mutually exclusive; no reads contained both alternative alleles on 23012 and 23013 (Figure 3.3b).

There were two major distinct mutant populations of S:E484K (0.68 on D111, 0.93 on D116) and S:E484A (0.26 on D111, 0.06 on D116), as well as a small wild-type population (0.06 on D111, 0.01 on D116). For K1, G22599A and C22605T on D38 were both at a frequency of 0.10 (Figure 3a), which could suggest co-occurrence of these mutations. However, when retrieving reads containing all three positions, we see that those substitutions belong to different viral populations (Figure 3.3c). These results highlight the importance of analyzing aligned reads to describe the intra-host population dynamics.

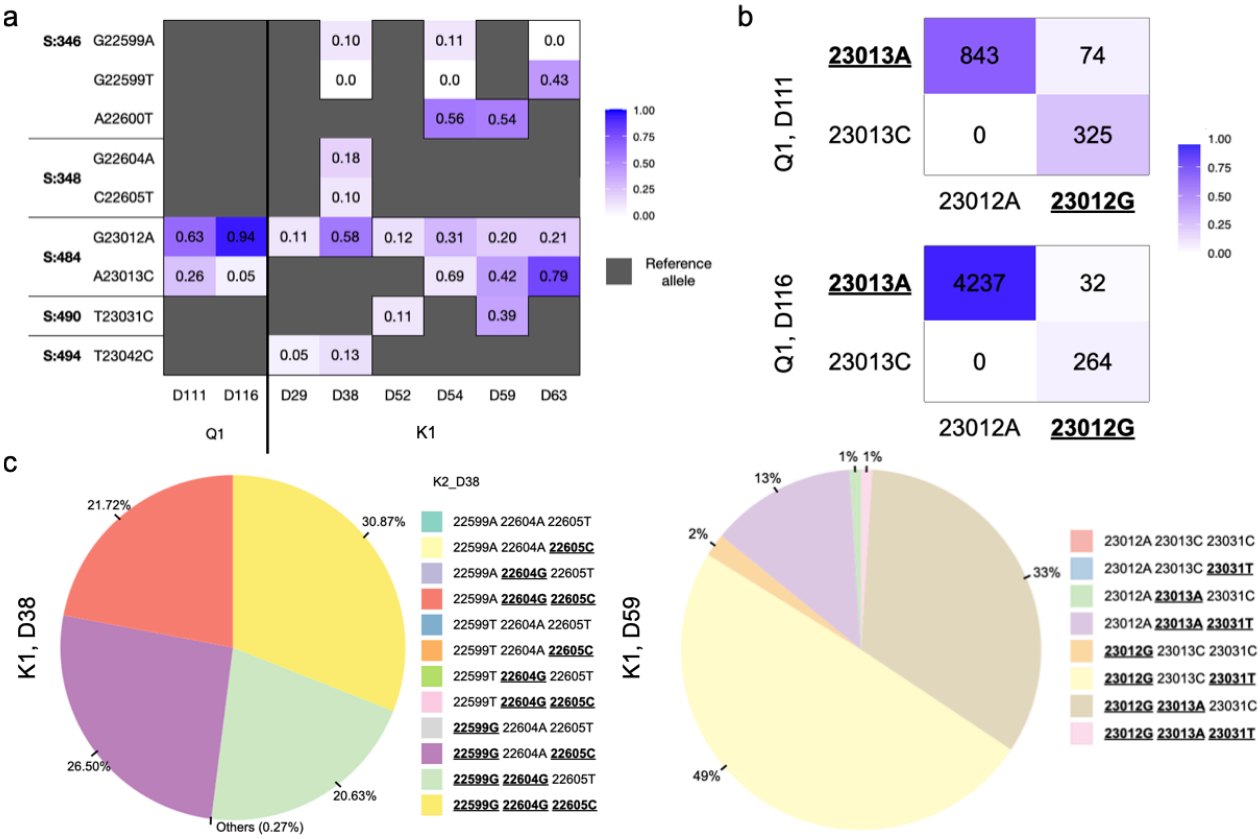


Figure 3.3 – Intra-host allelic frequencies for mutated positions in the S's RBD**.**
a: frequencies for the alternative allele per position for mutated positions at S:346, S:348, S:484, S:490, and S:494 in Q1 and K1. **b**: frequencies of the haplotypes present at codon S:484 on D111 and D116 for Q1. Reference alleles are underlined. **c**: Left - frequencies of the haplotypes present at codons S:346 and S:348 on D38 for K1 on reads encompassing all three positions. "Others" category includes 22599A/22604G/22605T, 22599G/22604A/22605T, and 22599T/22604G/22605C combinations (0.1%, 0.07%, 0.1%, respectively). Right - frequencies for the haplotypes present at codons S:484 and S:490 on D59 for K1 on reads containing all three positions.

3.4.4 Intra-host patterns at S:E484 in the general population infected by

SARS-CoV-2

Analysis of intra-host mutations in 149,013 SARS-CoV-2 sequencing libraries downloaded from NCBI revealed that no sequence had more than one mutation on codon S:346. Only four samples had more than one mutation on S:484 that led to distinct viral populations (SRR15258550 - Angola and SRR15061404, SRR17006835, SRR16298333 - South Africa). Unfortunately, there is no detail about how long after infection these samples were taken or if the patients were immunosuppressed, but their overall mutational burden was not characteristic of long-lasting infections. The small number of occurrences of the RBD mutational pattern found in the general population (at a frequency below 0.003%) highlights the uncommon nature of the mutational events identified here in immunosuppressed individuals.

3.5 Discussion

We performed intra-host analysis on serial SARS-CoV-2 sequences from two patients with hematologic cancers and compared the identified patterns with 147,537 sequencing libraries to look for intra-host populations of immune escaping mutations in the spike's RBD. We found evidence of multiple viral populations co-existing in this region within a single host in immunosuppressed patients. Furthermore, we found distinct populations of mutants for codon S:484 which is extremely rare in the general population, thus stressing the importance of studying immunosuppressed patients in a longitudinal design to get insights into key steps of viral evolution.

When comparing consensus mutations found at Q1's D1, D111/D116, and D172, we saw that all four time points share the mutations present at D1, demonstrating that this is a single long-lasting infection. However, D111/D116 and D172 accumulated 10 and 20 new mutations, respectively, that are not shared across time points. The only notable exception was C26728T, a synonymous mutation in the N protein present at frequencies of 0.29, 0.2, and 0.92 on D111, D116, and D172, respectively. This suggests that the substrains present at these time points evolved separately from the substrain at D1, and that C26728T may be a recurrent mutation. The substrain present at D111/D116 was cleared from the nasopharyngeal tissues and replaced by the one at D172. Given the lack of overlap of mutations between D111/D116 and D172, the substrain present at the last time point may have evolved in another location within the host's body, consistent with patterns observed in viral reservoirs. The theory of viral reservoirs as an explanation for long-COVID symptoms has been put forward because viral antigens or intermediate molecules of viral

replication have been detected in long-COVID cases despite negative PCR tests (101,103,248,249). However, to our knowledge, this is the first evidence of a viral reservoir based on intra-host phylogenetics. These results call for further investigation to determine whether SARS-CoV-2 viral reservoir can be found in immunocompetent patients.

Immune escape has been a growing concern in the past year due to the rise of Omicron and its multiple sublineages that escape natural immunity as well as available vaccines and several mAbs (78,79). Here we described two immunosuppressed patients that were not treated with mAbs but that developed de novo multiple viral populations with mutations known to cause resistance to different mAbs (73,90,239). We have shown that this pattern is extremely rare in the general population, making it very likely that those distinct populations in Q1 and K1 arose due to their condition, revealing SARS-CoV-2 escape strategies. This conclusion is supported by another recent case study with an immunocompromised patient, which also found multiple mutations on S:484 (96). As is the case for other viruses, our results suggest that combination strategies accelerating viral clearance may be required to clear viral populations with pre-existing mutations within vulnerable patients.

3.6 Acknowledgements

We thank members of the Hussin group for helpful discussions and paper revisions, specifically Matthew Scicluna as well as Simon Gravel and Vincent-Philippe Lavallée. We thank Floriane Point for complementary laboratory work, and members of the Quebec public health surveillance committee for SARS-CoV-2, specifically Judith Fafard, and Canadian COVID Genomics Network (CanCOGeN), specifically Eric Fournier and Paul Stretenowich. This work was completed thanks to computational resources provided by Calcul Quebec clusters Narval and Beluga. We acknowledge and thank GISAID and NCBI as well as all contributing laboratories for giving access to their SARS-CoV-2 genome sequences. This study was supported by funding from the Canadian Foundation for Innovation, IVADO COVID19 Rapid Response grant (CVD19-030), the National Sciences and Engineering Research Council (NSERC) (ALLRP 554923 – 20), and the Canadian Institutes of Health Research (CIHR) (#174924). D.F. is a BioTalent awardee. M.C. is a FRQS Junior 1 research scholar, J.G.H. is FRQS Junior 2 research scholar and D.E.K. is a FRQS Merit research scholar. This study was also supported by the CIHR operating grant to the Coronavirus Variants Rapid Response Network (CoVaRR-Net) and the Biobanque québécoise de la COVID-19 (BQC-19). Finally, we would like to thank members of

the Mila COVID19 Task Force for their camaraderie and valuable insight into data analysis strategies during the pandemic.

	ID	Date	Days since symptom onset	Tissue
Q1	L00409639001	08/01/2021	1	Nasopharyngeal swab
	L00350036001A	28/04/2021	111	Nasopharyngeal swab
	L00350839	03/05/2021	116	Nasopharyngeal swab
	L00363495001	28/06/2021	172	Nasopharyngeal swab
K1	SRR17793984	20/11/2020	0	Nasopharyngeal swab
	SRR17793983	02/12/2020	12	Nasopharyngeal swab
	SRR17793982	02/12/2020	12	Saliva
	SRR17793981	02/12/2020	12	Stool
	SRR17793980	07/12/2020	17	Throat swab
	SRR17793979	09/12/2020	19	Nasopharyngeal swab
	SRR17793978	09/12/2020	19	Saliva
	SRR17793977	09/12/2020	19	Stool
	SRR17793976	09/12/2020	19	Urine
	SRR17793975	19/12/2020	29	Nasopharyngeal swab
	SRR17793973	28/12/2020	38	Nasopharyngeal swab
	SRR17793972	11/01/2021	52	Nasopharyngeal swab
	SRR17793971	13/01/2021	54	Saliva
	SRR17793970	18/01/2021	59	Nasopharyngeal swab
	SRR17793969	22/01/2021	63	Nasopharyngeal swab
	SRR17793968	26/01/2021	67	Nasopharyngeal swab

Supplementary Table 3.1 – Description of analyzed samples

Chapitre 4 – Synthèse

4.1 Discussion

Des millions de séquences virales ont été collectées et rendues accessibles publiquement à des fins de recherche depuis le début de la pandémie de COVID-19. Ce partage de données sans précédent a permis une surveillance des variants du SARS-CoV-2 en temps réel. Il est devenu évident dès le début de la pandémie que certains acides aminés du génome viral mutaient de manière récurrente. Dans ce mémoire, je présente mes travaux sur l'étude de la diversité génétique de SARS-CoV-2 dans des échantillons du Québec et à travers le monde. En particulier, j'ai identifié des mutations convergentes présentes au Québec durant la deuxième vague, mesuré l'impact des mutations récurrentes sur la capacité des CMH à lier les épitopes viraux et caractérisé des mutations récurrentes apparaissant *de novo* à divers codons du RBD de la protéine S dans les populations virales intra-hôte de patients immunosupprimés. Mes résultats suggèrent que ces mutations convergentes surviennent à la suite de pressions de sélection.

4.1.1 Les mutations C > U jouent un rôle important dans le biais d'acides aminés

Dans le cadre d'études sur les mutations qui affectent la réponse immunitaire, on cherche souvent à déterminer quel impact le système immunitaire de l'hôte a sur les patrons de mutation de divers pathogènes (250,251). Une autre manière d'aborder ce problème est de mesurer l'impact des patrons de mutations globaux sur les changements d'acides aminés dans la composition d'un génome. On peut par la suite déterminer si ces changements affectent la réponse immunitaire en modifiant les épitopes du pathogène.

L'excès de mutation C > U dans les génomes du SARS-CoV-2 a été rapporté de nombreuses fois dans la littérature (58), nous avons montré que cet excès contribue de façon importante à l'introduction des résidus I, L, Y et F ainsi qu'au retrait des résidus A, P et T. Le biais de transition G > U, également documenté dans la littérature (59), contribue aussi au patron de mutations observé dans les données. La significativité statistique de ces résultats a été établie à l'aide de mes simulations génétiques du génome du SARS-CoV-2 sous évolution neutre. Le retrait des P est particulièrement lié à l'évasion immunitaire, puisque ce type de mutation compte pour 31% des

mutations qui abrogent les épitopes alors qu'elle ne représente que 9.1% des mutations globales. Il est à noter que dans les données réelles, l'excès de C > U et l'excès de G > U est présent en même temps, or je n'ai pas simulé de scénario dans lequel il y a un excès pour ces deux types de substitution. On remarque à la Figure 2.2 b qu'augmenter le taux d'un type de substitution fait baisser le taux de tous les autres types. Ceci est normal dans le contexte des simulations puisque le taux de mutation global est fixe. Il serait intéressant de faire l'expérience en combinant divers taux de substitution C > U et G > U et de voir comment ces deux biais interagissent ensemble.

Une théorie répandue pour expliquer l'enrichissement de mutations C > U est l'action de l'enzyme APOBEC (121). Des études sur l'effet d'APOBEC sur le VIH ont démontré que ces enzymes peuvent influencer la liaison des épitopes viraux au CHM et ainsi contribuer à l'évasion immunitaire (252,253). Nos résultats suggèrent que ce phénomène s'applique aussi dans le cas du SARS-CoV-2. Cependant, cela n'explique pas pourquoi ces mutations sont enrichies dans les régions des épitopes confirmés. Une explication pourrait être que les prolines ont une structure chimique différente des autres acides aminés, et que son introduction ou son retrait a le potentiel de grandement changer la structure 3D de la protéine (254). Ces changements pourraient avoir un effet délétère sur la *fitness* du virus sauf si celui-ci survient dans le cadre d'un épitope où il contribue à l'évasion immunitaire. Des résultats préliminaires obtenus par d'autres membres de mon groupe de recherche montrent que le taux de mutations C > U est plus bas dans la deuxième vague pour les positions pour lesquelles une mutation C > U n'a pas été rapportée durant la première vague. Ceci corrobore l'hypothèse des positions sous différentes contraintes évolutives dans le génome du SARS-CoV-2.

4.1.1.1 L'utilité des simulations génomiques dans l'étude de pathogènes

Mes résultats témoignent de la pertinence de l'utilisation des simulations génomiques virales dans le cadre de l'étude de l'évolution de pathogènes. Leur avantage est qu'elles permettent de générer des données rapidement et à faible coût. Grâce à elles, j'ai pu vérifier des hypothèses expérimentalement de manière *in silico*, ce qui aurait été difficile à faire *in vitro* ou *in vivo* parce qu'il aurait fallu créer des lignées de virus qui évoluent uniquement sous neutralité et avec des taux précis de substitution C > U et G > U.

Malgré son potentiel, cette méthodologie reste peu utilisée en génomique virale et cela peut s'expliquer en partie par les logiciels de simulation disponibles. Un logiciel de simulation vers le futur bien connu est SLiM dont la première version est publiée en 2013 (255) et qui en est aujourd'hui à sa troisième version (256). Ce dernier a été initialement développé pour effectuer des simulations génomiques sur des organismes diploïdes, ce qui rend son utilisation plus difficile pour l'étude de la majorité des virus qui sont haploïdes. SLiM a un module permettant d'accommoder les organismes haploïdes, mais ce dernier demande une bonne connaissance de base du logiciel ainsi que des principes de génétique de population virale afin de s'assurer de la justesse des paramètres utilisés. J'ai donc décidé d'utiliser SANTA-SIM, un nouveau logiciel dans le domaine de la simulation virale.

J'ai démontré que SANTA-SIM fonctionne comme prévu dans le cadre de simulations sous neutralité (Figures 2.1 et 2.2). Cependant, je n'ai pas été en mesure de réaliser des simulations dans des scénarios évolutifs plus complexes tels que la sélection négative et les expansions populationnelles. J'ai contacté les développeurs de l'outil qui m'ont annoncé que SANTA-SIM n'était plus en développement et que le logiciel n'était pas maintenu. Cette situation est dommage et est probablement causée en partie par le manque d'intérêt de la communauté pour les simulations génomiques virales ou alors une méconnaissance des outils disponibles. En conclusion, je recommande l'utilisation de SLiM pour la simulation de scénarios plus complexes d'évolution virale, à la suite de mes travaux des membres de mon laboratoire ont emprunté cette voie récemment.

4.1.2 La deuxième vague au Québec

La deuxième vague de la pandémie était une époque intéressante pour l'étude de l'évolution du SARS-CoV-2 puisque c'est à ce moment que les VOCs et VOIs ont commencé à émerger. Le premier VOC apparu était Alpha en décembre 2020, mais ce dernier n'a pas circulé de façon majoritaire au Québec avant la fin du mois de février 2021. Mes résultats montrent que la deuxième vague au Québec a plutôt été caractérisée par B.1.1.176, un variant qui circulait déjà dans la province durant la première vague, et par B.1.160, un variant introduit de France. Ce dernier a été catégorisé en tant que VOI au Québec. Mon analyse phylogénétique de B.1.160 au Québec montre que la grande majorité des infections par cette sous-lignée proviennent d'un ancêtre commun, c'est-à-dire d'un seul événement d'introduction. Ceci concorde avec les résultats de Murall et collègues

qui ont déterminé que la diversité génomique au Québec lors de la première vague était causée par un petit nombre d'introduction du virus (215). Mes résultats suggèrent également que les mesures de restriction de déplacements internationaux et entre les provinces ont été efficaces pour réduire l'introduction de nouveaux variants dans la province.

Une étude sur les deux premières vagues au Canada a déterminé que contrairement à la première vague, la deuxième vague était caractérisée par des introductions de nouveaux variants plutôt que par l'évolution de variants déjà présents sur le territoire (231). Bien que cela soit vrai dans le cas de B.1.160, et certainement aussi dans le cas d'Alpha et des VOCs lui succédant, le cas du Québec est différent parce que B.1.1.176 a continué d'y circuler. Ceci pourrait s'expliquer par un avantage que l'acquisition des mutations S:T20I, S:R357K et N:A211V lui aurait conféré comparativement aux autres variants de la première vague, et qui lui aurait permis de demeurer compétitif dans le contexte de la deuxième vague.

4.1.3 Description de l'infection de Q1

J'ai identifié un patron de mutation intra-hôte inédit chez une patiente atteinte d'un lymphome non-hodgkinien en plus d'une infection au SARS-CoV-2 de longue durée avec le variant B.1.160. La première caractéristique de son infection est la période entre D1 et D111 durant laquelle la patiente a reçu des résultats de tests PCR négatifs suite à un échantillonnage dans les voies nasopharyngiques. Ceci a mené à la conclusion que l'infection était résorbée alors qu'elle était encore active. Il n'est pas rare que le résultat d'un test PCR soit un faux négatif. Une étude a démontré que le taux de positivité de cette méthode diagnostique peut varier de 53.1%–85.3% lorsque l'échantillon est pris des voix nasopharyngiques selon la sévérité de l'infection et le moment où l'échantillon a été pris (257). Le taux de positivité est plus élevé entre les jours 0 et 7 de l'infection qu'entre les jours 8 et 14. Comme les infections au SARS-CoV-2 durent en moyenne 14 jours dans la population générale, peu de données existent sur les faux négatifs lors d'infections plus longues. Néanmoins, des études suggèrent que l'échantillonnage des voies respiratoires inférieures, par exemple du sputum, serait supérieur pour détecter les infections au-delà de la deuxième semaine (257,258).

Il a été supposé par les cliniciens que la séquence récoltée à D111 provenait d'une réinfection, entre autres parce que la mutation S:E484K a été détectée sur la séquence, potentiellement durant un criblage pour identifier les mutations d'intérêt à l'époque. J'ai déterminé à l'aide d'une analyse

phylogénétique que cette séquence provenait de la même infection qu'au D1. En tenant compte du faible nombre de séquences B.1.160 récoltées au Québec aux environs de D111 (le 28/04/2021) (Figure 2.5), il est peu probable que la patiente ait été exposée à nouveau au variant B.1.160 dans la communauté. De plus, en regardant les arbres de temps et de distance pour B.1.160 (figures 2.7 et 3.1), on remarque que ce variant avait acquis une grande diversité dans les mois durant lesquels Q1 a été infecté puisque plusieurs clades sont visibles sur l'arbre de distance. Ceci réduit encore plus le risque d'une réinfection à la même sous-lignée puisque plusieurs étaient en circulation au Québec durant les mois de janvier et février. Il serait toutefois difficile de quantifier la probabilité qu'il s'agisse de deux infections séparées parce que nous n'avons pas un portrait complet de la diversité virale au Québec pour cette période. Premièrement, tel que mentionné précédemment, seulement une partie des échantillons collectés ont été séquencés. Deuxièmement, à partir de février 2021, les échantillons étaient criblés pour détecter des mutations d'intérêt comme S:E484K par exemple, et seuls ceux identifiés comme était un possible VOC étaient séquencés. B.1.160 est donc probablement sous-représenté dans la base de données du LSPQ à partir du mois de mars puisque les efforts de séquençage étaient plutôt dirigés vers Alpha et Beta. En plus de ne pas pouvoir faire un ratio du nombre précis de B.1.160 sur le nombre total de cas de la province, nous ne disposons pas d'information sur le milieu de vie de Q1 et ne pouvons donc pas savoir s'il y avait une éclosion tardive de B.1.160 dans sa communauté.

Un facteur technique a grandement compliqué mes analyses intra-hôte pour cette patiente: les séquences D1 et D116 ont été séquencées avec Illumina et les séquences D111 et D172 avec Nanopore. Parce que les iSNVs peuvent se retrouver à basse fréquence dans un échantillon, il est pratique courante de réaliser les analyses intra-hôte avec des séquences générées par Illumina, qui a un taux d'erreur de moins de 1% (259,260). Le désavantage de Nanopore dans ce genre d'analyse est que son taux d'erreur peut aller jusqu'à 15%, il devient donc difficile de juger si une mutation à faible fréquence détectée par cette technologie est réelle ou le fruit d'erreurs de séquençage. Cette situation est illustrée à la Figure 3.2 dans laquelle on voit plusieurs iSNV à basse fréquence à D111 qu'on ne retrouve pas à D116. Deux scénarios peuvent expliquer ce résultat: le premier est qu'il y avait bel et bien une population virale présente à faible fréquence à D111 qui disparaît complètement à D116. Le deuxième est qu'un certain nombre de ces iSNV soient en fait dus à des erreurs de séquençage. En se fiant aux deux populations virales présentes au codon S:484 à D111, le deuxième scénario est plus probable. On remarque que certaines mutations qui ont des

fréquences similaires à A23013C à D111 (G11497A, C14407U, C16329U et C26270U) sont encore présentes à très faible fréquence à D116. On peut supposer que ces mutations font partie d'une population virale qui comprend A23013C. En revanche, il n'y a pas d'indications que les mutations présentes à des fréquences différentes que celles des deux populations virales bien établies à D111 et qui ne sont plus présentes à D116 ne soient pas des erreurs. Malheureusement, la distance génomique entre ces positions est trop grande pour nous permettre de vérifier si ces mutations sont sur les mêmes *reads*. La séquence à D172 comporte également des iSNV à basse fréquence, mais en l'absence d'une séquence Illumina provenant d'une date rapprochée, il est impossible de conclure sur la validité de ces mutations. Un autre exemple de position pour laquelle il est difficile de déterminer si on observe un phénomène réel ou une erreur de séquençage est à la position 6541, dont l'allèle alternative atteint ~60% à D1 et qui est complètement fixée à D172. La fréquence descend à 50% à D111 et remonte vers 70% à D116. Parce que les résultats d'illumina sont fiables, on sait que la fréquence de cet allèle augmente a augmenté à D116 par rapport à D1. Il est cependant impossible de dire si cette augmentation s'est faite de manière monotone ou non.

Malgré les difficultés d'interprétation des iSNV mentionnées ci-haut, j'ai été en mesure de déterminer que les séquences D111/D116 et D172 proviennent de populations virales qui ne descendent pas l'une de l'autre. Nos analyses montrent que D111 et D116 comportent deux populations virales distinctes, celle avec G23012A et celle avec G23013C, mais l'analyse phylogénétique permet d'affirmer que ces populations sont similaires parce qu'elles partagent plusieurs mutations consensus. Ce n'est pas le cas avec D111/D116 et D172, ces séquences n'ont pas de mutations consensus en commun autre que celles qui étaient déjà présentes à D1. Le scénario le plus probable menant à ce patron est que les populations virales à D111/D116 ont été éliminées des voies respiratoires supérieures où les échantillons ont été récoltés. La population virale parentale de D172 était possiblement présente ailleurs dans le corps et a migré vers les voies respiratoires supérieures entre D116 et D172. Ceci suggère que la population à D172 a été en mesure d'échapper au système immunitaire dans un autre organe, ce qu'on observe dans le cadre de réservoir viraux. La théorie du réservoir viral dans le cadre d'infections au SARS-CoV-2 a été proposée dans différentes études (101,103,248). Mes résultats apportent un nouvel élément de preuve pour ce phénomène, basé sur la phylogénétique et l'analyse des iSNV plutôt que sur la présence de molécule suggérant la réplication virale ou par histopathologie sur des tissus autopsiés.

4.1.5 Les patrons de mutation dans les patients immunosupprimés

Les patients immunosupprimés ont été proposés comme vecteurs d'évolution des VOCs et VOIs parce qu'ils développent *de novo* des mutations qui favorisent l'évasion immunitaire (90,95,96). Mes résultats montrent que deux patients souffrant de cancers hématologiques ont acquis plusieurs populations virales distinctes au sein d'un même échantillon. On peut donc dire que, même sans avoir été exposé à un traitement par anticorps monoclonaux (mAbs), ces patients étaient infectés par des quasiespèces à diverses propriétés d'évasion immunitaire.

Les implications cliniques de ces résultats est qu'il est primordial de traiter les infections de SARS-CoV-2 à l'aide de combinaison de mAbs pour être certains d'éradiquer toutes les populations virales. Les quasiespèces décrites ici proviennent d'échantillons récoltés dans les voies nasopharyngiques et dans la salive. Nous ne pouvons pas exclure la possibilité que d'autres tissus contiennent différentes mutations d'évasion immunitaire. Une prochaine question de recherche pourrait donc être de déterminer combien de mAbs différents administrer aux patients pour maximiser les chances qu'ils soient efficaces contre l'ensemble des populations virales déjà présentes. La modélisation mathématique est une méthode qui a fait ses preuves pour résoudre ce genre de problèmes dans les infections chroniques causées par le VIH et l'hépatite B et C. Alexander et Bonhoeffer ont développé un modèle basé sur les ODE décrits à la section 1.2.5.5 qui permet d'établir la probabilité qu'une population résistante soit présente au début d'un traitement antiviral ainsi que celle qu'une population résistante apparaisse suite au traitement (83). Il serait intéressant de l'appliquer aux infections de longue durée du SARS-CoV-2.

Au cours de nos analyses, nous nous sommes concentrés sur deux courtes régions du RBD de la S pour évaluer la présence de mutations sur le même *read* (position génomique 22580 à 22630 et 23000 à 23500). Malheureusement, la longueur des *reads* de nos séquences ne nous permet pas d'évaluer l'ensemble de la région de 22580 à 23500. Cela nous aurait permis de déterminer si les mutations sur S:346/S:348 et sur S:484/S:490/S:494 sont présentes dans les mêmes populations ou non. Oxford Nanopore Technologies a développé un protocole pour générer ce qu'ils appellent des *ultra-long reads* de plus de 100Kb (261). Une telle longueur de *reads* nous permettrait de caractériser les populations virales sur le génome du SARS-CoV-2 en entier. Tel que décrit à la section 4.1.4, le taux d'erreur élevé de Nanopore complique la réalisation d'analyse intra-hôte. C'est pourquoi une telle étude devra également séquencer les échantillons par Illumina afin de départager les mutations réelles des erreurs de séquençage. Cette combinaison des deux méthodes

est utilisée lorsqu'on veut tirer profit des avantages de la longueur des *reads* de Nanopore sans compromettre la qualité de la séquence (261).

Les résultats décrits ici se concentrent sur reconstruire les haplotypes une petite portion du génome du SARS-CoV-2 dans un nombre limité de séquences. Une prochaine étape serait faire le même exercice dans le génome au complet afin de pouvoir différencier tous les clades. Tel que mentionné précédemment, il faudrait reséquencer les échantillons avec un protocole permettant des ultra long *reads* pour faire cette analyse. Malheureusement, il ne reste pas toujours suffisamment de matériel génétique pour reséquencer un échantillon. Pour Q1 par exemple, nous aurions seulement pu faire reséquencer l'échantillon de D116. Un autre facteur limitant la possibilité d'expansion de cette méthode a une plus grande portion du génome est que l'outil permettant d'extraire les *reads* n'a pas été conçu ni optimisé pour l'analyse réalisée ici. Il a donc fallu tester toutes les paires de mutations d'intérêt pour recréer les haplotypes, ce qui serait très long et intensif computationnellement à faire pour tout le génome. Il serait pertinent de développer un outil capable d'automatiser ce processus pour analyser un génome au complet. Cet outil serait utile non seulement pour étudier la diversité intra-hôte dans le cadre d'une infection virale, mais aussi dans le cadre d'étude sur les tumeurs. Contrairement aux approches permettant de reconstruire des haplotypes intra-hôtes basée sur la phylogénétique, ce genre de méthode pourrait fonctionner avec un seul échantillon.

Mes résultats ne permettent cependant pas de confirmer ou d'infirmer la théorie selon laquelle les patients immunosupprimés seraient des vecteurs d'évolution virale favorisant l'émergence de VOCs ou VOIs. Bien que j'aie identifié plusieurs mutations présentes dans les VOCs, notamment S:484K et S:484A, ces mutations ne sont pas détectables de manière constante au fil de l'infection. Une étude sur la transmission de mutation intra-hôte a déterminé qu'il était rare qu'un iSNV soit transmis d'un membre d'une maisonnée à un autre, et encore plus rare qu'il soit transmis dans la communauté (262). Qui plus est, il n'est pas certain que les patients soient contagieux lors de toute la durée de leur infection, puisque le virus n'est pas tout le temps détectable par PCR ou par mesure de charge virale. Cependant, tout nouveau variant a certainement une origine intra-hôte. Si l'hypothèse des patients immunosupprimés comme vecteurs d'évolution virale est juste, mes analyses suggèrent que la transmission des mutations intra-hôte n'est possible qu'à certains moments durant le cours de l'infection.

Finalement, il serait pertinent de tenter de répliquer ces résultats dans une cohorte de patients

immunosupprimés plus grande en prenant soin de contrôler les traitements reçus par ces derniers. Q1 et K1 ont tous les deux été traités avec entre autres, rituximab, un agent anti-CD20 (263). P1, l'autre patient de l'étude de Lee et al. a plutôt été traité par blinatumomab (anti-CD3/CD19) et inotuzumab ozogamicin (anti-CD22) (86,264,265). Il serait intéressant d'étudier l'effet de divers agents immunosuppresseurs sur les patrons mutationnels des infections de SARS-CoV-2 chez les personnes immunosupprimées.

4.1.6 Les mutations récurrentes

Mes analyses réalisées dans le cadre de divers projets ont montré plusieurs types de mutations récurrentes dans l'évolution génomique du SARS-CoV-2. Premièrement, nous avons déterminé que certains acides aminés sont préférentiellement retirés ou introduits dans les régions des épitopes par le biais de mutation C > U et G > U récurrentes. Deuxièmement, les deux variants Québécois qui ont caractérisé la deuxième vague dans la province, soient B.1.160 et B.1.1.176 ont tous les deux acquis la mutation S:T20I de manière convergente. De plus, B.1.1.176 partageait une mutation convergente avec le variant Russe B.1.1.317 qui circulait au même moment. Finalement, nous avons identifié diverses populations virales intra-hôte qui mutent de manière récurrente au codon S:484 dans deux individus immunosupprimés, mais également dans quatre individus de la population générale. Ces résultats suggèrent que ces divers types de mutations récurrentes sont le fruit de processus évolutifs qui ont augmenté la capacité d'évasion immunitaire du SARS-CoV-2.

Beaucoup d'efforts de recherche sont mis sur l'analyse de mutations dans la protéine S, mais mes résultats ainsi que plusieurs autres rapportent des mutations récurrentes dans d'autres protéines, par exemple l'ORF3a (266) et la N (267). Puisque mes résultats suggèrent que les mutations récurrentes ont un impact sur la pathologie, on peut conclure que les mutations exceptionnellement récurrentes dans d'autres protéines du SARS-CoV-2 mériteraient d'être étudiées plus en détail.

Une hypothèse alternative en ce qui a trait aux mutations récurrentes est que ces mutations confèrent une meilleure adaptation du virus à diverses conditions environnementales. Morris et collègues ont rapporté que la transmissibilité du virus était affectée de manière non-linéaire par la température et à l'humidité relative (268). Une des toutes premières mutations à gagner rapidement en fréquence à travers le monde est S:D614G qui améliore la stabilité de la structure de la S lorsqu'elle est exposée à des températures chaudes et froides (269). On peut donc se demander si la mutation convergente S:T20I, présente sur deux variants distincts au Québec pourrait conférer

un avantage dans le contexte environnemental québécois. S:T20I est apparue en mars 2020 sur B.1.1.176 et cette sous-lignée a connu une expansion à l'automne 2020. La même mutation est apparue à l'automne 2020 sur B.1.160, mais cette sous-lignée est restée à une fréquence de moins de 20% (Figure 2.7 a). Il faudrait réaliser des expériences en laboratoire pour pouvoir affirmer que cette mutation augmente la *fitness* du virus selon la température ou l'humidité. Néanmoins, le contexte spatiotemporel de l'apparition de cette mutation sur B.1.160 et de l'expansion de la sous-lignée de B.1.1.176 + S:T20I est intéressant et mériterait plus d'investigation.

4.1.7 Les difficultés liées à l'analyse du SARS-CoV-2

Le SARS-CoV-2 a été l'objet de milliers de publications scientifiques depuis son apparition en décembre 2019. Chercher le terme "SARS-CoV-2" sur la base de données PubMed (https://pubmed.ncbi.nlm.nih.gov/) donne 175,085 résultats en date du 31/08/2022. Malgré le taux de publication élevé à son sujet, l'étude de ce virus comporte plusieurs défis qu'il est important de souligner.

Tout d'abord, il existe des biais importants dans les données disponibles publiquement. Sur les 12,925,445 séquences disponibles dans GISAID en date du 31/08/2022, 6,877,305 proviennent des États-Unis d'Amérique et du Royaume-Uni, ce qui correspond à 53.2% des séquences disponibles (191). Le total de cas combiné pour ces deux pays est cependant de 118,246,512 sur 602,857,847, soit 19.6% des cas de SARS-CoV-2 répertoriés à travers le monde (3). En plus de ne pas être représentatif de la diversité virale réelle au niveau planétaire, ce biais peut mener à des analyses phylogénétiques erronées (270,271).

Le caractère urgent de la pandémie a également accéléré le processus de publication pour les articles en lien avec le SARS-CoV-2. Par exemple, en 2020, le nombre de jours médian pour l'acceptation d'un manuscrit sur SARS-CoV-2 ou COVID-19 était de 8 jours, contre 92 jours pour un article sur l'influenza (272). Il y a un taux de rétraction par an plus élevé pour les articles portant sur le virus par rapport à d'autres sujets (272,273). De plus, le partage et la citation d'articles pré-publiés, ou *preprints*, a augmenté durant la pandémie. Une étude a déterminé que 57,9% des *preprints* sur le virus ont au moins une citation, comparativement à 21.5% de *preprints* qui ne sont pas en lien avec SARS-CoV-2 (274). D'une part, on peut se poser des questions sur la validité de cette pratique puisque ces articles n'ont pas été révisés par les pairs. D'autre part, le taux de publication est tellement rapide que de ne pas se fier aux *preprints* pourraient mener à des résultats désuets avant même leur publication.

Finalement, un point qui a particulièrement affecté mon travail dans le cadre de ces projets est l'accès aux données générées au Québec. Il faut comprendre que cette pandémie est sans précédent, et que par conséquent, il n'y avait pas de système de partage de données efficace en place en 2020.

Toutes les séquences virales de la province ont été rendues accessibles à notre groupe de recherche via une collaboration avec le LSPQ, mais certaines analyses dépendent d'informations sur les patients qui ne pouvaient être rendues disponibles aux chercheurs. Par exemple, pour réaliser une étude longitudinale sur un patient, il faut être en mesure de faire le lien entre les séquences anonymisées du LSPQ et le dossier médical du patient. Bien que nous fassions partie d'un groupe de recherche sur une cohorte de patients atteint de la COVID-19, et que nous avons accès aux séquences du LSPQ, il nous a été difficile d'avoir les informations nécessaires pour faire le lien entre les deux types de données. À ce jour, il nous manque plusieurs liens patients-séquences qui auraient potentiellement pu nous permettre d'ajouter plus de patients à notre étude de cas, et ainsi donner plus de puissance à notre analyse. La protection des données des patients est un impératif éthique en recherche. Il est donc important de réfléchir à de meilleurs systèmes pour le partage et maillage de données, qui serait sécuritaire mais efficace.

4.2 Perspectives

En conclusion, les résultats présentés dans ce mémoire démontrent qu'il y a des pressions de sélection sur divers types de mutations récurrentes dans le génome du SARS-CoV-2, et que ces mutations ont des fonctions d'évasion immunitaire. L'analyse du virus à travers le prisme de ces mutations nous a permis de déterminer les patrons de substitution responsables d'un grand nombre de mutations qui impactent la capacité d'un CMH à se lier à un épitope. L'utilisation des simulations génomiques sous neutralité s'est avérée être un outil important dans le cadre de cette étude. Cette méthodologie est portable à d'autres pathogènes et pourrait être utilisée pour l'étude d'autres pathogènes émergents.

Notre équipe a participé à la surveillance des variants du SARS-CoV-2 au Québec et j'ai caractérisé les deux variants les plus fréquents dans la province durant la deuxième vague. Mes résultats montrent que, contrairement à ce qui a été observé pour l'ensemble du Canada, un de ces variants, soit B.1.1.176, était déjà présent lors de la première vague.

Finalement, dans le cadre d'une étude longitudinale sur des patients immunosupprimés, j'ai rapporté pour la première fois un patron de mutation suggérant la présence d'un réservoir viral de

SARS-CoV-2, supporté par une analyse phylogénétique. Ce résultat démontre l'importance des études longitudinales dans notre compréhension de l'évolution virale. J'ai également été en mesure d'identifier différentes quasiespèces qui possèdent des mutations favorisant l'évasion immunitaire dans un même échantillon, un patron trouvé dans deux patients atteint de cancers hématologiques. Un point marquant de cette analyse est l'étude de la co-occurence des mutations sur les *reads* de séquençage qui permet de différencier la présence de différentes populations virales de celle de mutations présentes sur un même haplotype. J'espère que ce type d'analyse deviendra pratique courante lors d'études de iSNV. De plus, mes résultats démontrent l'importance de l'utilisation de thérapies combinatoires pour traiter le SARS-CoV-2 dans un contexte d'infection de longue durée.

Références

1. Zhu N, Zhang D, Wang W, Li X, Yang B, Song J, et al. A Novel Coronavirus from Patients with Pneumonia in China, 2019. N Engl J Med. 2020 Feb 20;382(8):727–33.

2. Cucinotta D, Vanelli M. WHO Declares COVID-19 a Pandemic. Acta Bio Medica Atenei Parm. 2020;91(1):157–60.

3. Dong E, Du H, Gardner L. An interactive web-based dashboard to track COVID-19 in real time. Lancet Infect Dis. 2020 May 1;20(5):533–4.

4. Davis HE, Assaf GS, McCorkell L, Wei H, Low RJ, Re'em Y, et al. Characterizing long COVID in an international cohort: 7 months of symptoms and their impact. eClinicalMedicine [Internet]. 2021 Aug 1 [cited 2022 Aug 6];38. Available from: https://www.thelancet.com/journals/eclinm/article/PIIS2589-5370(21)00299-6/fulltext

5. Mousavizadeh L, Ghasemi S. Genotype and phenotype of COVID-19: Their roles in pathogenesis. J Microbiol Immunol Infect. 2021 Apr 1;54(2):159–63.

6. Xiao K, Zhai J, Feng Y, Zhou N, Zhang X, Zou JJ, et al. Isolation of SARS-CoV-2-related coronavirus from Malayan pangolins. Nature. 2020 Jul 1;583(7815):286–9.

7. Chen Y, Liu Q, Guo D. Emerging coronaviruses: Genome structure, replication, and pathogenesis. J Med Virol. 2020 Apr 1;92(4):418–23.

8. Andersen KG, Rambaut A, Lipkin WI, Holmes EC, Garry RF. The proximal origin of SARS-CoV-2. Nat Med. 2020 Apr;26(4):450–2.

9. Drosten C, Günther S, Preiser W, van der Werf S, Brodt HR, Becker S, et al. Identification of a Novel Coronavirus in Patients with Severe Acute Respiratory Syndrome. N Engl J Med. 2003 May 15;348(20):1967–76.

10. Donnelly CA, Ghani AC, Leung GM, Hedley AJ, Fraser C, Riley S, et al. Epidemiological determinants of spread of causal agent of severe acute respiratory syndrome in Hong Kong. Lancet Lond Engl. 2003 May 24;361(9371):1761–6.

11. Lu L, Zhong W, Bian Z, Li Z, Zhang K, Liang B, et al. A comparison of mortality-related risk factors of COVID-19, SARS, and MERS: A systematic review and meta-analysis. J Infect. 2020 Oct;81(4):e18–25.

12. WHO. Summary of probable SARS cases with onset of illness from 1 November 2002 to 31 July 2003 [Internet]. [cited 2022 Aug 3]. Available from: https://www.who.int/publications/m/item/summary-of-probable-sars-cases-with-onset-of-illness-from-1-november-2002-to-31-july-2003

13. Mackay IM, Arden KE. MERS coronavirus: diagnostics, epidemiology and transmission. Virol J. 2015 Dec 22;12(1):222.

14. Donnelly CA, Malik MR, Elkholy A, Cauchemez S, Kerkhove MDV. Worldwide Reduction in MERS Cases and Deaths since 2016 - Volume 25, Number 9—September 2019 - Emerging Infectious Diseases journal - CDC. [cited 2022 Aug 3]; Available from:

https://wwwnc.cdc.gov/eid/article/25/9/19-0143_article

15. Abate SM, Ali SA, Mantfardo B, Basu B. Rate of Intensive Care Unit admission and outcomes among patients with coronavirus: A systematic review and Meta-analysis. PLOS ONE. 2020 Jul 10;15(7):e0235653.

16. Ebrahim SH, Maher AD, Kanagasabai U, Alfaraj SH, Alzahrani NA, Alqahtani SA, et al. MERS-CoV Confirmation among 6,873 suspected persons and relevant Epidemiologic and Clinical Features, Saudi Arabia — 2014 to 2019. eClinicalMedicine [Internet]. 2021 Nov 1 [cited 2022 Aug 4];41. Available from: https://www.thelancet.com/journals/eclinm/article/PIIS2589-5370(21)00472-7/fulltext

17. Al-Raddadi RM, Shabouni OI, Alraddadi ZM, Alzalabani AH, Al-Asmari AM, Ibrahim A, et al. Burden of Middle East respiratory syndrome coronavirus infection in Saudi Arabia. J Infect Public Health. 2020 May 1;13(5):692–6.

18. WHO. WHO-convened global study of origins of SARS-CoV-2: China part [Internet]. World Health Organization; 2021 Mar [cited 2022 Aug 4]. Available from: https://apo.org.au/node/311637

19. Wu F, Zhao S, Yu B, Chen YM, Wang W, Song ZG, et al. A new coronavirus associated with human respiratory disease in China. Nature. 2020 Mar 1;579(7798):265–9.

20. Worobey M, Levy JI, Malpica Serrano L, Crits-Christoph A, Pekar JE, Goldstein SA, et al. The Huanan Seafood Wholesale Market in Wuhan was the early epicenter of the COVID-19 pandemic. Science. 2022 Aug 26;377(6609):951–9.

21. Wang Y, Mao JM, Wang GD, Luo ZP, Yang L, Yao Q, et al. Human SARS-CoV-2 has evolved to reduce CG dinucleotide in its open reading frames. Sci Rep. 2020 Jul 23;10(1):12331.

22. Gulyaeva AA, Gorbalenya AE. A nidovirus perspective on SARS-CoV-2. COVID-19. 2021 Jan 29;538:24–34.

23. Jungreis I, Sealfon R, Kellis M. SARS-CoV-2 gene content and COVID-19 mutation impact by comparing 44 Sarbecovirus genomes. Nat Commun. 2021 May 11;12(1):2642.

24. Cao Y, Yang R, Lee I, Zhang W, Sun J, Wang W, et al. Characterization of the SARS-CoV-2 E Protein: Sequence, Structure, Viroporin, and Inhibitors. Protein Sci. 2021 Jun 1;30(6):1114–30.

25. Alharbi SN, Alrefaei AF. Comparison of the SARS-CoV-2 (2019-nCoV) M protein with its counterparts of SARS-CoV and MERS-CoV species. J King Saud Univ Sci. 2021 Mar;33(2):101335.

26. Bai Z, Cao Y, Liu W, Li J. The SARS-CoV-2 Nucleocapsid Protein and Its Role in Viral Structure, Biological Functions, and a Potential Target for Drug or Vaccine Mitigation. Viruses. 2021;13(6).

27. Finkel Y, Mizrahi O, Nachshon A, Weingarten-Gabbay S, Morgenstern D, Yahalom-Ronen Y, et al. The coding capacity of SARS-CoV-2. Nature. 2021 Jan 1;589(7840):125–30.

28. Zhang Q, Xiang R, Huo S, Zhou Y, Jiang S, Wang Q, et al. Molecular mechanism of interaction between SARS-CoV-2 and host cells and interventional therapy. Signal Transduct Target Ther. 2021 Jun 11;6(1):1–19.

29. 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. The Lancet. 2020 Feb

22;395(10224):565–74.

30. Zhou P, Yang XL, Wang XG, Hu B, Zhang L, Zhang W, et al. A pneumonia outbreak associated with a new coronavirus of probable bat origin. Nature. 2020 Mar 1;579(7798):270–3.

31. Frutos R, Serra-Cobo J, Chen T, Devaux CA. COVID-19: Time to exonerate the pangolin from the transmission of SARS-CoV-2 to humans. Infect Genet Evol. 2020 Oct;84:104493.

32. Andrews CA. The Hardy-Weinberg Principle. Nat Educ Knowl. 2010;3(10):65.

33. Loewe L. Negative Selection. Nat Educ. 2008;1(1):59.

34. Vallender EJ, Lahn BT. Positive selection on the human genome. Hum Mol Genet. 2004 Oct 1;13(suppl_2):R245–54.

35. Asthana S, Schmidt S, Sunyaev S. A limited role for balancing selection. Trends Genet. 2005 Jan 1;21(1):30–2.

36. Kitrinos Kathryn M., Nelson Julie A. E., Resch Wolfgang, Swanstrom Ronald. Effect of a Protease Inhibitor-Induced Genetic Bottleneck on Human Immunodeficiency Virus Type 1 env Gene Populations. J Virol. 2005 Aug 15;79(16):10627–37.

37. Kliman R. Genetic Drift and Effective Population Size | Learn Science at Scitable. Nat Educ. 2008;1(3):3.

38. Bomba L, Walter K, Soranzo N. The impact of rare and low-frequency genetic variants in common disease. Genome Biol. 2017 Apr 27;18(1):77.

39. Zwart MP, Elena SF. Matters of Size: Genetic Bottlenecks in Virus Infection and Their Potential Impact on Evolution. Annu Rev Virol. 2015 Nov 9;2(1):161–79.

40. Yewdell JW. Antigenic drift: Understanding COVID-19. Immunity. 2021 Dec 14;54(12):2681–7.

41. Nadeau SA, Vaughan TG, Scire J, Huisman JS, Stadler T. The origin and early spread of SARS-CoV-2 in Europe. Proc Natl Acad Sci. 2021 Mar 2;118(9):e2012008118.

42. Lai MM. RNA recombination in animal and plant viruses. Microbiol Rev. 1992 Mar;56(1):61–79.

43. Liu H, Fu Y, Jiang D, Li G, Xie J, Cheng J, et al. Widespread horizontal gene transfer from double-stranded RNA viruses to eukaryotic nuclear genomes. J Virol. 2010 Nov;84(22):11876–87.

44. Wertheim JO, Wang JC, Leelawong M, Martin DP, Havens JL, Chowdhury MA, et al. Detection of SARS-CoV-2 intra-host recombination during superinfection with Alpha and Epsilon variants in New York City. Nat Commun. 2022 Jun 25;13(1):3645.

45. Emilie Burel, Philippe Colson, Jean-Christophe Lagier, Anthony LEVASSEUR, Marielle Bedotto, Philippe Lavrard, et al. Sequential appearance and isolation of a SARS-CoV-2 recombinant between two major SARS-CoV-2 variants in a chronically infected immunocompromised patient. medRxiv [Internet]. 2022 Mar 23; Available from: https://medrxiv.org/cgi/content/short/2022.03.21.22272673

46. Karki R, Pandya D, Elston RC, Ferlini C. Defining “mutation” and “polymorphism” in the era of personal genomics. BMC Med Genomics. 2015 Jul 15;8(1):37.

47. Kuhn JH, Bao Y, Bavari S, Becker S, Bradfute S, Brister JR, et al. Virus nomenclature below the species level: a standardized nomenclature for natural variants of viruses assigned to the family Filoviridae. Arch Virol. 2013 Jan 1;158(1):301–11.

48. Chen J, Shang J, Wang J, Sun Y. A binning tool to reconstruct viral haplotypes from assembled contigs. BMC Bioinformatics. 2019 Nov 4;20(1):544.

49. Boucher B, Jenna S. Genetic interaction networks: better understand to better predict. Front Genet. 2013 Dec 17;4:290.

50. Gabora L. Convergent Evolution. In: Maloy S, Hughes K, editors. Brenner's Encyclopedia of Genetics (Second Edition) [Internet]. San Diego: Academic Press; 2013. p. 178–80. Available from: https://www.sciencedirect.com/science/article/pii/B9780123749840003363

51. Upadhyay V, Patrick C, Lucas A, Mallela KMG. Convergent Evolution of Multiple Mutations Improves the Viral Fitness of SARS-CoV-2 Variants by Balancing Positive and Negative Selection. Biochemistry. 2022 Jun 7;61(11):963–80.

52. Artesi Maria, Bontems Sébastien, Göbbels Paul, Franckh Marc, Maes Piet, Boreux Raphaël, et al. A Recurrent Mutation at Position 26340 of SARS-CoV-2 Is Associated with Failure of the E Gene Quantitative Reverse Transcription-PCR Utilized in a Commercial Dual-Target Diagnostic Assay. J Clin Microbiol. 2020 Sep 22;58(10):e01598-20.

53. Barton MI, MacGowan SA, Kutuzov MA, Dushek O, Barton GJ, van der Merwe PA. Effects of common mutations in the SARS-CoV-2 Spike RBD and its ligand, the human ACE2 receptor on binding affinity and kinetics. eLife. 2021;10:e70658.

54. Stoltzfus A, Norris RW. On the Causes of Evolutionary Transition:Transversion Bias. Mol Biol Evol. 2016 Mar 1;33(3):595–602.

55. Payne JL, Menardo F, Trauner A, Borrell S, Gygli SM, Loiseau C, et al. Transition bias influences the evolution of antibiotic resistance in Mycobacterium tuberculosis. PLOS Biol. 2019 May 13;17(5):e3000265.

56. Lyons DM, Lauring AS. Evidence for the Selective Basis of Transition-to-Transversion Substitution Bias in Two RNA Viruses. Mol Biol Evol. 2017 Dec;34(12):3205–15.

57. Simmonds P, Ansari MA. Extensive C->U transition biases in the genomes of a wide range of mammalian RNA viruses; potential associations with transcriptional mutations, damage- or host-mediated editing of viral RNA. PLoS Pathog. 2021 Jun 1;17(6):e1009596.

58. 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 Jun 24;5(3):e00408-20.

59. Klimczak LJ, Randall TA, Saini N, Li JL, 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 Oct 2;15(10):e0237689.

60. Volz EM, Koelle K, Bedford T. Viral Phylodynamics. PLOS Comput Biol. 2013 Mar 21;9(3):e1002947.

61. Hadfield J, Megill C, Bell SM, Huddleston J, Potter B, Callender C, et al. Nextstrain: real-time tracking of pathogen evolution. Bioinformatics. 2018 Dec 1;34(23):4121–3.

62. 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 Nov 1;5(11):1403–7.

63. Robertson DL, Anderson JP, Bradac JA, Carr JK, Foley B, Funkhouser RK, et al. HIV-1 Nomenclature Proposal. Science. 2000 Apr 7;288(5463):55–55.

64. WHO/OIE/FAO H5N1 Evolution Working Group. Continued evolution of highly pathogenic avian influenza A (H5N1): updated nomenclature. Influenza Other Respir Viruses. 2012 Jan 1;6(1):1–5.

65. Hodcroft EB, Hadfield J, Neher RA, Bedford T. Year-letter Genetic Clade Naming for SARS-CoV-2 on Nextstrain.org [Internet]. Nextstrain. [cited 2022 Aug 7]. Available from: https://nextstrain.org//blog/2020-06-02-SARSCoV2-clade-naming

66. O'Toole Á, Scher E, Rambaut A. pangoLEARN description [Internet]. cov-lineages. [cited 2022 Aug 7]. Available from: https://cov-lineages.org/resources/pangolin/pangolearn.html

67. Kwak C, Clayton-Matthews A. Multinomial Logistic Regression. Nurs Res [Internet]. 2002;51(6). Available from: https://journals.lww.com/nursingresearchonline/Fulltext/2002/11000/Multinomial_Logistic_Regression.9.aspx

68. Mostefai F, Gamache I, N'Guessan A, Pelletier J, Huang J, Murall CL, et al. Population Genomics Approaches for Genetic Characterization of SARS-CoV-2 Lineages. Front Med [Internet]. 2022 [cited 2022 Aug 7];9. Available from: https://www.frontiersin.org/articles/10.3389/fmed.2022.826746

69. WHO. Suivi des variants du SARS-CoV-2 [Internet]. [cited 2022 Aug 15]. Available from: https://www.who.int/fr/activities/tracking-SARS-CoV-2-variants

70. 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 Apr 9;372(6538):eabg3055.

71. Tegally H, Wilkinson E, Giovanetti M, Iranzadeh A, Fonseca V, Giandhari J, et al. Emergence and rapid spread of a new severe acute respiratory syndrome-related coronavirus 2 (SARS-CoV-2) lineage with multiple spike mutations in South Africa [Internet]. medRxiv; 2020 [cited 2022 Aug 22]. p. 2020.12.21.20248640. Available from: https://www.medrxiv.org/content/10.1101/2020.12.21.20248640v1

72. Moreira FRR, D'arc M, Mariani D, Herlinger AL, Schiffler FB, Rossi ÁD, et al. Epidemiological dynamics of SARS-CoV-2 VOC Gamma in Rio de Janeiro, Brazil. Virus Evol. 2021 Dec 1;7(2):veab087.

73. Nonaka CKV, Franco MM, Gräf T, de Lorenzo Barcia CA, de Ávila Mendonça RN, de Sousa KAF, et al. Genomic Evidence of SARS-CoV-2 Reinfection Involving E484K Spike Mutation, Brazil. Emerg Infect Dis. 2021 May;27(5):1522–4.

74. Parums DV. Editorial: World Health Organization (WHO) Variants of Concern Lineages Under Monitoring (VOC-LUM) in Response to the Global Spread of Lineages and Sublineages of Omicron,

or B.1.1.529, SARS-CoV-2. Med Sci Monit [Internet]. 2022 Jul 1 [cited 2022 Aug 15];28. Available from: https://medscimonit.com/abstract/full/idArt/937676

75. Lewnard JA, Hong VX, Patel MM, Kahn R, Lipsitch M, Tartof SY. Clinical outcomes associated with SARS-CoV-2 Omicron (B.1.1.529) variant and BA.1/BA.1.1 or BA.2 subvariant infection in Southern California. Nat Med. 2022 Jun 8;1–11.

76. Davies MA, Kassanjee R, Rousseau P, Morden E, Johnson L, Solomon W, et al. Outcomes of laboratory-confirmed SARS-CoV-2 infection in the Omicron-driven fourth wave compared with previous waves in the Western Cape Province, South Africa. Trop Med Int Health. 2022;27(6):564–73.

77. Mohsin M, Mahmud S. Omicron SARS-CoV-2 variant of concern: A review on its transmissibility, immune evasion, reinfection, and severity. Medicine (Baltimore). 2022 May 13;101(19):e29165.

78. Tao K, Tzou PL, Kosakovsky Pond SL, Ioannidis JPA, Shafer RW. Susceptibility of SARS-CoV-2 Omicron Variants to Therapeutic Monoclonal Antibodies: Systematic Review and Meta-analysis. Microbiol Spectr. 2022 Jun 14;e0092622.

79. Rössler A, Riepler L, Bante D, von Laer D, Kimpel J. SARS-CoV-2 Omicron Variant Neutralization in Serum from Vaccinated and Convalescent Persons. N Engl J Med. 2022 Feb 17;386(7):698–700.

80. Navas Sonia, Martín Julio, Quiroga Juan Antonio, Castillo Inmaculada, Carreño Vicente. Genetic Diversity and Tissue Compartmentalization of the Hepatitis C Virus Genome in Blood Mononuclear Cells, Liver, and Serum from Chronic Hepatitis C Patients. J Virol. 1998 Feb 1;72(2):1640–6.

81. Wang Y, Wang D, Zhang L, Sun W, Zhang Z, Chen W, et al. Intra-host variation and evolutionary dynamics of SARS-CoV-2 populations in COVID-19 patients. Genome Med. 2021 Feb 22;13(1):30.

82. Domingo E, Sheldon J, Perales C. Viral quasispecies evolution. Microbiol Mol Biol Rev MMBR. 2012 Jun;76(2):159–216.

83. Alexander HK, Bonhoeffer S. Pre-existence and emergence of drug resistance in a generalized model of intra-host viral dynamics. Epidemics. 2012 Dec 1;4(4):187–202.

84. Ribeiro RM, Bonhoeffer S, Nowak MA. The frequency of resistant mutant virus before antiviral therapy. AIDS Lond Engl. 1998 Mar 26;12(5):461–5.

85. Ghosh RK, Ghosh SM, Chawla S. Recent advances in antiretroviral drugs. Expert Opin Pharmacother. 2011 Jan 1;12(1):31–46.

86. Lee JS, Yun KW, Jeong H, Kim B, Kim MJ, Park JH, et al. SARS-CoV-2 shedding dynamics and transmission in immunosuppressed patients. Virulence. 2022 Dec 31;13(1):1242–51.

87. McCarthy KR, Rennick LJ, Nambulli S, Robinson-McCarthy LR, Bain WG, Haidar G, et al. Recurrent deletions in the SARS-CoV-2 spike glycoprotein drive antibody escape. Science. 2021 Mar 12;371(6534):1139–42.

88. Pretti MAM, Galvani RG, Scherer NM, Farias AS, Boroni M. In silico analysis of mutant epitopes in new SARS-CoV-2 lineages suggest global enhanced CD8+ T cell reactivity and also signs of immune response escape. Infect Genet Evol. 2022 Apr;99:105236.

89. Skidmore PT, Kaelin EA, Holland LA, Maqsood R, Wu LI, Mellor NJ, et al. Emergence of a SARS-CoV-2 E484K variant of interest in Arizona. medRxiv. 2021 Jan 1;2021.03.26.21254367.

90. Jensen B, Luebke N, Feldt T, Keitel V, Brandenburger T, Kindgen-Milles D, et al. Emergence of the E484K mutation in SARS-COV-2-infected immunocompromised patients treated with bamlanivimab in Germany. Lancet Reg Health – Eur [Internet]. 2021 Sep 1 [cited 2022 Aug 18];8. Available from: https://www.thelancet.com/journals/lanepe/article/PIIS2666-7762(21)00141-1/fulltext

91. Verghese M, Jiang B, Iwai N, Mar M, Sahoo MK, Yamamoto F, et al. A SARS-CoV-2 Variant with L452R and E484Q Neutralization Resistance Mutations. J Clin Microbiol. 2021 Jun 18;59(7):e00741-21.

92. Ferreira IATM, Kemp SA, Datir R, Saito A, Meng B, Rakshit P, et al. SARS-CoV-2 B.1.617 Mutations L452R and E484Q Are Not Synergistic for Antibody Evasion. J Infect Dis. 2021 Sep 17;224(6):989–94.

93. McCallum M, Czudnochowski N, Rosen LE, Zepeda SK, Bowen JE, Walls AC, et al. Structural basis of SARS-CoV-2 Omicron immune evasion and receptor engagement. Science. 2022 Feb 25;375(6583):864–8.

94. Liu Z, VanBlargan LA, Bloyet LM, 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 Jan 1;2020.11.06.372037.

95. Scherer EM, Babiker A, Adelman MW, Allman B, Key A, Kleinhenz JM, et al. SARS-CoV-2 Evolution and Immune Escape in Immunocompromised Patients. N Engl J Med. 2022 Jun 23;386(25):2436–8.

96. Quaranta EG, Fusaro A, Giussani E, D'Amico V, Varotto M, Pagliari M, et al. SARS-CoV-2 intra-host evolution during prolonged infection in an immunocompromised patient. Int J Infect Dis. 2022 Sep 1;122:444–8.

97. Chun TW, Justement JS, Moir S, Hallahan CW, Maenza J, Mullins JI, et al. Decay of the HIV Reservoir in Patients Receiving Antiretroviral Therapy for Extended Periods: Implications for Eradication of Virus. J Infect Dis. 2007 Jun 15;195(12):1762–4.

98. Shen L, Siliciano RF. Viral reservoirs, residual viremia, and the potential of highly active antiretroviral therapy to eradicate HIV infection. J Allergy Clin Immunol. 2008 Jul 1;122(1):22–8.

99. Chen T, Hudnall SD. Anatomical mapping of human herpesvirus reservoirs of infection. Mod Pathol Off J U S Can Acad Pathol Inc. 2006 May;19(5):726–37.

100. Marx V. Scientists set out to connect the dots on long COVID. Nat Methods. 2021 May;18(5):449–53.

101. Goh D, Lim JCT, Fernández SB, Lee JN, Joseph CR, Neo ZW, et al. Persistence of residual SARS-CoV-2 viral antigen and RNA in tissues of patients with long COVID-19 [Internet]. In Review; 2022 [cited 2022 Aug 27]. Available from: https://www.researchsquare.com/article/rs-1379777/v2

102. Neurath MF, Überla K, Ng SC. Gut as viral reservoir: lessons from gut viromes, HIV and COVID-19. Gut. 2021 Sep 1;70(9):1605–8.

103. Chertow D, Stein S, Ramelli S, Grazioli A, Chung JY, Singh M, et al. SARS-CoV-2 infection and

persistence throughout the human body and brain [Internet]. In Review; 2021 [cited 2022 Aug 27]. Available from: https://www.researchsquare.com/article/rs-1139035/v1

104. Yang J, Petitjean SJL, Koehler M, Zhang Q, Dumitru AC, Chen W, et al. Molecular interaction and inhibition of SARS-CoV-2 binding to the ACE2 receptor. Nat Commun. 2020 Sep 11;11(1):4541.

105. Hamming I, Timens W, Bulthuis M, Lely A, Navis G, van Goor H. Tissue distribution of ACE2 protein, the functional receptor for SARS coronavirus. A first step in understanding SARS pathogenesis. J Pathol. 2004 Jun 1;203(2):631–7.

106. 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 Apr 16;181(2):271-280.e8.

107. Cantuti-Castelvetri L, Ojha R, Pedro LD, Djannatian M, Franz J, Kuivanen S, et al. Neuropilin-1 facilitates SARS-CoV-2 cell entry and infectivity. Science. 2020 Nov 13;370(6518):856–60.

108. Daly JL, Simonetti B, Klein K, Chen KE, Williamson MK, Antón-Plágaro C, et al. Neuropilin-1 is a host factor for SARS-CoV-2 infection. Science. 2020 Nov 13;370(6518):861–5.

109. Song P, Li W, Xie J, Hou Y, You C. Cytokine storm induced by SARS-CoV-2. Clin Chim Acta Int J Clin Chem. 2020 Oct;509:280–7.

110. Ye Q, Wang B, Mao J. The pathogenesis and treatment of the `Cytokine Storm' in COVID-19. J Infect. 2020 Jun;80(6):607–13.

111. Fifi JT, Mocco J. COVID-19 related stroke in young individuals. Lancet Neurol. 2020 Sep 1;19(9):713–5.

112. Li J, Long X, Zhu C, Wang H, Wang T, Lin Z, et al. Olfactory Dysfunction in Recovered Coronavirus Disease 2019 (COVID-19) Patients. Mov Disord Off J Mov Disord Soc. 2020 Jul;35(7):1100–1.

113. Katsoularis I, Fonseca-Rodríguez O, Farrington P, Jerndal H, Lundevaller EH, Sund M, et al. Risks of deep vein thrombosis, pulmonary embolism, and bleeding after covid-19: nationwide self-controlled cases series and matched cohort study. BMJ. 2022 Apr 6;377:e069590.

114. Moresco EMY, LaVine D, Beutler B. Toll-like receptors. Curr Biol. 2011 Jul 12;21(13):R488–93.

115. Platanias LC. Mechanisms of type-I- and type-II-interferon-mediated signalling. Nat Rev Immunol. 2005 May;5(5):375–86.

116. Campbell N, Reece J. Les hormones et le système endocrinien. In: Biologie 4e éd. 4th ed. ERPI; 2012. p. 1131–56.

117. Vivier E, Raulet DH, Moretta A, Caligiuri MA, Zitvogel L, Lanier LL, et al. Innate or Adaptive Immunity? The Example of Natural Killer Cells. Science. 2011 Jan 7;331(6013):44–9.

118. Fauriat C, Long EO, Ljunggren HG, Bryceson YT. Regulation of human NK-cell cytokine and chemokine production by target cell recognition. Blood. 2010 Mar 18;115(11):2167–76.

119. Eisenberg E, Levanon EY. A-to-I RNA editing - immune protector and transcriptome diversifier. Nat Rev Genet. 2018 Aug;19(8):473–90.

120. Sola I, Almazán F, Zúñiga S, Enjuanes L. Continuous and Discontinuous RNA Synthesis in Coronaviruses. Annu Rev Virol. 2015;2(1):265–88.

121. Wedekind JE, Dance GSC, Sowden MP, Smith HC. Messenger RNA editing in mammals: new members of the APOBEC family seeking roles in the family business. Trends Genet TIG. 2003 Apr;19(4):207–16.

122. Janahi EM, McGarvey MJ. The inhibition of hepatitis B virus by APOBEC cytidine deaminases. J Viral Hepat. 2013;20(12):821–8.

123. Malim MH. Natural resistance to HIV infection: The Vif–APOBEC interaction. C R Biol. 2006 Nov 1;329(11):871–5.

124. Katz DH, Benacerraf B. The Regulatory Influence of Activated T Cells on B Cell Responses to Antigen. In: Dixon FJ, Kunkel HG, editors. Advances in Immunology [Internet]. Academic Press; 1972 [cited 2022 Sep 1]. p. 1–94. Available from: https://www.sciencedirect.com/science/article/pii/S0065277608606835

125. Wieczorek M, Abualrous ET, Sticht J, Álvaro-Benito M, Stolzenberg S, Noé F, et al. Major Histocompatibility Complex (MHC) Class I and MHC Class II Proteins: Conformational Plasticity in Antigen Presentation. Front Immunol [Internet]. 2017 [cited 2022 Sep 1];8. Available from: https://www.frontiersin.org/articles/10.3389/fimmu.2017.00292

126. Reche PA, Reinherz EL. Definition of MHC supertypes through clustering of MHC peptide-binding repertoires. Methods Mol Biol Clifton NJ. 2007;409:163–73.

127. Sepil I, Lachish S, Hinks AE, Sheldon BC. Mhc supertypes confer both qualitative and quantitative resistance to avian malaria infections in a wild bird population. Proc R Soc B Biol Sci. 2013 May 22;280(1759):20130134.

128. Schwensow N, Fietz J, Dausmann KH, Sommer S. Neutral versus adaptive genetic variation in parasite resistance: importance of major histocompatibility complex supertypes in a free-ranging primate. Heredity. 2007 Sep;99(3):265–77.

129. 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):325–35.

130. Mauri C, Bosma A. Immune regulatory function of B cells. Annu Rev Immunol. 2012;30:221–41.

131. Finelli L, Gupta V, Petigara T, Yu K, Bauer KA, Puzniak LA. Mortality Among US Patients Hospitalized With SARS-CoV-2 Infection in 2020. JAMA Netw Open. 2021 Apr 8;4(4):e216556.

132. Dupont L, Snell LB, Graham C, Seow J, Merrick B, Lechmere T, et al. Neutralizing antibody activity in convalescent sera from infection in humans with SARS-CoV-2 and variants of concern. Nat Microbiol. 2021 Nov;6(11):1433–42.

133. Dougan M, Azizad M, Mocherla B, Gottlieb RL, Chen P, Hebert C, et al. A randomized, placebo-controlled clinical trial of bamlanivimab and etesevimab together in high-risk ambulatory patients with COVID-19 and validation of the prognostic value of persistently high viral load. Clin Infect Dis Off Publ Infect Dis Soc Am. 2021 Oct 28;ciab912.

134. Weinreich DM, Sivapalasingam S, Norton T, Ali S, Gao H, Bhore R, et al. REGEN-COV Antibody Combination and Outcomes in Outpatients with Covid-19. N Engl J Med. 2021 Dec 2;385(23):e81.

135. Gupta A, Gonzalez-Rojas Y, Juarez E, Crespo Casal M, Moya J, Rodrigues Falci D, et al. Effect of Sotrovimab on Hospitalization or Death Among High-risk Patients With Mild to Moderate COVID-19: A Randomized Clinical Trial. JAMA. 2022 Apr 5;327(13):1236–46.

136. 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 Feb;602(7898):657–63.

137. 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 Feb;602(7898):664–70.

138. Han X, Xu P, Ye Q. Analysis of COVID-19 vaccines: Types, thoughts, and application. J Clin Lab Anal. 2021;35(9):e23937.

139. Khobragade A, Bhate S, Ramaiah V, Deshpande S, Giri K, Phophle H, et al. Efficacy, safety, and immunogenicity of the DNA SARS-CoV-2 vaccine (ZyCoV-D): the interim efficacy results of a phase 3, randomised, double-blind, placebo-controlled study in India. The Lancet. 2022 Apr 2;399(10332):1313–21.

140. Hou X, Zaks T, Langer R, Dong Y. Lipid nanoparticles for mRNA delivery. Nat Rev Mater. 2021 Dec;6(12):1078–94.

141. Halperin SA, Ye L, MacKinnon-Cameron D, Smith B, Cahn PE, Ruiz-Palacios GM, et al. Final efficacy analysis, interim safety analysis, and immunogenicity of a single dose of recombinant novel coronavirus vaccine (adenovirus type 5 vector) in adults 18 years and older: an international, multicentre, randomised, double-blinded, placebo-controlled phase 3 trial. The Lancet. 2022 Jan 15;399(10321):237–48.

142. Ammerman NC, Beier-Sexton M, Azad AF. Growth and Maintenance of Vero Cell Lines. Curr Protoc Microbiol. 2008 Nov;APPENDIX:Appendix-4E.

143. Xia S, Duan K, Zhang Y, Zhao D, Zhang H, Xie Z, et al. Effect of an Inactivated Vaccine Against SARS-CoV-2 on Safety and Immunogenicity Outcomes: Interim Analysis of 2 Randomized Clinical Trials. JAMA. 2020 Sep 8;324(10):951–60.

144. Heidary M, Kaviar VH, Shirani M, Ghanavati R, Motahar M, Sholeh M, et al. A Comprehensive Review of the Protein Subunit Vaccines Against COVID-19. Front Microbiol [Internet]. 2022 [cited 2022 Aug 19];13. Available from: https://www.frontiersin.org/articles/10.3389/fmicb.2022.927306

145. Moghadas SM, Vilches TN, Zhang K, Wells CR, Shoukat A, Singer BH, et al. The Impact of Vaccination on Coronavirus Disease 2019 (COVID-19) Outbreaks in the United States. Clin Infect Dis Off Publ Infect Dis Soc Am. 2021 Jan 30;73(12):2257–64.

146. Abbasi J. Widespread Misinformation About Infertility Continues to Create COVID-19 Vaccine Hesitancy. JAMA. 2022 Mar 15;327(11):1013–5.

147. Loomba S, de Figueiredo A, Piatek SJ, de Graaf K, Larson HJ. Measuring the impact of COVID-19 vaccine misinformation on vaccination intent in the UK and USA. Nat Hum Behav. 2021 Mar;5(3):337–48.

148. Alakija A. Global North and South must work hand in glove to stop COVID-19. Nat Hum Behav. 2022 Feb;6(2):171–171.

149. Sow SO. Global South cannot just live with COVID-19. Nat Hum Behav. 2022 Feb;6(2):170–170.

150. Kausar S, Said Khan F, Ishaq Mujeeb Ur Rehman M, Akram M, Riaz M, Rasool G, et al. A review: Mechanism of action of antiviral drugs. Int J Immunopathol Pharmacol. 2021 Dec;35:20587384211002620.

151. Beigel JH, Tomashek KM, Dodd LE, Mehta AK, Zingman BS, Kalil AC, et al. Remdesivir for the Treatment of Covid-19 — Final Report. N Engl J Med. 2020 Oct 8;NEJMoa2007764.

152. Gottlieb RL, Vaca CE, Paredes R, Mera J, Webb BJ, Perez G, et al. Early Remdesivir to Prevent Progression to Severe Covid-19 in Outpatients. N Engl J Med. 2022 Jan 27;386(4):305–15.

153. Webb SA, Higgins AM, McArthur CJ. Glucocorticoid Dose in COVID-19: Lessons for Clinical Trials During a Pandemic. JAMA. 2021 Nov 9;326(18):1801–2.

154. Mizumoto K, Kagaya K, Zarebski A, Chowell G. Estimating the asymptomatic proportion of coronavirus disease 2019 (COVID-19) cases on board the Diamond Princess cruise ship, Yokohama, Japan, 2020. Eurosurveillance. 2020 Mar 12;25(10):2000180.

155. Stockman LJ, Massoudi MS, Helfand R, Erdman D, Siwek AM, Anderson LJ, et al. Severe acute respiratory syndrome in children. Pediatr Infect Dis J. 2007 Jan;26(1):68–74.

156. Payne AB, Gilani Z, Godfred-Cato S, Belay ED, Feldstein LR, Patel MM, et al. Incidence of Multisystem Inflammatory Syndrome in Children Among US Persons Infected With SARS-CoV-2. JAMA Netw Open. 2021 Jun 10;4(6):e2116420.

157. Zhang H, Wu Y, He Y, Liu X, Liu M, Tang Y, et al. Age-Related Risk Factors and Complications of Patients With COVID-19: A Population-Based Retrospective Study. Front Med. 2022 Jan 11;8:757459.

158. Davies NG, Klepac P, Liu Y, Prem K, Jit M, Pearson CAB, et al. Age-dependent effects in the transmission and control of COVID-19 epidemics. Nat Med. 2020 Aug 1;26(8):1205–11.

159. Zatz M, Silva MVR, de Castro MV, Naslavsky MS. The 90 plus: longevity and COVID-19 survival. Mol Psychiatry. 2022 Apr;27(4):1936–44.

160. CDC. Risk for COVID-19 Infection, Hospitalization, and Death By Age Group [Internet]. Centers for Disease Control and Prevention. 2020 [cited 2022 Sep 1]. Available from: https://www.cdc.gov/coronavirus/2019-ncov/covid-data/investigations-discovery/hospitalization-death-by-age.html

161. Nolen SC, Evans MA, Fischer A, Corrada MM, Kawas CH, Bota DA. Cancer—Incidence, prevalence and mortality in the oldest-old. A comprehensive review. Mech Ageing Dev. 2017 Jun;164:113–26.

162. Rebello CJ, Kirwan JP, Greenway FL. Obesity, the most common comorbidity in SARS-CoV-2: is leptin the link? Int J Obes. 2020 Sep 1;44(9):1810–7.

163. Shi Y, Yu X, Zhao H, Wang H, Zhao R, Sheng J. Host susceptibility to severe COVID-19 and establishment of a host risk score: findings of 487 cases outside Wuhan. Crit Care Lond Engl. 2020

Mar 18;24(1):108.

164. Divo MJ, Martinez CH, Mannino DM. Ageing and the epidemiology of multimorbidity. Eur Respir J. 2014 Oct;44(4):1055–68.

165. Thng ZX, De Smet MD, Lee CS, Gupta V, Smith JR, McCluskey PJ, et al. COVID-19 and immunosuppression: a review of current clinical experiences and implications for ophthalmology patients taking immunosuppressive drugs. Br J Ophthalmol. 2021 Mar 1;105(3):306.

166. Corey L, Beyrer C, Cohen MS, Michael NL, Bedford T, Rolland M. SARS-CoV-2 Variants in Patients with Immunosuppression. N Engl J Med. 2021 Aug 5;385(6):562–6.

167. Byrne AW, McEvoy D, Collins AB, Hunt K, Casey M, Barber A, et al. Inferred duration of infectious period of SARS-CoV-2: rapid scoping review and analysis of available evidence for asymptomatic and symptomatic COVID-19 cases. BMJ Open. 2020 Aug 1;10(8):e039856.

168. Pairo-Castineira E, Clohisey S, Klaric L, Bretherick AD, Rawlik K, Pasko D, et al. Genetic mechanisms of critical illness in COVID-19. Nature. 2021 Mar 1;591(7848):92–8.

169. Downes DJ, Cross AR, Hua P, Roberts N, Schwessinger R, Cutler AJ, et al. Identification of LZTFL1 as a candidate effector gene at a COVID-19 risk locus. Nat Genet. 2021 Nov;53(11):1606–15.

170. Kasela S, Daniloski Z, Bollepalli S, Jordan TX, tenOever BR, Sanjana NE, et al. Integrative approach identifies SLC6A20 and CXCR6 as putative causal genes for the COVID-19 GWAS signal in the 3p21.31 locus. Genome Biol. 2021 Aug 23;22:242.

171. He J, Feng D, de Vlas SJ, Wang H, Fontanet A, Zhang P, et al. Association of SARS susceptibility with single nucleic acid polymorphisms of OAS1 and MxA genes: a case-control study. BMC Infect Dis. 2006 Jul 6;6(1):106.

172. Zeberg H, Pääbo S. A genomic region associated with protection against severe COVID-19 is inherited from Neandertals. Proc Natl Acad Sci. 2021 Mar 2;118(9):e2026309118.

173. Smieszek SP, Polymeropoulos VM, Xiao C, Polymeropoulos CM, Polymeropoulos MH. Loss-of-function mutations in IFNAR2 in COVID-19 severe infection susceptibility. J Glob Antimicrob Resist. 2021 Sep;26:239–40.

174. Cruz R, Almeida SD de, Heredia ML, Quintela I, Ceballos FC, Pita G, et al. Novel genes and sex differences in COVID-19 severity. Hum Mol Genet. 2022 Jun 16;ddac132.

175. Kousathanas A, Pairo-Castineira E, Rawlik K, Stuckey A, Odhams CA, Walker S, et al. Whole-genome sequencing reveals host factors underlying critical COVID-19. Nature. 2022;607(7917):97–103.

176. Garcia-Vidal C, Moreno-García E, Hernández-Meneses M, Puerta-Alcalde P, Chumbita M, Garcia-Pouton N, et al. Personalized Therapy Approach for Hospitalized Patients with Coronavirus Disease 2019. Clin Infect Dis Off Publ Infect Dis Soc Am. 2022 Jan 7;74(1):127–32.

177. Butowt R, von Bartheld CS. Anosmia in COVID-19: Underlying Mechanisms and Assessment of an Olfactory Route to Brain Infection. The Neuroscientist. 2021 Dec 1;27(6):582–603.

178. Hampshire A, Trender W, Chamberlain SR, Jolly AE, Grant JE, Patrick F, et al. Cognitive deficits in

people who have recovered from COVID-19. eClinicalMedicine [Internet]. 2021 Sep 1 [cited 2022 Aug 6];39. Available from: https://www.thelancet.com/journals/eclinm/article/PIIS2589-5370(21)00324-2/fulltext

179. Bouziat R, Hinterleitner R, Brown JJ, Stencel-Baerenwald JE, Ikizler M, Mayassi T, et al. Reovirus infection triggers inflammatory responses to dietary antigens and development of celiac disease. Science. 2017 Apr 7;356(6333):44–50.

180. Chen X, Leach D, A. Hunter D, Sanfelippo D, J. Buell E, J. Zemple S, et al. Characterization of Intestinal Dendritic Cells in Murine Norovirus Infection. Open Immunol J [Internet]. 2011 Oct 20 [cited 2022 Aug 6];4(1). Available from: https://openimmunologyjournal.com/VOLUME/4/PAGE/22/ABSTRACT/

181. Afrin LB, Weinstock LB, Molderings GJ. Covid-19 hyperinflammation and post-Covid-19 illness may be rooted in mast cell activation syndrome. Int J Infect Dis. 2020 Nov;100:327–32.

182. Sanger F, Nicklen S, Coulson AR. DNA sequencing with chain-terminating inhibitors. Proc Natl Acad Sci. 1977 Dec;74(12):5463–7.

183. Morozova O, Marra MA. Applications of next-generation sequencing technologies in functional genomics. Genomics. 2008 Nov 1;92(5):255–64.

184. Shendure J, Mitra RD, Varma C, Church GM. Advanced sequencing technologies: methods and goals. Nat Rev Genet. 2004 May;5(5):335–44.

185. van Dijk EL, Auger H, Jaszczyszyn Y, Thermes C. Ten years of next-generation sequencing technology. Trends Genet. 2014 Sep 1;30(9):418–26.

186. Xiao T, Zhou W. The third generation sequencing: the advanced approach to genetic diseases. Transl Pediatr. 2020 Apr;9(2):163–73.

187. Muzzey D, Evans EA, Lieber C. Understanding the Basics of NGS: From Mechanism to Variant Calling. Curr Genet Med Rep. 2015;3(4):158–65.

188. Lin B, Hui J, Mao H. Nanopore Technology and Its Applications in Gene Sequencing. Biosensors. 2021 Jun 30;11(7):214.

189. Leggett RM, Clark MD. A world of opportunities with nanopore sequencing. J Exp Bot. 2017 Nov 28;68(20):5419–29.

190. Rang FJ, Kloosterman WP, de Ridder J. From squiggle to basepair: computational approaches for improving nanopore sequencing read accuracy. Genome Biol. 2018 Jul 13;19(1):90.

191. Khare S, Gurry C, Freitas L, Schultz MB, Bach G, Diallo A, et al. GISAID's Role in Pandemic Response. China CDC Wkly. 2021 Dec 3;3(49):1049–51.

192. Yang Z, Rannala B. Molecular phylogenetics: principles and practice. Nat Rev Genet. 2012 May 1;13(5):303–14.

193. Liò P, Goldman N. Models of Molecular Evolution and Phylogeny. Genome Res. 1998 Jan 12;8(12):1233–44.

194. Makarenkov V, Kevorkov D, Legendre P. 3 - Phylogenetic Network Construction Approaches. In:

Arora DK, Berka RM, Singh GB, editors. Applied Mycology and Biotechnology [Internet]. Elsevier; 2006 [cited 2022 Aug 24]. p. 61–97. (Applied Mycology and Biotechnology; vol. 6). Available from: https://www.sciencedirect.com/science/article/pii/S1874533406800067

195. Ritchie AM, Ho SYW. Influence of the tree prior and sampling scale on Bayesian phylogenetic estimates of the origin times of language families. J Lang Evol. 2019 Jul 1;4(2):108–23.

196. Fitch WM. On the Problem of Discovering the Most Parsimonious Tree. Am Nat. 1977 Mar;111(978):223–57.

197. van Dorp L, Acman M, Richard D, Shaw LP, Ford CE, Ormond L, et al. Emergence of genomic diversity and recurrent mutations in SARS-CoV-2. Infect Genet Evol. 2020 Sep 1;83:104351.

198. Morel B, Barbera P, Czech L, Bettisworth B, Hübner L, Lutteropp S, et al. Phylogenetic Analysis of SARS-CoV-2 Data Is Difficult. Mol Biol Evol. 2021 May 4;38(5):1777–91.

199. Katoh K, Misawa K, Kuma K, Miyata T. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic Acids Res. 2002 Jul 15;30(14):3059–66.

200. Nguyen LT, Schmidt HA, von Haeseler A, Minh BQ. IQ-TREE: A Fast and Effective Stochastic Algorithm for Estimating Maximum-Likelihood Phylogenies. Mol Biol Evol. 2015 Jan 1;32(1):268–74.

201. Lanave C, Preparata G, Saccone C, Serio G. A new method for calculating evolutionary substitution rates. J Mol Evol. 1984;20(1):86–93.

202. Kelleher J, Etheridge AM, McVean G. Efficient Coalescent Simulation and Genealogical Analysis for Large Sample Sizes. PLOS Comput Biol. 2016 May 4;12(5):e1004842.

203. Pekar JE, Magee A, Parker E, Moshiri N, Izhikevich K, Havens JL, et al. The molecular epidemiology of multiple zoonotic origins of SARS-CoV-2. Science. 2022 Jul 26;0(0):eabp8337.

204. Carvajal-Rodríguez A. Simulation of genes and genomes forward in time. Curr Genomics. 2010 Mar;11(1):58–61.

205. 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 Jan;5(1):vez003.

206. Fenton A. Editorial: Mathematical modelling of infectious diseases. Parasitology. 2016 Jun;143(7):801–4.

207. Boerlijst MC, Bonhoeffer S, Nowak MA. Viral Quasi-Species and Recombination. Proc Biol Sci. 1996;263(1376):1577–84.

208. Best K, Perelson AS. Mathematical modeling of within-host Zika virus dynamics. Immunol Rev. 2018 Sep;285(1):81–96.

209. Jenner AL, Aogo RA, Alfonso S, Crowe V, Deng X, Smith AP, et al. COVID-19 virtual patient cohort suggests immune mechanisms driving disease outcomes. PLOS Pathog. 2021 Jul 14;17(7):e1009753.

210. Deo RC. Machine Learning in Medicine. Circulation. 2015 Nov 17;132(20):1920–30.

211. Pesaranghader A, Pelletier J, Grenier JC, Poujol R, Hussin J. ImputeCoVNet: 2D ResNet Autoencoder for Imputation of SARS-CoV-2 Sequences [Internet]. bioRxiv; 2022 [cited 2022 Aug 24]. p. 2021.08.13.456305. Available from: https://www.biorxiv.org/content/10.1101/2021.08.13.456305v3

212. 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 Jul 2;48(W1):W449–54.

213. Nielsen M, Connelley T, Ternette N. Improved Prediction of Bovine Leucocyte Antigens (BoLA) Presented Ligands by Use of Mass-Spectrometry-Determined Ligand and in Vitro Binding Data. J Proteome Res. 2018 Jan 5;17(1):559–67.

214. INSPQ. Données COVID-19 au Québec [Internet]. INSPQ. [cited 2022 Sep 1]. Available from: https://www.inspq.qc.ca/covid-19/donnees

215. Murall CL, Fournier E, Galvez JH, N'Guessan A, Reiling SJ, Quirion PO, et al. A small number of early introductions seeded widespread transmission of SARS-CoV-2 in Québec, Canada. Genome Med. 2021 Oct 28;13(1):169.

216. Godin A, Xia Y, Buckeridge DL, Mishra S, Douwes-Schultz D, Shen Y, et al. The role of case importation in explaining differences in early SARS-CoV-2 transmission dynamics in Canada—A mathematical modeling study of surveillance data. Int J Infect Dis. 2021 Jan 1;102:254–9.

217. Moreira S. Le séquençage génomique pour la surveillance du SRAS-CoV-2 [Internet]. https://www.aipi.qc.ca/. 2021 [cited 2022 Sep 1]. Available from: https://www.aipi.qc.ca/wp-content/uploads/2021/06/Pleniere-9-Sandrine-Moreira-1.pdf

218. Hamelin DJ, Fournelle D, Grenier JC, Schockaert J, Kovalchik KA, Kubiniok P, et al. The mutational landscape of SARS-CoV-2 variants diversifies T cell targets in an HLA-supertype-dependent manner. Cell Syst. 2022 Feb 16;13(2):143-157.e3.

219. 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. Cooper VS, Perry GH, editors. eLife. 2020 Sep 2;9:e60067.

220. Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet.journal. 2011 May 2;17(1):10–2.

221. Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM [Internet]. arXiv; 2013 [cited 2022 Aug 15]. Available from: http://arxiv.org/abs/1303.3997

222. Tarasov A, Vilella AJ, Cuppen E, Nijman IJ, Prins P. Sambamba: fast processing of NGS alignment formats. Bioinforma Oxf Engl. 2015 Jun 15;31(12):2032–4.

223. Grubaugh ND, Gangavarapu K, Quick J, Matteson NL, De Jesus JG, Main BJ, et al. An amplicon-based sequencing framework for accurately measuring intrahost virus diversity using PrimalSeq and iVar. bioRxiv. 2018 Jan 1;383513.

224. Danecek P, Bonfield JK, Liddle J, Marshall J, Ohan V, Pollard MO, et al. Twelve years of SAMtools and BCFtools. GigaScience. 2021 Feb 16;10(2):giab008.

225. Garrison E, Marth G. Haplotype-based variant detection from short-read sequencing [Internet]. arXiv;

2012 [cited 2022 Aug 15]. Available from: http://arxiv.org/abs/1207.3907

226. Ueno Y, Arita M, Kumagai T, Asai K. Processing sequence annotation data using the Lua programming language. Genome Inform Int Conf Genome Inform. 2003;14:154–63.

227. Li H. Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics. 2018 Sep 15;34(18):3094–100.

228. Loman NJ, Quick J, Simpson JT. A complete bacterial genome assembled de novo using only nanopore sequencing data. Nat Methods. 2015 Aug 1;12(8):733–5.

229. 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 Feb 5;371(6529):eabf4063.

230. 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 Rep Med. 2021 Jun 15;2(6):100312.

231. McLaughlin A, Montoya V, Miller RL, Mordecai GJ, Canadian COVID-19 Genomics Network (CanCOGen) Consortium, Worobey M, et al. Genomic epidemiology of the first two waves of SARS-CoV-2 in Canada. Cobey SE, Sawyer SL, Gutierrez B, editors. eLife. 2022 Aug 2;11:e73896.

232. Fournier PE, Colson P, Levasseur A, Devaux CA, Gautret P, Bedotto M, et al. Emergence and outcomes of the SARS-CoV-2 'Marseille-4' variant. Int J Infect Dis. 2021 May 1;106:228–36.

233. Hodcroft EB, Zuber M, Nadeau S, Vaughan TG, Crawford KHD, Althaus CL, et al. Spread of a SARS-CoV-2 variant through Europe in the summer of 2020. Nature. 2021 Jul;595(7869):707–12.

234. Wang H xin, Zhang L, Liang Z teng, Nie J hui, Wu J jing, Li Q qian, et al. Infectivity and antigenicity of pseudoviruses with high-frequency mutations of SARS-CoV-2 identified in Portugal. Arch Virol. 2022 Feb 1;167(2):459–70.

235. Klink GV, Safina KR, Garushyants SK, Moldovan M, Nabieva E, CoRGI consortium, et al. Spread of endemic SARS-CoV-2 lineages in Russia - SARS-CoV-2 coronavirus / nCoV-2019 Genomic Epidemiology [Internet]. Virological. 2021 [cited 2022 Sep 1]. Available from: https://virological.org/t/spread-of-endemic-sars-cov-2-lineages-in-russia/689

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

237. Burra P, Soto-Díaz K, Chalen I, Gonzalez-Ricon RJ, Istanto D, Caetano-Anollés G. Temperature and Latitude Correlate with SARS-CoV-2 Epidemiological Variables but not with Genomic Change Worldwide. Evol Bioinforma. 2021 Jan 1;17:1176934321989695.

238. Caetano-Anollés K, Hernandez N, Mughal F, Tomaszewski T, Caetano-Anollés G. Chapter 2 - The seasonal behaviour of COVID-19 and its galectin-like culprit of the viral spike. In: Pavia CS, Gurtler V, editors. Methods in Microbiology [Internet]. Academic Press; 2022. p. 27–81. Available from: https://www.sciencedirect.com/science/article/pii/S0580951721000350

239. Shah M, Ahmad B, Choi S, Woo HG. Mutations in the SARS-CoV-2 spike RBD are responsible for stronger ACE2 binding and poor anti-SARS-CoV mAbs cross-neutralization. Comput Struct Biotechnol J. 2020 Jan 1;18:3402–14.

240. VanBlargan LA, Errico JM, Halfmann PJ, Zost SJ, Crowe JE, Purcell LA, et al. An infectious SARS-CoV-2 B.1.1.529 Omicron virus escapes neutralization by therapeutic monoclonal antibodies. Nat Med. 2022 Mar;28(3):490–5.

241. Focosi D, McConnell S, Casadevall A, Cappello E, Valdiserra G, Tuccori M. Monoclonal antibody therapies against SARS-CoV-2. Lancet Infect Dis [Internet]. 2022 Jul 5 [cited 2022 Oct 12];0(0). Available from: https://www.thelancet.com/journals/laninf/article/PIIS1473-3099(22)00311-5/fulltext

242. Bourgey M, Dali R, Eveleigh R, Chen KC, Letourneau L, Fillon J, et al. GenPipes: an open-source framework for distributed and scalable genomic analyses. GigaScience. 2019 Jun 1;8(6):giz037.

243. Grubaugh ND, Gangavarapu K, Quick J, Matteson NL, De Jesus JG, Main BJ, et al. An amplicon-based sequencing framework for accurately measuring intrahost virus diversity using PrimalSeq and iVar. Genome Biol. 2019 Jan 8;20(1):8.

244. O'Toole Á, Scher E, Underwood A, Jackson B, Hill V, McCrone JT, et al. Assignment of epidemiological lineages in an emerging pandemic using the pangolin tool. Virus Evol. 2021 Dec 1;7(2):veab064.

245. Alkodsi A, Meriranta L, Pasanen A, Leppä S. ctDNAtools: An R package to work with sequencing data of circulating tumor DNA. bioRxiv. 2020 Jan 1;2020.01.27.912790.

246. Weisblum Y, Schmidt F, Zhang F, DaSilva J, Poston D, Lorenzi JC, et al. Escape from neutralizing antibodies by SARS-CoV-2 spike protein variants. eLife. 9:e61312.

247. Wang R, Chen J, Gao K, Hozumi Y, Yin C, Wei GW. Characterizing SARS-CoV-2 mutations in the United States. Res Sq. 2020 Aug 11;rs.3.rs-49671.

248. Brodin P, Casari G, Townsend L, O'Farrelly C, Tancevski I, Löffler-Ragg J, et al. Studying severe long COVID to understand post-infectious disorders beyond COVID-19. Nat Med. 2022 May 1;28(5):879–82.

249. Martínez-Colón GJ, Ratnasiri K, Chen H, Jiang S, Zanley E, Rustagi A, et al. SARS-CoV-2 infection drives an inflammatory response in human adipose tissue through infection of adipocytes and macrophages. Sci Transl Med. 0(0):eabm9151.

250. Woolthuis RG, van Dorp CH, Keşmir C, de Boer RJ, van Boven M. Long-term adaptation of the influenza A virus by escaping cytotoxic T-cell recognition. Sci Rep. 2016 Sep 15;6:33334.

251. 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 Apr;458(7238):641–5.

252. Grant M, Larijani M. Evasion of adaptive immunity by HIV through the action of host APOBEC3G/F enzymes. AIDS Res Ther. 2017 Sep 12;14(1):44.

253. 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 Med. 2009 Dec 28;207(1):39–49.

254. Bajaj K, Madhusudhan MS, Adkar BV, Chakrabarti P, Ramakrishnan C, Sali A, et al. Stereochemical criteria for prediction of the effects of proline mutations on protein stability. PLoS Comput Biol. 2007 Dec;3(12):e241.

255. Messer PW. SLiM: Simulating Evolution with Selection and Linkage. Genetics. 2013 Aug 1;194(4):1037–9.

256. Haller BC, Messer PW. SLiM 3: Forward Genetic Simulations Beyond the Wright–Fisher Model. Mol Biol Evol. 2019 Mar 1;36(3):632–7.

257. Yang Y, Yang M, Yuan J, Wang F, Wang Z, Li J, et al. Laboratory Diagnosis and Monitoring the Viral Shedding of SARS-CoV-2 Infection. The Innovation [Internet]. 2020 Nov 25 [cited 2022 Sep 2];1(3). Available from: https://www.cell.com/the-innovation/abstract/S2666-6758(20)30064-3

258. Sethuraman N, Jeremiah SS, Ryo A. Interpreting Diagnostic Tests for SARS-CoV-2. JAMA. 2020 Jun 9;323(22):2249–51.

259. Stoler N, Nekrutenko A. Sequencing error profiles of Illumina sequencing instruments. NAR Genomics Bioinforma. 2021 Mar 1;3(1):lqab019.

260. Manley LJ, Ma D, Levine SS. Monitoring Error Rates In Illumina Sequencing. J Biomol Tech JBT. 2016 Dec;27(4):125–8.

261. Jain M, Koren S, Miga KH, Quick J, Rand AC, Sasani TA, et al. Nanopore sequencing and assembly of a human genome with ultra-long reads. Nat Biotechnol. 2018 Apr;36(4):338–45.

262. Braun KM, Moreno GK, Wagner C, Accola MA, Rehrauer WM, Baker DA, et al. Acute SARS-CoV-2 infections harbor limited within-host diversity and transmit via tight transmission bottlenecks. PLOS Pathog. 2021 Aug 23;17(8):e1009849.

263. Onrust SV, Lamb HM, Balfour JA. Rituximab. Drugs. 1999 Jul;58(1):79–88; discussion 89-90.

264. Dahl J, Marx K, Jabbour E. Inotuzumab ozogamicin in the treatment of acute lymphoblastic leukemia. Expert Rev Hematol. 2016;9(4):329–34.

265. Locatelli F, Zugmaier G, Rizzari C, Morris JD, Gruhn B, Klingebiel T, et al. Effect of Blinatumomab vs Chemotherapy on Event-Free Survival Among Children With High-risk First-Relapse B-Cell Acute Lymphoblastic Leukemia: A Randomized Clinical Trial. JAMA. 2021 Mar 2;325(9):843–54.

266. Hassan SkS, Attrish D, Ghosh S, Choudhury PP, Roy B. Pathogenic perspective of missense mutations of ORF3a protein of SARS-CoV-2. Virus Res. 2021 Jul 15;300:198441.

267. Rahman MS, Islam MR, Alam ASMRU, Islam I, Hoque MN, Akter S, et al. Evolutionary dynamics of SARS-CoV-2 nucleocapsid protein and its consequences. J Med Virol. 2021;93(4):2177–95.

268. Morris DH, Yinda KC, Gamble A, Rossine FW, Huang Q, Bushmaker T, et al. Mechanistic theory predicts the effects of temperature and humidity on inactivation of SARS-CoV-2 and other enveloped viruses. Garrett WS, Ogbunugafor CB, Handel A, editors. eLife. 2021 Apr 27;10:e65902.

269. Yang TJ, Yu PY, Chang YC, Hsu STD. D614G mutation in the SARS-CoV-2 spike protein enhances viral fitness by desensitizing it to temperature-dependent denaturation. J Biol Chem [Internet]. 2021 Oct 1 [cited 2022 Sep 2];297(4). Available from: https://www.jbc.org/article/S0021-9258(21)01041-3/abstract

270. Lemey P, Hong SL, Hill V, Baele G, Poletto C, Colizza V, et al. Accommodating individual travel history and unsampled diversity in Bayesian phylogeographic inference of SARS-CoV-2. Nat

Commun. 2020 Oct 9;11(1):5110.

271. Mavian C, Pond SK, Marini S, Magalis BR, Vandamme AM, Dellicour S, et al. Sampling bias and incorrect rooting make phylogenetic network tracing of SARS-COV-2 infections unreliable. Proc Natl Acad Sci. 2020 Jun 9;117(23):12522–3.

272. Schonhaut L, Costa-Roldan I, Oppenheimer I, Pizarro V, Han D, Díaz F. Scientific publication speed and retractions of COVID-19 pandemic original articles. Rev Panam Salud Pública. 2022 Apr 12;46:e25.

273. Cortegiani A, Catalisano G, Ippolito M, Giarratano A, Absalom AR, Einav S. Retracted papers on SARS-CoV-2 and COVID-19. BJA Br J Anaesth. 2021 Apr;126(4):e155–6.

274. Fraser N, Brierley L, Dey G, Polka JK, Pálfy M, Nanni F, et al. The evolving role of preprints in the dissemination of COVID-19 research and their impact on the science communication landscape. PLOS Biol. 2021 Apr 2;19(4):e3000959.

Annexe A

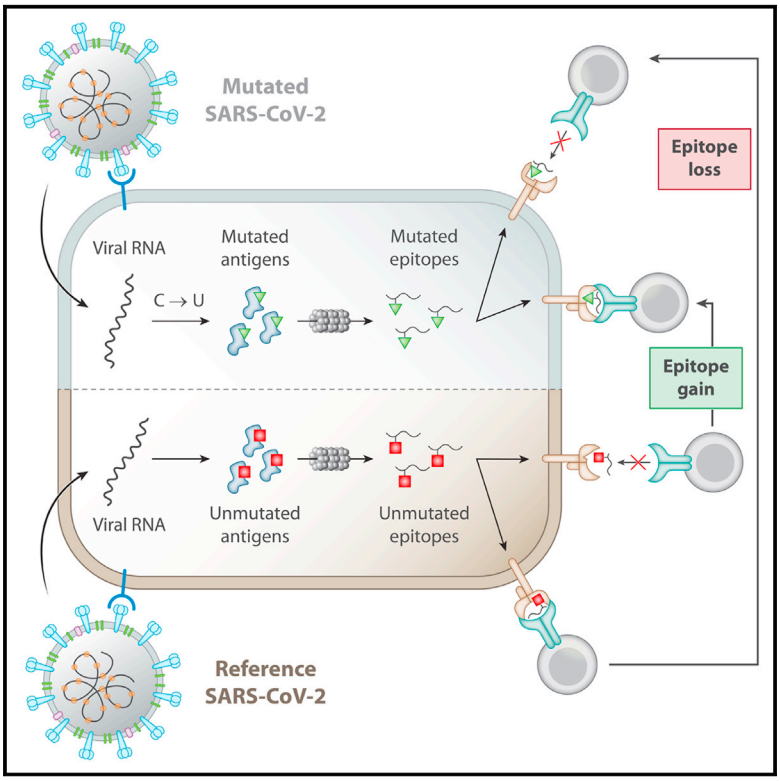
Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID-19. The COVID-19 resource centre is hosted on Elsevier Connect, the company's public news and information website.

Article

Cell Systems

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

Graphical abstract



Highlights

- Link between SARS-COV-2 mutation biases, HLA alleles, and immune escape
- Dominant C→U SARS-CoV-2 mutations diversify the CD8$^+$ T cell epitope repertoire
- Mutation biases modulate epitope presentation in an HLA-supertype-dependent manner
- Preferential loss of epitopes in B7 HLA supertype due to prevalent loss of proline

Authors

David J. Hamelin, Dominique Fournelle, Jean-Christophe Grenier, ..., Hélène Decaluwe, Julie Hussin, Etienne Caron

Correspondence

julie.hussin@umontreal.ca (J.H.), etienne.caron@umontreal.ca (E.C.)

In brief

Hamelin et al. investigated the global mutation landscape of SARS-CoV-2 by interrogating 330,246 SARS-CoV-2 sequences from GISAID. The dominant C→U mutation type was found to diversify the repertoire of experimentally validated SARS-CoV-2 CD8$^+$ T cell epitopes in an HLA-supertype-dependent manner. Notably, the prevalent removal of proline was predicted to preferentially abrogate epitopes presented by the B7 HLA supertype. This model lays a foundation for testing the impact of SARS-COV-2 mutants on T cell escape in an HLA-dependent manner.

Hamelin et al., 2022, Cell Systems *13*, 143–157
February 16, 2022
https://doi.org/10.1016/j.cels.2021.09.013

CellPress

Cell Systems

CellPress

Article

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

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,*] and Etienne Caron[1,7,8,*]

[1]CHU Sainte-Justine Research Center, Montréal, 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, 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, Montréal, QC, Canada
[8]Lead contact
*Correspondence: julie.hussin@umontreal.ca (J.H.), etienne.caron@umontreal.ca (E.C.)
https://doi.org/10.1016/j.cels.2021.09.013

SUMMARY

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.

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 upward 4.6 million deaths and 222 million confirmed cases worldwide (https://coronavirus.jhu.edu/map.html), making vaccine development and deployment an urgent necessity (Callaway, 2020). As a result of unprecedent efforts, vaccines have been developed and licensed within a 1-year time frame and are currently being widely distributed for mass vaccination (Krammer, 2020).

A clear understanding of the natural protective immune response against SARS-CoV-2 is essential for the development of vaccines that can trigger lifelong immunologic memory to prevent COVID-19 (Sette and Crotty, 2021; Stephens and McElrath, 2020). Since the start of the pandemic, numerous studies have investigated the association between COVID-19 clinical outcomes and SARS-CoV-2-specific antibodies and T cell immunity (Altmann and Boyton, 2020; Le Bert et al., 2020; Braun et al., 2020; Grifoni et al., 2020a; Long et al., 2020a, 2020b; Meckiff et al., 2020; Moderbacher et al., 2020; Sekine et al., 2020; Weiskopf et al., 2020). Memory may be a concern for SARS-CoV-2-specific antibodies, as they were recently shown to be present in convalescent COVID-19 patients in a highly heterogeneous manner (Dan et al., 2021) and, in some cases, observed to be undetectable just a few months post-infection (Seow et al., 2020). In contrast, an increasing number of studies point $CD4^+$ and $CD8^+$ T cells as key regulators of disease severity (Liao et al., 2020; Moderbacher et al., 2020; Schub et al., 2020; Weiskopf et al., 2020; Zhou et al., 2020). Studies of convalescent COVID-19 patients have also shown broad and strong $CD4^+$ and $CD8^+$ memory T cells induced by SARS-COV-2, suggesting that T cells may provide robust and long-term protection (Dan et al., 2021; Peng et al., 2020). Similar observations have been made for the most closely related human coronavirus, SARS-CoV, for which T cells have been detected 11 years (Ng et al., 2016) and 17 years (Le Bert et al., 2020) after the initial infection, whereas antibodies were noted to be undetectable after 2–3 years (Liu et al., 2006; Tang et al., 2011; Wu et al., 2007). Thus, vaccines designed to produce robust T cell responses are likely to be important for eliciting lifelong immunity against COVID-19 in the general population.

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

T cell epitopes that have been mapped across the entire SARS-CoV-2 viral proteome are reference peptides that are unmutated because they have been predicted from the sequence of the original SARS-CoV-2 that emerged from Wuhan, China (Grifoni et al., 2020b). However, analyses of unprecedented numbers of SARS-CoV-2 genome assemblies available from large-scale efforts have shown that SARS-CoV-2 is accumulating an array of mutations across the world, leading to the circulation and transmission of thousands of variants around the globe at various frequencies, and hence, contributing to the global genomic diversification of SARS-CoV-2 (van Dorp et al., 2020a; Korber et al., 2020; Laamarti et al., 2020; Mercatelli and Giorgi, 2020; Mercatelli et al., 2021; Popa et al., 2020). This extensive diversification has resulted in widespread variants such as B.1.1.7 (alpha), B.1.351 (beta), and B.1.617.2 (delta) (Cherian et al., 2021; Frampton et al., 2021; Tegally et al., 2021). 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 (Tarke et al., 2021b). 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 (van Dorp et al., 2020a; Di Giorgio et al., 2020; Klimczak et al., 2020; Kosuge et al., 2020; Li et al., 2020; Matyášek and Kovařík, 2020; Rice et al., 2020; Simmonds, 2020; Wang et al., 2020). Since shown for other viruses (Grant and Larijani, 2017; Monajemi et al., 2014), we reasoned that the putative action of such host enzymes during the first year of the pandemic could lead to the large-scale escape from immunodominant and protective SARS-CoV-2-specific T cell responses, thereby potentially compromising their effectiveness to control the virus at the population scale.

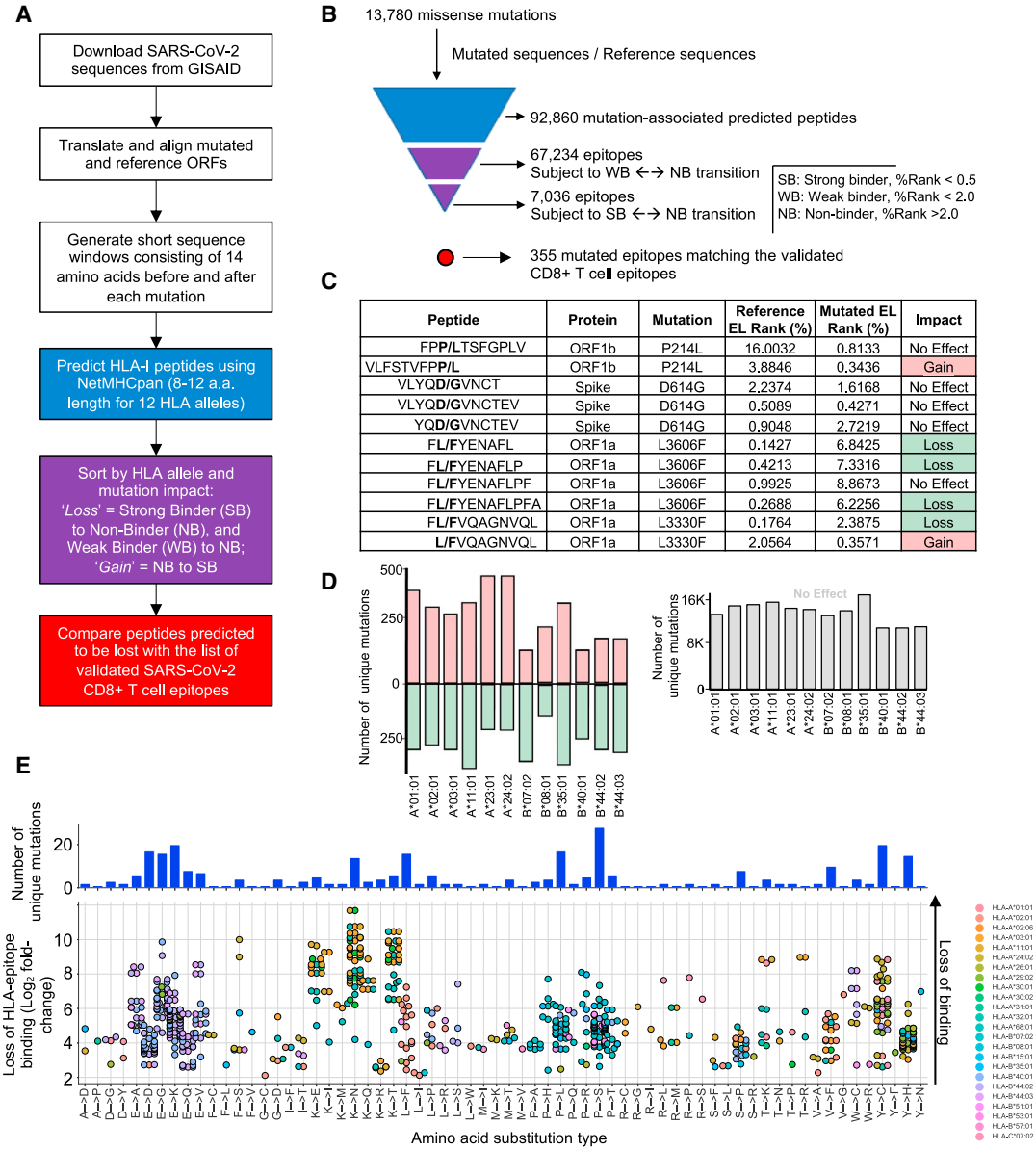
In this study, we report a comprehensive study of the global genetic diversity of SARS-CoV-2 to expose the impact of mutation bias on epitope presentation and HLA-restricted T cell response within the first year of the pandemic, from December 2019 to December 2020. More specifically, we asked the following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

RESULTS

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

As of May 2021, nearly 1.7 million complete SARS-CoV-2 genome assemblies are publicly available via the Global Initiative on Sharing All Influenza Data (GISAID) repository. In the context of this large-scale effort, we performed a global analysis of SARS-CoV-2 genomes to assess whether mutations that emerged during the first year of the pandemic could disrupt HLA binding of clinically relevant SARS-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 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%) (Korber et al., 2020) (Tables S1 and S2). Next, we implemented a bioinformatics pipeline to assess the impact of these mutations on HLA binding for 620 unique SARS-CoV-2 HLA class-I epitopes that were recently reported to trigger a $CD8^+$ T cell response in acute or convalescent COVID-19 patients (Quadeer et al., 2021; Tarke et al., 2021a) (see STAR 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; Figure 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 (Figures 2A and 2B). 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 2B). Epitopes located in the RBD vaccine locus were also impacted by mutations (Figure 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 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 with the reference/unmutated epitope YQDVNCTEV (Figures S1A and S1B). Our analysis also identified a mutation in the HLA-B*07:02-restricted N105 epitope SPRWYFYYL, which is one of the most immunodominant SARS-CoV-2 epitope (Ferretti et al., 2020; Kared et al., 2021; Saini et al., 2021; Schulien et al., 2021; Sekine et al., 2020; Tarke et al., 2021a). Although relatively



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 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 (1) missense mutations in SARS-CoV-2 genomes, (2) predicted class I mutated peptides, (3) predicted class I peptides subject to Weak Binder (WB) to non-binder (NB) and strong binder (SB) to NB transition (epitope loss category), and (4) 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.

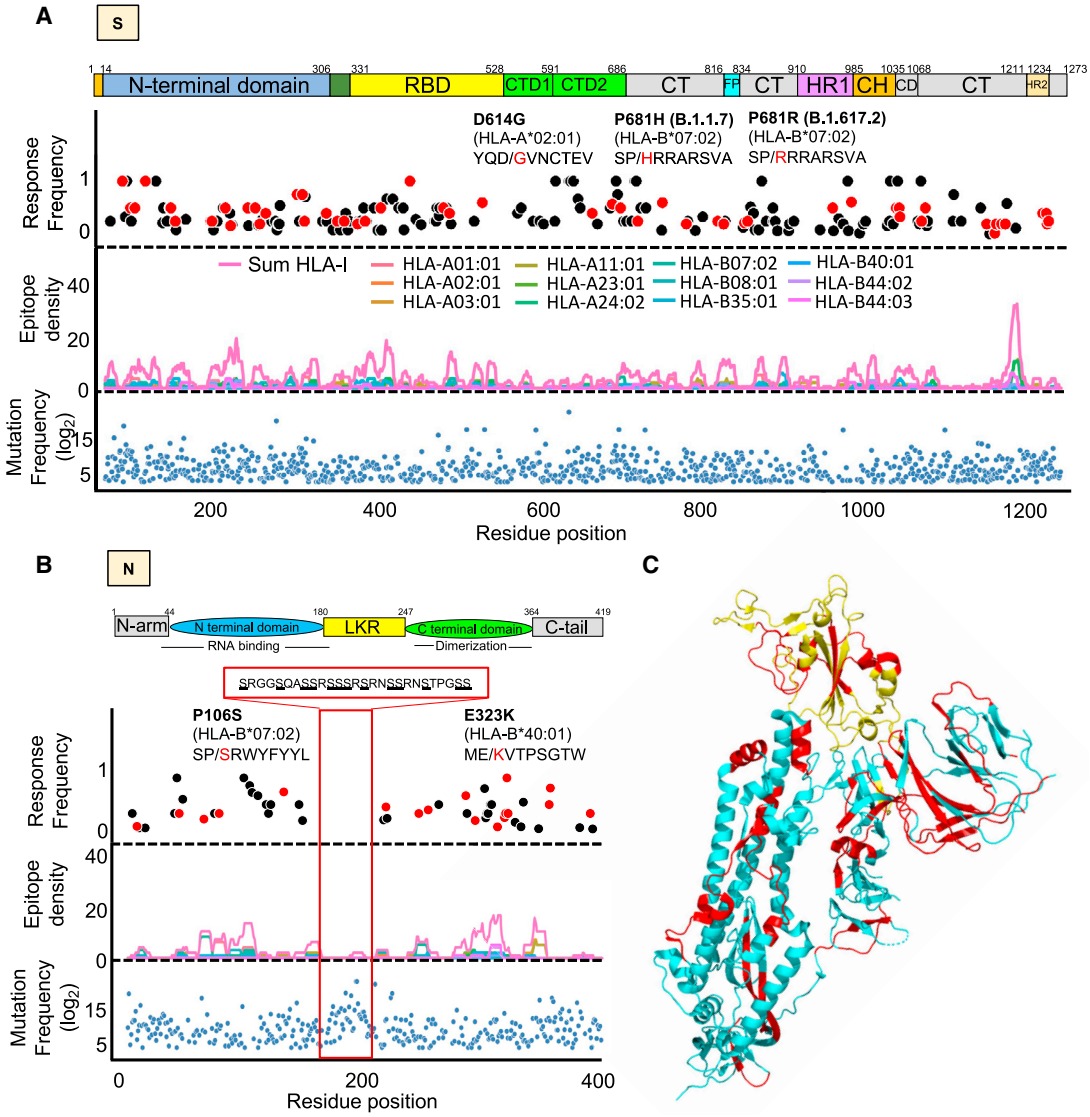


Figure 2. Distribution of CD8⁺ T cell epitopes and their mutated variants across the immunodominant spike (S) and nucleocapsid (N) antigens

(A and 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.

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 2B) and is predicted to decrease HLA epitope binding by 47-fold (Figure 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 S2) (Agerer et al., 2021). Together, these results demonstrate that mutations driving the global genomic diversity of SARS-CoV-2 can drastically disrupt HLA binding of clinically relevant CD8⁺ T cell epitopes, including epitopes encoded by the immunodominant S and N antigens, therefore affecting epitope-specific T cell responses in COVID-19 patients.

In addition to mutations leading to a loss of HLA epitope binding, we identified a significant number of mutations predicted to enhance the presentation of peptides by their respective HLA molecules, leading to a "gain" of binding (Figures 1C, 1D, and

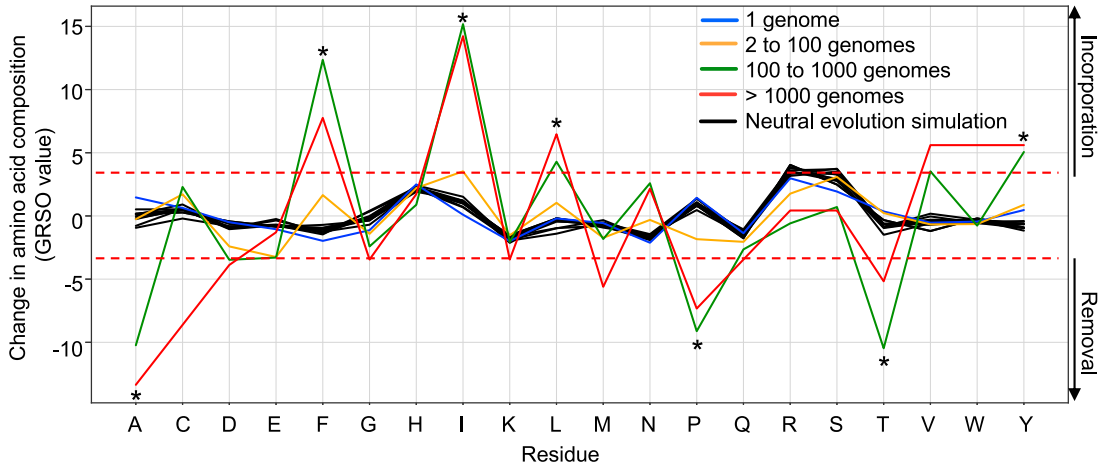


Figure 3. Global amino acid mutational biases in SARS-CoV-2 proteomes

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 percentage 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 1,000 genomes (green line), and more than 1,000 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 < 1×10^{-11}) that were used to define the residues that were preferentially removed or introduced (asterisk).

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.

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

While analyzing 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 (Figures 1E, S1C, and S1D). 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 S1C). These observations led to the hypothesis that the disproportionate presence of certain mutation types among 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 percentage difference in overall amino acid composition for individual amino acids (see STAR 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 1,000 genomes, and >1,000 genomes) (Figure 3). Distinct mutational patterns at the amino acid level were observed among mutations detected in more than 100 genomes/individuals (Figure 3), referred to in this study as "prevalent mutations" (see STAR Methods and Table S2). Among those mutations, the amino acids alanine (A), proline (P), and threonine (T) were preferentially removed by 10.2% ($p = 1.2 \times 10^{-13}$), 9.1% ($p = 1.6 \times 10^{-15}$), and 10.5% ($p = 1.3 \times 10^{-14}$), respectively. In contrast, phenylalanine (F), isoleucine (I), leucine (L), and tyrosine (Y) were preferentially introduced by 13.4% ($p = 2.0 \times 10^{-17}$), 15.2% ($p = 2.4 \times 10^{-17}$), 4.3% ($p = 6.3 \times 10^{-11}$), and 5.0% ($p = 7.0 \times 10^{-14}$), respectively (Figure 3). Statistical significance of these GRSO values was assessed by generating simulated samples of 1,000 SARS-CoV-2 genomes evolving under neutrality (n = 10 replicates) using the SANTA-SIM algorithm (Jariani et al., 2019) (see STAR 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 cutoff values (fold-change >4; p value < 1×10^{-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 S2.

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

The association of peptides with the binding groove of HLA molecules largely relies on the presence of anchor residues, also known as peptide-binding motifs (Falk et al., 1991). Hundreds of different peptide-binding motifs have been reported over the last decades (Gfeller and Bassani-Sternberg, 2018). Overlapping binding motifs are qualified as "HLA supertypes" on the basis of their main anchor specificity (Greenbaum et al., 2011; Sidney et al., 2008). Of relevance here, proline acts as a critical anchor residue at position P2 for epitopes presented by HLA-B7 (B7) supertype molecules, which include a wide range of commonly expressed HLA-B alleles in humans, i.e., HLA-B*07, -B*15, -B*35, -B*42, -B*51, -B*53, -B*54, -B*55, -B*56, -B*67, and B*78 (Sidney et al., 2008). In fact, the B7 supertype covers ~35% of the human population (Franciscodos et al., 2015). Hence, we reasoned that the global removal of proline residues observed in the proteome of prevalent SARS-CoV-2 mutants (Figure 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: (1) we applied NetMHCpan 4.1 (Reynisson et al., 2020) using the reference and mutated SARS-CoV-2 genomes to generate a list of all possible reference/mutated peptide pairs (8–11 mers) predicted to bind 16 common HLA-B types that belong to the B7 supertype family (Figure S4B). (2) 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]. (3) We categorized all peptide pairs based on the mutation type (amino acid X → amino acid Y) and the position of the mutation within the peptide sequence. (4) Lastly, we quantified the number of reference/mutated peptide pairs and the associated fold-change in predicted binding affinity for each category. Our results show that prevalent mutations predicted to impact the presentation of peptides by the B7 supertype are dominated by P→L ($p = 8.6 \times 10^{-35}$) and P→S ($p = 3.4 \times 10^{-24}$) substitutions at anchor residue position P2 (Figures 4A and 4B). Reference/mutated peptide pairs from these categories were the most abundant, with >250 mutated peptides per category (Figure 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 4D).

In addition to the dominant P→S/L substitution type, other P→X 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 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.

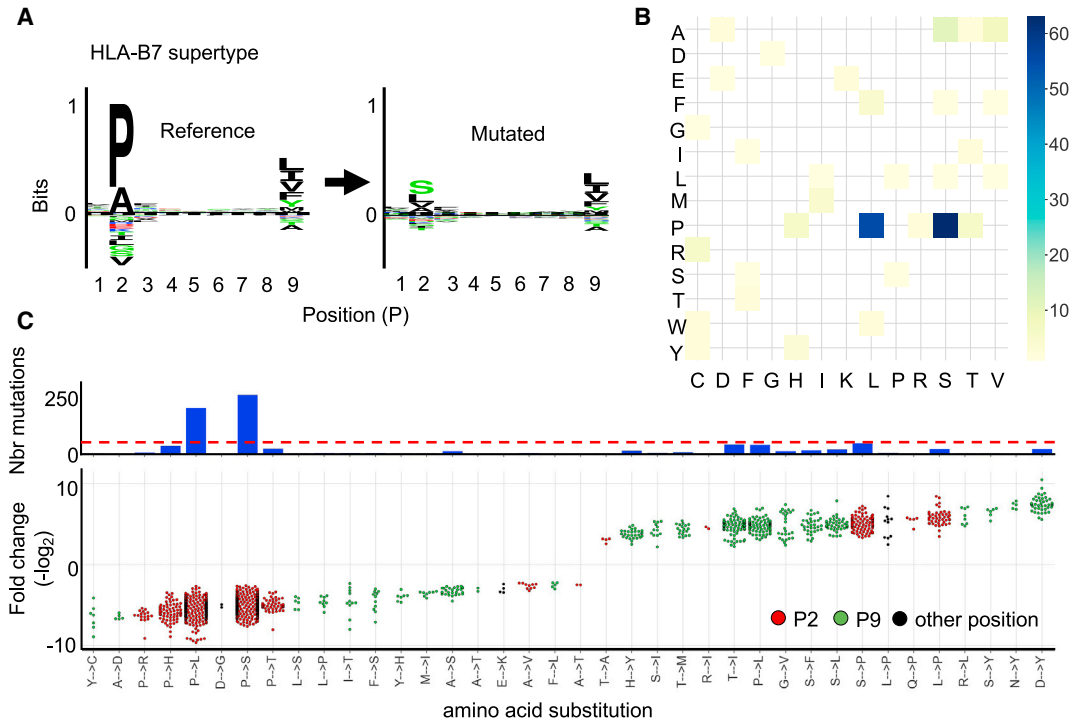
The mutational landscape of SARS-CoV-2 enables disruption or enhancement of epitope presentation in an HLA-supertype-dependent manner

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 3). Most of these amino acids act as key epitope anchor residues for multiple HLA class-I supertypes (Figure 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 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.

In order to compare supertypes with each other, we generated a "gain/loss plot" for each supertype assessed (Figure 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 STAR 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 ($p = 4.5 \times 10^{-11}$), A2 ($p = 0.001$), A24 ($p = 1.0 \times 10^{-26}$), B8 ($p = 2.4 \times 10^{-14}$), B27 ($p = 2.5 \times 10^{-6}$). 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 5A). In fact, omitting motifs containing isoleucine or phenylalanine increased the significance of epitope lost versus gained ($p = 2.6 \times 10^{-7}$) (Figure 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.

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



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 4. Mutation of 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, and -B*67:01). (Right panel) Motif view of the corresponding mutated peptides.

(B) Heatmap 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 S, N, open reading frame (ORF) 1a, 1b, and 3a are shown as representative examples. 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).

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 S5A), in line with observations made by others (Di Giorgio et al., 2020; Klimczak et al., 2020; Kosuge et al., 2020; Li et al., 2020; Matyášek and Kovařík, 2020; Rice et al., 2020; Simmonds, 2020; Wang et al., 2020).

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 1,000 SARS-CoV-2 genomes using SANTA-SIM (Jariani et al., 2019), applying various extents of mutational biases corresponding to the two most

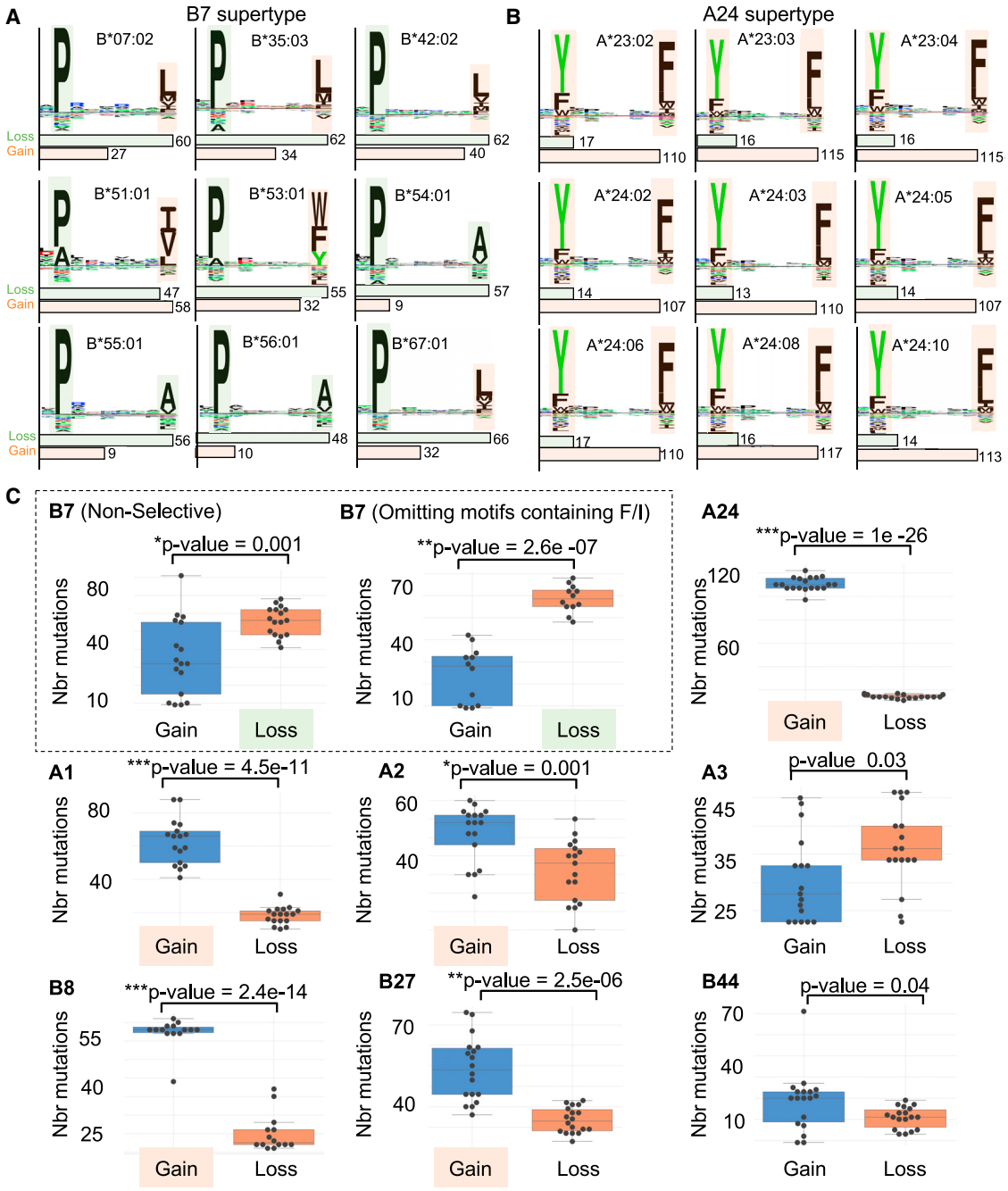


Figure 5. Loss or gain of SARS-CoV-2 mutated epitopes for different HLA class-I supertypes

(A and 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−5, ***p < 1e−10).

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 6A). Namely, the *in silico* introduction

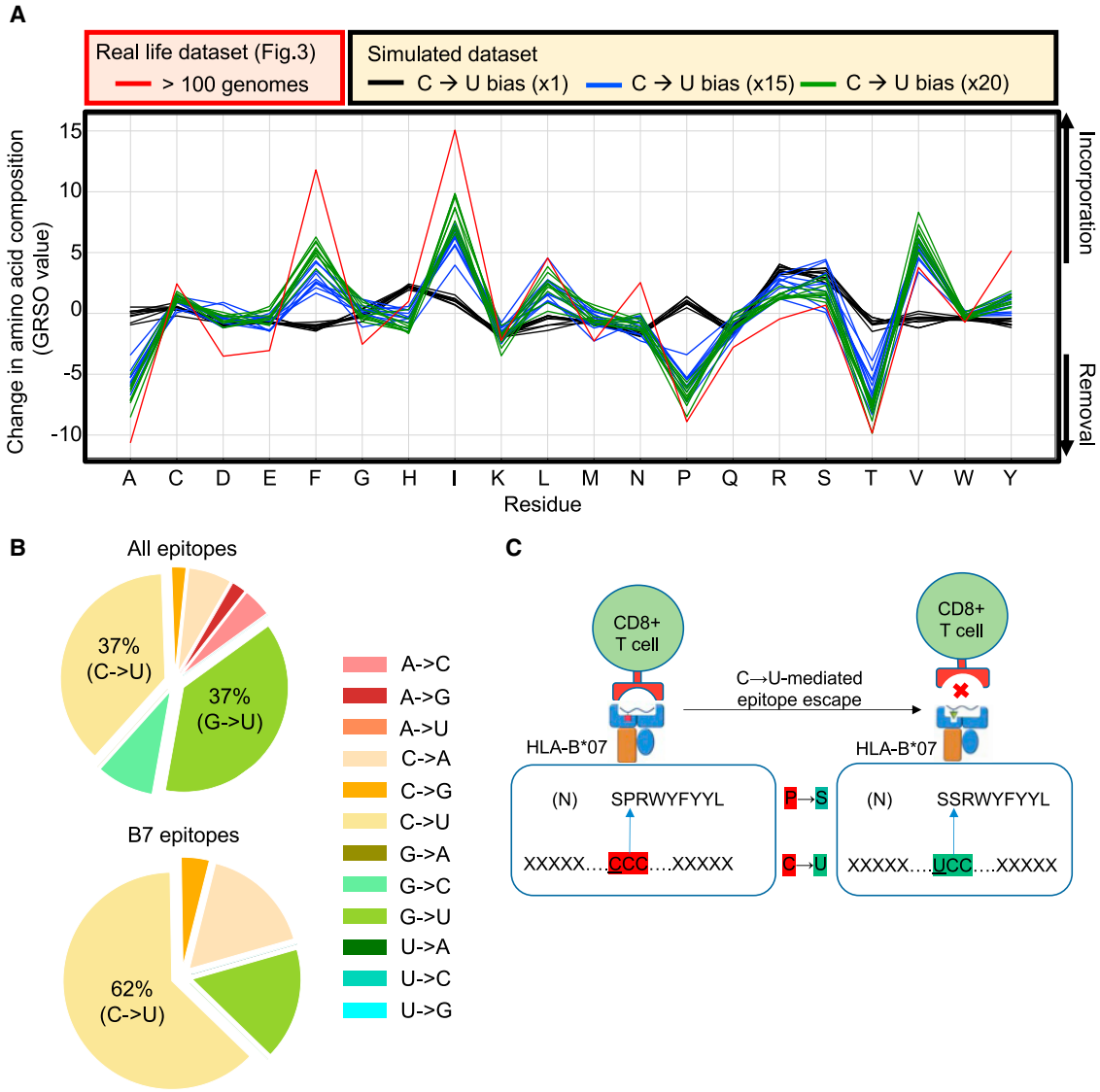


Figure 6. The C-to-U point mutation bias largely drives the diversity of SARS-CoV-2 proteomes and CD8$^+$ T cell epitopes

(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 percentage 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., ×1, ×15, and ×20 (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 N antigen is shown as an example.

of a C-to-U mutation bias resulted in the preferential removal of alanine, proline, and threonine, by 6.7% ($p = 5.1 \times 10^{-11}$), 6.9% ($p = 1.2 \times 10^{-11}$), and 8% ($p = 4.8 \times 10^{-12}$), respectively, as well as the introduction of isoleucine and phenylalanine by 8.2% ($p = 1.3 \times 10^{-8}$) and 5.2% ($p = 4.3 \times 10^{-11}$), respectively (Figure 6A). The G-to-U mutation bias also contributed to the introduction of isoleucine and phenylalanine (Figure 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.

Given the significant impact of the C-to-U point mutation bias on the amino acid content of SARS-CoV-2 proteomes, we reasoned that C-to-U could be the main driver shaping the repertoire and diversification of SARS-CoV-2 T cell targets in human

populations, including targets presented by the particularly interesting B7 supertype molecules. To investigate this, we used all the SARS-CoV-2 CD8$^+$ T cell epitopes that were experimentally validated using peripheral blood mononuclear cells (PBMC) of acute and convalescent COVID-19 patients (Quadeer et al., 2021; Tarke et al., 2021a) 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 among prevalent mutations (>100 individuals) (Figure 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 6B). These results strongly suggest that the dominant C-to-U point mutation bias found among 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 6C), and hence, may contribute to molding T cell immunity against COVID-19 at the population level.

DISCUSSION

Mutations contribute to the genetic diversity of SARS-CoV-2 and shape the progression of the COVID-19 pandemic (van Dorp et al., 2020b, 2020a; Popa et al., 2020). T cells are key players controlling COVID-19 disease severity. Therefore, determining whether and how the mutational landscape of SARS-CoV-2 shapes 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 (Brumme et al., 2007; Kawashima et al., 2009) and influenza (Woolthuis et al., 2016). 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 epidemic or pandemic 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 (Di Giorgio et al., 2020; Klimczak et al., 2020; Kosuge et al., 2020; Li et al., 2020; Matyášek and Kovařík, 2020; Rice et al., 2020; Simmonds, 2020; Wang et al., 2020). 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 (Franciscodos et al., 2015). Therefore, the C-to-U mutational bias observed among 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 among 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-dependent 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 (van Dorp et al., 2020a; Di Giorgio et al., 2020; Kosuge et al., 2020; Olson et al., 2018; Salter et al., 2016). In this regard, APOBEC activity has been shown to broadly drive viral evolution and diversity, including in human immunodeficiency virus (HIV) (Albin et al., 2010; Cuevas et al., 2015; Haché et al., 2008; Jern et al., 2009; Peretti et al., 2018; Sadler et al., 2010; Wood et al., 2009). In fact, APOBEC-induced mutations driving the evolution and diversification of HIV-1 were shown to have an intimate relationship with T cell immunity (Kim et al., 2014; Wood et al., 2009). Those studies have shown that the impact of APOBEC-induced mutations may result in either a decrease or increase of $CD8^+$ T cell recognition and that the direction of this response is dictated by the HLA context (Casartelli et al., 2010; Grant and Larijani, 2017; Kim et al., 2014; Monajemi et al., 2014; Squires et al., 2015; Wood et al., 2009). This is very much in line with our findings. Indeed, we showed that amino acid mutation biases in SARS-CoV-2 proteomes generally positively affect epitope binding for various HLA class-I supertypes, and most strikingly for A24, whereas B7 is the only supertype that is consistently negatively affected by the mutation biases given the markable loss of proline residues in SARS-CoV-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 with 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 with individuals expressing the B7 supertype (i.e., ~35% of the human population) (Franciscodos et al., 2015). Interestingly, a recent computational study corroborated the propensity of HLA-B*07:02 to lose epitopes due to SARS-CoV-2 variants (Nersisyan et al., 2021). 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 (Naemi et al., 2021; Pisanti et al., 2020; Tomita et al., 2020) as well as mutations and T cell evasion (Agerer et al., 2021; Geers et al., 2021; Motozono et al., 2021). 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 (Grifoni et al., 2021), another interesting avenue would be to study the mutational patterns of those emerging variants and assess whether and how the potential loss of B7-associated epitopes in those specific variants impact T cell response in infected patients. Understanding the impact of losing several subdominant B7-associated epitopes versus one single immunodominant epitope could also be investigated in the context of those variants. In this regard, a particular attention was allocated in our study to the B*07:02-restricted N105 epitope SPRWYFYYL. This epitope is of high interest as its immunodominance was experimentally demonstrated in many independent studies (Ferretti et al., 2020; Kared et al., 2021; Saini et al., 2021; Schulien et al., 2021; Sekine et al., 2020; Tarke et al., 2021a). Precisely, we found a rare mutation consisting of P→S at P2 of this epitope (SPRWYFYYL → SSRWYFYYL). Its occurrence was predicted to result in the complete abrogation of binding of the epitope to B*07:02, thereby likely reducing the breadth of the immune response in individuals carrying this mutation. As such, we advise the community to carefully monitor this mutation in subsequent months. Moreover, it is also possible that B7 individuals

respond less efficiently to the currently available vaccines, as genetic variants promoting B7 escape might favorably emerge in the future. The B7 supertype could therefore potentially represent a biomarker of vaccine resistance.

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

Limitations and future directions

Our analyses focused on class-I molecules for which predictors are established to be more accurate in comparison with class II. HLA-C and non-classical HLA were not included in this study. Predictions were performed on the most common HLA class-I alleles and rare HLA alleles were not included. Study has been performed using the GISAID dataset available in December 31, 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.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - Lead contact
  - Materials availability
  - Data and code availability
- METHOD DETAILS
  - Identification of SARS-CoV-2 mutations
  - Prediction of mutated and reference CD8+ T-cell epitopes
  - In vitro HLA-peptide binding assays
  - SANTA-SIM simulations
  - Determination of amino acid mutational patterns
  - Prediction of mutation impacts on peptide presentation in the context of HLA supertypes
  - Assessing the contribution of nucleic acid mutation types to the global amino acid mutational patterns
  - Statistical analysis

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.cels.2021.09.013.

ACKNOWLEDGMENTS

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, the Charles-Bruneau Foundations, Canada Foundation for Innovation, IVADO COVID19 Rapid Response grant (CVD19-030), the 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.A.K. is a recipient of IVADO's postdoctoral scholarship (#4879287150). D.F. is a BioTalent awardee. E.C. and J.H. are FRQS Junior 1 research scholars.

AUTHOR CONTRIBUTIONS

Conceptualization, D.J.H., J.H., and E.C.; data curation and bioinformatic analysis, D.J.H., D.F., J.-C.G., F.M., K.A.K., and P.K.; formal analysis, D.J.H. and D.F.; investigation, D.J.H., D.F., J.S., J.-C.G., K.A.K., J.D.D., F.S., P.K., I.S., H.D., S.P., J.H., and E.C.; writing – original draft, D.J.H. and E.C.; writing – review & editing, D.J.H., D.F., J.S., J.-C.G., F.M., K.A.K., P.K., J.D.D., F.S., I.S., M.A.S., H.S., H.D., S.P., J.H., and E.C.; supervision, J.H. and E.C.; funding acquisition, J.H. and E.C.

DECLARATION OF INTERESTS

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

Received: February 8, 2021
Revised: June 3, 2021
Accepted: September 23, 2021
Published: October 5, 2021

REFERENCES

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

Albin, J.S., Haché, G., Hultquist, J.F., Brown, W.L., and Harris, R.S. (2010). Long-term restriction by APOBEC3F selects human immunodeficiency virus type 1 variants with restored Vif function. J. Virol. *84*, 10209–10219.

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

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

Brumme, Z.L., Brumme, C.J., Heckerman, D., Korber, B.T., Daniels, M., Carlson, J., Kadie, C., Bhattacharya, T., Chui, C., Szinger, J., et al. (2007). Evidence of differential HLA class I-mediated viral evolution in functional and accessory/regulatory genes of HIV-1. PLoS Pathog. *3*, e94.

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

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

Cherian, S., Potdar, V., Jadhav, S., Yadav, P., Gupta, N., Das, M., Rakshit, P., Singh, S., Abraham, P., and Panda, S.; NIC Team (2021). SARS-CoV-2 spike mutations, L452R, T478K, E484Q and P681R, in the second wave of COVID-19 in Maharashtra, India. Microorganisms *9*, 1542.

Cuevas, J.M., Geller, R., Garijo, R., López-Aldeguer, J., and Sanjuán, R. (2015). Extremely high mutation rate of HIV-1 in vivo. PLoS Biol. *13*, e1002251.

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

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

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

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

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

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

Ferretti, A.P., Kula, T., Wang, Y., Nguyen, D.M.V., Weinheimer, A., Dunlap, G.S., Xu, Q., Nabilsi, N., Perullo, C.R., Cristofaro, A.W., et al. (2020). 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 *53*, 1095–1107.e3.

Frampton, D., Rampling, T., Cross, A., Bailey, H., Heaney, J., Byott, M., Scott, R., Sconza, R., Price, J., Margaritis, M., et al. (2021). 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. *21*, 1246–1256.

Geers, D., Shamier, M.C., Bogers, S., Hartog, G. den, Gommers, L., Nieuwkoop, N.N., Schmitz, K.S., Rijsbergen, L.C., van Osch, J.A.T., Dijkhuizen, E., et al. (2021). SARS-CoV-2 variants of concern partially escape humoral but not T-cell responses in COVID-19 convalescent donors and vaccinees. Sci. Immunol. *6*, eabj1750.

Gfeller, D., and Bassani-Sternberg, M. (2018). Predicting antigen presentation-what could we learn From a million peptides? Front. Immunol. *9*, 1716.

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

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

Grifoni, A., Sidney, J., Vita, R., Peters, B., Crotty, S., Weiskopf, D., and Sette, A. (2021). SARS-CoV-2 human T cell Epitopes: adaptive immune response against COVID-19. Cell Host Microbe *29*, 1076–1092.

Grifoni, A., Weiskopf, D., Ramirez, S.I., Mateus, J., Dan, J.M., Moderbacher, C.R., Rawlings, S.A., Sutherland, A., Premkumar, L., Jadi, R.S., et al. (2020a). Targets of T cell responses to SARS-CoV-2 coronavirus in humans with COVID-19 disease and unexposed individuals. Cell *181*, 1489–1501.e15.

Grifoni, A., Sidney, J., Zhang, Y., Scheuermann, R.H., Peters, B., and Sette, A. (2020b). A sequence homology and bioinformatic approach can predict candidate targets for immune responses to SARS-CoV-2. Cell Host Microbe *27*, 671–680.e2.

Haché, G., Shindo, K., Albin, J.S., and Harris, R.S. (2008). Evolution of HIV-1 isolates that use a novel Vif-independent mechanism to resist restriction by human APOBEC3G. Curr. Biol. *18*, 819–824.

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

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

Jern, P., Russell, R.A., Pathak, V.K., and Coffin, J.M. (2009). Likely role of APOBEC3G-mediated G-to-A mutations in HIV-1 evolution and drug resistance. PLoS Pathog. *5*, e1000367.

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

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

Kim, E.-Y., Lorenzo-Redondo, R., Little, S.J., Chung, Y.-S., Phalora, P.K., Maljkovic Berry, I.M., Archer, J., Penugonda, S., Fischer, W., Richman, D.D., et al. (2014). Human APOBEC3 induced mutation of human immunodeficiency virus Type-1 contributes to adaptation and evolution in natural infection. PLoS Pathog *10*, e1004281.

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

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

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

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

Laamarti, M., Alouane, T., Kartti, S., Chemao-Elfihri, M.W., Hakmi, M., Essabbar, A., Laamarti, M., Hlali, H., Bendani, H., Boumajdi, N., et al. (2020). 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 *15*, e0240345.

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

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

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

Liu, G., Carter, B., Bricken, T., Jain, S., Viard, M., Carrington, M., and Gifford, D.K. (2020). Computationally optimized SARS-CoV-2 MHC class I and II vaccine formulations predicted to target human haplotype distributions. Cell Syst *11*, 131–144.e6.

Liu, W., Fontanet, A., Zhang, P.H., Zhan, L., Xin, Z.T., Baril, L., Tang, F., Lv, H., and Cao, W.-C. (2006). Two-year prospective study of the humoral immune response of patients with severe acute respiratory syndrome. J. Infect. Dis. *193*, 792–795.

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

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

Matyášek, R., and Kovařík, A. (2020). 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) *11*, 761.

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

Mercatelli, D., and Giorgi, F.M. (2020). Geographic and genomic distribution of SARS-CoV-2 mutations. Front. Microbiol. *11*, 1800.

Mercatelli, D., Triboli, L., Fornasari, E., Ray, F., and Giorgi, F.M. (2021). Coronapp: a web application to annotate and monitor SARS-CoV-2 mutations. J. Med. Virol. *93*, 3238–3245.

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

Monajemi, M., Woodworth, C.F., Zipperlen, K., Gallant, M., Grant, M.D., and Larijani, M. (2014). Positioning of APOBEC3G/F mutational hotspots in the human immunodeficiency virus genome favors reduced recognition by CD8+ T cells. PLoS One *9*, e93428.

Motozono, C., Toyoda, M., Zahradnik, J., Saito, A., Nasser, H., Tan, T.S., Ngare, I., Kimura, I., Uriu, K., Kosugi, Y., et al. (2021). SARS-CoV-2 spike L452R variant evades cellular immunity and increases infectivity. Cell Host Microbe *29*, 1124–1136.e11.

Naemi, F.M.A., Al-Adwani, S., Al-Khatabi, H., and Al-Nazawi, A. (2021). Association between the HLA genotype and the severity of COVID-19 infection among South Asians. J. Med. Virol. *93*, 4430–4437.

Nersisyan, S., Zhiyanov, A., Shkurnikov, M., and Tonevitsky, A. (2021). T-CoV: a comprehensive portal of HLA-peptide interactions affected by SARS-CoV-2 mutations. Nucleic Acids Res. https://doi.org/10.1093/nar/gkab701.

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

Olson, M.E., Harris, R.S., and Harki, D.A. (2018). APOBEC enzymes as targets for virus and cancer therapy. Cell Chem. Biol. *25*, 36–49.

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

Peretti, A., Geoghegan, E.M., Pastrana, D.V., Smola, S., Feld, P., Sauter, M., Lohse, S., Ramesh, M., Lim, E.S., Wang, D., et al. (2018). Characterization of BK polyomaviruses from kidney transplant recipients suggests a role for APOBEC3 in driving in-host virus evolution. Cell Host Microbe *23*, 628–635.e7.

Pisanti, S., Deelen, J., Gallina, A.M., Caputo, M., Citro, M., Abate, M., Sacchi, N., Vecchione, C., and Martinelli, R. (2020). Correlation of the two most frequent HLA haplotypes in the Italian population to the differential regional incidence of Covid-19. J. Transl. Med. *18*, 352.

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

Quadeer, A.A., Ahmed, S.F., and McKay, M.R. (2021). Landscape of epitopes targeted by T cells in 852 individuals recovered from COVID-19: Meta-analysis, immunoprevalence, and web platform. Cell Rep. Med. *2*, 100312.

Reynisson, B., Alvarez, B., Paul, S., Peters, B., and Nielsen, M. (2020). 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 *48*, W449–W454.

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

Sadler, H.A., Stenglein, M.D., Harris, R.S., and Mansky, L.M. (2010). APOBEC3G contributes to HIV-1 variation through sublethal mutagenesis. J. Virol. *84*, 7396–7404.

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

Salter, J.D., Bennett, R.P., and Smith, H.C. (2016). The APOBEC protein family: united by structure, divergent in function. Trends Biochem. Sci. *41*, 578–594.

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

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

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

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

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

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

Sidney, J., Southwood, S., Moore, C., Oseroff, C., Pinilla, C., Grey, H.M., and Sette, A. (2013). Measurement of MHC/peptide interactions by gel filtration or monoclonal antibody capture. Curr. Protoc. Immunol. *100*, 18.3.1–18.3.36.

Simmonds, P. (2020). 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 *5*, e00408.

Squires, K.D., Monajemi, M., Woodworth, C.F., Grant, M.D., and Larijani, M. (2015). Impact of APOBEC mutations on CD8+ T cell recognition of HIV epitopes varies depending on the restricting HLA. J. Acquir. Immune Defic. Syndr. *70*, 172–178.

Stephens, D.S., and McElrath, M.J. (2020). COVID-19 and the path to immunity. JAMA *324*, 1279–1281.

Tang, F., Quan, Y., Xin, Z.-T., Wrammert, J., Ma, M.-J., Lv, H., Wang, T.-B., Yang, H., Richardus, J.H., Liu, W., and Cao, W.-C. (2011). Lack of peripheral memory B cell responses in recovered patients with severe acute respiratory syndrome: A six-year follow-up study. J. Immunol. *186*, 7264–7268.

Tarke, A., Sidney, J., Kidd, C.K., Dan, J.M., Ramirez, S.I., Yu, E.D., Mateus, J., da Silva Antunes, R.da S., Moore, E., Rubiro, P., et al. (2021a). Comprehensive analysis of T cell immunodominance and immunoprevalence of SARS-CoV-2 epitopes in COVID-19 cases. Cell Rep. Med. *2*, 100204.

Tarke, A., Sidney, J., Methot, N., Yu, E.D., Zhang, Y., Dan, J.M., Goodwin, B., Rubiro, P., Sutherland, A., Wang, E., et al. (2021b). Impact of SARS-CoV-2 variants on the total CD4+ and CD8+ T cell reactivity in infected or vaccinated individuals. Cell Rep. Med. *2*, 100355.

Tegally, H., Wilkinson, E., Giovanetti, M., Iranzadeh, A., Fonseca, V., Giandhari, J., Doolabh, D., Pillay, S., San, E.J., Msomi, N., et al. (2021). Detection of a SARS-CoV-2 variant of concern in South Africa. Nature *592*, 438–443.

Tomita, Y., Ikeda, T., Sato, R., and Sakagami, T. (2020). Association between HLA gene polymorphisms and mortality of COVID-19: an in silico analysis. Immun. Inflamm. Dis. *8*, 684–694.

Wang, R., Hozumi, Y., Zheng, Y.-H., Yin, C., and Wei, G.-W. (2020). Host immune response driving SARS-CoV-2 evolution. Viruses *12*, 1095.

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

Wood, N., Bhattacharya, T., Keele, B.F., Giorgi, E., Liu, M., Gaschen, B., Daniels, M., Ferrari, G., Haynes, B.F., McMichael, A., et al. (2009). HIV evolution in early infection: selection pressures, patterns of insertion and deletion, and the impact of APOBEC. PLoS Pathog. *5*, e1000414.

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

Wu, L.-P., Wang, N.-C., Chang, Y.-H., Tian, X.-Y., Na, D.-Y., Zhang, L.-Y., Zheng, L., Lan, T., Wang, L.-F., and Liang, G.-D. (2007). Duration of antibody responses after severe acute respiratory syndrome. Emerg. Infect. Dis. *13*, 1562–1564.

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

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Chemicals, peptides, and recombinant proteins		
Synthetic peptides	TC Peptide Lab	tcpeptidelab.com
Deposited data		
Wuhan-Hu-1 RNA isolate	NCBI nuccore database	NCBI: NC_045512.2
Structure of SARS-CoV-2 Spike Protein Trimer	Xiong et al., 2020	PDB: 6ZP2
GISAID	Freunde von GISAID e.V.	https://www.gisaid.org/
Experimentally Validated SARS-CoV-2 T cell epitopes	Quadeer et al., 2021	https://www.mckayspcb.com/SARS2TcellEpitopes/
Supplemental information	Hamelin et al.	DOI: 10.5281/zenodo.5520066
Software and algorithms		
netMHCpan 4.1	Reynisson et al., 2020	https://services.healthtech.dtu.dk/service.php?NetMHCpan-4.1
Python (v3.7)	Python Software Foundation	https://www.python.org/
Santa-Sim	Jariani et al., 2019	https://github.com/santa-dev/santa-sim
CoVescape	In-house algorithm	DOI: 10.5281/zenodo.5493359

RESOURCE AVAILABILITY

Lead contact

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

Materials availability

This study did not generate new 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 as Table S5.
- 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 (Sidney et al., 2013). 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 $\geq$ [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. (Huddleston et al., 2020) who used SANTA-SIM to simulate influenza A/H3N2 that has a yearly substitution rate approximately twice as high as SARS-CoV-2 [~48,824 substitutions/year (https://nextstrain.org/flu/seasonal/h3n2/ha/2y?l=clock) vs. ~24.5 substitution/year (https://nextstrain.org/ncov/global?l=clock)], we chose 400 generations/year, with the mutation rate per position per generation set to 2.04E-6 (yearly substitution rate/(generations in one year * genome size)). The transition bias was set to 3.0 for baseline simulations. To evaluate the impact of specific substitution biases, additional simulations were conducted using a substitution matrix with scores set to 1.0 of transversions, 3.0 for transitions, and biases ranging from 4.0 to 20.0 for the targeted substitution. We generated 10 replicates for all simulated scenarios, except for C-to-U where we made 100 replicates to better assess statistical significance.

Determination of amino acid mutational patterns

Mutational biases were identified by calculating the overall change in amino acid composition caused by the mutational landscape of SARS-CoV-2 for each individual amino acid, referred in the main text as 'global residue substitution output' (GRSO). For this analysis, all mutations found globally in at least 4 GISAID entries were analysed together. Preferential introduction or removal of amino acids was determined by comparing the overall amino acid composition in reference residues vs mutated residues throughout the mutation pool, resulting in a percentile difference in amino acid composition. As such, for amino acid *X*, the % difference was calculated according to the following formula:

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

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

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

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

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

To assess the contribution of various nucleic acid mutation types to the observed amino acid mutational patterns, we first determined the respective contributions of each nucleic acid mutation type to the global mutation landscape. We then selected the five most abundant mutation types [C→U (41%), G→U (18%), A→G, G→A, U→C (9.7-11.6%)] and assessed their individual impacts on amino acid mutational patterns using the simulation algorithm SANTA SIM as follows:

For each mutation type, we simulated the evolution of 1000 SARS-CoV-2 genomes over 1000 generations (N = 10 replicates) with varying degrees of biases (the coefficient used to determine the extent of the biases was exploratively set to 'x4', 'x8', 'x15', and 'x20') (Figure 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* → *Y* nucleic acid mutation type to all other mutation types was generated using the following formula:

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

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.