amU
Aix Marseille Université

NNT : 0000AIXM0000

THÈSE DE DOCTORAT

Soutenue à Aix-Marseille Université
le 16 juin 2025 par

Emilie BUREL

Évolution virale au cours des infections chroniques : Contributions de la génomique à la compréhension de la COVID-19

Discipline
Recherches Biomédicales

Spécialité
Ingénierie biomécanique et biomédicale et thérapeutiques innovantes

École doctorale
Recherches Biomédicales – ED659

Laboratoire
UMR RITMES

Composition du jury

Astrid VABRET Université de Caen Normandie	Rapporteuse
Christophe RODRIGUEZ Université Paris-Est	Rapporteur
Bruno CANARD Université d'Aix-Marseille	Examinateur
Dominique SALMON Université de Paris-Descartes	Examinatrice
Florence FENOLLAR Université d'Aix-Marseille	Présidente du jury
Pierre-Edouard FOURNIER Université d'Aix-Marseille	Directeur de thèse

RITMES
MÉDITERRANÉE INFECTION
AP-HM Assistance Publique Hôpitaux de Marseille

Affidavit

Je soussigné, Emilie Burel, déclare par la présente que le travail présenté dans ce manuscrit est mon propre travail, réalisé sous la direction scientifique de Pr. Pierre-Edouard Fournier, dans le respect des principes d'honnêteté, d'intégrité et de responsabilité inhérents à la mission de recherche. Les travaux de recherche et la rédaction de ce manuscrit ont été réalisés dans le respect à la fois de la charte nationale de déontologie des métiers de la recherche et de la charte d'Aix-Marseille Université relative à la lutte contre le plagiat.

Ce travail n'a pas été précédemment soumis en France ou à l'étranger dans une version identique ou similaire à un organisme examinateur.

Fait à Marseille, le 15 mars 2025

Emilie Burel

Cette œuvre est mise à disposition selon les termes de la Licence Creative Commons Attribution - Pas d'Utilisation Commerciale - Pas de Modification 4.0 International.

Remerciements

Au terme de ce travail, je souhaite exprimer ma gratitude envers les personnes qui ont contribué de près ou de loin à son aboutissement. En premier lieu, je tiens à remercier chaleureusement mon directeur de thèse, le Professeur Pierre-Edouard Fournier, pour avoir accepté de m'accompagner dans cette aventure scientifique, pour sa patience et sa gentillesse.

Je remercie vivement les membres de mon jury, les Prs. Astrid Vabret et Christophe Rodriguez pour avoir accepté d'être rapporteurs, les Prs. Dominique Salmon et Bruno Canard en qualité d'examinateurs, ainsi que le Pr. Florence Fenollar qui m'honore de sa présence en qualité de présidente.

Je remercie le Pr. Didier Raoult, qui m'a permis de débuter cette thèse, le Pr. Pierre Pontarotti, vos perspectives ont enrichi mon travail et alimenté ma réflexion, ainsi que le Pr. Philippe Colson.

Je remercie particulièrement Nathalie Caniparoli et le Pr. Jean-Louis Mege, pour m'avoir donné l'impulsion fondamentale à l'aboutissement de ce travail. Trouvez ici ma reconnaissance absolue.

Je remercie l'ensemble des médecins avec lesquels j'ai pu collaborer et échanger, spécialement Estelle Menu et Philippe Lavrard-Meyer.

Je remercie l'équipe bio-informatique de l'IHU, en particulier Aurélia, merci pour ta bonne humeur et ta présence, et Mamadou, merci à vous deux pour ce que vous m'aurez appris !

Une pensée pour les ingénieurs, techniciens et cadres, Priscilla, Fabrizio, Richard, Line et les personnes que j'oublie, Claudia, merci pour ta bienveillance.

Je remercie également Tomaso Antonacci et mes amis prefaciens, vous avez été un moteur tout au long de cette aventure !

Enfin, je souhaite exprimer ma reconnaissance et mon amour à ma famille, à mes amis, et à Jonathan. Merci d’être là !

Table des matières

Affidavit 1
Remerciements 2
Publications 5
Liste des abréviations 6
Résumé 8
Abstract 10
Introduction générale 12
1. Contexte de la recherche 12
Présentation de la pandémie de COVID-19 13
Classification et structure 14
Physiopathologie 16
Diagnostic et séquençage du SARS-CoV-2 19
Hypothèses sur l'origine des nouveaux variants 20
Article 1 (revue de littérature) : Contribution des variants minoritaires dans les études du SARS-CoV-2 22
Avant-propos 23
Article 2 : Evolution intra-hôte et souches recombinantes 71
Avant-propos 72
Article 3 : Evolution intra-hôte et adaptation virale à l'hôte 103
Avant-propos 104
Conclusion et perspectives 175
Annexe 1 Etude d'une épidémie à Trichosporon dans un hôpital de Marseille 178
Avant-propos 179
Annexe 2 : Etude génomique sur la circulation de HCoV-229e dans le sud de la France 211
Avant-propos 212
Références 238

Publications

Liste des publications réalisées dans le cadre du projet de thèse :

Emilie Burel, Pierre Edouard Fournier**. "SARS-CoV-2 Low-Frequency Genetic variants: accurate detection and advancing understanding in evolution and epidemiology".** (Soumis à Archive of Virology)

Emilie Burel 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**." *Viruses* vol. 14,6 1266. 10 Jun. 2022, doi:10.3390/v14061266

Emilie Burel et al. **"Long-Term Intra-Host Evolution of SARS-CoV-2 in an Immunocompromised Patient: Recombination and de novo Mutations Driving Viral Adaptation".** (preprint, https://doi.org/10.1101/2025.03.25.645210, soumis à Virus Evolution)

Annexes :

Emilie Burel et al**. "A genome-based investigation of emerging *Trichosporon austroamericanum* infections at the University Hospital of Marseille"** (soumis à Emerging Infectious Disease)

Houmadi Hikmat, Justine Py, Céline Boschi, Emilie Burel, Lorlane Le Targa, Matthieu Million, Lucile Lesage, Aurélie Morand, Bernard La Scola, Philippe Colson. **"Implementation of a multiplex PCR amplification system combined with next-generation genome sequencing to decipher the circulation of Human coronavirus 229E lineages in Southern France"** (preprint https://doi.org/10.1101/2025.03.12.639729)

Participation aux conférences et écoles d'été au cours de la période de thèse :

Présentation orale : International Conference on Infectious Diseases (20-21 mars 2023-Rome).

Liste des abréviations

ACE-2 : Angiotensin-Converting Enzyme 2

Ct: Cycle threshold

E: Enveloppe

GISAID: Global Initiative on Sharing Avian Influenza Data

hCoV-229E : Human coronavirus 229E

ICTV : International Committee on Taxonomy of Viruses

iSNVs: intra host Single Nucleotide Variations

LFVs : Low Frequency variants

MAF : Minor Allele Frequency

Mafft: Multiple Alignment Using Fast Fourier Transform

MALDI-TOF : Matrix-Assisted Laser Desorption/Ionization Time-of-Flight

M: Membrane

MERS-CoV: Middle East respiratory syndrome-related coronavirus

MT : Mitochondrie

N : Nucléocapside

NGS : Next Generation Sequencing

NSP : Non-structural proteins

NTD : N- terminal domain

OMS : Organisation Mondiale de la Santé

ONT : Oxford Nanopore Technologies

ORF: Open Reading Frame

PangoLIN : Phylogenetic Assignment of NamedGlobal Outbreak Lineages

PCR: Polymerase Chain Reaction

RBD: Recepteur Binding Domain

RdRp: RNA dependent ARN polymerase

RT-PCR: Reverse Transcriptase- Polymerase Chain Reaction

S: Spike

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

SNP: Single nucleotide polymorphism

TMPRSS2 : Transmembrane protease serine 2

UTR: Untranslated Region

VOC: Variant of Concern

VOI : Variant of Interest

VUM: Variant Under Monitoring

Résumé

Depuis l'émergence du SARS-CoV-2 fin 2019, la pandémie de COVID-19 a fortement mobilisé la communauté scientifique mondiale. Le recours massif au séquençage génomique a notamment conduit au dépôt de plus de 16 millions de séquences dans les bases de données internationales, permettant un suivi en temps réel de l'évolution du virus ainsi que de ses dynamiques de transmission. Cette thèse s'inscrit dans ce contexte et porte sur l'étude de l'évolution intra-hôte du SARS-CoV-2, un processus par lequel le virus accumule des mutations et potentiellement des recombinaisons au sein d'un même individu. Les infections persistantes chez les patients immunodéprimés offrent un cadre d'étude particulièrement pertinent, du fait d'une réplication virale prolongée dans un environnement immunologiquement permissif, favorisant l'adaptation virale.

Le travail présenté s'articule autour d'un cas d'évolution intra-hôte du SARS-CoV-2 chez un patient immunodéprimé atteint de lymphome, porteur chronique pendant 440 jours. Le premier article rapporte l'émergence de recombinaisons *in vivo* entre deux variants majeurs du SARS-CoV-2 (B.1.160 et B.1.1.7), confirmée par des PCR ciblées aux points de cassure, la mise en culture des souches isolées, et leur séquençage complet. Cette étude constitue la première démonstration expérimentale d'une recombinaison entre deux variants préoccupants avec preuve de leur capacité de réplication. Dans la continuité du précédent, le deuxième article approfondit les mécanismes évolutifs intra-hôtes, en s'intéressant aux mutations *de novo*, aux variants minoritaires et aux signaux de sélection positive. L'étude révèle un taux d'évolution intra-hôte 1,54 fois supérieur au taux inter-hôte, une forte diversité génétique avec des mutations qui émergent, se fixent, ou disparaissent, ainsi qu'un enrichissement marqué en substitutions C>T, compatible avec une activité d'édition de l'ARN par les enzymes APOBEC. Plusieurs gènes viraux clés présentent des signes de sélection positive, notamment Spike, N,

ORF9b et nsp13, traduisant leur rôle dans l'adaptation virale. Une délétion de 126 nucléotides dans le gène ORF8 a également été observée. Enfin, l'identification de mutations convergentes, telles que S:D796H ou S:L5F, renforce l'hypothèse de pressions de sélection similaires dans différents contextes évolutifs. L'ensemble de ces travaux illustre la complexité des trajectoires évolutives du SARS-CoV-2 en contexte d'immunosuppression, où peuvent coexister recombinaisons, mutations adaptatives et diversité sub-consensus. Cette thèse illustre et souligne l'importance des infections chroniques comme foyers de biodiversité virale.

Mots clés : SARS-CoV-2, iSNVs, intra-host evolution, recombination, genomics, bioinformatics, LFVs

Abstract

Since the emergence of SARS-CoV-2 in late 2019, the COVID-19 pandemic has mobilized the global scientific community. The massive use of genomic sequencing has led to the deposit of over 16 million sequences in international databases, enabling real-time monitoring of the virus' evolution and transmission dynamics. In this context, this thesis focuses on the intra-host evolution of SARS-CoV-2, a process by which the virus accumulates mutations and potentially recombination within the same individual. Persistent infections in immunocompromised patients offer a particularly relevant framework for study, due to prolonged viral replication in an immunologically permissive environment, favoring viral adaptation.

The work presented here revolves around a case of intra-host evolution of SARS-CoV-2 in an immunocompromised patient with lymphoma, who carried the virus for 440 days. The first article reports the emergence of recombination between two major variants of SARS-CoV-2 (B.1.160 and B.1.1.7) within the host, confirmed by targeted PCR at breakpoints, culture of isolated strains, and their complete sequencing. This study constitutes the first experimental demonstration of recombination between two variants of concern, with proof of their replication capacity. Following on from the previous article, the second paper takes a closer look at intra-host evolutionary mechanisms, focusing on *de novo* mutations, minority variants and positive selection signals. The study reveals an intra-host evolution rate 1.54 times higher than the inter-host rate, high genetic diversity with mutations that emerge, become fixed or disappear, and a marked enrichment in C>T substitutions, compatible with RNA editing activity by APOBEC enzymes. Several key viral genes show signs of positive selection, including Spike, N, ORF9b and nsp13, reflecting their role in viral adaptation. A 126-nucleotide deletion in the ORF8 gene was also observed. Finally, the identification of convergent mutations, such as S:D796H or S:L5F, reinforces the hypothesis of similar selection pressures in different evolutionary

contexts. Taken together, this work illustrates the complexity of SARS-CoV-2 evolutionary trajectories in a context of immunosuppression, where recombination, adaptive mutations and sub-consensus diversity can coexist. This thesis illustrates and underlines the importance of chronic infections as hotbeds of viral biodiversity.

Keywords: SARS-CoV-2, iSNVs, intra-host evolution, recombination, genomics, bioinformatics, LFVs

Introduction générale

1. Contexte de la recherche

La pandémie de COVID-19 a profondément bouleversé la santé publique mondiale et stimulé dans un temps réduit une production scientifique sans précédent. Comprendre les mécanismes évolutifs du virus SARS-CoV-2 (Severe Acute Respiratory Syndrome Coronavirus 2), est essentiel pour anticiper l'émergence de nouveaux variants et adapter les stratégies de surveillance. Cette thèse s'inscrit ainsi dans un effort de recherche visant à approfondir la compréhension de la dynamique évolutive intra-hôte du SARS-CoV-2, en particulier à travers l'étude des variants minoritaires ou low-frequency variants (LFVs).

La première partie de ce travail s'ouvre sur une brève présentation de la pandémie COVID-19 et du SARS-CoV-2, complétée par une revue de la littérature dédiée à la contribution des variants minoritaires dans l'étude du SARS-CoV-2.

Le second volet de cette recherche porte sur la mise en évidence d'événements de recombinaison survenus *in vivo* entre deux variants du SARS-CoV-2, à la suite du suivi d'un patient immunodéprimé long porteur. Lors du génotypage de routine, l'identification d'un profil génomique atypique a suscité notre intérêt et conduit à la mise en place d'une étude rétrospective approfondie. Ces événements de recombinaison ont été validés par des approches complémentaires, notamment des PCR ciblant les régions de recombinaison afin d'identifier des reads chimériques contigus, ainsi que par la mise en culture des virus isolés, confirmant leur existence biologique. Le séquençage complet de leur génome a permis de corroborer les profils observés. Dans un troisième travail, l'accent a été mis sur l'étude des mutations *de novo,* y compris celles présentes à faible fréquence allélique. Cette analyse a permis de suivre leur devenir au cours de l'infection : certaines de ces mutations finissent par s'imposer, d'autres sont

éliminées, conférant également des signes de sélection positive ou négative en fonction des gènes. Un travail collaboratif de modélisation moléculaire a été conduit en parallèle afin d'explorer les conséquences structurales et fonctionnelles de ces mutations au niveau de la protéine Spike, notamment en lien avec les événements de recombinaison ou d'adaptation intra-hôte.

Enfin, en annexe de cette thèse, deux contributions complémentaires sont présentées :

Une étude consacrée à une épidémie hospitalière impliquant *Trichosporon austroamericanum*, une espèce fongique nouvellement décrite, initialement identifiée comme *T. inkin*. Mon travail a porté sur la détermination du caractère polyclonal de l'épidémie et a débouché sur une publication soumise dans Emerging Infectious Disease. Enfin, une analyse génomique des souches de HCoV-229E, un coronavirus humain responsable du rhume commun, ayant circulé dans le sud-est de la France sur une période donnée, dans le cadre d'un travail de génotypage réalisé à Marseille. Ma contribution a porté sur l'aspect analytique en fournissant les génomes consensus et participé aux travaux préliminaires de génotypage.

2. Présentation de la pandémie de COVID-19

La pandémie de COVID-19 [1] a été induite par le virus à ARN SARS-CoV-2, qui émergea dans la province de Wuhan en décembre 2019 [2]. Cette pandémie s'est imposée comme un événement marquant de l'histoire récente, un choc épidémique majeur [3]. D'après l'Organisation Mondiale de la Santé, plus de 777 millions de personnes ont contracté le SARS-CoV-2 durant les quatre premières années de la pandémie, causant plus de 7 millions de morts [4]. L'impact social, économique et politique s'est également révélé colossal [5].

Les virus à ARN ont par le passé occasionné des épidémies dévastatrices, parmi lesquelles, la grippe espagnole qui sévit de 1918 à 1920 et fit entre 50 et 100 millions de morts à travers le monde, le virus responsable, *Myxovirus influenzae*, ne fut identifié chez l'Homme qu'en 1933 [6]. Avec l'épidémie du syndrome respiratoire aigu sévère (SARS-CoV) en 2002–2003 [7] et celle

du syndrome respiratoire du Moyen-Orient (MERS-CoV) en 2012 [8], la pandémie de COVID-19 constitue la troisième grande épidémie due à un coronavirus. Le SARS-CoV-2 s'est toutefois révélé moins mortel que le MERS-CoV et le SARS-CoV [9], mais avec un taux de transmission plus élevé. Le taux de reproduction (R_0), évalue le potentiel de propagation d'un pathogène au cours d'une épidémie en se basant sur le nombre moyen d'infections secondaires causées par une personne infectée. Dans une population naïve en l'absence de mesures sanitaires de contrôle, il est qualifié de basique, et d'effectif dans les conditions réelles. Pour le SARS-CoV-2, le R_0 basique a été estimé entre 2.43 et 3.10 au début de la pandémie [10], puis a augmenté selon les variants circulants, passant à 5.08 en moyenne pour le variant Delta [11], et à 8.2 et 3.6 pour le variant Omicron [12] (basic et effectif respectivement). Sa transmission par les cas asymptomatiques a participé à rendre difficile le contrôle de l'épidémie, et la surcharge hospitalière a incité les gouvernements à instaurer des mesures sanitaires globales, comme la distanciation sociale, le port du masque obligatoire dans les lieux publiques, le recours massif aux tests diagnostiques [13] ou encore le confinement à domicile [14]. Enfin, une politique vaccinale internationale soutenue a été menée, ainsi mi-2021, trois billions de doses avaient été administrées de par le monde [15], à l'origine d'une diminution des formes graves, montrant une efficacité variable en fonction des différents variants [16], variants qui seront décrits ultérieurement.

3. Classification et structure

Au sein de l'ordre des *Nidovirales*, regroupant les virus à ARN monocaténaire de polarité positive (+ssRNA) [17] , la famille des *Coronaviridae*, décrite en 1975 [18], compte le plus de représentants . Parmi lesquels la sous famille des *Orthocoronavirinae,* qui regroupe les Alpha-, Bêta-, Gamma- et Deltacoronavirus, en fonction de la classification ICTV (International Committee on Taxonomy of Viruses) révisée [19]. Le genre des Bêtacoronavirus est divisé en 5 clades, parmi lesquels le sous-genre *Sarbecovirus*, comprenant le SARS-CoV et le SARS-CoV-2.

Le SARS-CoV-2, d'un diamètre compris entre 60 et 140 nm [20], se compose d'une enveloppe lipidique proche d'une sphère (Figure 1) associée à la protéine Spike (S), lui conférant sa forme caractéristique en couronne, reliée également aux protéines structurales de Membrane (M) et d'Enveloppe (E) (Figure 2).

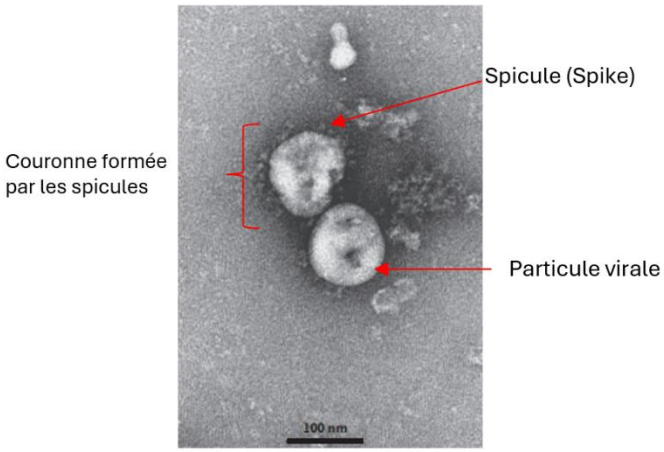


Figure 1 Observation du SARS-CoV-2 au microscope électronique à transmission, adapté de Zhu et al. ;2020 [20]

Le génome de référence du SARS-CoV-2 a été publié pour la première fois en 2020 [2], son organisation est schématisé en figure 2. Sa structure génomique est typique des coronavirus et se compose d'environ 30 000 nucléotides. A l'extrémité 5' -UTR, la séquence promotrice (210 pb), est commune aux 14 ORFs subséquents. ORF1a et ORF1ab, codent pour une poly-protéine précurseur, clivée en 16 protéines non structurales (NSP1 à NSP16). Le génome code également pour 4 protéines structurelles : la protéine Spike (S), impliquée dans l'attachement et l'entrée dans les cellules médiées par le récepteur ACE2 (Angiotensin Converting Enzyme 2), la protéine d'enveloppe (E), impliquée dans l'assemblage des virions produits, la protéine de Membrane (M), assurant l'intégrité structurelle du virus et la Nucléocapside (N), qui enrobe l'ARN viral et participe à sa stabilisation et à son transport. Les ORF accessoires (ORF3a, ORF6, ORF7a, ORF7b, ORF8,ORF9b, ORF10) quant à elles, codent pour des protéines dites

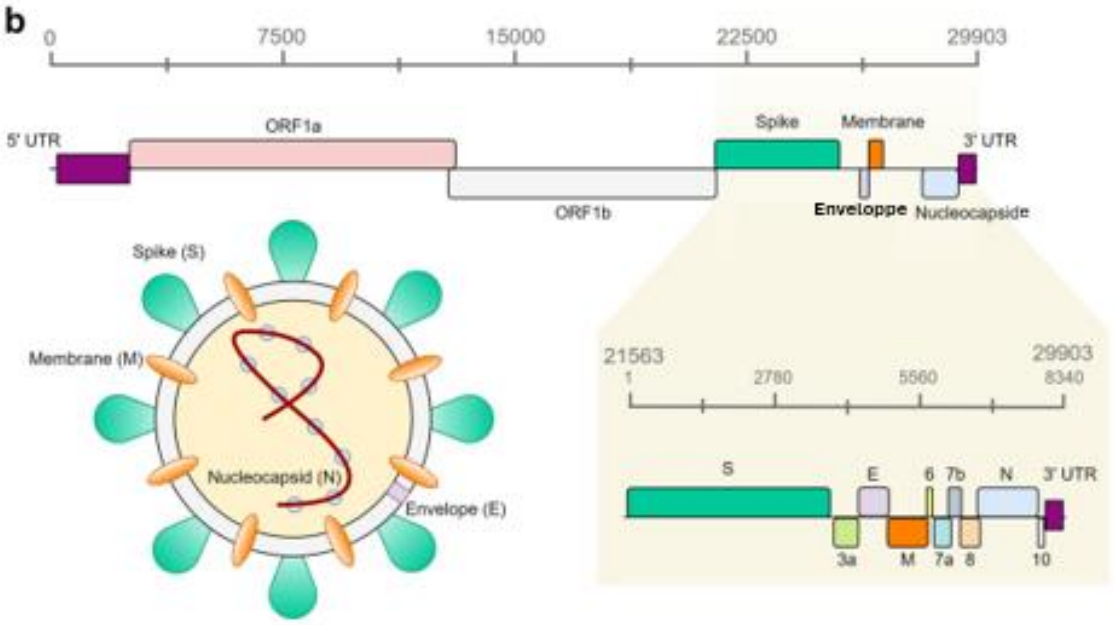


Figure 2 Structure génomique du SARS-CoV-2, Adaptée de Singh et Yi.;2021 (21)

accessoires dont l'ensemble des activités restent encore à décrire, comportant parmi d'autres, des activités immunomodulatrices [21,22].

4. Physiopathologie

La transmission se fait principalement par contact direct, *via* les gouttelettes et aérosols émises lors de l'infection [23,24], la voie de transmission indirecte *via* des surfaces contaminées ou la transmission féco-orale est considérée comme mineure en raison des faibles niveaux de matériel génétique retrouvés dans ces échantillons au regard des fluides nasopharyngés[17].

Le virus pénètre dans la cellule après fixation au récepteur ACE2 (Figure 3,4), empruntant deux voies distinctes, dépendante et indépendante de la protéase transmembranaire TMPRSS2 : les voies directe et endosomale respectivement [24] (Figure 4).

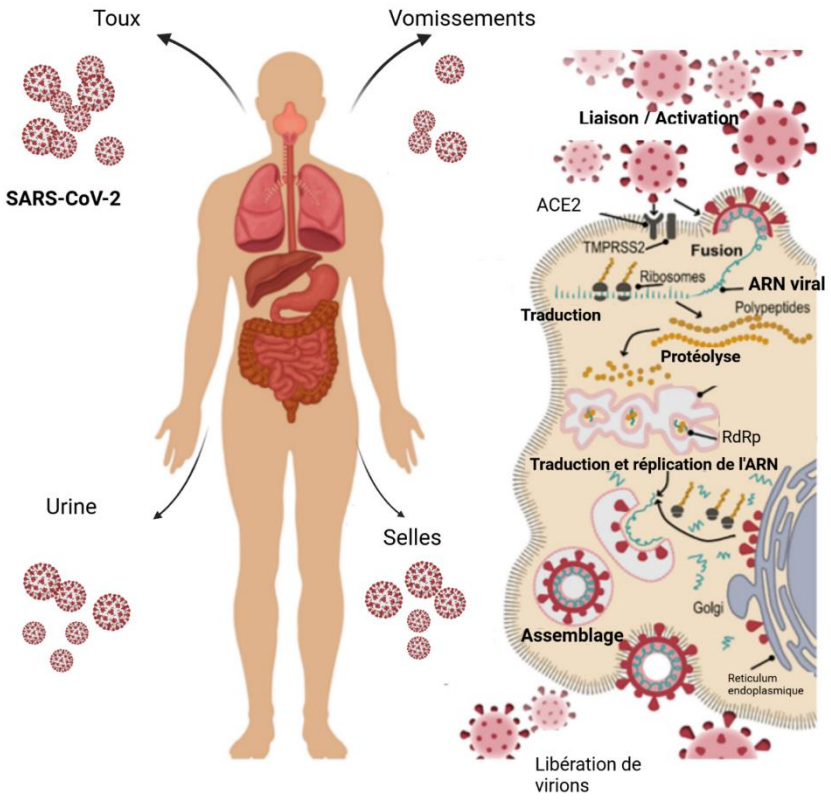


Figure 3 Voies de transmission du SARS-CoV-2 et d'entrée intra-cellulaire. Adaptée de Jones et al.,2020 [24]

En l'absence de TMPRSS2**,** le SARS-CoV-2 pénètre par endocytose, cette voie d'entrée se fait de manière cathepsine-dépendante. À la suite de l'acidification de l'endosome mature, les protéases lysosomales cathepsines activent alors la Spike en clivant la sous-unité S2, ce qui a pour conséquence d'exposer le peptide de fusion (FP) qui devient alors fonctionnel, induisant la fusion entre les membranes du SARS-CoV-2 et endosomale, déversant le génome viral dans le cytoplasme.

En présence de la protéase transmembranaire TMPRSS2, l'ARN viral pénètre directement dans la cellule après la fixation de la sous-unité S1 au récepteur ACE2. La protéase TMPRSS2 clive la protéine Spike au niveau de la sous-unité S2, ce qui active la fusion des membranes virale et cellulaire et la libération du matériel génétique dans le cytoplasme[31].

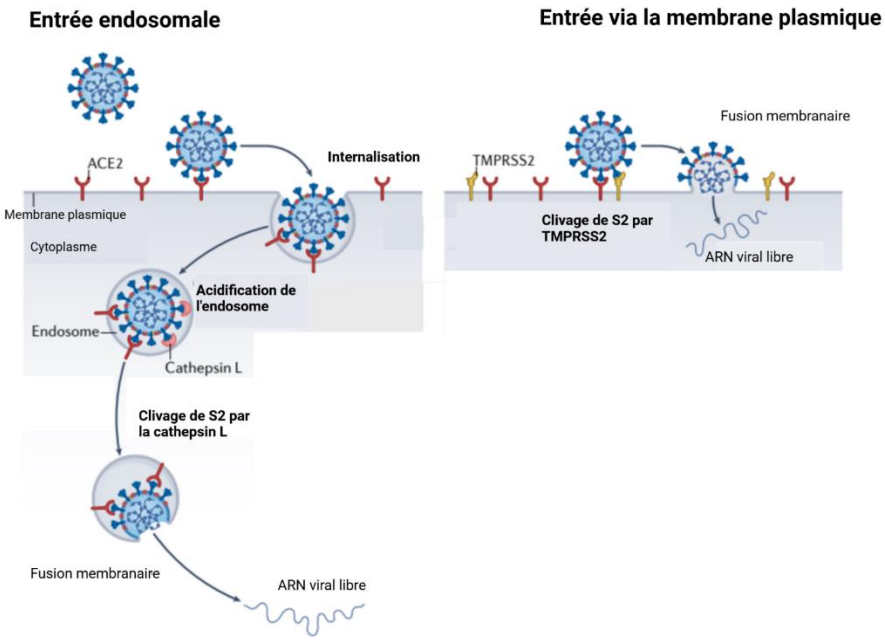


Figure 4 Voies d'entrée cellulaire de SARS-CoV2, adaptée de Jackson et al.;2022(31)

Concernant la symptomatologie, les cas décrits dans la littérature incluent des formes asymptomatiques, bénignes, graves à critiques pouvant mener à un syndrome de détresse respiratoire aiguë, voire au décès[25]. Les manifestations cliniques peuvent être multi systémiques, au-delà d'une atteinte du système respiratoire, les systèmes cardiologiques et neurologiques entre autres peuvent être atteints[16].

Le récepteur ACE2 est exprimé en grande quantité à la surface des cellules ciliées de la muqueuse nasale, ce taux diminue en progressant dans l'appareil respiratoire. Le récepteur ACE2 est également retrouvé au niveau du système digestif, du cœur et des reins[26].

Concernant la durée de portage, une méta-analyse comprenant 79 études a indiqué un portage nasopharyngé moyen de 17 jours[27] dans les voies respiratoires supérieurs avec une durée maximale de 83 jours, puis dans les voies respiratoires inférieures un portage de 14,6 jours en moyenne, et maximal de 59 jours, puis dans les selles, une durée moyenne de 17,2 jours ainsi qu'une durée maximale de 126 jours. Les portages viraux supérieurs à 17 jours ont été classés comme persistants, 30% des personnes infectées par le variant original ont été concernés[22]. Il a

également été démontré que certains facteurs, tels que la gravité de la maladie et un état d'immunosuppression, étaient associés à une excrétion prolongée de virus viables[21].

5. Diagnostic et séquençage du SARS-CoV-2

La RT-PCR en temps réel est considérée comme la méthode de référence pour diagnostiquer le SARS-CoV-2[23]. Cependant, chez les personnes infectées, ce test peut rester positif pendant des semaines, voire des mois, après la résolution des symptômes cliniques, sans permettre de distinguer entre l'excrétion de virus viables et potentiellement infectieux et celle de fragments viraux non infectieux[21]. Actuellement, la culture virale du SARS-CoV-2 reste la seule méthode fiable pour confirmer la présence de virus infectieux et évaluer le risque de transmission[24].

Depuis le 10 janvier 2020, 219 pays ont partagé 16 843 380 séquences du SARS-CoV-2, provenant d'infections humaines par le COVID-19, *via* la plateforme GISAID (Global initiative on sharing all influenza data) [28] (données au 25 février 2025). Le partage rapide du premier génome complet de SARS-CoV-2[2] a ainsi introduit la génomique au cœur de la gestion de cette pandémie.

L'étude de ces séquences a permis de mettre en évidence l'évolution en temps réelle du SARS-CoV-2, à travers l'étude des différents variants circulants, catégorisés comme suit : variants d'intérêts (VOI), variant préoccupants (VOC) et variants sous surveillance (VUM) [29]. Les variants du SARS-CoV-2 présentant des modifications génétiques pouvant affecter la transmissibilité, la sévérité de la maladie, ou la résistance aux traitements et vaccins ont été qualifiés de VOI. Les VOC remplissent les critères des VOI mais constituent immédiatement un risque pour la santé publique en raison de leur prévalence.

Ainsi, en 2022, l'OMS comptabilisait 5 variants préoccupants : Alpha (B.1.1.7), Beta (B.1.351), Gamma (P1), Delta (B.1.617.2) et Omicron (B.1.1.529), puis la lignée Omicron a généré une multitude de sous-variants, allant de BA.1 à KP.3.1.1 en 2025, selon la nomenclature de l'OMS (Figure 5)

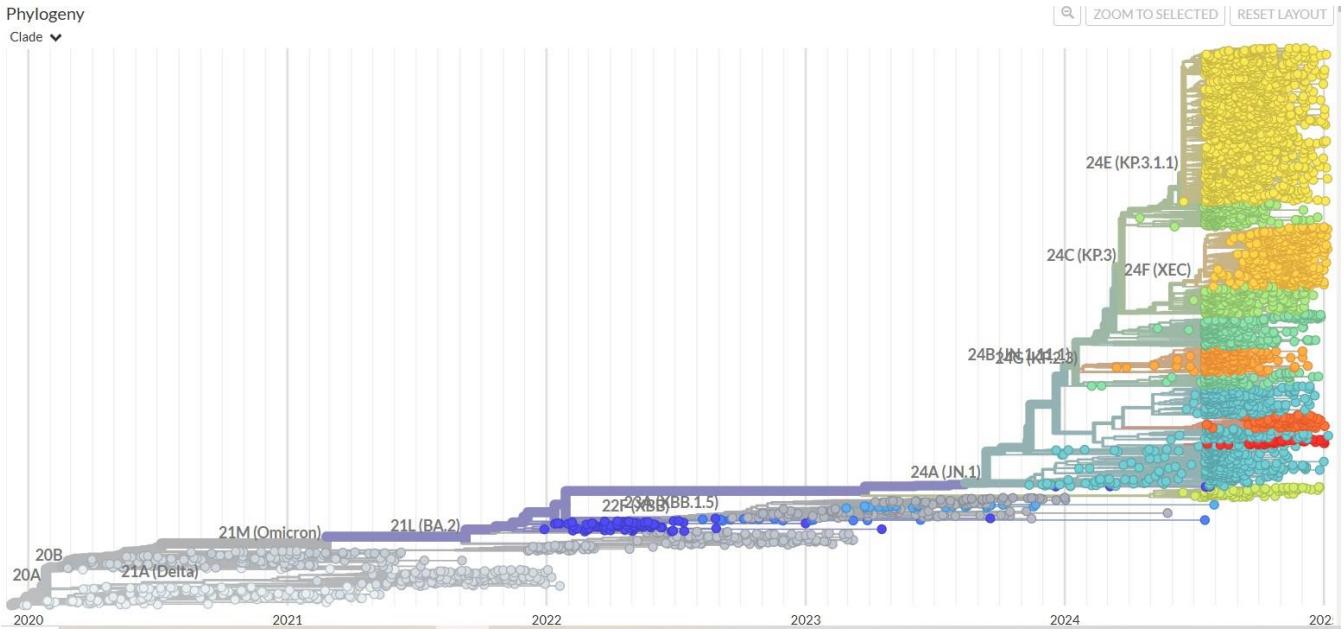


Figure 5 : Illustration des différents variants de SARS-CoV-2 détectés sur la période 2020-2025 (Nextstrain, https://nextstrain.org/ [30])

6. Hypothèses sur l'origine des nouveaux variants

Deux hypothèses principales sont aujourd'hui avancées pour expliquer l'émergence de nouveaux variants [31]. La transmission zoonotique, de l'Homme à l'animal, avec l'acquisition d'un nombre important de mutations, puis la retransmission à l'Homme. Des cas documentés de transmission entre humains et visons ont notamment été décrits [32]. La deuxième possibilité est l'infection persistante, notamment chez des individus immunodéprimés. Cette hypothèse, la plus répandue, suggère que chez les personnes immunodéprimées, le virus peut persister pendant plusieurs mois, offrant un environnement propice pour évoluer. L'équilibre entre une réponse immunitaire insuffisante et la réplication virale active favoriserait ainsi l'émergence de nouveaux variants[33]. Des études ont démontré que le SARS-CoV-2 pouvait évoluer chez des patients immunodéprimés long porteurs, donnant lieu à l'apparition de nouvelles mutations, qui dans certains cas ont conféré un échappement immunitaire. Une forte pression sélective exercée

sur le SARS-CoV-2 a été observée lors de l'administration de plasma convalescent dans un cas d'infection persistante chez un patient immunodéprimé[34]. Cette intervention a été associée à l'émergence de variants présentant une sensibilité réduite aux anticorps neutralisants [34]. Chez des individus immunodéprimés infectés par le VIH, des taux élevés de mutations ont également été observés [35]. Ces conditions favorisent non seulement l'évolution du virus au sein d'un hôte mais aussi, potentiellement, une "évolution inter-hôte" lorsque ces variants se propagent dans la population.

Article 1 (revue de littérature) : Contribution des variants minoritaires dans les études du SARS-CoV-2

Avant-propos

Durant la pandémie, les analyses et le partage des génomes consensus a permis de rendre compte rapidement des génotypes circulants. En revanche, le fait de se focaliser sur les séquences consensus peut réduire la complexité des données, en masquant par exemple les possibles co-infections, les évènements de recombinaisons en cours, ou encore l'émergence de mutations conférant des résistances aux traitements ou un échappement immunitaire. L'exploitation des données brutes, comprenant les mutations minoritaires, se révèle essentielle pour comprendre les mécanismes d'adaptation et de transmission virale notamment. Cela nécessite de mettre en place des paramètres expérimentaux et analytiques rigoureux, tenant compte, entre autres, de la profondeur minimale de lectures requise et des seuils minimums adaptés pour la fréquence allélique. L'absence de standardisation dans les pipelines d'analyse NGS pose des défis pour les agences réglementaires comme la FDA, qui doit tenir compte des résultats soumis par divers acteurs[36].

L'objectif de cette revue est de proposer une synthèse des connaissances actuelles sur la détection précise des mutations à faible fréquence (Low-Frequency Variants, LFVs) du SARS-CoV-2 et leur contribution dans différentes disciplines : en épidémiologie, en évolution virale et dans le domaine des interactions hôte-pathogène. Ce travail aspire également à renforcer le besoin de standardisation des approches de détection des LFVs, concernant les protocoles de séquençage et les analyses bio-informatiques subséquentes.

Cette revue a été soumise dans Archive of Virology le 28 mars 2025.

Draft Manuscript

Archives of Virology

SARS-CoV-2 Low-Frequency Genetic variants: accurate detection, advancing understanding in epidemiology, evolution, and host-pathogen interactions

--Manuscript Draft--

Manuscript Number:	ARVI-D-25-00254
Full Title:	SARS-CoV-2 Low-Frequency Genetic variants: accurate detection, advancing understanding in epidemiology, evolution, and host-pathogen interactions
Article Type:	Review

Manuscript

Click here to access/download;Manuscript;Figures.pdf

Figure 1

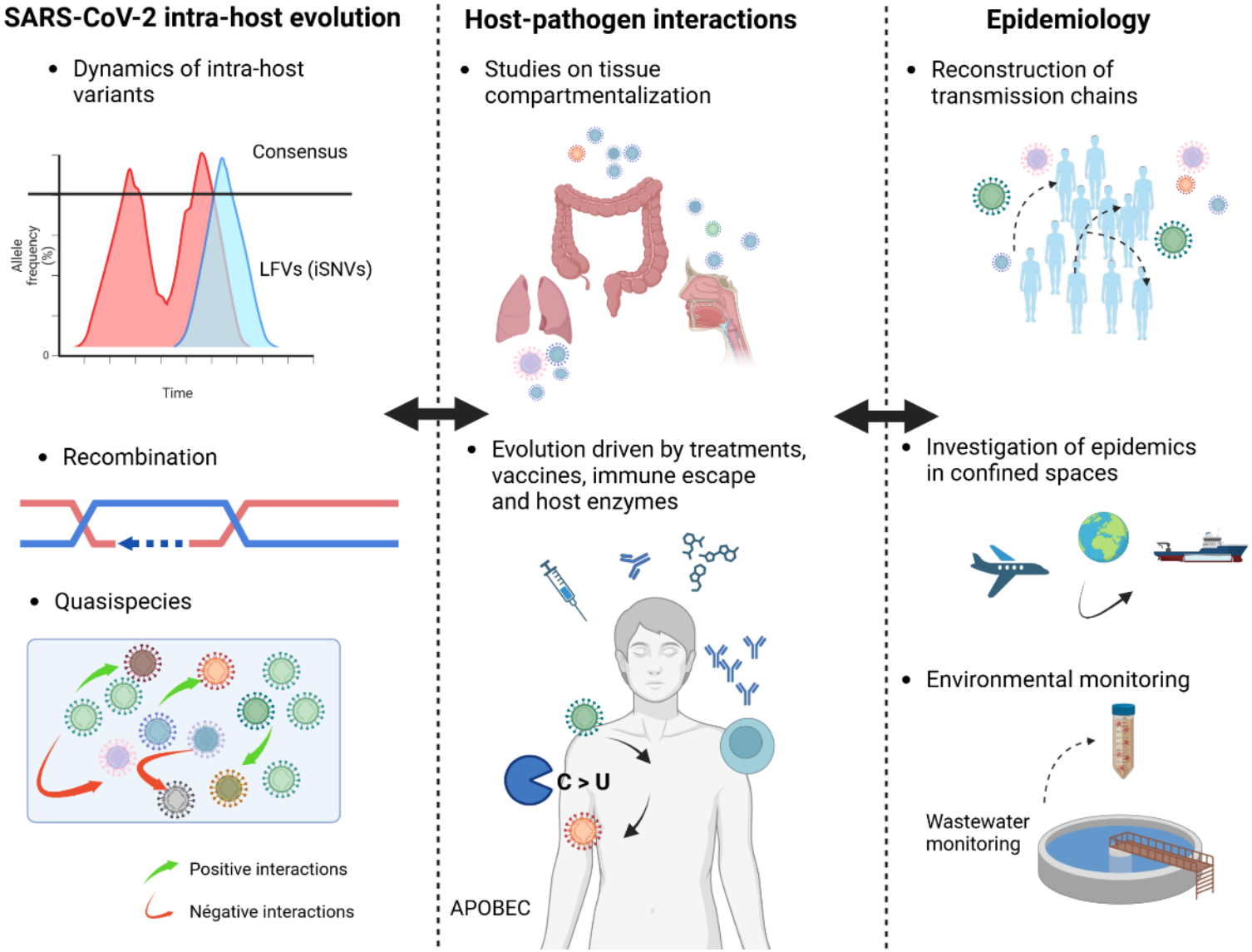
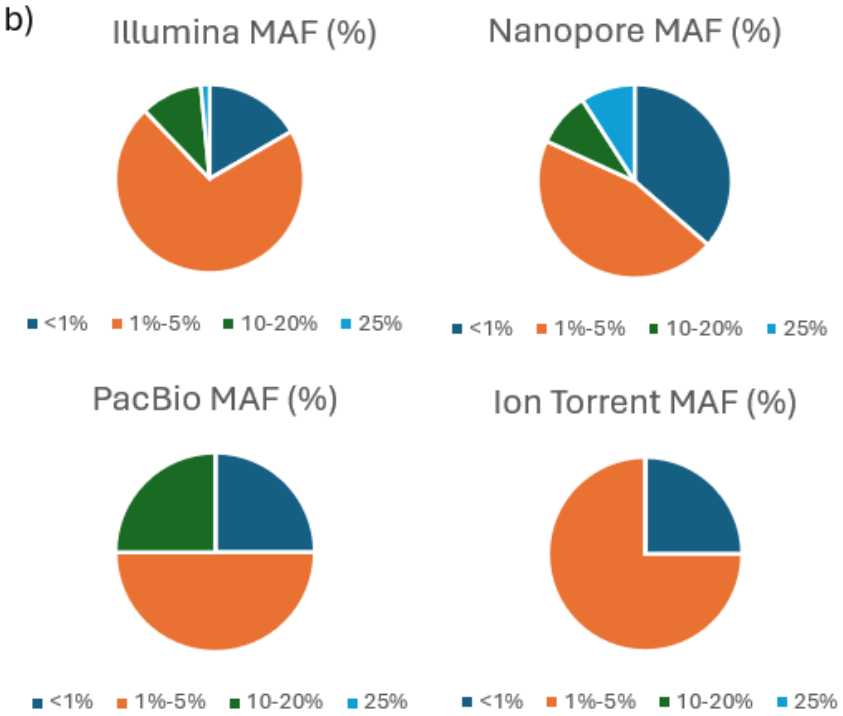


Figure 2

a)

	Illumina	Oxford Nanopore	PacBio	MGI	Sanger
Number of articles listed	67	9	6	3	1

b)



c)

iVar	LoFreq	GATK	VarScan
23%	13%	5.70%	5%

d)

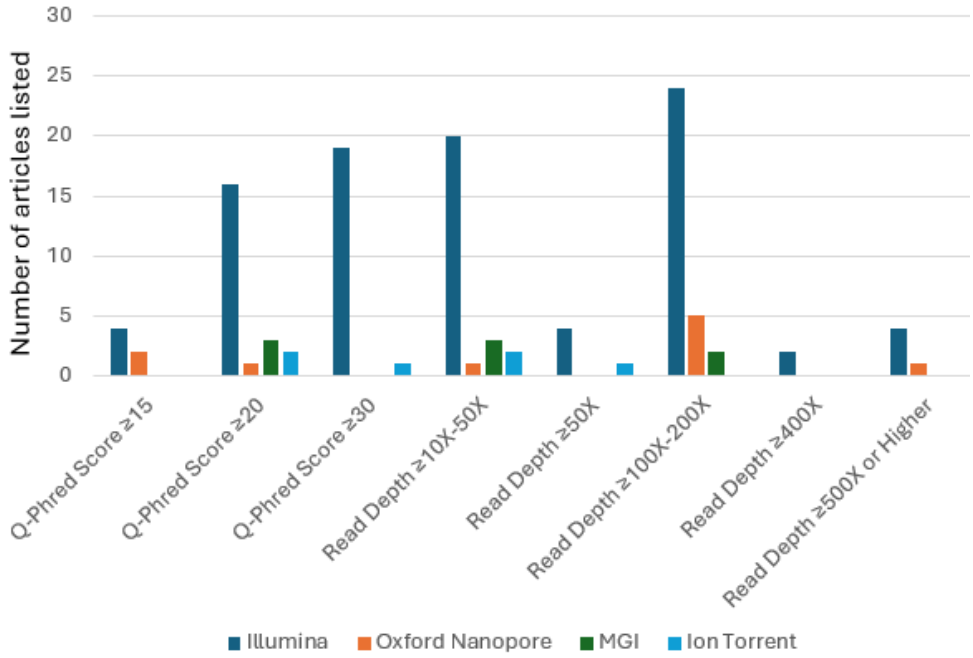
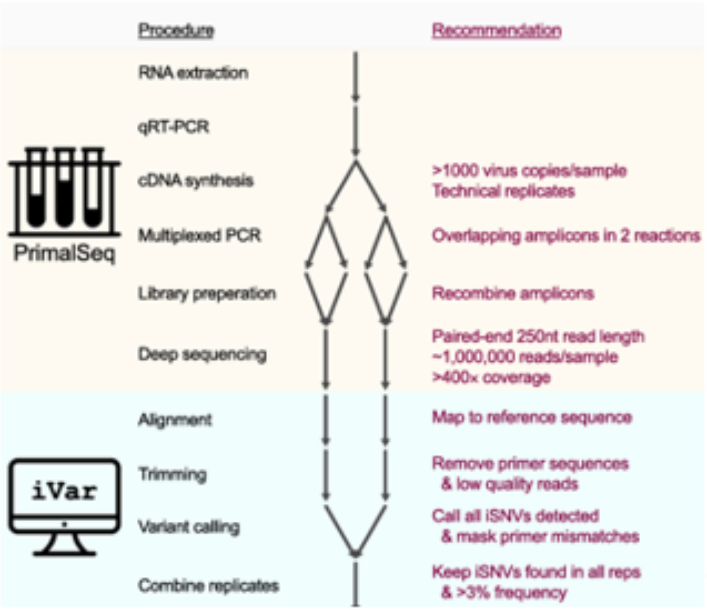


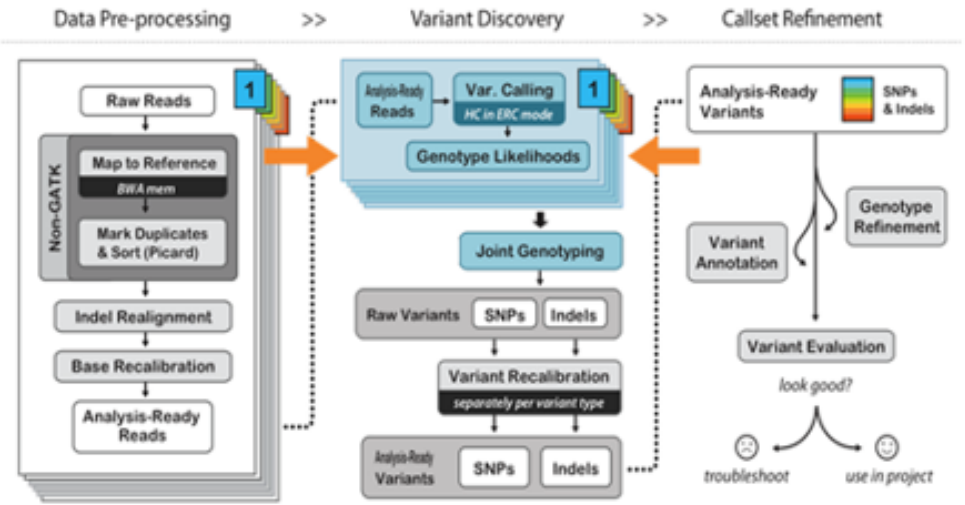
Figure 3



iVar
Grubaugh et al.;2019

•**Pre-Processing**:
•Quality trimming and filtering of raw sequencing reads to remove low-quality bases and adapters.
•Align filtered reads to a reference genome,
•**Quality-Aware Error Modeling**:
•Model sequencing errors using Phred-scaled quality scores for each base.
•Apply a Poisson-binomial distribution to calculate probabilities of observing non-reference bases, adjusting for position-specific error rates.
•**Variant Calling**:
•Perform exact statistical testing for each position to distinguish true variants from sequencing errors.
•Compute P-values for each candidate variant
•Use a Bonferroni-corrected P-value threshold to filter statistically significant SNVs.
•**Filtering and Refinement**:
•Remove low-quality variant calls based on coverage thresholds and strand bias testing (via Fisher's exact test).
•**Output**:
•Generate a list of high-confidence single nucleotide variants (SNVs) (AF ≥0,05%)

Lofreq
Data from Wilm et al.;2012



GATK
Poplin et al.2018

Pre-Processing:
•Quality trimming and filtering of raw sequencing reads to remove low-quality bases and sequencing adapters.
•Align high-quality reads to a reference genome
•Generate an mpileup file from the aligned BAM files using SAMtools, ensuring proper handling of sequencing depth and quality metrics.
Variant Calling:
•Use VarScan to process the mpileup file and identify variants.
•**Filtering**
•Apply filtering thresholds to reduce false-positive calls:
- Minimum read depth (e.g., ≥10 for normal, ≥20 for tumor).
- Minimum variant allele frequency (e.g., ≥0.01 for rare variants).

Refinement:
- Remove strand-biased or low-confidence variants.

Output:
•Generate a categorized list of variants

VarScan2
Koboldt et al.;2012

Figures Click here to access/download;Figure;Manuscript.docx

Archive of Virology

Title: SARS-CoV-2 Low-Frequency Genetic variants: accurate detection, advancing understanding in epidemiology, evolution, and host-pathogen interactions

Authors: Emilie Burel[1,2,3]*, Pierre Edouard Fournier[1,2,3]

1) IHU Méditerranée Infection, 19-21 Bd Jean Moulin, 13005 Marseille, France

2) Assistance Publique-Hôpitaux de Marseille, Marseille, France

3) RITMES, Aix-Marseille Université, Service de Santé des Armées, IHU Méditerranée, Infection, Marseille, France

***Corresponding author:**

emilie.burel@etu.univ-amu.fr

ABSTRACT

The study of SARS-CoV-2 low-frequency variants (LFVs), resulting from genetic mutations present at minor allele frequencies within a host, has emerged as a key approach for understanding viral evolution, transmission dynamics, and immune escape mechanisms. LFVs reflect the genetic diversity inherent in viral populations, often masked by consensus genome analyses, and can reveal early-stage adaptive mutations, recombination events, and tissue-specific compartmentalization of the virus. These variants are also essential for tracing infection dynamics, including reinfections and superinfections, and for enhancing epidemiological surveillance through wastewater sequencing. This review synthesizes findings from 86 studies leveraging deep sequencing

to detect SARS-CoV-2 LFVs, highlighting the diversity in sequencing technologies, protocols and bioinformatics tools used. Experimental parameters such as cycle threshold (Ct) values, sequencing depth, genome coverage, and base quality metrics were reported for their impact on the detection of LFVs. The review also addresses the semantic variability and lack of standardization in LFV detection methodologies, which challenge reproducibility and comparability between studies.

Keywords: SARS-CoV-2, COVID-19, low-frequency variants, iSNV, quasispecies, ultra-deep sequencing, viral evolution, epidemiology.

INTRODUCTION

Since the onset of the COVID-19 pandemic in December 2019, the study of the SARS-CoV-2 genome has been crucial in understanding the virus's evolutionary dynamics through the study of circulating variants[1,2,3]. However, reliance solely on consensus genomes, which represent dominant nucleotide sequences at each position, overlooks intra-host diversity[4,5].

SARS-CoV-2 low-frequency variants (LFV) refer to genetic mutations detected at low allele frequencies, eventually below the threshold of 90% as specified by Ramazzotti et al.[6]. LFVs arise due to the high mutation rates of SARS-CoV-2, driven by the error-prone RNA-dependent RNA polymerase[7] which generates significant genomic plasticity, and eventually recombination events derived from co-infection[8]. Accurate detection of SARS-CoV-2 LFVs is essential for understanding intra-host viral diversity, as they can provide insights into better understanding viral evolution, transmission dynamics, immune escape and antiviral resistance (Ramazzotti et al.;2022[9]). LFVs are identified using deep sequencing techniques and bioinformatics tools designed to detect minor

genetic variations within background noise, as highlighted by Van Poelvoorde et al.[10] concerning the Influenza virus. Their detection in SARS-CoV-2 genomes is particularly challenging due to the lack of standardized guidelines and the variability in sequencing technologies and protocols, and bioinformatics strategies used (Roder et al., 2023[11], Walter et al.[12]).

Within the host, SARS-CoV-2 diversity is captured within the quasispecies concept, introduced in the 1970s (Eigen and Schuster, 1977[13]) and defined as " a given distribution of macromolecular species with closely interrelated sequences", subject to positive or negative interactions (Ojosnegros et al;2011.[14]) where a dominant clone or major variant prevails. As evidenced for some RNA viruses (Seligman et al.;2007[15]; Domingo et al.;2018[16]; Singh et al.;2023[17]), within host SARS-CoV-2 populations are not homogeneous entities but complex mixtures of interdependent mutants or called subpopulations, contributing together to the virus's overall evolutionary path (Jary et al.;2020[18]; Nyayanit et al.;2020[19]; Capobianchi et al.;2020[20]; Karamitros et al. ;2020[21]; Sun et al.;2021[22]; Quasispecies : Subramoney et al.;2023[23]). This intricate process seems to be driven by host-pathogen interactions, including mechanisms such as RNA editing by enzymes like APOBEC, which are hypothesized to play a role in the generation of quasispecies and are further influenced by cellular senescence induced during SARS-CoV-2 infection (Karakasiliotis et al.;2021[24]). Indels (insertions and deletions) are also considered low-frequency variants (LFVs) and contribute significantly to the composition of viral quasispecies (Andrés et al.;2020[25]).

In the environment, SARS-CoV-2 LFVs have been identified in wastewater and have also proved to be valuable epidemiological indicators (Lawal et al.;2022[26]). These samples lack true quasispecies in the strict sense but serve as a valuable snapshot of circulating

variants within a community. This approach hence acts as an early warning system for emerging variants, without relying on individual testing (Rios et al.;2021[27]).

This review focuses on the contributions of LFVs to understanding viral evolution, host-pathogen interactions and epidemiological surveillance. Moreover, this review aims to synthesize current knowledge on the accurate detection of SARS-CoV-2 LFVs, with a focus on molecular and bioinformatics strategies. Based on an analysis of 86 articles, we highlighted significant technical heterogeneity in sequencing technologies, protocols, metrics, and bioinformatics tools used for detecting LFVs, emphasizing the need for standardized guidelines to ensure consistency and comparability across studies. The reliable detection of LFVs and their integration into genomic studies provide essential insights into viral diversity, evolutionary dynamics, and transmission pathways, ultimately supporting more effective strategies for pandemic management and prevention.

I) Intrahost evolution

The study of SARS-CoV-2's LFVs provides critical insights into intra-host evolution (Figure 1), particularly in immunocompromised patients experiencing prolonged infections. While intra-host diversity of SARS-CoV-2 has been shown in immunocompetent patients (Goldswain et al.;2024), it has been shown that immunocompromised patients can be the reservoir of viral diversification (Shapira et al.;2023[28]), often including functionally significant mutations associated with immune or therapeutic escape (Avanzato et al.;2020[29]; Choi et al.;2020[30]; Siqueira et al.;2021[31]; Scherer et al.;2022[32] ; Caccuri et al.;2022[33], Quarantana et al.;2022[34]; Hettle et al.;2022[35]; Chaguza et al.;2023[36]; Raglow et al.;2024[37]). The study by Manuto et al.[38] suggested that

a compromised immune system alone does not influence within-host SARS-CoV-2 genomic variability, highlighting the potential influence of additional factors such as infection duration. By representing early-stage mutations that can either remain as minor variants or rise to dominance under selective pressures (Kemp et al.;2020[39]), intra host LFVs contribute to the quasispecies dynamics and the potential emergence of clinically significant variants (Sonnleitner et al.;2022[40]). The nature of these mutations can further shed light on early viral adaptation processes (Markov et al.;2023[41]).

Under selective pressures such as antiviral drugs like Remdesivir (Heyer et al.;2022[42], Nirmalarajah et al.;2023[43]), immune responses leading to escape from T cells, as shown by Stanevich et al.[44], or monoclonal antibodies therapies like Sotrovimab (Palomino Cabrera et al.;2023[45]), SARS-CoV-2 LFVs have been shown to play a critical role in shaping the viral population (Karamitros et al.;2020[21]). While many intra-host LFVs are filtered out by bottleneck effects during transmission (Armero et al.;2021[46]), their potential to develop into fixed mutations (single nucleotide polymorphisms, SNPs) highlights their critical role in viral evolution (Agius et al.;2022[47]). These variants frequently emerge in fluctuating patterns rather than a linear progression, as noted by Sonnleitner et al.[40], with new adaptive mutations arising and being replaced by the wild type or other variants. A vaccine-induced evolution has also been highlighted with reduced sensitivity to the BNT162b2 vaccine developed during the course of infection (Khateeb et al.;2022[48]).

Morevover, the study of low-frequency variants (LFVs) in SARS-CoV-2 provides valuable insights into the relationship between intra-host viral diversity and disease severity. According to Zhang et al.[49], a higher number of iSNVs (intra host single nucleotide variations) with specific mutational patterns, such as C>U and G>A

substitutions, are associated with severe COVID-19 cases, reflecting strong immune-driven selective pressures during infection. Martínez-González et al.[50] found that patients with mild disease displayed a more diverse mutant spectrum compared to those with moderate or severe disease, suggesting that a higher diversity of LFVs may indicate a less severe pathogenic trajectory. Meanwhile, Li et al.[51] highlighted a two-step fitness selection process for iSNVs, where nonsynonymous mutations are positively selected within the host, potentially contributing to immune evasion and impacting clinical outcomes as the disease progresses.

However, intra-host diversity should not be overestimated, as reported by Xi et al.[52], with an average of 1.90 iSNVs per sample observed in the cases studied, this low diversity was attributed to the virus's short replication period, the proofreading mechanisms of nsp14, and purifying selection against deleterious mutations.

II)Tissue-specific compartmentalization of SARS-CoV-2

The analysis of low-frequency variants (LFVs) has been pivotal in uncovering tissue-specific compartmentalization of SARS-CoV-2 (Figure 1), revealing how the virus evolves and adapts differently across various body sites. Sun et al.[22] advanced this understanding by demonstrating that quasispecies complexity varies significantly across tissues, with sputum samples exhibiting lower genetic diversity compared to nasopharyngeal swabs or stool samples. By comparing iSNVs across tissue sites, Farjo et al.[53] observed significant genetic compartmentalization between nasal and saliva samples during acute infection. They observed discrepancies in both variant frequencies and the presence or absence of specific iSNVs, indicating that distinct microenvironments drive differential viral replication dynamics. Similarly, Gaiarsa et al.[54] documented distinct quasispecies profiles between the upper and lower respiratory tracts, identifying a higher

frequency of non-synonymous mutations and indels in the lower tract, possibly driven by interactions with immune cells like macrophages. Rueca et al.[55] further supported this, showing profound differences in the minority variant patterns between the upper and lower respiratory tracts too, suggesting compartmentalized replication within these distinct environments. These works emphasized the importance of considering tissue-specific LFVs when studying SARS-CoV-2 evolution, taking into account localized immune responses and microenvironmental pressures.

III) Co-infection, reinfection, superinfection and recombination events.

The study of SARS-CoV-2 LFVs has proven essential for distinguishing reinfections from ongoing viral evolution, as well as identifying coinfections and superinfections that can lead to recombination events. Naveca et al.[56] demonstrated that LFVs can identify distinct viral lineages between episodes of infection, providing clear evidence of reinfection rather than the continued evolution of an initial strain. This was particularly significant in confirming reinfections with the Gamma variant in Brazil, shedding light on the epidemiological role of reinfections in areas with high transmission rates. Lythgoe et al.[57] revealed evidence of coinfection or contamination in 1–2% of the 1313 sequenced SARS-CoV-2, including a case of a dual infection with the OC43 coronavirus. In the same way, Alisoltani et al.[58] described a unique case of superinfection in an immunocompromised pediatric patient, where an initial Delta variant infection was followed by a transient Omicron superinfection and a subsequent resurgence of Delta. This case illustrates the complex interplay between coexisting strains within a host, as detected through LFVs. Garcia Lopéz et al.[59] further emphasized the importance of LFVs in identifying coinfections by detecting distinct lineages within individuals during periods of variant turnover. This evidence supports the idea that coinfections create opportunities for lineage

interactions such as recombinations events where distinct viral strains exchange genetic material (Figure 1). Markov et al.[41] highlighted the role of LFVs in tracing how recombination and convergent evolution generate haplotypes that later define variants of concern (VOCs), emphasizing their importance in understanding SARS-CoV-2 evolution. Sayama et al.[60] analyzed a patient co-infected with Delta and Omicron variants, identifying nine distinct recombinant sequences in the spike gene using long-read sequencing. These recombinants were linked to the co-circulation period of these variants, showcasing how LFVs can reveal complex evolutionary interactions between variants (Wertheim et al.;2022[61]; Sayama et al.;2023[60]; Pipek et al.;2024[62] ; Ravi et al.;2024[63]).

IV) The role of low-frequency variants of SARS-CoV-2 in improving epidemiological knowledge

SARS-CoV-2 LFVs are essential for detecting subtle transmission patterns that would otherwise be overlooked in the consensus genome and still need to be studied intensively (Delgado et al.;2024[64]). Chan et al.[4] observed the frequent occurrence of multiple biallelic sequence polymorphisms among the 360 analyzed samples and integrated ambiguity codes at positions where the alternate allele frequency (AAF) exceeded 5%, while advocating for this use in genomic databases. They emphasized that incorporating such diversity into epidemiological reconstructions can uncover complexities in SARS-CoV-2 transmission dynamics (Figure 1). They reported that many nucleotide changes classified as LFVs have later been identified as lineage-defining mutations in emerging variants of concern (VOCs), as noticed by others (Pathak et al.;2022[65], Chan et al.;2022[4]). Bendall et al.[66] highlighted how minority variants can occasionaly resolve transmission ambiguities in household outbreaks, moreover, the inclusion of LFVs in phylogenetic

reconstructions, traditionally based on consensus genomes, allowed to refine the resolution of the SARS-CoV-2 evolutionary and transmission patterns, notably using the VERSO framework (Ramazzotti et al.;2021[6]). In the same way, Fiam et al.[67] introduced a new method to preserve all nucleotide frequencies at each genomic position by constructing a "genome matrix", enabling a comprehensive variation profile analysis. This approach improved accuracy by 20% compared to consensus-based methods. Similarly, Ortiz et al.[68] showed that taking intra-host diversity into account improved phylogenetic inference.

Regarding the contribution of LFVs to epidemiological studies in confined environments, He et al.[69] reported rapid dynamic changes in intra-host diversity with the emergence of new site mutations during a flight-related cluster outbreak. In contrast, Hannon et al.[70] found that SARS-CoV-2 transmission in a fishing boat outbreak was characterized by narrow transmission bottlenecks (Markov et al.;2023[41]), significantly reducing intra-host viral diversity. They observed that mutations either became fixed during single transmission events or were lost, limiting the utility of iSNVs for detailed transmission reconstruction in such tightly constrained settings.

The study of LFVs in wastewater has emerged as a complementary tool for epidemiological surveillance, providing early warnings of emerging variants at the community level. De Llanos et al.[71] conducted a pilot project on a university campus in Spain, demonstrating that wastewater analysis could detect mutation profiles of SARS-CoV-2 variants well before their peak circulation. This approach allowed for proactive public health decisions to mitigate outbreaks. Izquierdo-Lara et al.[72] extended this finding, showing that wastewater sequencing could monitor the diversity and circulation of SARS-CoV-2 at a community level in Belgium and the Netherlands, offering critical

insights into variant introduction. A similar work by Rios et al.[27] in France, demonstrated the ability of nanopore sequencing to assess lineage-specific dynamics in Nice's neighborhoods, offering detailed insight into localized epidemics. In Portugal, Reis et al.[73] demonstrated that systematic sequencing of the S-gene in wastewater samples enabled early lineage detection, identifying rare mutations and VOCs ahead of clinical cases. This highlights the ability of wastewater surveillance to complement patient-based epidemiology.

V) SARS-CoV-2 LFVs: terminology variability

The terminology and thresholds associated with SARS-CoV-2 LFVs are marked by variability, which may introduce challenges in comparing studies and interpreting results. Various studies adopt different criteria for defining LFVs, minor variants, and consensus-level mutations, often overlapping but not equivalent, which complicates the harmonization of findings across the field. Intra-host single nucleotide variants (iSNVs) are often defined as variants with an allele frequency (AF) between 2% and 50%, while those exceeding 50% are considered consensus-level SNPs, reflecting standard thresholds in genome assembly pipelines (Sapoval et al.[74]). The minimum allele frequency threshold can be set as low as 0.1% (Martinez-Gonzalez et al.[50]). In contrast, thresholds for LFVs can be set higher, such as positions with nucleotides representing 10% to 70% of mapped reads, while variants below 10% are excluded from consideration (Jacot et al.[75]). Additional classifications introduce categories like minority variants (<20%), intermediate variants (20–80%), and fixed variants (>80%) to describe varying mutation frequencies (Pérez-Lago et al.[76]). Some studies classify mutations into three frequency categories—low (≤10%), medium (11–50%), and high (>50%)—further complicating comparisons (Ravi et al.[63]). Other approaches adopt distinct consensus thresholds, with

consensus SNPs defined as supported by >50% of reads (Stanevich et al.[44]), requiring ≥80% AF for consensus calling (Xiao et al.[77]), or considering minor variants detected at frequencies below 90% (Ramazzotti et al.[9]), widening their allele frequency range to overlap with the consensus thresholds used in other studies.

VI) Accurate detection of SARS-CoV-2 LFVs

Van Poelvoorde et al.[10], in their work on influenza, proposed a general strategy to standardize low-frequency variant detection methodologies, showing the importance of defining reliable thresholds for variants, including recommendations for minimal genome copy numbers and allele frequency cutoffs. To address the critical challenges in detecting SARS-CoV-2 low-frequency variants (LFVs) (Jacot et al.[75]), this review compiles 86 studies that have leveraged deep sequencing data for SARS-CoV-2 LFVs detection. We analyzed the methodological approaches employed, with a focus on experimental parameters such as cycle threshold (Ct) values, sequencing technologies and protocols, bioinformatics tools, genome coverage, sequencing depth, and base quality scores.

a) Viral load

The cycle threshold (Ct) value, which serves as a proxy for viral load, was explicitly mentioned for 47 out of 86 articles and ranged from 19.5 to 40 with an average value of around 30. Low Ct values were not overrepresented. These variations suggest a significant diversity in viral loads across samples used for sequencing, which strongly influence sequencing depth and quality, as reported by Voloch et al.[78], that more variability was found at higher Ct values, while Kubik et al.[79] highlighted the higher proportion of mapped reads obtained with samples with Ct values <26, and samples (with high Ct values were associated with an apparent increase in the number of detected iSNVs

at any given threshold while samples with high viral loads, corresponding to a Ct of ~22, exhibited highly reproducible iSNVs with a minimum MAF of 3% underscoring the importance of accounting for viral load when interpreting intra-host variability, due to stochastic sampling effects and reduced sequencing coverage. Finally, statistical models have been developed to account for variability related to viral load in longitudinal studies (Heyer et al.[42]).

In particular, the advantages of using PacBio SMRT (Single Molecule Real-Time): (the SMRT technology does not require any PCR amplification), sequencing for detecting LFVs and identifying recombination events in SARS-CoV-2 were studied. Lhomme et al.[80] demonstrated that PacBio and Illumina technologies can detect minority variants at comparable levels, in the Spike gene based on three samples, but PacBio exceled in preserving sequence continuity due to its longer reads. Trémeaux et al.[81] revealed that PacBio's long-read capability can distinguish recombinant haplotypes in co-infections, where strains exchange genetic material to form novel variants. Recombinants were identified in 24% of co-infected patients, highlighting the precision of PacBio in resolving these events.

b) Sequencing technologies and protocols

Among the 86 articles considered, Illumina stands out as the most used sequencing technology, with a prevalence of 77% of the studies mentioned (Fig 2a). This technology offers a high sequencing depth and low error rate, estimated at 0.1% under controlled conditions and supported by studies such as that by Stoler and Nekrutenko[82], which reports a median error rate as low as 0.087% for specific instruments such as the HiSeq X Ten and NovaSeq 6000. These features make Illumina technology particularly suitable to detect low-frequency variants (LFV), which explains its predominant adoption in

288 research focused on this area. Other technologies are used to a lesser extent in this corpus
289 of articles. Oxford Nanopore is cited in 9 articles (10.5%), PacBio in 6 articles (6.5%),
290 MGISeq in 3 articles (3.5%), and Ion Torrent in 3 articles (3.5%). Their marginal
291 representation likely reflects their limited adoption for detecting LFVs in the contexts
292 studied. Dudouet et al. (2022)[83] mentionned the limitations of Oxford Nanopore
293 Technology (ONT) in this application due to its higher error rate, which was mitigated by
294 using a high sequencing coverage and focusing on variants with a minor allele frequency
295 (MAF) higher than 20%. In the same way, PacBio represents a small proportion of the
296 sequencing technologies used in this corpus (6.5%), although several studies have
297 demonstrated its value in the analysis of viral quasispecies (Trémeaux et al.;2023[81];
298 Lhomme et al.(2021)[80]). Similarly, Sun et al. (2021)[22] emphasized the utility of PacBio
299 for directly generating full-length haplotypes. They noted that this method provides a
300 comprehensive view of the quasispecies structure, capturing even low-abundance minor
301 variants. In the same way, Xiao et al. (2020)[77] who conducted a comparative study
302 between sequencing technologies and protocols, highlighted the PacBio capacity for high
303 accuracy and sensitivity in capturing full-length SARS-CoV-2 haplotypes and minimizing
304 sequencing errors. The use of MGISeq technology for the detection of low-frequency
305 variants (LFVs) has been valued for its capability to deliver robust sequencing depth,
306 particularly when applied with hybrid capture methods. Hou et al.;2023[84] demonstrated
307 that MGISeq platform provided reliable identification of intrahost single nucleotide
308 variants (iSNVs), even in samples with low viral loads. Oxford Nanopore Technology
309 (ONT) is recognized for its ability to generate long sequencing reads, making it
310 particularly advantageous for structural analyses like recombination detection (Fournelle
311 et al.[85]; Ravi et al.[63]). Finally, Ion Torrent sequencing effectively identifies treatment-

emergent mutations and quasispecies complexity in SARS-CoV-2 genomes during antiviral therapy (Gruber et al.[86]).

Regarding sequencing protocols, the majority of studies included employed amplicon-based sequencing (90%), which need to be processed using appropriate bioinformatics. Hybrid capture RNA sequencing, which was used in 3% of studies, offers unbiased enrichment of target regions and typically relies on tools such as ViralFlow or custom pipelines. Direct RNA sequencing (5%) and metatranscriptomic sequencing (2%) need robust bioinformatics tools such as LoFreq or in-house scripts to manage the inherent complexity and variability of their datasets, but are exempt from the biases inherent to amplicons.

Lam et al.[87] evaluated three different methods: Random Viral Oligonucleotide Priming (RVOP), ARTIC v3, and long-amplicon sequencing protocols. They showed that reproducibility and consistency in SARS-CoV-2 LFVs detection varied significantly across methods, particularly in low-viral-load contexts. The RVOP protocol demonstrated superior efficiency in detecting and replicating iSNVs, compared to traditional methods like Long-amp and ARTIC v3, which exhibited lower variant reproducibility

c) Metric parameters

To ensure accurate detection and analysis of low-frequency variants (LFVs), specific quality thresholds and metrics have been established across sequencing studies. Commonly accepted thresholds include read depth (≥10 for Illumina, ≥20 for ONT), Q-phred score >25, and a genome coverage ≥90% (Zufan et al.[88]; Xiaoli et al.[89]). Jacot et al.[75] proposed 11 quality metrics for SARS-CoV-2 low-frequency variant detection,

emphasizing that LFVs, with AF<70%, often arise from PCR amplification errors or sample contamination.

c-1) Minor allele frequency

Among the 86 studied articles, the percentage distribution of MAF class for each sequencing technology is presented in Figure 2b. Across all platforms, a common observation is the predominance of variants with a MAF between 1% and 5%. Illumina is associated with smaller proportion of variants with higher MAF thresholds (>10%) while Oxford Nanopore and PacBio are associated with greater proportion of MAFs in the 10–20% and >25% ranges. This likely reflects the need to mitigate their higher error rates. For instance, the use of Oxford Nanopore often applies a higher MAF threshold to distinguish true variants from ONT sequencing errors (5-15%) (Dudouet et al.;2022[83], Fournelle et al.;2024[85]).

c-2) Q-Phred

The Q-Phred score thresholds reported in the 86 studies is presented in Figure 2d. Illumina, the most widely used platform, was consistently associated with Q-Phred ≥20 (error rate ~1%) and Q-Phred ≥30 (error rate ~0.1%). Specifically, 16 studies used Q-Phred ≥20, while 19 studies adopted Q-Phred ≥30, underscoring Illumina's reliability and dominance in high-quality sequencing, these results align with Jacot et al.[75]'s recommendations. In contrast, for Oxford Nanopore, the few available studies were consistently equal to or below 20, reflecting the technology's higher error rates and the challenges in achieving higher quality scores. Finally, for the limited represented MGI and Ion Torrent platform, they reported Q-Phred ≥20 but minimal adoption of higher thresholds.

c-3) Minimum Read Depths

The analysis of minimum read depth requirements across the 86 studies pointed distinct technological preferences. Illumina was dominantly associated with the 100X–200X range and appeared to be the preferred threshold for high-quality and reliable sequencing data. A significant number of studies (20 articles) reported using the 10X–50X range, demonstrating Illumina's versatility for applications where lower coverage is sufficient. At ≥500X or higher, Illumina remains the dominant technology, particularly for studies integrating ultra-rare variant detection (AF>1%) such that performed by Jary et al.[18]

A preference for the 50X–100X range was associated with ONT, for which lower depths (e.g., 10X–50X) were less represented, likely due to ONT high error rates, which require greater coverage for reliable detection of low-frequency variants (LFVs). A threshold set at ≥500X or higher was noticed once (Choga et al.;2024[90]).

PacBio studies primarily reported depths in the 10X–50X range, consistent with its niche application for generating long reads where ultra-high depths are not typically required. Finally, both MGI and Ion Torrent were underrepresented in the studies analyzed. However, when used, these technologies predominantly appeared in the 10X–50X range.

Overall, Illumina emerged as the most versatile and dominant platform used, particularly at 100X–200X, while Oxford Nanopore, PacBio, and MGI/Ion Torrent technologies were preferentially used in lower or intermediate depth ranges, and sporadically high.

d) Bioinformatic tools

The choice of variant-calling tools and analysis methods significantly impacts single nucleotide variations (SNVs) detection accuracy and depends on selecting appropriate bioinformatic workflows and parameter thresholds (Zufan et al.;2023[88],Xiaoli et

al.;2022[89]). What applies to the detection of SNVs is even more critical for low-frequency variants (LFVs), as the challenges in distinguishing true variants from sequencing errors are amplified at lower allele frequencies. In human papillomavirus (HPV) studies, Mishra et al.([91]) proposed innovative approaches to improve LFVs detection. They implemented statistical models that incorporate data coverage, quality parameters, and trinucleotide context, while additionally applying supervised machine learning techniques, which effectively discriminated true intra-host single nucleotide variants (iSNVs) from false positives. Similarly, in studies on the foot-and-mouth disease virus (FMDV), Cacciabue et al.([92]) demonstrated the effectiveness of combining ultra-sensitive variant detection tools like LoFreq[93] with haplotype reconstruction algorithms such as CliqueSNV[94], QuRe[95], and ViQuaS[96]. These tools were calibrated using both highly variable and low-diversity samples to ensure robust detection of sub-consensus variants and accurate reconstruction of viral quasispecies, emphasizing the need for tailored bioinformatics pipelines for reliable LFV analysis.

On SARS-CoV-2 studies, Roder et al.[11], using synthetic RNA controls, emphasized that combining multiple variant callers with stringent cutoffs is necessary to mitigate false-positive and false-negative rates, particularly when replicates are unavailable. A notable example was provided by Kimbrel et al.[97], who employed both LoFreq[93] and iVar[92] to ensure a comprehensive and accurate detection of intra-host single nucleotide variants (iSNVs). The combined use of these two tools leverages their complementary strengths: the highly sensitive tool LoFreq, capable of detecting LFVs at extremely low allele frequencies (≥0.05%), and iVar[92], adapted to amplicon-based sequencing data. The algorithm PySNV (Li et al.[98]) was designed to detect intra-host variants, including insertions, deletions, and linked mutations, with high performance at low allele

frequencies (<4%), which enabled identification of up to 21% more variants than Lofreq. Moreover, bioinformatic tools can be specifically adapted to the specific challenges of each sequencing technology. In the case of Oxford Nanopore Technology (ONT), the majority of sequencing errors are deletions, which can occur in homopolymer regions – where the same nucleotide is repeated – or in low-complexity genomic regions. (O'Donnell et al.;2013[99]). To solve this drawback, Liu et al.[100] developped a bioinformatic tool, Variabel, tested on SARS-CoV-2, which utilizes variant frequency information and entropy-based filtering to differentiate true intra-host variants from sequencing errors inherent to ONT data.

Among the bioinformatics tools listed in the 86 collected articles, 4 were the most represented (Figure 3): iVar most frequently used (23;%), followed by LoFreq (13%), GATK[101] (5.7%) and VarScan2[102] (5%). A wide variety of bioinformatic tools and approaches were observed across the 86 studies (Supplementary table). These included popular tools like Geneious, FreeBayes, VirVarSeq, and V-Phaser2, as well as in-house scripts and pipelines such as VICOS and Mappgene, illustrating the diversity in workflows for detecting low-frequency variants (LFVs). The workflows of iVar, LoFreq, GATK, and VarScan2 demonstrate their distinct strengths and specialized applications for variant detection. iVar, designed for amplicon-based sequencing workflows like those used for SARS-CoV-2, integrates primer trimming and replicate merging to enhance the detection of intra-host single nucleotide variants (iSNVs) at frequencies ≥3%. LoFreq stands out for its exceptional sensitivity in identifying low-frequency variants (LFVs, ≥0.05%) using quality-aware statistical models, making it particularly well-suited for studying viral quasispecies and their evolutionary dynamics. GATK, with its highly flexible and modular pipeline, uses advanced probabilistic models via HaplotypeCaller to

reconstruct haplotypes, improving the detection of LFVs even in complex samples. Combined with features like base quality score recalibration (BQSR) and Variant Quality Score Recalibration (VQSR), GATK ensures reliable detection of rare mutations while minimizing false positives, VarScan2, although originally developed for tumor-normal comparisons, is also highly effective for viral genomics, providing customizable thresholds for allele frequency and read depth that allow for the reliable detection of LFVs. Its robust filtering capabilities minimize false positives, making it invaluable for exploring intra-host diversity and quasispecies populations.

e) Replicate analysis

Finally, to improve the reliability of low-frequency variant (LFV) detection, many studies recommended performing duplicate analyses or replicates. This approach not only accounts for potential sequencing errors but also increases confidence in the observed results. For instance, Marques et al.[103] highlighted the importance of generating at least two independent viral genome sequences per patient. By ensuring replicates from at least two time points, the authors observed high consistency in the identified mutations and their proportional occurrence, reinforcing the reproducibility of the results. However, the authors caution against biases introduced by sampling, emphasizing that the vast pool of diversity within viral populations might not always be fully captured. Similarly, Gao et al.[104] required that minor alleles be sequenced at least twice to validate their presence and reduce false-positive calls. Li et al.[51] further refined this approach in their study where they re-sequenced libraries in 62 samples and validated their findings using independent methods such as PCR and Sanger sequencing. The results demonstrated that over 90% of the LFVs identified in the first round of sequencing were also detected in the second round, with a high correlation in mutation allele frequencies ($R^2 = 0.89$).

Conclusion

The study of SARS-CoV-2 low-frequency variants (LFVs) has proven indispensable for advancing our understanding of viral evolution, transmission, and host-pathogen interactions. By highlighting methodological challenges and synthesizing insights from 86 studies, this review emphasizes the need for standardized protocols in sequencing and bioinformatics to ensure reproducibility and comparability of LFVs analyses. LFVs provide a unique lens into the early stages of viral adaptation, immune escape, and epidemiological dynamics, offering critical opportunities for refining genomic surveillance and public health strategies. Continued investment in accurate detection and interpretation of LFVs will enhance our capacity to manage current and future viral pandemics effectively.

Material and Methods

Articles were searched in the PubMed database using specific keywords as follows: "Intra-host variants" “SARS-CoV-2", "iSNV" and “SARS-CoV-2", "quasispecies" and “SARS-CoV-2", and "Low-frequency variants" and “SARS-CoV-2". These keywords were chosen to capture studies addressing low-frequency variants and intra-host mutations associated with SARS-CoV-2, the same keywords were used with COVID-19. The initial search identified a total of 185 articles. These articles were then screened based on their methodological relevance and the content of reported data. To be included in the review, articles needed to provide specific data and detailed methodologies used to detect SARS-CoV-2 iSNVs or low-frequency variants, from human samples. 86 articles were included in the final literature review. Figure 1 was produced using Biorender (https://www.biorender.com/).

Acknowledgments

The authors thanks the Institut Hospitalo-Universitaire Méditerranée Infection and the French National Research Agency.

Funding

The study was supported by the Institut Hospitalo-Universitaire Méditerranée Infection and the French National Research Agency under the program “investissements d'avenir” (reference ANR-10-IAHU-03).

Competing interests

The authors declare no competing interests.

Author’s contribution

PEF was involved in the design and supervision of this study, EB collected and analysed the data and drafted the manuscript; EB performed data analysis. Both authors were involved in writing the paper and have approved the final version.

REFERENCES

1. Abdool Karim, S. S. & de Oliveira, T. New SARS-CoV-2 Variants — Clinical, Public Health, and Vaccine Implications. *N. Engl. J. Med.* **384**, 1866–1868 (2021).
2. Tao, K. *et al.* The biological and clinical significance of emerging SARS-CoV-2 variants. *Nat. Rev. Genet.* **22**, 757–773 (2021).
3. Boehm, E. *et al.* Novel SARS-CoV-2 variants: the pandemics within the pandemic Erik. *Clin. Microbiol. Infect.* **27**, 1109–117 (2021).

4. Chan, E. R. *et al.* COVID-19 infection and transmission includes complex sequence diversity. *PLOS Genet.* **18**, e1010200 (2022).
5. Colson, P. *et al.* From viral democratic genomes to viral wild bunch of quasispecies. *J. Med. Virol.* **95**, e29209 (2023).
6. Ramazzotti, D. *et al.* VERSO: A comprehensive framework for the inference of robust phylogenies and the quantification of intra-host genomic diversity of viral samples. *Patterns* **2**, 100212 (2021).
7. Banerjee, A., Mossman, K. & Grandvaux, N. Molecular Determinants of SARS-CoV-2 Variants. *Trends Microbiol.* **29**, 871–873 (2021).
8. Focosi, D. & Maggi, F. Recombination in Coronaviruses, with a Focus on SARS-CoV-2. *Viruses* **14**, 1239 (2022).
9. Ramazzotti, D. *et al.* Large-scale analysis of SARS-CoV-2 synonymous mutations reveals the adaptation to the human codon usage during the virus evolution. *Virus Evol.* **8**, veac026 (2022).
10. Van Poelvoorde, L. A. E. *et al.* A general approach to identify low-frequency variants within influenza samples collected during routine surveillance. *Microb. Genomics* **8**, (2022).
11. Roder, A. E. *et al.* Optimized quantification of intra-host viral diversity in SARS-CoV-2 and influenza virus sequence data. *mBio* e01046-23 (2023) doi:10.1128/mbio.01046-23.
12. Walter, K. S. *et al.* Challenges in Harnessing Shared Within-Host Severe Acute Respiratory Syndrome Coronavirus 2 Variation for Transmission Inference. *Open Forum Infect. Dis.* **10**, ofad001 (2023).

13. Eigen, M. & Schuster, P. A principle of natural self-organization: Part A: Emergence of the hypercycle. *Naturwissenschaften* **64**, 541–565 (1977).
14. Ojosnegros, S., Perales, C., Mas, A. & Domingo, E. Quasispecies as a matter of fact: Viruses and beyond. *Virus Res.* **162**, 203–215 (2011).
15. Seligman, S. J. Evidence for Quasi Species in Severe Acute Respiratory Syndrome–Associated Coronavirus Deletion Mutants. *J. Infect. Dis.* **195**, 609–609 (2007).
16. Domingo, E. & Perales, C. Quasispecies and virus. *Eur. Biophys. J.* **47**, 443–457 (2018).
17. Singh, K., Mehta, D., Dumka, S., Chauhan, A. S. & Kumar, S. Quasispecies Nature of RNA Viruses: Lessons from the Past. *Vaccines* **11**, 308 (2023).
18. Jary, A. *et al.* Evolution of viral quasispecies during SARS-CoV-2 infection. *Clin. Microbiol. Infect.* **26**, 1560.e1-1560.e4 (2020).
19. Nyayanit, D., Yadav, P., Kharde, R. & Shete-Aich, A. Quasispecies analysis of the SARS-CoV-2 from representative clinical samples: A preliminary analysis. *Indian J. Med. Res.* **152**, 105 (2020).
20. Capobianchi, M. R. *et al.* Molecular characterization of SARS-CoV-2 from the first case of COVID-19 in Italy. *Clin. Microbiol. Infect.* **26**, 954–956 (2020).
21. Karamitros, T. *et al.* SARS-CoV-2 exhibits intra-host genomic plasticity and low-frequency polymorphic quasispecies. *J. Clin. Virol.* **131**, (2020).
22. Sun, F. *et al.* SARS-CoV-2 Quasispecies Provides an Advantage Mutation Pool for the Epidemic Variants. *Microbiol. Spectr.* **9**, e00261-21 (2021).
23. Subramoney, K. *et al.* SARS-CoV-2 spike protein diversity at an intra-host level, among SARS-CoV-2 infected individuals in South Africa, 2020 to 2022. *PLOS ONE* **18**, e0286373 (2023).

24. Karakasiliotis, I., Lagopati, N., Evangelou, K. & Gorgoulis, V. G. Cellular senescence as a source of SARS-CoV-2 quasispecies. *FEBS J.* **290**, 1384–1392 (2023).
25. Andrés, C. *et al.* Naturally occurring SARS-CoV-2 gene deletions close to the spike S1/S2 cleavage site in the viral quasispecies of COVID19 patients. *Emerg. Microbes Infect.* **9**, 1900–1911 (2020).
26. Lawal, O. U. *et al.* Metagenomics of Wastewater Influent from Wastewater Treatment Facilities across Ontario in the Era of Emerging SARS-CoV-2 Variants of Concern. *Microbiol. Resour. Announc.* **11**, e00362-22 (2022).
27. Rios, G. *et al.* Monitoring SARS-CoV-2 variants alterations in Nice neighborhoods by wastewater nanopore sequencing. *Lancet Reg. Health - Eur.* **10**, 100202 (2021).
28. Shapira, G., Patalon, T., Gazit, S. & Shomron, N. Immunosuppression as a Hub for SARS-CoV-2 Mutational Drift. *Viruses* **15**, 855 (2023).
29. Avanzato, V. A. *et al.* Case Study: Prolonged Infectious SARS-CoV-2 Shedding from an Asymptomatic Immunocompromised Individual with Cancer. *Cell* **183**, 1901-1912.e9 (2020).
30. Choi, B. *et al.* Persistence and Evolution of SARS-CoV-2 in an Immunocompromised Host. *N. Engl. J. Med.* **383**, 2291–2293 (2020).
31. Siqueira, J. D. *et al.* SARS-CoV-2 genomic analyses in cancer patients reveal elevated intrahost genetic diversity. *Virus Evol.* **7**, (2021).
32. Scherer, E. M. *et al.* SARS-CoV-2 Evolution and Immune Escape in Immunocompromised Patients. *N. Engl. J. Med.* **386**, 2436–2438 (2022).
33. Caccuri, F. *et al.* Competition for dominance within replicating quasispecies during prolonged SARS-CoV-2 infection in an immunocompromised host. *Virus Evol.* **8**, veac042 (2022).

34. Quaranta, E. G. *et al.* SARS-CoV-2 intra-host evolution during prolonged infection in an immunocompromised patient. *Int. J. Infect. Dis.* (2022) doi:10.1016/j.ijid.2022.06.023.

35. Hettle, D., Hutchings, S., Muir, P. & Moran, E. Persistent SARS-CoV-2 infection in immunocompromised patients facilitates rapid viral evolution: Retrospective cohort study and literature review. *Clin. Infect. Pract.* **16**, 100210 (2022).

36. Chaguza, C. *et al.* Accelerated SARS-CoV-2 intrahost evolution leading to distinct genotypes during chronic infection. *Cell Rep. Med.* **4**, 100943 (2023).

37. Raglow, Z. *et al.* SARS-CoV-2 shedding and evolution in patients who were immunocompromised during the omicron period: a multicentre, prospective analysis. *Lancet Microbe* **5**, e235–e246 (2024).

38. Manuto, L. *et al.* Immune System Deficiencies Do Not Alter SARS-CoV-2 Evolutionary Rate but Favour the Emergence of Mutations by Extending Viral Persistence. *Viruses* **16**, 447 (2024).

39. Kemp, S. A. *et al.* SARS-CoV-2 evolution during treatment of chronic infection. *Nature* **592**, 277–282 (2021).

40. Sonnleitner, S. T. *et al.* Cumulative SARS-CoV-2 mutations and corresponding changes in immunity in an immunocompromised patient indicate viral evolution within the host. *Nat. Commun.* **13**, 2560 (2022).

41. Markov, P. V. *et al.* The evolution of SARS-CoV-2. *Nat. Rev. Microbiol.* **21**, 361–379 (2023).

42. Heyer, A. *et al.* Remdesivir-induced emergence of SARS-CoV2 variants in patients with prolonged infection. *Cell Rep. Med.* **3**, 100735 (2022).

43. Nirmalarajah, K. *et al.* Use of whole genome sequencing to identify low-frequency mutations in SARS-CoV-2 patients treated with remdesivir. *Influenza Other Respir. Viruses* **17**, e13179 (2023).
44. Stanevich, O. V. *et al.* SARS-CoV-2 escape from cytotoxic T cells during long-term COVID-19. *Nat. Commun.* **14**, 149 (2023).
45. Palomino-Cabrera, R. *et al.* Frequent Emergence of Resistance Mutations Following Complex Intra-Host Genomic Dynamics in SARS-CoV-2 Patients Receiving Sotrovimab. *Antimicrob. Agents Chemother.* **67**, e00266-23 (2023).
46. Armero, A., Berthet, N. & Avarre, J.-C. Intra-Host Diversity of SARS-Cov-2 Should Not Be Neglected: Case of the State of Victoria, Australia. *Viruses* **13**, 133 (2021).
47. Agius, J. E. *et al.* SARS-CoV-2 Within-Host and in vitro Genomic Variability and Sub-Genomic RNA Levels Indicate Differences in Viral Expression Between Clinical Cohorts and in vitro Culture. *Front. Microbiol.* **13**, 824217 (2022).
48. Khateeb, D. *et al.* SARS-CoV-2 variants with reduced infectivity and varied sensitivity to the BNT162b2 vaccine are developed during the course of infection. *PLOS Pathog.* **18**, e1010242 (2022).
49. Zhang, Y. *et al.* SARS-CoV-2 intra-host single-nucleotide variants associated with disease severity. *Virus Evol.* **8**, veac106 (2022).
50. Martínez-González, B. *et al.* SARS-CoV-2 Point Mutation and Deletion Spectra and Their Association with Different Disease Outcomes. *Microbiol. Spectr.* **10**, e00221-22 (2022).
51. Li, J. *et al.* Two-step fitness selection for intra-host variations in SARS-CoV-2. *Cell Rep.* **38**, 110205 (2022).

617 52. Xi, B. *et al.* SARS-CoV-2 within-host diversity of human hosts and its implications
618 for viral immune evasion. *mBio* e00679-23 (2023) doi:10.1128/mbio.00679-23.
619 53. Farjo, M. *et al.* Within-host evolutionary dynamics and tissue
620 compartmentalization during acute SARS-CoV-2 infection. *J. Virol.* **98**, e01618-23
621 (2024).
622 54. Gaiarsa, S. *et al.* Comparative analysis of SARS-CoV-2 quasispecies in the upper
623 and lower respiratory tract shows an ongoing evolution in the spike cleavage site.
624 *Virus Res.* **315**, 198786 (2022).
625 55. Rueca, M. *et al.* Compartmentalized Replication of SARS-Cov-2 in Upper vs. Lower
626 Respiratory Tract Assessed by Whole Genome Quasispecies Analysis.
627 *Microorganisms* **8**, 1302 (2020).
628 56. Naveca, F. G. *et al.* SARS-CoV-2 intra-host diversity, antibody response, and
629 disease severity after reinfection by the variant of concern Gamma in Brazil. *Sci.*
630 *Rep.* **13**, 7306 (2023).
631 57. Lythgoe, K. A. *et al.* SARS-CoV-2 within-host diversity and transmission. *Science*
632 **372**, eabg0821 (2021).
633 58. Alisoltani, A. *et al.* Resurgence of SARS-CoV-2 Delta after Omicron variant
634 superinfection in an immunocompromised pediatric patient. *Virol. J.* **20**, 246
635 (2023).
636 59. García-López, R. *et al.* Exploration of low-frequency allelic variants of SARS-CoV-2
637 genomes reveals coinfections in Mexico occurred during periods of VOCs
638 turnover. *Microb. Genomics* **10**, (2024).

60. Sayama, Y. *et al.* Identification of Various Recombinants in a Patient Coinfected With the Different SARS-CoV-2 Variants. *Influenza Other Respir. Viruses* **18**, e13340 (2024).

61. Wertheim, J. O. *et al.* Detection of SARS-CoV-2 intra-host recombination during superinfection with Alpha and Epsilon variants in New York City. *Nat. Commun.* **13**, 3645 (2022).

62. Pipek, O. A. *et al.* Systematic detection of co-infection and intra-host recombination in more than 2 million global SARS-CoV-2 samples. *Nat. Commun.* **15**, 517 (2024).

63. Ravi, V. *et al.* Unraveling the genetic evolution of SARS-CoV-2 Recombinants using mutational dynamics across the different lineages. *Front. Med.* **10**, 1294699 (2024).

64. Delgado, S. *et al.* Incipient functional SARS-CoV-2 diversification identified through neural network haplotype maps. *Proc. Natl. Acad. Sci.* **121**, e2317851121 (2024).

65. Pathak, A. K. *et al.* Spatio-temporal dynamics of intra-host variability in SARS-CoV-2 genomes. *Nucleic Acids Res.* **50**, 1551–1561 (2022).

66. Bendall, E. E. *et al.* SARS-CoV-2 Genomic Diversity in Households Highlights the Challenges of Sequence-Based Transmission Inference. *mSphere* **7**, e00400-22 (2022).

67. Fiam, R. N., István, C. & Norbert, S. Comparing full variation profile analysis with the conventional consensus method in SARS-CoV-2 phylogeny. *Brief. Bioinform.* **25**, bbae296 (2024).

662 68. Torres Ortiz, A. *et al.* Within-host diversity improves phylogenetic and
663 transmission reconstruction of SARS-CoV-2 outbreaks. *eLife* **12**, e84384 (2023).
664 69. He, Y. *et al.* Temporal dynamics of SARS-CoV-2 genome mutations that occurred
665 in vivo on an aircraft. *Biosaf. Health* **5**, 62–67 (2023).
666 70. Hannon, W. W. *et al.* Narrow transmission bottlenecks and limited within-host
667 viral diversity during a SARS-CoV-2 outbreak on a fishing boat. *Virus Evol.* **8**,
668 veac052 (2022).
669 71. De Llanos, R. *et al.* Monitoring the evolution of SARS-CoV-2 on a Spanish
670 university campus through wastewater analysis: A pilot project for the reopening
671 strategy. *Sci. Total Environ.* **845**, 157370 (2022).
672 72. Izquierdo-Lara, R. *et al.* Monitoring SARS-CoV-2 Circulation and Diversity through
673 Community Wastewater Sequencing, the Netherlands and Belgium. *Emerg. Infect.*
674 *Dis.* **27**, 1405–1415 (2021).
675 73. Reis, A. C. *et al.* Systematic SARS-CoV-2 S-gene sequencing in wastewater samples
676 enables early lineage detection and uncovers rare mutations in Portugal. *Sci.*
677 *Total Environ.* **921**, 170961 (2024).
678 74. Sapoval, N. *et al.* SARS-CoV-2 genomic diversity and the implications for qRT-PCR
679 diagnostics and transmission. *Genome Res.* **31**, 635–644 (2021).
680 75. Jacot, D., Pillonel, T., Greub, G. & Bertelli, C. Assessment of SARS-CoV-2 Genome
681 Sequencing: Quality Criteria and Low-Frequency Variants. *J. Clin. Microbiol.* **59**,
682 e00944-21 (2021).
683 76. Pérez-Lago, L. *et al.* Different within-host viral evolution dynamics in severely
684 immunosuppressed cases with persistent sars-cov-2. *Biomedicines* **9**, (2021).

77. Xiao, M. *et al.* Multiple approaches for massively parallel sequencing of SARS-CoV-2 genomes directly from clinical samples. *Genome Med.* **12**, 57 (2020).
78. Voloch, C. M. *et al.* Intra-host evolution during SARS-CoV-2 prolonged infection. *Virus Evol.* **7**, (2021).
79. Kubik, S. *et al.* Recommendations for accurate genotyping of SARS-CoV-2 using amplicon-based sequencing of clinical samples. *Clin. Microbiol. Infect.* **27**, 1036.e1-1036.e8 (2021).
80. Lhomme, S. *et al.* Prediction of SARS-CoV-2 Variant Lineages Using the S1-Encoding Region Sequence Obtained by PacBio Single-Molecule Real-Time Sequencing. *Viruses* **13**, 2544 (2021).
81. Trémeaux, P. *et al.* SARS-CoV-2 Co-Infections and Recombinations Identified by Long-Read Single-Molecule Real-Time Sequencing. *Microbiol. Spectr.* **11**, e00493-23 (2023).
82. Stoler, N. & Nekrutenko, A. Sequencing error profiles of Illumina sequencing instruments. *NAR Genomics Bioinforma.* **3**, lqab019 (2021).
83. Dudouet, P. *et al.* SARS-CoV-2 quasi-species analysis from patients with persistent nasopharyngeal shedding. *Sci. Rep.* **12**, 18721 (2022).
84. Hou, M. Intra- vs. Interhost Evolution of SARS-CoV-2 Driven by Uncorrelated Selection—The Evolution Thwarted.
85. Fournelle, D. *et al.* Intra-Host Evolution Analyses in an Immunosuppressed Patient Supports SARS-CoV-2 Viral Reservoir Hypothesis. *Viruses* **16**, 342 (2024).
86. Gruber, C. E. M. *et al.* Molnupiravir increases SARS-CoV-2 genome diversity and complexity: A case-control cohort study. *J. Med. Virol.* **96**, e29642 (2024).

87. Lam, C. *et al.* SARS-CoV-2 Genome Sequencing Methods Differ in Their Abilities To Detect Variants from Low-Viral-Load Samples. *J. Clin. Microbiol.* **59**, e01046-21 (2021).

88. Zufan, S. E. *et al.* Bioinformatic investigation of discordant sequence data for SARS-CoV-2: insights for robust genomic analysis during pandemic surveillance. *Microb. Genomics* **9**, (2023).

89. Xiaoli, L. *et al.* Benchmark datasets for SARS-CoV-2 surveillance bioinformatics. *PeerJ* **10**, e13821 (2022).

90. Choga, W. T. *et al.* Rapid dynamic changes of FL.2 variant: A case report of COVID-19 breakthrough infection. *Int. J. Infect. Dis.* **138**, 91–96 (2024).

91. Mishra, S. K. *et al.* Improved detection of low-frequency within-host variants from deep sequencing: A case study with human papillomavirus. *Virus Evol.* **10**, veae013 (2024).

92. Cacciabue, M., Currá, A., Carrillo, E., König, G. & Gismondi, M. I. A beginner's guide for FMDV quasispecies analysis: sub-consensus variant detection and haplotype reconstruction using next-generation sequencing. *Brief. Bioinform.* **21**, 1766–1775 (2020).

93. Wilm, A. *et al.* LoFreq: a sequence-quality aware, ultra-sensitive variant caller for uncovering cell-population heterogeneity from high-throughput sequencing datasets. *Nucleic Acids Res.* **40**, 11189–11201 (2012).

94. Knyazev, S. *et al.* Accurate assembly of minority viral haplotypes from next-generation sequencing through efficient noise reduction. *Nucleic Acids Res.* **49**, e102–e102 (2021).

731 95. Prosperi, M. C. F. & Salemi, M. QuRe: software for viral quasispecies
732 reconstruction from next-generation sequencing data. *Bioinformatics* **28**, 132–
733 133 (2012).
734 96. Jayasundara, D. *et al.* ViQuaS: an improved reconstruction pipeline for viral
735 quasispecies spectra generated by next-generation sequencing. *Bioinformatics*
736 **31**, 886–896 (2015).
737 97. Kimbrel, J. *et al.* Multiple Mutations Associated with Emergent Variants Can Be
738 Detected as Low-Frequency Mutations in Early SARS-CoV-2 Pandemic Clinical
739 Samples. *Viruses* **14**, 2775 (2022).
740 98. Li, L., Fu, H., Ma, W. & Li, M. PySNV for complex intra-host variation detection.
741 *Bioinformatics* **40**, btae116 (2024).
742 99. O'Donnell, C. R., Wang, H. & Dunbar, W. B. Error analysis of idealized nanopore
743 sequencing. *ELECTROPHORESIS* **34**, 2137–2144 (2013).
744 100. Liu, Y. *et al.* Rescuing low frequency variants within intra-host viral populations
745 directly from Oxford Nanopore sequencing data. *Nat. Commun.* **13**, 1321 (2022).
746 101. Poplin, R. *et al.* Scaling accurate genetic variant discovery to tens of thousands of
747 samples. Preprint at https://doi.org/10.1101/201178 (2017).
748 102. Koboldt, D. C. *et al.* VarScan 2: Somatic mutation and copy number alteration
749 discovery in cancer by exome sequencing. *Genome Res.* **22**, 568–576 (2012).
750 103. Marques, A. D. *et al.* SARS-CoV-2 evolution during prolonged infection in
751 immunocompromised patients. *mBio* **15**, e00110-24 (2024).
752 104. Gao, S. J., Guo, H. & Luo, G. Omicron variant (B.1.1.529) of SARS-CoV-2, a global
753 urgent public health alert! *J. Med. Virol.* **94**, 1255–1256 (2022).

105. Goes, L. R. *et al.* Evidence of recurrent selection of mutations commonly found in SARS-CoV-2 variants of concern in viruses infecting immunocompromised patients. *Front. Microbiol.* **13**, 946549 (2022).
106. Al Khatib, H. A. *et al.* Within-Host Diversity of SARS-CoV-2 in COVID-19 Patients With Variable Disease Severities. *Front. Cell. Infect. Microbiol.* **10**, 575613 (2020).
107. Bader, W., Delerce, J., Aherfi, S., La Scola, B. & Colson, P. Quasispecies Analysis of SARS-CoV-2 of 15 Different Lineages during the First Year of the Pandemic Prompts Scratching under the Surface of Consensus Genome Sequences. *Int. J. Mol. Sci.* **23**, 15658 (2022).
108. Bendall, E. E. *et al.* Rapid transmission and tight bottlenecks constrain the evolution of highly transmissible SARS-CoV-2 variants. *Nat. Commun.* **14**, 272 (2023).
109. Bhattacharjee, M. J. *et al.* Genome analysis of SARS-CoV-2 isolates from a population reveals the rapid selective sweep of a haplotype carrying many pre-existing and new mutations. *Virol. J.* **20**, 201 (2023).
110. Lau, B. T. *et al.* Profiling SARS-CoV-2 mutation fingerprints that range from the viral pangenome to individual infection quasispecies. *Genome Med.* **13**, 62 (2021).
111. Cancela, F. *et al.* Wastewater surveillance of SARS-CoV-2 genomic populations on a country-wide scale through targeted sequencing. *PLOS ONE* **18**, e0284483 (2023).
112. Dezordi, F. Z. *et al.* Unusual SARS-CoV-2 intrahost diversity reveals lineage superinfection. *Microb. Genomics* **8**, (2022).
113. Valesano, A. L. *et al.* Temporal dynamics of SARS-CoV-2 mutation accumulation within and across infected hosts. *PLoS Pathog.* **17**, (2021).

114. Colson, P. *et al.* Resistance-associated mutations to the anti-SARS-CoV-2 agent nirmatrelvir: Selection not induction. *J. Med. Virol.* **96**, e29462 (2024).
115. Gao, R. *et al.* Quasispecies of SARS-CoV-2 revealed by single nucleotide polymorphisms (SNPs) analysis. *Virulence* **12**, 1209–1226 (2021).
116. Goldswain, H. *et al.* SARS-CoV-2 population dynamics in immunocompetent individuals in a closed transmission chain shows genomic diversity over the course of infection. *Genome Med.* **16**, 89 (2024).
117. Gonzalez-Reiche, A. S. *et al.* Sequential intrahost evolution and onward transmission of SARS-CoV-2 variants. *Nat. Commun.* **14**, 3235 (2023).
118. Goya, S. *et al.* Assessing the hidden diversity underlying consensus sequences of SARS-CoV-2 using VICOS, a novel bioinformatic pipeline for identification of mixed viral populations. *Virus Res.* **325**, 199035 (2023).
119. Graudenzi, A., Maspero, D., Angaroni, F., Piazza, R. & Ramazzotti, D. Mutational signatures and heterogeneous host response revealed via large-scale characterization of SARS-CoV-2 genomic diversity. *iScience* **24**, 102116 (2021).
120. Gruber, C. E. M. *et al.* Treatment-Emergent Cilgavimab Resistance Was Uncommon in Vaccinated Omicron BA.4/5 Outpatients. *Biomolecules* **13**, 1538 (2023).
121. Jin, B. *et al.* Investigation of the individual genetic evolution of SARS-CoV-2 in a small cluster during the rapid spread of the BF.5 lineage in Tokyo, Japan. *Front. Microbiol.* **14**, 1229234 (2023).
122. Ko, S. H. *et al.* Rapid Emergence and Evolution of SARS-CoV-2 Variants in Advanced HIV Infection. Preprint at https://doi.org/10.1101/2024.01.05.574420 (2024).

123. Lau, B. T. *et al.* Profiling SARS-CoV-2 mutation fingerprints that range from the viral pangenome to individual infection quasispecies. *Genome Med.* **13**, 62 (2021).
124. Laubscher, F. *et al.* SARS-CoV-2 Evolution among Oncological Population: In-Depth Virological Analysis of a Clinical Cohort. *Microorganisms* **9**, 2145 (2021).
125. Liu, P. *et al.* Intra-host genetic diversity of SARS-CoV-2 Omicron variants in children. *J. Infect.* **88**, 106132 (2024).
126. Ramazzotti, D. *et al.* Early detection and improved genomic surveillance of SARS-CoV-2 variants from deep sequencing data. *iScience* **25**, 104487 (2022).
127. Vellas, C. *et al.* Impact of Casirivimab-Imdevimab on Severe Acute Respiratory Syndrome Coronavirus 2 Delta Variant Nasopharyngeal Virus Load and Spike Quasispecies. *Open Forum Infect. Dis.* **9**, ofac093 (2022).

TABLE

Reference	Ct value	Genome coverage	Bioinformatic tools for LFVs detection	MAF	Q-Phred score	Minimum read depth	Sequencing technology	Sequencing protocol
Jary et al. [18]	≤30	Partial (8257 nt)	VarScan2	≥1%	Not mentionned	≥1000x	Illumina MiSeq	gene amplified
Siqueira et al.[31]	≤20	Near full length (97.9%)	Geneious R11	≥2%	≥30	≥500x	Illumina MiSeq	Amplicon-based sequencing (Artic V.3 primers)
Goes et al.[105]	≤31	Near full length	LoFreq v.2.1.5	≥2%	≥30	≥75x	Sanger	Amplicon-based sequencing
Andrés et al.[25]	Not mentionned	Partial (S gene)	in-house scripts	>0.1% on both strands	≥30	≥20x	Illumina MiSeq	S-gene amplified
Dudouet et al.[83]	≤34	Near full length	CLC Genomic workbench v.7, in-house Excel matrix	≥25%	≥7	≥150x	Oxford Nanopore MinION	Amplicon-based sequencing
Armero et al.[46]	Not mentionned	Near full length (90%)	VirVarSeq, V-Phaser2	≥5%	≥20	≥30x	Illumina HiSeq	Not mentionned
Agius et al.[47]	Not mentionned	>80%	iVar version 1.2.1	≥5%	≥20	≥ 100X	Illumina MiniSeqTM or iSeq	hybrid capture-enrichment
Al Khatib et al.[106]	Not mentionned	Near full length (mean 98%)	GATK	≥10%	≥35	≥ 100X	Illumina MiSeq	Amplicon-based sequencing
Avanzato et al.[29]	Not mentionned	Near full length	GATK, version 4.1.2	≥1%	Not mentionned	≥20X	Illumina MiSeq	Amplicon-based sequencing (Artic V.3 primers)
Bader et al.[107]	Not mentionned	Near full length (> 99%)	In-house Excel file ("QuasiS")	≥4%	Not mentionned	≥169X	Illumina MiSeq	direct RNA sequencing
Bendall et al.[108]	≤30	Near full length	iVar 1.2.1	≥2%	Not mentionned	≥ 400X	Illumina MiSeq	Amplicon-based sequencing (Artic V.3 primers)

Bendall et al.[66]	≤30	Near full length (> 97%)	iVar 1.2.1	≥5%	≥30	≥ 400X	Illumina MiSeq	Amplicon-based sequencing (Artic V.3 primers)
Bhattacharjee et al.[109]	Not mentionned	Not mentionned	GATK	≥1%	≥20	Not mentionned	Illumina Hi-SeqX	Amplicon-based sequencing
Lau et al.[110]	Not mentionned	Near full length (> 95%)	Sentieon (v201808.08)	≥1%	≥30	Not mentionned	Illumina Hi-Seq	Amplicon-based sequencing
Caccuri et al.[33]	Ct: 17.8 - 33.3	Near full length (> 90%)	Geneious	>0%	≥25	≥ 4000X	Illumina MiSeq	Amplicon-based sequencing
Cancela et al.[111]	Not mentionned	Partial (RBD region)	LoFreq	≥5%	>25	≥ 100X	Illumina iSeq	RBD region amplified
Chaguza et al.[36]	≤35	Near full length (> 70%)	bcftools (version 1.11–99-g5105724)	≥3%	Not mentionned	≥ 100X	Illumina MiSeq	Amplicon-based sequencing
Chan et al.[4]	Not mentionned	Near full length (> 80%)	iVar 1.2.1	≥5%	≥20	≥ 100X	Illumina MiSeq	Amplicon-based sequencing
Dezordi et al.[112]	Not mentionned	Near full length (> 95%)	iVar 1.2.1	≥5%	≥30	≥ 100X	Illumina	Amplicon-based sequencing
Heyer et al.[42]	≤33	Not mentionned	FreeBayes Bayesian haplotype caller v1.3.1	≥1%	≥20	≥10X	Illumina NextSeq 550	Amplicon-based sequencing
Marques et al.[103]	Not mentionned	Near full length (> 98%)	Bcftools package (v1.10.2–34)	≥1%	≥20	≥ 200X	Illumina NextSeq 550	Amplicon-based sequencing
Pérez-Lago et al.[76]	≤30	Near full length (> 96%)	iVar 1.2.3	≥10%	Not mentionned	Not mentionned	Illumina MiSeq	Amplicon-based sequencing
Quarantana et al.[34]	Not mentionned	Near full length (> 99%)	LoFreq v.2.1.5	>0%	Not mentionned	Not mentionned	Illumina	Amplicon-based sequencing
Raglow et al.[37]	≤40	Not mentionned	iVar 1.2.1	≥2%	≥35	≥10X	Illumina NextSeq 1000	Amplicon-based sequencing
Valesano et al.[113]	Not mentionned	Near full length (> 80%)	iVar 1.2.1	≥2%	≥35	≥10X	Illumina MiSeq	Amplicon-based sequencing
Voloch et al.[78]	≤35	Near full length (> 97%)	LoFreq v 2.1.5 and GATK v 4.1.2	≥5%	≥15	≥100X	Illumina MiSeq	Amplicon-based sequencing

Choga et al.[90]	Not mentionned	Near full length (> 93%)	SeqPanther	≥1%	Not mentionned	≥ 1000x	Oxford Nanopore GridION	Amplicon-based sequencing
Colson et al.[114]	Not mentionned	Near full length (> 99%)	in-house Python script, CLC genomics workbench software v7	≥4%	≥13	≥50X	Illumina MiSeq, NovaSeq 6000, Oxford Nanopore GridION	direct RNA sequencing
De Llanos et al.[71]	≤33	Near full length (> 80%)	iVar 1.2.1	>0%	≥30	≥50X	Illumina MiSeq	Amplicon-based sequencing
Farjo et al.[53]	≤26	Not mentionned	iVar (v1.3, v1.3.1)	≥3%	Not mentionned	≥1000X	Illumina NovaSeq	Amplicon-based sequencing
Fournelle et al[85].	Not mentionned	Near full length (> 90%)	iVar 1.2.1	≥5% (Illumina), ≥10% (ONT)	Not mentionned	≥50X (Illumina), ≥16X (ONT)	Oxford Nanopore, Illumina	Amplicon-based sequencing
Gao et al.[115]	Not mentionned	Not mentionned	GATK, v4.1.4.1	≥1%	≥25	≥10X	DNBSEQ-T7 platform (MGI)	metatranscriptomic sequencing
García-López et al.[59]	≤25	Near full length (> 90%)	Samtools v0.1.18	≥3%	≥20	≥20X	Illumina MiniSeq, NextSeq 500	Amplicon-based sequencing
Goldswain et al.[116]	≤40	Near full length (> 85%)	LoFreq v.2.1.5	≥5%	≥20	≥10X	Illumina NovaSeq 6000	Amplicon-based sequencing
Gonzalez-Reiche et al.[117]	≤35	Near full length (> 95%)	iVar: Lofreq (customized)	≥10%	≥20	≥200X	Illumina MiSeq	Amplicon-based sequencing
Goya et al.[118]	Not mentionned	>57%	VICOS pipeline	≥20%	≥30	≥10X	Illumina MiSeq, Illumina NextSeq 500	Amplicon-based sequencing
Graudenzi et al.[119]	Not mentionned	>75%	VarScan2	≥5%	≥20	≥20X	Illumina	Amplicon-based sequencing

Gruber et al.[120]	≤34	Not mentionned	Geneious Prime v.2019.2.3	Not mentionned	Not mentionned	Not mentionned	Ion Torrent Gene Studio S5 Prime (GSS5 Prime)	Amplicon-based sequencing
Gruber et al.[86]	≤23	Not mentionned	VarScan2	≥5%	≥20	≥50X	Ion Torrent Gene Studio S5 Prime (GSS5 Prime)	Amplicon-based sequencing
Hannon et al.[70]	≤20	Not mentionned	s, ivar, varscan2, and lofreq	≥2%	≥25	≥100X	Illumina NextSeq	direct RNA sequencing
He et al.[69]	≤35	Not mentionned	VarScan2	≥5%	≥20	≥50X	Illumina MiSeq	direct RNA sequencing
Hou et al.[84]	Not mentionned	Not mentionned	iVar v1.3.1	≥5%	≥20	≥100X	MGISEQ-2000	Amplicon-based sequencing
Jin et al.[121]	Not mentionned	Not mentionned	Torrent Suite 5.14.0, COVID19AnnotateSnpEff (v.1.3.0)	≥2,5%	≥30	≥45X	Ion GeneStudio S5 System	Amplicon-based sequencing
Karamitros et al.[21]	≤25	Near full length (> 90%)	LoFreq v 2.1.5	Not mentionned	≥30	≥40X	Illumina MiSeq	Not mentionned
Kemp et al.[39]	≤36	Not mentionned	AnCovMulti package	≥0%	≥30	≥20X	Illumina MiSeq, NextSeq, ONT MinION	Amplicon-based sequencing
Khateeb et al.[48]	Not mentionned	Not mentionned	GATK	Not mentionned	Not mentionned	Not mentionned	Illumina MiSeq	Amplicon-based sequencing
Kimbrel et al.[97]	≤28	Near full length	Mappgene pipeline	≥0%	≥20	Not mentionned	Illumina NovaSeq XP	Amplicon-based sequencing
Ko et al.[122]	≤30	Not mentionned	CLC Genomics Workbench v22.0.3	≥1%	Not mentionned	≥10X	PacBio Sequel II	Amplicon-based sequencing

Lau et al.[123]	≤30	Not mentionned	Sentieon	Not mentionned	Not mentionned	Not mentionned	Illumina	Amplicon-based sequencing
Laubscher et al.[124]	≤25	Not mentionned	in-house scripts	≥20%	Not mentionned	≥20X	Illumina HiSeq 4000	direct RNA sequencing
Lawal et al.[26]	≤35	Near full length (mean 98%)	iVar v1.3.1	≥10%	≥15	≥10X	Illumina MiniSeq	Amplicon-based sequencing
Lhomme et al.[80]	≤37	>75%	in-house scripts	Not mentionned	Not mentionned	≥10X	PacBio Sequel II, Illumina Novaseq 6000	Amplicon-based sequencing
Li et al.[51]	≤25	>50%	iSNV-calling scripts	≥5%	≥20	≥100X	Illumina NextSeq500	Hybrid capture-enrichment
Liu et al.[125]	Not mentionned	Near full length (mean 90%)	iVar 1.2.1	≥5%	≥20	≥200X	Illumina NovaSeq	Amplicon-based sequencing
Liu et al.[100]	Not mentionned	Near full length (mean 90%)	Clair version3, LoFreq (v2.1.4), Variabel	≥2%	≥15	≥100X	Oxford Nanopore MinION, Illumina	Amplicon-based sequencing
Lythgoe et al.[57]	≤22	>75%	Shiver	≥3%	Not mentionned	≥100X	Illumina MiSeq, NovaSeq, Oxford Nanopore	Hybrid capture-enrichment
Manuto et al.[38]	Not mentionned	Not mentionned	LoFreq v 2.1.5	≥1%	≥30	≥600X	Illumina NextSeq	Hybrid capture-enrichment
Marques et al.[103]	Not mentionned	Near full length (>98%)	bbtools	≥3%	≥30	Not mentionned	Illumina NextSeq	Amplicon-based sequencing
Martínez-González et al.[50]	≤27	Not mentionned	Not mentionned	≥0.1%	Not mentionned	Not mentionned	Illumina MiSeq	Amplicon-based sequencing
Naveca et al.[56]	≤25	Near full length (>97%)	ViralFlow v.0.0.6	≥5%	≥20	≥100X	Illumina MiSeq	Amplicon-based sequencing
Nirmalarajah et al.[43]	Not mentionned	>75%	Breseq v035.6	≥5%	Not mentionned	≥2X	Illumina MiniSeq	Amplicon-based sequencing
Ortiz et al.[68]	≤30	Not mentionned	Python script	≥2%	≥30	≥100X	Illumina NextSeq	Amplicon-based sequencing

Palomino Cabrera et al.[45]	≤32	Not mentionned	provaryota script	>0%	≥30	≥10X	Illumina MiSeq	Amplicon-based sequencing
Pathak et al.[65]	Not mentionned	Not mentionned	REDItools2	≥1%	≥30	≥20X	Illumina NextSeq	Amplicon-based sequencing
Pérez-Lago et al.[76]	Not mentionned	Near full length (>96%)	iVar v1.3.1	≥10%	Not mentionned	≥100X	Illumina MiSeq	Amplicon-based sequencing
Ramazzotti et al.[126]	Not mentionned	Near full length (>90%)	iVar v1.3.1	≥5%	≥20	≥100X	Illumina	Amplicon-based sequencing
Ravi et al.[63]	Not mentionned	Near full length (>90%)	Bcftools package	≥0.05%	Not mentionned	≥100X	Illumina , Oxford Nanopore	Not mentionned
Rios et al.[27]	Not mentionned	Not mentionned	iVar v1.3.1	>0%	≥20	≥100X	Oxford Nanopore PromethION	Amplicon-based sequencing
Rueca et al.[55]	≤30	Near full length	ViVOfinder	≥1%	≥20	≥10X	Ion Torrent Platform	Amplicon-based sequencing
Sapoval et al.[74]	≤32.5	Not mentionned	LoFreq v 2.1.5	≥2%	≥20	≥10X	Illumina NovaSeq 6000	Amplicon-based sequencing
Sayama et al.[60]	Not mentionned	Not mentionned	Not mentionned	≥5%	≥20	≥5X	Illumina MiniSeq , PcBio Sequel IIe	Amplicon-based sequencing
Shapira et al.[28]	Not mentionned	Near full length (>85%)	iVar v1.3.1	>0%	Not mentionned	≥20X	Illumina NestSeq	Amplicon-based sequencing
Sonnleitner et al.[40]	≤34.4	Near full length	breseq 0.35.5	>0%	Not mentionned	Not mentionned	Ion Torrent S5 Plus	Amplicon-based sequencing
Stanevich et al.[44]	Not mentionned	Not mentionned	LoFreq version 2.1.5	>0%	Not mentionned	≥4X	Illumina MiSeq	Amplicon-based sequencing
Subramoney et al.[23]	≤37	Not mentionned	ARTICPEdata(V0.5), ARTIC ONT (V0.3.1)	≥1%	Not mentionned	Not mentionned	Illumina MiSeq , Oxford Nanopore MinION	Amplicon-based sequencing
Sun et al.[22]	Not mentionned	Not mentionned	PERL script	Not mentionned	Not mentionned	Not mentionned	PacBio Sequel I	Amplicon-based sequencing

Trémeaux et al.[81]	≤19.5	length (>36%)	Juliet (v1.12.0)	≥15%	≥20	≥100X	PacBio Sequel II	Amplicon-based sequencing
Valesano et al.[113]	≤24.5	Near full length (>80%)	iVar 1.2.1	≥2%	≥35	≥200X	Illumina MiSeq	Amplicon-based sequencing
Vellas et al.[127]	≤25	Not mentionned	Not mentionned	≥0.05%	Not mentionned	Not mentionned	PacBio Sequel II	Amplicon-based sequencing
Voloch et al.[78]	≤35	Near full length	LoFreq v 2.1.5, GATK v4.1.7.0	≥5%	≥15	≥100X	Illumina MiSeq	Amplicon-based sequencing
Wertheim et al.[61]	≤32	Not mentionned	LoFreq (Galaxy)	≥5%	Not mentionned	Not mentionned	Illumina MiSeq	Amplicon-based sequencing
Xi et al.[52]	Not mentionned	Not mentionned	Freebayes v1.3.6	≥5%	≥20	≥100X	Illumina NovaSeq 6000	Amplicon-based sequencing
Xiao et al.[77]	≤32	Near full length (mean 99%)	Freebayes (v1.3.1)	≥5%	≥20	≥100X (A), ≥ 10X (M), ≥20X (C)	MGISEQ 2000	amplicon-based (A), hybrid capture-based (C), and metatranscriptomic(M) sequencing
Zhang et al.[49]	Not mentionned	Not mentionned	Annovar	≥5%	≥20	≥200X	Illumina NovaSeq 6000	Amplicon-based sequencing
Delgado et al.[64]	≤28.5	nsp12- and spike (S)-coding regions	SeekDeep	>0.1%	>30	Not mentionned	Illumina MiSeq	nsp12 and S-gene amplified

Supplementary Table: SARS-CoV-2 genome characteristics and sequencing methods in the various studied articles

LEGENDS

Figure 1: <u>Key contributions of low-frequency variants (LFVs) to SARS-CoV-2 evolution, host interactions, and epidemiological studies.</u>

Composite figure illustrating the roles of LFVs in intra-host evolution, host-pathogen interactions, and epidemiology. It shows LFVs shaping genetic diversity, recombination, and quasispecies dynamics, influenced by treatments, vaccines, and host enzymes. Tissue compartmentalization studies reveal site-specific viral evolution, while LFVs aid in reconstructing transmission chains, tracking outbreaks in confined spaces, and monitoring variants through environmental surveillance, such as wastewater analysis.

Figure 2 Overview of sequencing technologies, minimum allele frequencies, bioinformatics tools, and quality metrics for LFVs detection across 87 studies.

(a) Distribution of sequencing technologies used in 87 studies for the detection of low-frequency variants (LFVs).

(b) Proportion (%) of the different minimum allele frequency (MAF) used depending on the four majority sequencing technologies. Most variants detected fall within the 1%-5% MAF range across all platforms.

(c) Proportions of the four major bioinformatics tools employed for LFVs. (d) distribution of Q-Phred quality scores and minimum read depths employed across studies categorized by sequencing technologies.

Figure 3: Comparative workflows of the four most commonly used variant calling tools.

Summarized workflows and key features of iVar, LoFreq, GATK, and VarScan2. Each tool demonstrates distinct capabilities tailored for specific applications and shared steps.

Article 2 : Evolution intra-hôte et souches recombinantes

Avant-propos

Lors de la rédaction de cet article, début 2022, les phénomènes de recombinaison entre variants du SARS-CoV-2, bien qu'attendus d'un point de vue évolutif, demeuraient encore peu documentés dans la littérature, et concernaient majoritairement des études sur l'origine potentielle du SARS-CoV-2 [37]. Cette rareté peut ainsi s'expliquer par la faible divergence génétique entre les variants circulants au cours des premières années de la pandémie (2020-2021), ce qui limitait la détection de recombinaisons. En effet, la mise en évidence d'une recombinaison repose sur l'existence de différences substantielles entre les fragments d'origines distinctes ; or, en partageant une origine commune récente, les lignages se distinguaient alors par un nombre réduit de mutations.

Sur le plan biologique, la recombinaison nécessite une co-infection active de la même cellule par deux variants distincts. La probabilité de recombinaison augmente ainsi avec la durée de l'infection, et sont donc plus susceptibles de survenir lors d'infections persistantes, notamment chez des patients immunodéprimés, qu'un cas d'infections aigues.

Lors du séquençage et génotypage des prélèvements de routine, durant mon activité d'ingénieur hospitalier sur la plateforme bio-informatique de l'institut, le profil atypique d'une séquence, comprenant un gène spike de profil B.1.1.7 dans une séquence globale assignée B.1.160 a attiré notre attention. Nous avons alors entamé une étude rétrospective sur les prélèvements du patient, après avoir éliminé la piste de la contamination. L'article présenté ici décrit ainsi l'émergence séquentielle et l'isolement de virus recombinants issus de deux variants majeurs (B.1.160 et Alpha (B.1.160)), chez un patient atteint d'un lymphome, ayant présenté une infection prolongée de 440 jours. Une étude approfondie a été menée à partir de prélèvements nasopharyngés archivés. Nous avons eu recours aux technologies de séquençage Illumina

et Nanopore, nous avons complété les analyses par la réalisation de PCR ciblant les points de cassure génomiques putatifs, nous avons mis les prélèvements en culture cellulaire, ces souches isolées cultivées et leur séquençage ont permis de confirmer la nature biologique des recombinants, validant ainsi la survenue de la recombinaison *in vivo*.

Lors de ce travail, j'ai pu contribuer à la conceptualisation de la méthodologie pour confirmer les évènements de recombinaison, mener les analyses bio-informatiques, designer les amorces encadrantes, interpréter les résultats, et corédiger l'article. Ce travail a donné lieu à une publication dans la revue *Viruses* en 2022[38], acceptée le 8 juin 2022. La version présentée dans cette thèse correspond au manuscrit publié.

viruses

MDPI

Article

Sequential Appearance and Isolation of a SARS-CoV-2 Recombinant between Two Major SARS-CoV-2 Variants in a Chronically Infected Immunocompromised Patient

Emilie Burel [1,2,†], Philippe Colson [1,2,3,†], Jean-Christophe Lagier [1,2,3], Anthony Levasseur [1,2], Marielle Bedotto [1], Philippe Lavrard-Meyer [1,2,3], Pierre-Edouard Fournier [1,2,4], Bernard La Scola [1,2,3,*] and Didier Raoult [1,2,*]

1 IHU Méditerranée Infection, 19-21 Boulevard Jean Moulin, 13005 Marseille, France; burel.emilie@hotmail.fr (E.B.); philippe.colson@univ-amu.fr (P.C.); jean-christophe.lagier@univ-amu.fr (J.-C.L.); anthony.levasseur@univ-amu.fr (A.L.); marielle.bedotto@gmail.com (M.B.); philippe.lavrard-meyer@ap-hm.fr (P.L.-M.); pierre-edouard.fournier@univ-amu.fr (P.-E.F.)
2 Microbes Evolution Phylogeny and Infections (MEPHI), Institut de Recherche Pour le Développement (IRD), Aix-Marseille University, 27 Boulevard Jean Moulin, 13005 Marseille, France
3 Assistance Publique-Hôpitaux de Marseille (AP-HM), 264 rue Saint-Pierre, 13005 Marseille, France
4 Vecteurs-Infections Tropicales et Méditerranéennes (VITROME), Institut de Recherche Pour le Développement (IRD), Aix-Marseille University, 27 Boulevard Jean Moulin, 13005 Marseille, France
* Correspondence: bernard.la-scola@univ-amu.fr (B.L.S.); didier.raoult@univ-amu.fr (D.R.); Tel.: +33-413-732-401 (B.L.S. & D.R.)
† These authors contributed equally to this work.

check for updates

Citation: Burel, E.; Colson, P.; Lagier, J.-C.; Levasseur, A.; Bedotto, M.; Lavrard-Meyer, P.; Fournier, P.-E.; La Scola, B.; Raoult, D. Sequential Appearance and Isolation of a SARS-CoV-2 Recombinant between Two Major SARS-CoV-2 Variants in a Chronically Infected Immunocompromised Patient. *Viruses* **2022**, *14*, 1266. https://doi.org/10.3390/v14061266

Academic Editors: Judd F. Hultquist, Ramón Lorenzo-Redondo and Egon Anderson Ozer

Received: 21 March 2022
Accepted: 8 June 2022
Published: 10 June 2022

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Copyright: © 2022 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

Abstract: Genetic recombination is a major evolutionary mechanism among RNA viruses, and it is common in coronaviruses, including those infecting humans. A few SARS-CoV-2 recombinants have been reported to date whose genome harbored combinations of mutations from different mutants or variants, but only a single patient's sample was analyzed, and the virus was not isolated. Here, we report the gradual emergence of a hybrid genome of B.1.160 and Alpha variants in a lymphoma patient chronically infected for 14 months, and we isolated the recombinant virus. The hybrid genome was obtained by next-generation sequencing, and the recombination sites were confirmed by PCR. This consisted of a parental B.1.160 backbone interspersed with two fragments, including the spike gene, from an Alpha variant. An analysis of seven sequential samples from the patient decoded the recombination steps, including the initial infection with a B.1.160 variant, then a concurrent infection with this variant and an Alpha variant, the generation of hybrid genomes, and eventually the emergence of a predominant recombinant virus isolated at the end of the patient's follow-up. This case exemplifies the recombination process of SARS-CoV-2 in real life, and it calls for intensifying the genomic surveillance in patients coinfected with different SARS-CoV-2 variants, and more generally with several RNA viruses, as this may lead to the appearance of new viruses.

Keywords: SARS-CoV-2; variant; recombination; chronic infection; immunosuppression

1. Introduction

A major evolutionary mechanism of RNA viruses is genetic recombination [1,2]. Recombinations are extremely common in coronaviruses and have been implicated in the emergence of several genotypes, including endemic human coronaviruses [3–6]. The involvement of genetic recombination in the origin of SARS-CoV-2 is also suspected [7]. Regarding SARS-CoV-2, coinfection in the same patient with distinct variants has been reported [8–14]. In addition, several studies have described or suspected genetic recombinations for this virus [10,13–25]. However, most of these recombinants have relied solely on the coexistence of signature mutations of different SARS-CoV-2 variants in genomes obtained from a single patient's sample, and they were not isolated in culture. Since

January 2020, our laboratory has screened more than one million respiratory specimens for SARS-CoV-2 infection by real-time reverse transcription-PCR (qPCR) without interruption or limited capacity, including for all patients sampled in our institute and in the Marseille public hospitals [26,27]. This has provided us with sequential samples from multiple patients, and enabled us to detect reinfections, and prolonged or even chronic infections in severely immunocompromised patients [28–30]. Here, we report the gradual emergence of a recombinant SARS-CoV-2 involving two variants in a lymphoma patient chronically infected over a period of 14 months, and the isolation of the recombinant virus in culture.

2. Results

2.1. Chronic SARS-CoV-2-Infection in a Severely Immunocompromised Patient

A 56-year-old immunocompromised male had an uncontrolled SARS-CoV-2 infection for 14 months until death (Supplementary Material: Supplementary Methods and Results). He was diagnosed in 2017 with mixed Hodgkin and follicular lymphoma and was in complete remission of the Hodgkin lymphoma following chemotherapy, and on maintenance therapy. Follicular lymphoma progression led to the administration of lenalidomide plus rituximab in 2021, and pembrolizumab was also administered. Progressive multifocal leukoencephalitis was diagnosed in May 2021. In August 2020, the patient developed severe SARS-CoV-2-associated pneumonia, leading to admission in the intensive care unit. He improved clinically but viral clearance did not occur, and SARS-CoV-2 RNA remained detectable by qPCR on most nasopharyngeal samples collected between September 2020 and December 2021. qPCR was negative in February 2021 but positive when re-tested in April 2021, and then only transiently negative for $\leq$3 days. COVID-19 patient convalescent plasma was administered during hospitalization outside our institute in October 2020 and April 2021, and hydroxychloroquine was administered in July 2021. The history of SARS-CoV-2 infection did not lead to vaccine administration. The patient died unfortunately in December 2021 from the complications of his hematological and neurological diseases.

2.2. Evidence of Hybrids of Variants

After 14 months of infection, we identified a virus whose genome was a hybrid of two known variants, B.1.160 (according to PANGOLIN (Phylogenetic Assignment of Named Global Outbreak Lineages) lineage (https://cov-lineages.org/resources/pangolin.html; accessed on 20 October 2021) [31]) (a.k.a. Nextstrain clade (https://nextstrain.org/; accessed on 20 October 2021) [32] 20A.EU2, or Marseille-4 [27]) and Alpha (according to the WHO denomination (https://www.who.int/fr/activities/tracking-SARS-CoV-2-variants; accessed on 20 October 2021) (a.k.a. 20I or B.1.1.7)) (Figures 1 and 2).

This hybrid genome was obtained from the respiratory samples by next-generation sequencing, as previously described [27]. In addition, the hybrid virus was isolated in culture, as previously described [33]. The hybrid genome sequence consisted of a B.1.160 variant matrix, of which two regions, the first one being located at the 5′ tip of the genome and containing the synonymous mutation C913U, and the second one spanning from positions 17,109–18,877 to positions 25,710–27,972, were replaced by those of an Alpha variant (Figures 1 and 2; Supplementary Material: Figures S1 and S2). All eight signature mutations of the Alpha variant were detected in the spike gene in the absence of the S477N mutation that is a signature of the B.1.160 variant. Nucleotide diversity at the 35 positions harboring signature mutations of the Alpha or B.1.160 variants was low (mean ($\pm$standard deviation) value of 3.1 $\pm$ 6.8%) (Figure 2; Supplementary Material: Figures S1 and S2), indicating that the hybrid content of the genome was not explained by a co-infection from the two variants or by contamination. These findings indicated that this mosaic genome was the result of recombinations between parental genomes of the B.1.160 and Alpha variants.

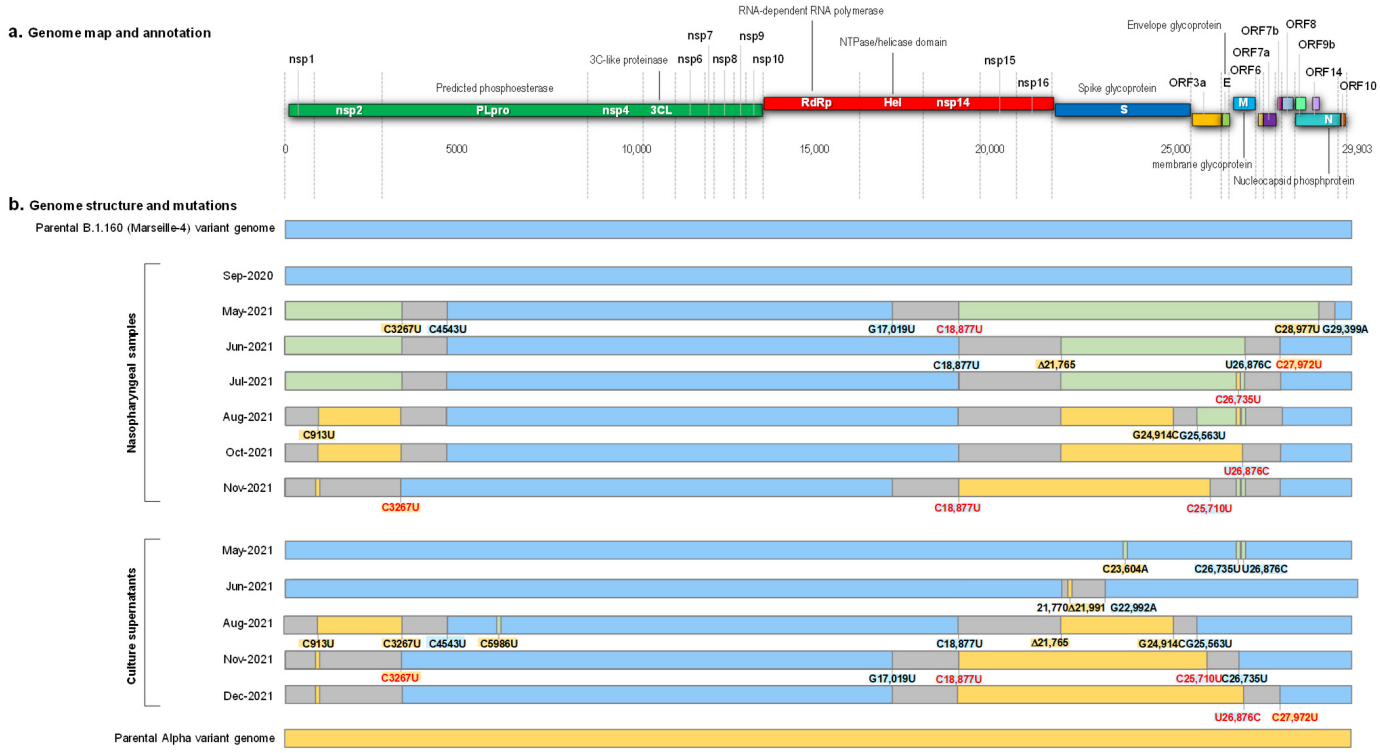


Figure 1. Schematic representation of the structure of the SARS-CoV-2 genomes obtained from the nasopharyngeal samples and from the culture supernatants, as well as of the recombination events over time, in reference to parental genomes of the B.1.160 and Alpha variants. (**a**) Genome map and annotation; (**b**) Genome structure and mutations. Blue color of rectangles indicates sequences from a B.1.160 variant; yellow color indicates sequences from an Alpha variant; green color indicates co-detection of sequences from a B.1.160 variant and from an Alpha variant; grey color indicates sequences from indeterminate origin. Signature mutations from the B.1.160 and Alpha variants are indicated by a blue background and a yellow background, respectively. Signature mutations that are absent are indicated by a red font. Δ21,765: -6 nucleotides; Δ21,991: -3 nucleotides. Nsp, nonstructural protein; ORF, open reading frame.

By analyzing the sequential samples available from this patient, we were able to determine that he was first infected with the B.1.160 variant, which was epidemic at the time of diagnosis of his infection in September 2020. This variant predominated in our region from August 2020 until January 2021 and was replaced by the Alpha variant, which emerged in December 2020 [27]. SARS-CoV-2 could not be isolated retrospectively from this sample, but its genome was typical of a B.1.160 variant and displayed no significant nucleotide diversity (mean, 0.2 ± 0.5%). It was classified by phylogeny as a B.1.160 variant (Figure 3).

Unfortunately, although SARS-CoV-2 qPCR was still positive in another laboratory in January 2021, the sample was unavailable. Therefore, we were not able to confirm whether Marseille-4 and Alpha variants were co-infecting the patient at this time nor the duration of co-infection, and we did not obtain the genome nor an isolate of the Alpha variant. The closest sample in time to the initial one dated from May 2021, 8 months after the first SARS-CoV-2 diagnosis, and it already demonstrated a mosaicism between genomes of the B.1.160 and Alpha variants (Figures 1 and 2).

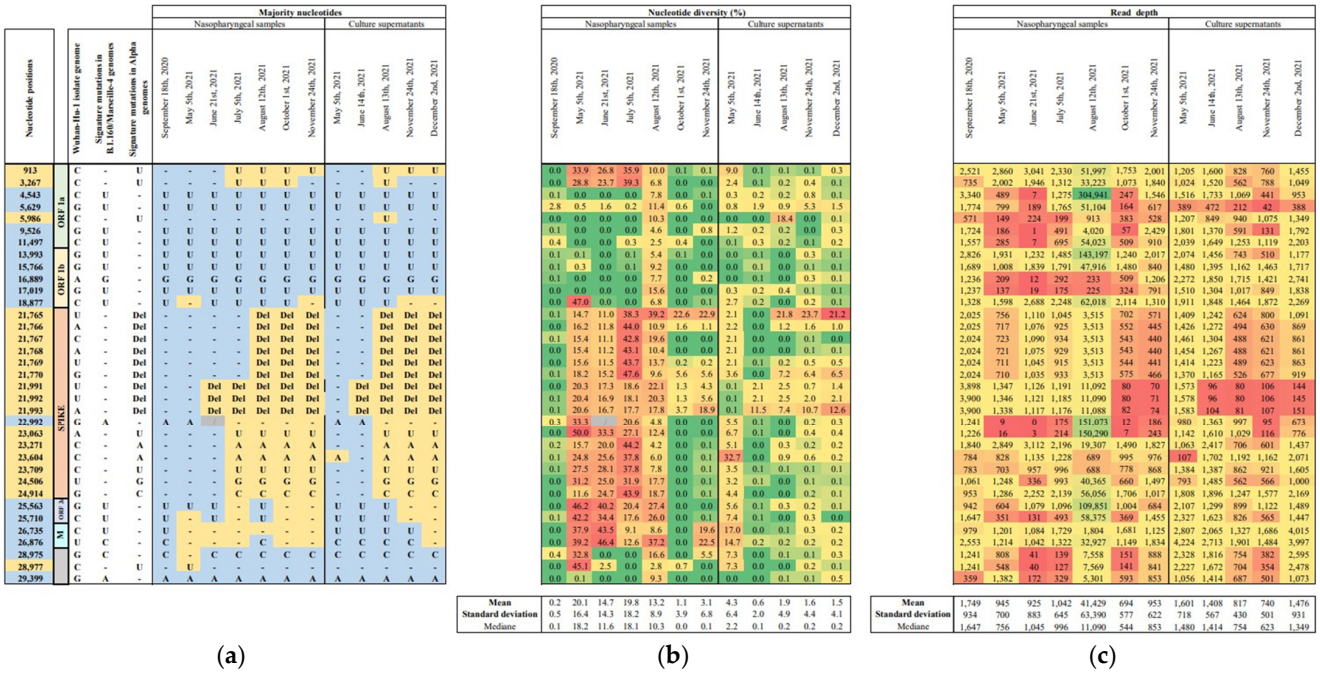


Figure 2. Majority nucleotides (**a**) nucleotide diversity (**b**) and sequencing depth (**c**) for sequences obtained from the respiratory samples and the culture supernatant at nucleotide positions of the SARS-CoV-2 genome that harbor signature mutations of the B.1.160 or Alpha variants. Del, nucleotide deletion. Nucleotide positions are in reference to the genome of the Wuhan-Hu-1 isolate GenBank accession no. NC_045512.2. (**b**) Nucleotide diversity is the proportion of sequence reads that do not harbor the consensus (majority) nucleotide. (**c**) Read depth is the number of reads covering a given nucleotide position.

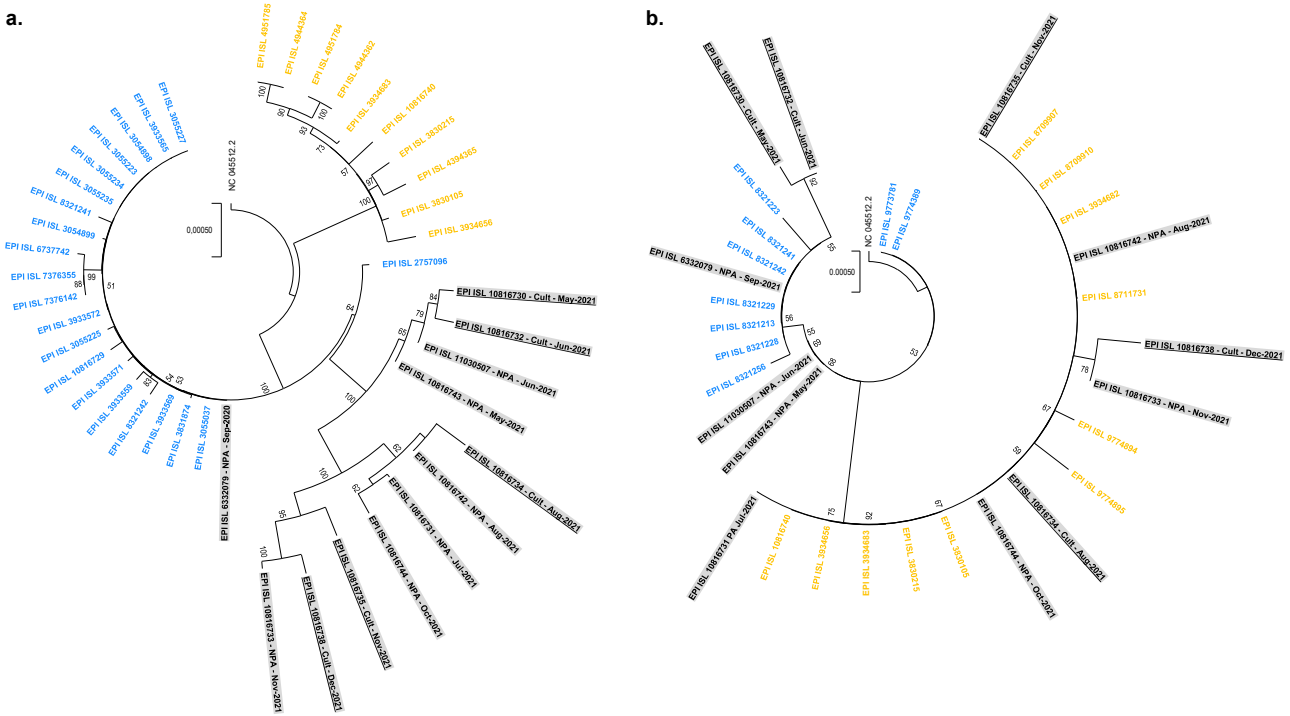


Figure 3. Phylogenetic analyses based on SARS-CoV-2 genomes (**a**) and spike gene sequences (**b**). Sequences obtained from the case-patient are indicated by a grey background, and those obtained

from cultures are underlined. Other sequences from our SARS-CoV-2 sequence database are indicated by a blue font when classified as of the B.1.160 variant, and by a yellow font when classified as of the Alpha variant. Sequences are labeled with their GISAID (https://www.gisaid.org/; accessed on 18 May 2022) [34] identifiers. Trees are rooted with the genome of the Wuhan-Hu-1 isolate GenBank accession no. NC_045512.2.2.3. Genome of the initial virus.

2.3. Steps in Generation of the Recombinant

We used three procedures to characterize the different recombination steps by analyzing seven sequential respiratory samples collected from the patient (Tables 1 and 2). First, through sequencing from the respiratory samples of the viral genomes; second, sequencing from the respiratory samples of PCR products overlapping the putative recombination sites; third, a viral culture with sequencing of the genomes of the isolates. These approaches allowed us to evidence that several viruses and recombinant forms had coexisted in the sequential samples, as signature mutations of the two variants were co-detected at multiple positions, with a nucleotide diversity that reached high levels and that evolved over time (Figure 2; Supplementary Material: Supplementary Results and Figure S1). We observed an evolution towards the genome sequence of the recombinant virus that predominated at the end of the patient's follow-up, following recombination events at three sites between parental genomes of B.1.160 and Alpha variants, with a low level of nucleotide diversity observed at that time at the positions harboring signature mutations of these variants.

Table 1. Genome sequences obtained from the sequential nasopharyngeal samples of the case-patient.

GISAID Identifier	Sampling Date	Time from Diagnosis (Days)	Next-Generation Sequencing Technology, Instrument
EPI_ISL_6332079	18 September 2020	0	Illumina, NovaSeq
EPI_ISL_10816743	5 May 2021	229	Illumina, NovaSeq
EPI_ISL_11030507	21 June 2021	276	Illumina, NovaSeq
EPI_ISL_10816731	5 July 2021	290	Illumina, NovaSeq
EPI_ISL_10816742	12 August 2021	328	Nanopore, GridION
EPI_ISL_10816744	1 October 2021	378	Illumina, NovaSeq
EPI_ISL_10816733	24 November 2021	432	Illumina, NovaSeq

See also Supplementary Table S3a.

Table 2. Genome sequences obtained from the culture supernatants.

GISAID Identifier	Sampling Date of the Nasopharyngeal Sample	Time to Cytopathic Effect (Days)	Next-Generation Sequencing Technology, Instrument
EPI_ISL_10816730	5 May 2021	8	Illumina, NovaSeq
EPI_ISL_10816732	14 June 2021	4	Illumina, NovaSeq
EPI_ISL_10816734	12 August 2021	4	Illumina, NovaSeq
EPI_ISL_10816735	24 November 2021	5	Illumina, NovaSeq
EPI_ISL_10816738	2 December 2021	7	Illumina, NovaSeq

See also Supplementary Table S3b.

To further support the breakpoints identified between the parental genomes of the Alpha and Marseille-4 variants, we generated Marseille-4 variant/Alpha variant chimeric sequences of the regions overlapping the three putative recombination sites by PCR amplification (Supplementary Material: Figures S3–S5), followed by next-generation sequencing with Nanopore Technology on a gridION instrument, as previously described [27]. For the amplicon corresponding to positions 3100–4570 of the genome, 56 and 87% of the reads were chimeras harboring both Alpha (C3267U) and Marseille-4 (C4543U) signature mutations from the respiratory samples collected in 5 May and in 12 August 2021, respectively (Supplementary Material: Figures S3 and S4). For the amplicon corresponding to positions 24,880–29,010, 70% of the reads were chimeras harboring the Alpha mutation G24914C and the Marseille-4 mutations G25563U, C25710U, C26735U, U26876C and G28975C, and 24% were chimeras harboring the Alpha mutation G24914C and the Marseille-4 mutation

G28975C (Supplementary Material: Figure S5). No PCR amplification was obtained for the region between positions 21,417–23,901. Next-generation sequencing was also carried out using a metagenomic approach with the Nanopore technology to obtain large reads and detect additional Marseille-4 variant/Alpha variant chimeras. For the first previously identified recombination region, four reads with a length ranging between 1855 and 11,979 nucleotides were obtained that harbored signature mutations of the Alpha (C3267U) and Marseille-4 (C4543U) variants. For the third recombination region, four reads with a length between 1486–8143 nucleotides were obtained that harbored signature mutations of the Alpha (G24914C) and Marseille-4 (G25563U) variants (available from URL: https://www.mediterranee-infection.com/sars-cov-2-recombinant/; accessed on 18 May 2022). No reads were obtained that covered the Marseille-4 mutation C18877U and the Alpha deletion UACAUG21765. Therefore, several breakpoints between the parental genomes of Alpha and Marseille-4 variants were supported by the presence of chimeric reads.

3. Discussion

We highlight here, in an immunocompromised lymphoma patient chronically infected with SARS-CoV-2 and who received several treatments, the presence of a virus hybrid of two known variants, B.1.160 and Alpha, which successively predominated in our region during the follow-up period of this patient [27,35]. The absence of available samples covering the period between the diagnosis of infection by the B.1.160 variant and first evidence of a hybrid genome did not allow us to date the superinfection by the Alpha variant. The signature mutations of the Alpha variant observed in the hybrid genomes between 8 and 14 months cannot have occurred randomly considering their number, distribution along the genome, and their majority presence, and their location indicates recombinations in three regions. In addition, genomic analyses carried out for sequential respiratory samples and viral cultures demonstrate the successive presence of several viruses with hybrid genomes, one of them having established itself in this patient and the one that continued circulating until his death.

We believe that this observation, which sheds light on the recombination mechanism of RNA viruses, is significant, and to our knowledge, this is the first study describing, through the analysis of sequential samples over more than a year, the generation of recombinant SARS-CoV-2 and its isolation in culture. Sixteen interlineage recombinants between the Alpha variant and non-Alpha viruses were reported in 2021 in the UK, of 279,000 genomes analyzed [8]. In addition, 1175 (0.2%) putative recombinant genomes were identified among 537,360 genomes, and it was reported that up to 5% of SARS-CoV-2 that circulated in the USA and UK might be recombinants [18]. Moreover, the number of cases that capture detection of recombinant genomes is growing [10,13–25], including with recombinant events involving or between Omicron variants [36–39], which highlights the importance of recombination in the evolution of SARS-CoV-2. Besides recombination between SARS-CoV-2 infecting the same human cells, other evolutionary pathways may exist [39]. For instance, different evolutionary trajectories in distinct cell types of the same infected host have been reported [40], and coronaviruses have been reported to harbor a sequence shared with four different families of positive-sense single-stranded RNA viruses and that is putatively shared with insects [41].

Such natural mosaicisms in these viruses make it possible to understand the emergence of RNA viruses and should lead to a strengthening of genomic surveillance in patients, especially in immunocompromised long-term viral carriers, presenting with coinfections by several RNA viruses, as observed in patients infected with several respiratory viruses including SARS-CoV-2, endemic human coronaviruses, influenza viruses, or rhinoviruses [42]. Such infectious episodes could perhaps lead to the emergence of new emerging viruses, as has been, for instance, reported for enteroviruses of humans and great apes [43].

4. Materials and Methods

4.1. SARS-CoV-2 Genome Sequencing

SARS-CoV-2 genome sequencing was performed as previously described. Briefly, viral RNA was extracted from 200 µL of nasopharyngeal swab fluid using the EZ1 Virus Mini kit v2.0 on an EZ1 Advanced XL instrument (Qiagen, Courtaboeuf, France) or using the MagMax Viral/Pathogen Nucleic Acid Isolation kit on the KingFisher Flex system (Thermo Fisher Scientific, Waltham, MA, USA), following the manufacturer's instructions. SARS-CoV-2 genome sequences were obtained by next-generation sequencing with various procedures with the Illumina COVIDSeq protocol on a NovaSeq 6000 instrument (Illumina Inc., San Diego, CA, USA), or by multiplex PCR with ARTIC nCoV-2019 V3 Panel primers (IDT, Coralville, IA, USA) were combined with the Oxford Nanopore technology (ONT) on a GridION instrument (Oxford Nanopore Technologies Ltd., Oxford, UK), as previously described [14,27]. After its extraction, viral RNA was reverse-transcribed according to the COVIDSeq protocol (Illumina Inc., San Diego, CA, USA) or by using the LunaScript RT SuperMix kit (New England Biolabs, Ipswich, MA, USA) when performing next-generation sequencing with the Nanopore technology, following the manufacturer's recommendations.

4.2. Genome Analysis

After using the Nanopore technology combined with the ARTIC protocol, fastq files were processed using the ARTIC field bioinformatics pipeline (ARTIC-nCoV-bioinformatics-SOP-v1.1.0, Nick Loman, Will Rowe, Andrew Rambaut, the ZiBRA Project and the ARTIC project, University of Birmingham, UK, https://artic.network/ncov-2019/ncov2019-bioinformatics-sop.html; https://github.com/artic-network/fieldbioinformatics; accessed on 18 May 2022), as previously described [14,27]. Next-generation sequencing reads were base-called using Guppy (4.0.14) and were aligned to the genome of the Wuhan-Hu-1isolate, GenBank accession No. MN908947.3, using minimap2 (v2.17-r941) (https://github.com/lh3/minimap2; accessed on 18 May 2022, Dana-Farber Cancer Institute of Boston, USA) [44]. The ARTIC tool align_trim was used to softmask primers from read alignment and cap sequencing depth at a maximum of 400-fold coverage. Consensus-level variant candidates were identified using a threshold of 70% and the Medaka (v.0.11.5) workflow, developed by ARTIC (Nick Loman, Will Rowe, Andrew Rambaut, the ZiBRA Project and the ARTIC project, University of Birmingham, UK, https://github.com/artic-network/artic-ncov2019; accessed on 18 May 2022). From the unique sequence obtained with the ARTIC-Nanopore technology, a sorted bam file was loaded on the CLC Genomics workbench v7 software and a tsv file was then exported. NovaSeq reads were basecalled using the Dragen Bcl Convert pipeline (v3.9.3; https://emea.support.illumina.com/sequencing/sequencing_software/bcl-convert.html; accessed on 18 May 2022 (Illumina Inc., San Diego, CA, USA). Mapping was carried out on the Wuhan-Hu-1 isolate genome with the bwa-mem2 tool (v2.2.1; https://github.com/bwa-mem2/bwa-mem2; accessed on 18 May 2022, Heng Li, Harvard, USA) and was cleaned with Samtools (v1.13; https://www.htslib.org/; accessed on 18 May 2022) [45]. Variant calling was performed with freebayes (v1.3.5; https://github.com/freebayes/freebayes; accessed on 18 May 2022, Erik Garrison, Gabor Marth, University of Tennessee and Utah) [46] and consensus genomes were built with Bcftools (v1.13; https://samtools.github.io/bcftools/bcftools.html; accessed on 18 May 2022, 2012–2021 Genome Research Ltd.). Freebayes results were filtered with a threshold of 70% for the majority nucleotide. A tsv file was generated using an in-house Python script. The clade was designated at the consensus level with the Nextclade online tool (https://clades.nextstrain.org/; accessed on 18 May 2022, 2020 Nextstrain developers, Swiss Institute of Bioinformatics) [32,47] and an in-house Python script allowed for the detection of variants and the hybrids of variants. At the sub-consensus level, variant frequencies that were compared to the Wuhan-Hu-1 isolate genome were calculated from tsv files. Nucleotide diversity at genomic positions was calculated using the Microsoft Excel software (https://www.microsoft.com/en-us/microsoft-365/excel; accessed on 18 May 2022, Microsoft Corporation, Redmond, WA, USA) with an in-house built

file. It corresponded to the proportion of sequence reads that do not harbor the consensus (majority) nucleotide. Genome sequences obtained in the present study were submitted to the GISAID sequence database (https://www.gisaid.org/; accessed on 18 May 2022, 2008–2022, Freunde von GISAID e.V. Munich, Germany) [34] (see Supplementary Material: Supplementary Table S1).

4.3. Generation of Additional Sequence Reads

Sequencing of reverse-transcription-PCR-targeted regions: extraction of the RNA samples was carried out using the EZ1 Virus kit with the EZ1 Advanced XL instrument (Qiagen) following the manufacturer's recommendations. PCR amplification of the 3 regions was amplified in a 25 µL total volume using the SuperScript III One-Step RT-PCR Kit (Invitrogen, Carlsbad, CA, USA), using primer concentrations of 200 nM per reaction. PCR were performed with following conditions: 50 °C for 25 min, 95 °C for 2 min, then 40 cycles including 15 s at 95 °C, 45 s at 60 °C, and 2 min at 68 °C. Sequences of PCR primers are provided in the Supplementary Material. Amplicons were sequenced with Nanopore technology on a GridION instrument (Oxford Nanopore Technologies Ltd., Oxford, UK), following the manufacturer's instructions. Fastq files were processed as described above. Continuous reads overlapping signature mutations of distinct variants were filtered using SAMtools (v1.13; https://github.com/samtools/2008-2022 Genome Research Ltd., accessed on 18 May 2022) [45] combined with an in-house awk script. Reads were then filtered according to variant-specific nucleotide patterns using SAMtools combined with an in-house awk script. Groups of reads with same patterns of mutations were then visualized using the IGV software (https://software.broadinstitute.org/software/igv/; accessed on 18 May 2022, Broad Institute and the Regents of the University of California, USA) [48].

Metagenomic sequencing: nucleic acid extraction was performed with the EZ1 Virus kit with the EZ1 Advanced XL instrument (Qiagen) following the manufacturer's recommendations, using 200 µL of sample and eluting in 60 µL of elution buffer. Reverse transcription was performed with all 60 µL of this solution using the TaqMan Reverse Transcription Reagent kit (Applied-Biosystems, Foster City, CA, USA), according to the manufacturer's protocol under the following conditions: 10 min at 25 °C, 30 min at 48 °C, and 5 min at 95 °C. Then, 300 µL of obtained cDNA was transferred to a 1.5 mL Eppendorf LoBind tube (Eppendorf, Le Pecq, France). Second DNA strand was synthesized by adding a mix of 24 µL of Klenow Fragment DNA polymerase (New England Biolabs, Beverly, MA, USA), 66 µL of nuclease-free water, 45 µL of NEB Buffer 2 (New England Biolabs, Ipswich, MA, USA), and 15 µL of dNTPs working solution produced with 10 µL of each dNTP at a 100 mM concentration, and 60 µL of nuclease-free water (New England Biolabs, Ipswich, MA, USA). This mix was kept at 37 °C for one hour. A purification step consisted of adding 450 µL of magnetic CleanNGS beads for a 1:1 volume ratio (CleanNA, Waddinxveen, the Netherlands) then incubating them for 5 min in a magnetic support, and washing with 1000 µL of ethanol at 70%, before elution of the beads was performed in 50 µL of Tris EDTA 1X with centrifugation for 10 min at 300 rpm at room temperature. Subsequently, a DNA library was prepared with the DNA ligation sequencing kit SQK-LSK109 (Oxford Nanopore Technologies Ltd., Oxford, UK), and next-generation sequencing was performed with the Nanopore technology on a PromethION instrument (Oxford Nanopore Technologies Ltd., Oxford, UK). Each sample was sequenced on a different PromethION Flow cell R10.4 (Oxford Nanopore Technologies Ltd., Oxford, UK).

4.4. Phylogenetic Analysis Based on Whole and Partial Genome Sequences

Phylogenetic analyses were performed separately for the twelve genome sequences and the twelve spike gene sequences obtained from the nasopharyngeal samples or the culture supernatants. Sequences were aligned using MAFFT v.7 (Kazutaka Katoh and Daron M Standley, Osaka University, Japan) [49] with their 20 most similar hits identified with the BLAST tool (National Library of Medicine, Rockville Pike, MD, USA) [50] among SARS-CoV-2 genomes from our database, which contains sequences obtained

from clinical samples collected between February 2020 and February 2022 [14,27]. Phylogeny reconstruction was performed using the IQ-TREE software with the GTR Model and 1000 ultrafast bootstrap repetitions (http://www.iqtree.org, Free Software Foundation, Boston, MA, USA) [51], and trees were visualized with iTOL (Interactive Tree of Life) (https://itol.embl.de/; accessed on 18 May 2022, EMBL 2022) [52] and MEGA X (v10.2.6; https://www.megasoftware.net/; accessed on 18 May 2022, 2007 Free Software Foundation) [53] softwares.

4.5. Virus Culture Isolation

Culture isolation was performed on Vero E6 cells, as previously described [33]. Briefly, 500 µL of nasopharyngeal swab fluid was passed through a centrifugal filter with a pore size of 0.22-µm (Merck Millipore, Darmstadt, Germany) before the inoculation of 100 µL of filtrate was performed in 4 wells of culture microplates with 96 wells that contained Vero E6 cells (ATCC CRL-1586) in Minimum Essential Medium culture medium, comprising 4% fetal calf serum and 1% glutamine. After a centrifugation step at 4000× *g*, culture microplates were incubated at 37 °C and were observed daily with an inverted microscope for evidence of cytopathogenic effect. We attempted, without success, to isolate the parental Alpha variant from respiratory samples collected in 2021 from the case-patient by selecting this virus through inoculating the respiratory samples with neutralizing serum from a convalescent individual previously infected by the B.1.160 variant, as previously reported [54], at equal volumes (50 µL of filtered respiratory sample and 50 µL of serum).

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/v14061266/s1. Additional Methods, Results, Figure S1: Virological follow-up by real-time reverse transcription PCR (qPCR) targeting SARS-CoV-2 RNA for the patient, Figure S2: Majority nucleotides and nucleotide diversity for sequences obtained from the respiratory samples and the culture supernatant at nucleotide positions of the SARS-CoV-2 genome that harbor signature mutations of the B.1.160 or Alpha variants and at any other positions that harbor mutations (a), and sequencing depth at nucleotide positions (b), Figure S3: Detailed content of amplicons corresponding to positions 3100–4570 of the SARS-CoV-2 genome and retrieved from respiratory samples collected on 5 May 2021, Figure S4: Detailed content of amplicons corresponding to positions 3100–4570 of the SARS-CoV-2 genome and retrieved from respiratory samples collected on 12 August 2021, Figure S5: Detailed content of amplicons corresponding to positions 24,880–29,010 of the SARS-CoV-2 genome and retrieved from respiratory samples collected on 12 August 2021, Table S1: List of GISAID identifiers for sequences used in the present study, and Table S2: Sampling time and location for sequences used in the present study, Table S3: Identifiers of raw data deposited in the European Bioinformatics Institute (EMBL-EBI) sequence database (https://www.ebi.ac.uk/; accessed on 18 May 2022), and References, are provided in the Supplementary Material.

Author Contributions: P.C., B.L.S. and D.R. conceived the project. E.B., P.C., J.-C.L., A.L., M.B. and P.L.-M. provided materials or performed analyses. E.B., P.C., J.-C.L., A.L., M.B., P.-E.F., B.L.S. and D.R. analyzed the data. E.B., P.C., P.-E.F., B.L.S. and D.R. drafted the paper. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the French Government under the "Investments for the Future" program managed by the National Agency for Research (ANR) (Méditerranée-Infection 10-IAHU-03), by the Région Provence Alpes Côte d'Azur and European funding FEDER PRIMMI (Fonds Européen de Développement Régional-Plateformes de Recherche et d'Innovation Mutualisées Méditerranée Infection) (FEDER PA 0000320 PRIMMI), and by the French Ministry of Higher Education, Research and Innovation (ministère de l'Enseignement supérieur, de la Recherche et de l'Innovation) and the French Ministry of Solidarity and Health (Ministère des Solidarités et de la Santé).

Institutional Review Board Statement: This study has been approved by the ethics committee of University Hospital Institute (IHU) Méditerranée Infection (No. 2022-008). Access to the patients' biological and registry data issued from the hospital information system was approved by the data protection committee of Assistance Publique-Hôpitaux de Marseille (APHM) and was recorded in the European General Data Protection Regulation registry under number RGPD/APHM 2019-73.

Informed Consent Statement: This study has been approved by the ethics committee of University Hospital Institute (IHU) Méditerranée Infection (No. 2022-008). Access to the patients' biological and registry data issued from the hospital information system was approved by the data protection committee of Assistance Publique-Hôpitaux de Marseille (APHM) and was recorded in the European General Data Protection Regulation registry under number RGPD/APHM 2019-73.

Data Availability Statement: The dataset generated and analyzed during the current study is available in the GISAID sequence database (https://www.gisaid.org/; accessed on 18 May 2022, 2008–2022, Freunde von GISAID e.V. Munich, Germany) [34].

Acknowledgments: We are thankful to Ludivine Brechard, Claudia Andrieu, Raphael Tola, Jeremy Delerce for their technical help; to Sophie Amrane and Justine Raclot for their help with the clinical management; and to the medical biology laboratory Alphabio, hôpital Européen, Marseille, France, for providing some results of SARS-CoV-2 qPCR testing.

Conflicts of Interest: The authors have no conflict of interest to declare relative to the present study. Didier Raoult was a consultant for the Hitachi High-Technologies Corporation, Tokyo, Japan from 2018 to 2020. He is a scientific board member of the Eurofins company and a founder of a microbial culture company (Culture Top). Funding sources had no role in the design and conduct of the study, the collection, management, analysis, and interpretation of the data, and the preparation, review, or approval of the manuscript.

References

1. Xiao, Y.; Rouzine, I.M.; Bianco, S.; Acevedo, A.; Goldstein, E.F.; Farkov, M.; Andino, R. RNA Recombination Enhances Adaptability and Is Required for Virus Spread and Virulence. *Cell Host Microbe* **2017**, *22*, 420. [CrossRef] [PubMed]
2. Bentley, K.; Evans, D.J. Mechanisms and consequences of positive-strand RNA virus recombination. *J. Gen. Virol.* **2018**, *99*, 1345–1356. [CrossRef] [PubMed]
3. Lai, M.M.C. Recombination in large RNA viruses: Coronaviruses. *Semin. Virol.* **1996**, *7*, 381–388. [CrossRef]
4. Zhang, Y.; Li, J.; Xiao, Y.; Zhang, J.; Wang, Y.; Chen, L.; Paranhos-Baccalà, G.; Ren, L.; Wang, J. Genotype shift in human coronavirus OC43 and emergence of a novel genotype by natural recombination. *J. Infect.* **2015**, *70*, 641–650. [CrossRef] [PubMed]
5. So, R.T.; Chu, D.K.; Miguel, E.; Perera, R.A.; Oladipo, J.O.; Fassi-Fihri, O.; Peiris, M. Diversity of Dromedary Camel Coronavirus HKU23 in African Camels Revealed Multiple Recombination Events among Closely Related Betacoronaviruses of the Subgenus Embecovirus. *J. Virol.* **2019**, *93*, e01236-19. [CrossRef] [PubMed]
6. Gribble, J.; Stevens, L.J.; Agostini, M.L.; Anderson-Daniels, J.; Chappell, J.D.; Lu, X.; Pruijssers, A.J.; Routh, A.L.; Denison, M.R. The coronavirus proofreading exoribonuclease mediates extensive viral recombination. *PLoS Pathog.* **2021**, *17*, e1009226. [CrossRef] [PubMed]
7. Zhu, Z.; Meng, K.; Meng, G. Genomic recombination events may reveal the evolution of coronavirus and the origin of SARS-CoV-2. *Sci. Rep.* **2020**, *10*, 21617. [CrossRef]
8. Jackson, B.; Boni, M.F.; Bull, M.J.; Colleran, A.; Colquhoun, R.M.; Darby, A.C.; Rambaut, A. Generation and transmission of interlineage recombinants in the SARS-CoV-2 pandemic. *Cell* **2021**, *184*, 5179–5188.e8. [CrossRef]
9. Francisco, R.D.S., Jr.; Benites, L.F.; Lamarca, A.P.; de Almeida, L.G.; Hansen, A.W.; Gularte, J.S.; Demoliner, M.; Gerber, A.L.; Guimarães, A.P.D.C.; Antunes, A.K.E.; et al. Pervasive transmission of E484K and emergence of VUI-NP13L with evidence of SARS-CoV-2 co-infection events by two different lineages in Rio Grande do Sul, Brazil. *Virus Res.* **2021**, *296*, 198345. [CrossRef]
10. Taghizadeh, P.; Salehi, S.; Heshmati, A.; Houshmand, S.M.; InanlooRahatloo, K.; Mahjoubi, F.; Sanati, M.H.; Yari, H.; Alavi, A.; Jamehdar, S.A.; et al. Study on SARS-CoV-2 strains in Iran reveals potential contribution of co-infection with and recombination between different strains to the emergence of new strains. *Virology* **2021**, *562*, 63–73. [CrossRef]
11. Rockett, R.J.; Draper, J.; Gall, M.; Sim, E.M.; Arnott, A.; Agius, J.E.; Sintchenko, V. Co-infection with SARS-CoV-2 Omicron and Delta Variants revealed by genomic surveillance. *Nat. Commun.* **2022**, *13*, 2745. [CrossRef] [PubMed]
12. Musso, N.; Maugeri, J.G.; Bongiorno, D.; Stracquadanio, S.; Bartoloni, G.; Stefani, S.; Di Stefano, E.D. SARS-CoV-2's high rate of genetic mutation under immune selective pressure: From oropharyngeal B.1.1.7 to intrapulmonary B.1.533 in a post-vaccine patient. *Int. J. Infect. Dis.* **2022**, *118*, 169–172. [CrossRef] [PubMed]
13. He, Y.; Ma, W.; Dang, S.; Chen, L.; Zhang, R.; Mei, S.; Wei, X.; Lv, Q.; Peng, B.; Chen, J.; et al. Possible recombination between two variants of concern in a COVID-19 patient. *Emerg. Microbes Infect.* **2022**, *11*, 552–555. [CrossRef] [PubMed]
14. Colson, P.; Fournier, P.E.; Delerce, J.; Million, M.; Bedotto, M.; Houhamdi, L.; La Scola, B. Culture and identification of a "Deltamicron" SARS-CoV-2 in a three cases cluster in southern France. *J. Med. Virol.* 2022; Online ahead of print. [CrossRef]
15. Yi, H. 2019 Novel Coronavirus Is Undergoing Active Recombination. *Clin. Infect. Dis.* **2020**, *71*, 884–887. [CrossRef]
16. Yeh, T.Y.; Contreras, G.P. Emerging viral mutants in Australia suggest RNA recombination event in the SARS-CoV-2 genome. *Med. J. Aust.* **2020**, *213*, 44. [CrossRef]
17. Gallaher, W.R. A palindromic RNA sequence as a common breakpoint contributor to copy-choice recombination in SARS-COV-2. *Arch. Virol.* **2020**, *165*, 2341–2348. [CrossRef]

18. VanInsberghe, D.; Neish, A.S.; Lowen, A.C.; Koelle, K. Recombinant SARS-CoV-2 genomes are currently circulating at low levels. *bioRxiv* **2021**. [CrossRef]
19. Haddad, D.; John, S.E.; Mohammad, A.; Hammad, M.M.; Hebbar, P.; Channanath, A.; Al-Mulla, F. SARS-CoV-2: Possible recombination and emergence of potentially more virulent strains. *PLoS ONE* **2021**, *16*, e0251368. [CrossRef]
20. Varabyou, A.; Pockrandt, C.; Salzberg, S.L.; Pertea, M. Rapid detection of inter-clade recombination in SARS-CoV-2 with Bolotie. *Genetics* **2021**, *218*, iyab074. [CrossRef]
21. Leary, S.; Gaudieri, S.; Parker, M.D.; Chopra, A.; James, I.; Pakala, S.; Sette, A. Generation of a novel SARS-CoV-2 sub-genomic RNA due to the R203K/G204R variant in nucleocapsid: Homologous recombination has potential to change SARS-CoV-2 at both protein and RNA level. *bioRxiv* **2021**. [CrossRef]
22. Lohrasbi-Nejad, A. Detection of homologous recombination events in SARS-CoV-2. *Biotechnol. Lett.* **2022**, *17*, 399–414. [CrossRef] [PubMed]
23. Kreier, F. Deltacron: The story of the variant that wasn't. *Nature* **2022**, *602*, 19. [CrossRef] [PubMed]
24. Ignatieva, A.; Hein, J.; Jenkins, P.A. Ongoing Recombination in SARS-CoV-2 Revealed through Genealogical Reconstruction. *Mol. Biol. Evol.* **2022**, *39*, msac028. [CrossRef] [PubMed]
25. Sekizuka, T.; Itokawa, K.; Saito, M.; Shimatani, M.; Matsuyama, S.; Hasegawa, H.; Saito, T.; Kuroda, M. Genome Recombination between Delta and Alpha Variants of Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2). *Jpn. J. Infect. Dis.* 2022; Online ahead of print. [CrossRef] [PubMed]
26. Million, M.; Lagier, J.-C.; Tissot-Dupont, H.; Ravaux, I.; Dhiver, C.; Tomei, C.; Cassir, N.; Delorme, L.; Cortaredona, S.; Amrane, S.; et al. Early combination therapy with hydroxychloroquine and azithromycin reduces mortality in 10,429 COVID-19 outpatients. *Rev. Cardiovasc. Med.* **2021**, *22*, 1063–1072. [CrossRef] [PubMed]
27. Colson, P.; Fournier, P.-E.; Chaudet, H.; Delerce, J.; Giraud-Gatineau, A.; Houhamdi, L.; Andrieu, C.; Brechard, L.; Bedotto, M.; Prudent, E.; et al. Analysis of SARS-CoV-2 Variants From 24,181 Patients Exemplifies the Role of Globalization and Zoonosis in Pandemics. *Front. Microbiol.* **2022**, *12*, 786233. [CrossRef]
28. Brouqui, P.; Colson, P.; Melenotte, C.; Houhamdi, L.; Bedotto, M.; Devaux, C.; Gautret, P.; Million, M.; Parola, P.; Stoupan, D.; et al. COVID-19 re-infection. *Eur. J. Clin. Investig.* **2021**, *51*, e13537. [CrossRef]
29. Drancourt, M.; Cortaredona, S.; Melenotte, C.; Amrane, S.; Eldin, C.; La Scola, B.; Parola, P.; Million, M.; Lagier, J.-C.; Raoult, D.; et al. SARS-CoV-2 Persistent Viral Shedding in the Context of Hydroxychloroquine-Azithromycin Treatment. *Viruses* **2021**, *13*, 890. [CrossRef]
30. Colson, P.; Devaux, C.A.; Lagier, J.C.; Gautret, P.; Raoult, D. A Possible Role of Remdesivir and Plasma Therapy in the Selective Sweep and Emergence of New SARS-CoV-2 Variants. *J. Clin. Med.* **2021**, *10*, 3276. [CrossRef]
31. Rambaut, A.; Holmes, E.C.; O'Toole, Á.; Hill, V.; McCrone, J.T.; Ruis, C.; du Plessis, L.; Pybus, O.G. A dynamic nomenclature proposal for SARS-CoV-2 lineages to assist genomic epidemiology. *Nat. Microbiol.* **2020**, *5*, 1403–1407. [CrossRef]
32. Aksamentov, I.; Roemer, C.; Hodcroft, E.B.; Neher, R.A. Nextclade: Clade assignment, mutation calling and quality control for viral genomes. *J. Open Source Softw.* **2021**, *6*, 3773. [CrossRef]
33. La Scola, B.; Le Bideau, M.; Andreani, J.; Hoang, V.T.; Grimaldier, C.; Colson, P.; Raoult, D. Viral RNA load as determined by cell culture as a management tool for discharge of SARS-CoV-2 patients from infectious disease wards. *Eur. J. Clin. Microbiol. Infect. Dis.* **2020**, *39*, 1059–1061. [CrossRef] [PubMed]
34. Alm, E.; Broberg, E.K.; Connor, T.; Hodcroft, E.B.; Komissarov, A.B.; Maurer-Stroh, S.; Pereyaslov, D. Geographical and temporal distribution of SARS-CoV-2 clades in the WHO European Region, January to June 2020. *Euro. Surveill.* **2020**, *25*, 2001410. [CrossRef]
35. Fournier, P.E.; Colson, P.; Levasseur, A.; Devaux, C.A.; Gautret, P.; Bedotto, M.; Raoult, D. Emergence and outcomes of the SARS-CoV-2 'Marseille-4' variant. *Int. J. Infect. Dis.* **2021**, *106*, 228–236. [CrossRef] [PubMed]
36. Ou, J.; Lan, W.; Wu, X.; Zhao, T.; Duan, B.; Yang, P.; Ren, Y.; Quan, L.; Zhao, W.; Seto, D.; et al. Tracking SARS-CoV-2 Omicron diverse spike gene mutations identifies multiple inter-variant recombination events. *Signal Transduct. Target. Ther.* **2022**, *7*, 138. [CrossRef] [PubMed]
37. Gu, H.; Ng, D.Y.; Liu, G.Y.; Cheng, S.S.; Krishnan, P.; Chang, L.D.; Poon, L. Samuel SM Cheng, Pavithra Krishnan, Lydia DJ Chang, Sammi SY Cheuk, Mani MY Hui, Tommy TY Lam, Malik Peiris, Leo Poon. Detection of a BA.1/BA.2 recombinant in travelers arriving in Hong Kong, February 2022. *medRxiv* **2022**. [CrossRef]
38. Colson, P.; Delerce, J.; Marion-Paris, E.; Lagier, J.C.; Levasseur, A.; Fournier, P.E.; Raoult, D. A 21L/BA.2-21K/BA.1 "MixOmicron" SARS-CoV-2 hybrid undetected by qPCR that screen for variant in routine diagnosis. *medRxiv* **2022**. [CrossRef]
39. Perez, J.C.; Lounnas, V.; Montagnier, L. The Omicron variant breaks the evolutionary lineage of SARS-CoV-2 variants. *Int. J. Res. Granthaalayah* **2021**, *9*, 108–132. [CrossRef]
40. Gupta, K.; Toelzer, C.; Williamson, M.K.; Shoemark, D.K.; Oliveira, A.S.F.; Matthews, D.A.; Berger, I. Structural insights in cell-type specific evolution of intra-host diversity by SARS-CoV-2. *Nat. Commun.* **2022**, *13*, 222. [CrossRef]
41. Tengs, T.; Delwiche, C.F.; Monceyron Jonassen, C. A genetic element in the SARS-CoV-2 genome is shared with multiple insect species. *J. Gen. Virol.* **2021**, *102*, 001551. [CrossRef]
42. Le Glass, E.; Hoang, V.T.; Boschi, C.; Ninove, L.; Zandotti, C.; Boutin, A.; Colson, P. Incidence and Outcome of Coinfections with SARS-CoV-2 and Rhinovirus. *Viruses* **2021**, *13*, 2528. [CrossRef] [PubMed]

43. Amona, I.; Medkour, H.; Akiana, J.; Davoust, B.; Tall, M.L.; Grimaldier, C.; Mediannikov, O. Enteroviruses from Humans and Great Apes in the Republic of Congo: Recombination within Enterovirus C Serotypes. *Microorganisms* **2020**, *8*, 1779. [CrossRef] [PubMed]
44. Li, H. Minimap2: Pairwise alignment for nucleotide sequences. *Bioinformatics* **2018**, *34*, 3094–3100. [CrossRef] [PubMed]
45. Danecek, P.; Bonfield, J.K.; Liddle, J.; Marshall, J.; Ohan, V.; Pollard, M.O.; Whitwham, A.; Keane, T.; A. McCarthy, S.; Davies, R.M.; et al. Twelve years of SAMtools and BCFtools. *GigaScience* **2021**, *10*, giab008. [CrossRef]
46. Garrison, E.; Marth, G. Haplotype-Based variant detection from short-read sequencing. *arXiv* **2012**, arXiv:1207.3907.
47. Hadfield, J.; Megill, C.; Bell, S.M.; Huddleston, J.; Potter, B.; Callender, C. Nextstrain: Real-time tracking of pathogen evolution. *Bioinformatics* **2018**, *34*, 4121–4123. [CrossRef]
48. Robinson, J.T.; Thorvaldsdottir, H.; Wenger, A.M.; Zehir, A.; Mesirov, J.P. Variant Review with the Integrative Genomics Viewer (IGV). *Cancer Res.* **2017**, *77*, 31–34. [CrossRef]
49. Katoh, K.; Standley, D.M. MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. *Mol. Biol. Evol.* **2013**, *30*, 772–780. [CrossRef]
50. Altschul, S.F.; Gish, W.; Miller, W.; Myers, E.W.; Lipman, D.J. Basic local alignment search tool. *J. Mol. Biol.* **1990**, *215*, 403–410. [CrossRef]
51. Minh, B.Q.; Schmidt, H.A.; Chernomor, O.; Schrempf, D.; Woodhams, M.D.; von Haeseler, A.; Lanfear, R. IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. *Mol. Biol. Evol.* **2020**, *37*, 1530–1534. [CrossRef]
52. Letunic, I.; Bork, P. Interactive tree of life (iTOL) v3: An online tool for the display and annotation of phylogenetic and other trees. *Nucleic Acids Res.* **2016**, *44*, W242–W245. [CrossRef] [PubMed]
53. Kumar, S.; Stecher, G.; Li, M.; Knyaz, C.; Tamura, K. MEGA X: Molecular Evolutionary Genetics Analysis across Computing Platforms. *Mol. Biol. Evol.* **2018**, *35*, 1547–1549. [CrossRef] [PubMed]
54. Jaafar, R.; Boschi, C.; Aherfi, S.; Bancod, A.; Le Bideau, M.; Edouard, S.; La Scola, B. High individual heterogeneity of neutralizing activities against the original strain and nine different variants of SARS-CoV-2. *Viruses* **2021**, *13*, 2177. [CrossRef] [PubMed]

Article

Sequential Appearance and Isolation of a SARS-CoV-2 Recombinant between Two Major SARS-CoV-2 Variants in a Chronically Infected Immunocompromised Patient

Emilie Burel [1,2,†], Philippe Colson [1,2,3,†], Jean-Christophe Lagier [1,2,3], Anthony Levasseur [1,2], Marielle Bedotto [1], Philippe Lavrard-Meyer [1,2,3], Pierre-Edouard Fournier [1,2,4], Bernard La Scola [1,2,3,*] and Didier Raoult [1,2,*]

[1] IHU Méditerranée Infection, 19-21 Boulevard Jean Moulin, 13005 Marseille, France; burel.emilie@hotmail.fr (E.B.); philippe.colson@univ-amu.fr (P.C.); jean-christophe.lagier@univ-amu.fr (J.-C.L.); anthony.levasseur@univ-amu.fr (A.L.); marielle.bedotto@gmail.com (M.B.); philippe.lavrard-meyer@ap-hm.fr (P.L.-M.); pierre-edouard.fournier@univ-amu.fr (P.-E.F.)
[2] Microbes Evolution Phylogeny and Infections (MEPHI), Institut de Recherche Pour le Développement (IRD), Aix-Marseille University, 27 Boulevard Jean Moulin, 13005 Marseille, France
[3] Assistance Publique-Hôpitaux de Marseille (AP-HM), 264 rue Saint-Pierre, 13005 Marseille, France
[4] Vecteurs-Infections Tropicales et Méditerranéennes (VITROME), Institut de Recherche Pour le Développement (IRD), Aix-Marseille University, 27 Boulevard Jean Moulin, 13005 Marseille, France
* Correspondence: bernard.la-scola@univ-amu.fr (B.L.S.); didier.raoult@univ-amu.fr (D.R.); Tel.: +33-413-732-401 (B.L.S. & D.R.)
† These authors contributed equally to this work.

1. SUPPLEMENTARY METHODS

Nasopharyngeal samples from the immunocompromised patient were tested in our university hospital institute (Méditerranée Infection; https://www.mediterranee-infection.com/) from September 2020 to December 2021 for SARS-CoV-2 infection by real-time reverse-transcription-PCR (qPCR) using the BGI real-time fluorescent RT-PCR assay (BGI Genomics, Shanghai Fosun Long March Medical Science Co., Ltd., Shenzhen, China) or the NeuMoDx SARS-CoV-2 assay (NeuMoDx, Ann Arbor, Michigan). Cycle threshold values for qPCR positivity was 35. Outside our institute (medical biology laboratory Alphabio, hôpital Européen, France), SARS-CoV-2 qPCR was performed using the BD MAX System (Becton Dickinson, Sparks, MD, USA).

2. SUPPLEMENTARY RESULTS

2.1. Virological follow-up by SARS-CoV-2 real-time reverse-transcription-PCR (qPCR)

SARS-CoV-2 qPCR performed on nasopharyngeal samples was positive between September 2020 and January 2021, then negative once in February 2021 but positive again on the next sample tested in April 2021. Between May 2021 and August 2021, qPCR remained almost always positive being only transiently negative for ≤3 days. Then qPCR was consistently positive when performed between October and December 2021.

2.2. Nucleotide diversity at positions harboring signature mutations of the B.1.160 or Alpha variants

Regarding the SARS-CoV-2 genomes obtained from the respiratory samples, which were obtained with the Illumina technology except for the sample collected in August 2021 for which the genome was obtained with the Nanopore technology, nucleotide diversity at the 35 positions harboring signature mutations of the B.1.160 or Alpha variants differed according to the time of sample collection. It was low in September 2020, with a mean (±standard deviation) value of 0.2±0.5%, increased to reach high values in May 2021 (20.1±16.4%), June 2021 (14.7±14.3%), July 2021 (19.8±18.2%), and August 2021

(13.2±8.9%), then decreased to low values again in October 2021 (1.1±3.9%) and November 2021 (3.1±6.8%) (Figure 2 of the main text). The low nucleotide diversity observed for genomes obtained in October and November 2021 indicated that the mosaicism consisting in the presence of signature mutations of the two variants was not explained by a co-infection by these two variants or by a contamination of the samples prior or during the next-generation sequencing procedure.

Regarding the genomes obtained from the culture supernatants, which were all obtained with the Illumina technology, nucleotide diversity at positions harboring signature mutations of the B.1.160 or Alpha variants was low, indicating that these genomes, including those recombinants, were generated from a single viral isolate. Indeed, mean nucleotide diversity ranged between 0.6±2.0% and 4.3±6.4% (Figure 2 of the main text). The viral genomes obtained from the culture of the respiratory samples collected in November and December 2021 were highly similar to the viral genome obtained directly from the respiratory sample collected in November 2021.

2.3. Generation of additional sequence reads

Sequencing of reverse-transcription-PCR-targeted regions: We attempted to generate sequence reads that are hybrids of the B.1.160 and Alpha variants from the respiratory sample collected in August 2021 by PCR amplification of the regions overlapping recombination sites then next-generation sequencing with the Oxford Nanopore Technology (ONT) on a GridION instrument (Oxford Nanopore Technologies Ltd., Oxford, United Kingdom). Amplicons corresponding to positions 3,100-4,570 (in reference to the genome of the Wuhan-Hu-1 isolate GenBank accession no. NC_045512.2) were amplified using the following primers (in 5′-3′ orientation): Forward: TTCAACCTGAAGAAGAGCAA; reverse: TGGCATTGTAACAAGAGTTT. Amplicons corresponding to positions 24,813-29,074 were amplified using the following primers: Forward: ATGGAAAAGCACACTTTCCT; reverse: GCTTTAGTGGCAGTACGTTT. For the amplicons corresponding to positions 3,100-4,570, 56% and 87% of the reads were chimeras harboring both Alpha (C3267U) and Marseille-4 (C4543U) signature mutations from the respiratory samples collected in May 5th and in August 12th, 2021, respectively (Supplementary Material: Figures S3, S4). For the amplicon corresponding to positions 24,813-29,074, 70% of the reads were hybrids harboring Alpha (G24914C) and B.1.160 (G25563U, C25710U, C26735U, U26876C and G28975C) mutations, and 24% were hybrids harboring Alpha mutation G24914C and B.1.160 mutation G28975C (Supplementary Material: Figure S5).

Metagenomic sequencing: Next-generation sequencing was also carried out using a metagenomic approach with the Nanopore technology on a GridION instrument (Oxford Nanopore Technologies Ltd.) to attempt obtaining long sequence reads and detect additional reads that are hybrids of the B.1.160 and Alpha variants. For the first recombination site, four reads with a length ranging between 1,855-11,979 nucleotides were obtained, which harbored signature mutations of the Alpha (C3267U) and B.1.160 (C4543U) variants. For the third recombination site, four reads with a length comprised between 1,486-8,143 nucleotides were obtained that harbored signature mutations of the Alpha (G24914C) and B.1.160 (C25563U) variants (reads are available from URL: https://www.mediterranee-infection.com/sars-cov-2-recombinant/).

2.4. Viral culture

Cytopathic effects were observed between 4 and 8 days after the inoculation of the nasopharyngeal samples on Vero E6 cells as previously described [1] (Table 2 of the main text).

2.5. Phylogenetic analyses based on SARS-CoV-2 genomes and spike gene sequences

Phylogeny reconstructions were performed based on viral genomes or spike genes, which included sequences obtained directly from the nasopharyngeal samples and from the cultures, and their best BLAST hits from the IHU Méditerranée Infection sequence database that were classified as B.1.160 or Alpha variants [2]. The genome obtained from

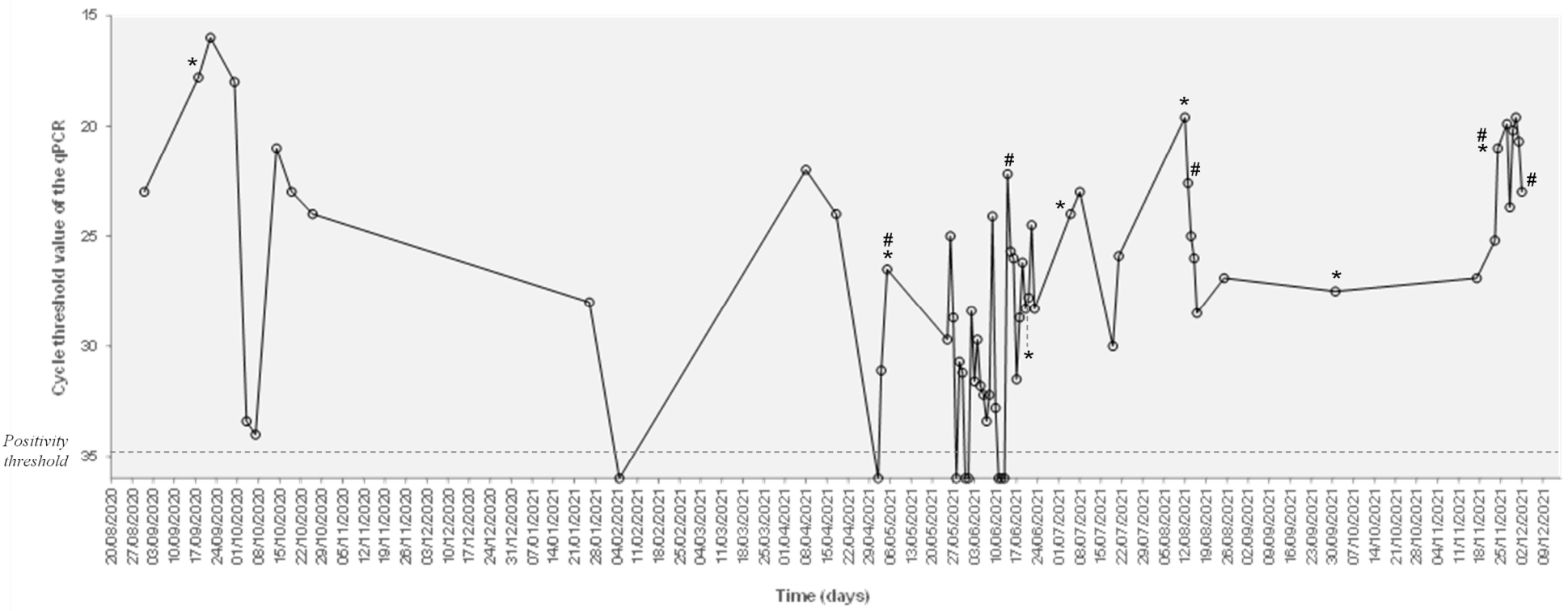
the nasopharyngeal sample collected in September 2020 was clustered with genomes of the B.1.160 variant, apart from the other genomes obtained from the case-patient (Figure 3a of the main text). Other genomes obtained from the nasopharyngeal samples collected since May 2021 and those obtained from cultures of the nasopharyngeal samples collected since May 2021 were also clustered with genomes of the B.1.160 variant, but they were clustered together, apart from the other genomes. The Alpha variant genomes were clustered apart from all other genomes. Spike gene phylogeny showed the clusterisation with sequences of the B.1.160 variant of the sequences obtained from the case-patient either directly or post-culture from samples collected between September 2020 and June 2021 (Figure 3b of the main text). In contrast, sequences obtained from the case-patient either directly or post-culture from samples collected since August 2021 were clustered with Alpha variant sequences.

3. SUPPLEMENTARY FIGURES

Supplementary Figure S1. Virological follow-up by real-time reverse transcription PCR (qPCR) targeting SARS-CoV-2 RNA for the patient.

Time points of collection of nasopharyngeal samples used for direct next-generation sequencing are indicated by *; time points of collection of samples from which positive culture was obtained leading to next-generation sequencing from the culture supernatant are indicated by #.

Nasopharyngeal samples from the immunocompromised patient were tested in our university hospital institute (Méditerranée Infection; https://www.mediterranee-infection.com/) from September 2020 to December 2021 for SARS-CoV-2 infection by real-time reverse-transcription-PCR (qPCR) using the BGI real-time fluorescent RT-PCR assay (BGI Genomics, Shanghai Fosun Long March Medical Science Co., Ltd., Shenzhen, China) or the NeuMoDx SARS-CoV-2 assay (NeuMoDx, Ann Arbor, Michigan). Cycle threshold values for qPCR positivity was 35. Outside our institute (medical biology laboratory Alphabio, hôpital Européen, France), SARS-CoV-2 qPCR was performed using the BD MAX System (Becton Dickinson, Sparks, MD, USA).


Cycle threshold value of the qPCR
15
20
25
30
35
Positivity threshold
Time (days)

Supplementary Figure S2. Majority nucleotides and nucleotide diversity for sequences obtained from the respiratory samples and the culture supernatant at nucleotide positions of the SARS-CoV-2 genome that harbor signature mutations of the B.1.160 or Alpha variants and at any other positions that harbor mutations (a), and sequencing depth at nucleotide positions (b).

Del, nucleotide deletion.

Nucleotide positions are in reference to the genome of the Wuhan-Hu-1 isolate GenBank accession no. NC_045512.2. Nucleotide diversity is the proportion of sequence reads that do not harbor the consensus (majority) nucleotide.

Supplementary Figure S2a.

Nucleotide positions	Genes	Wuhan-Hu-1 isolate genome	Signature mutations in B.1.160/Marseille-4 genomes	Signature mutations in Alpha genomes	Majority nucleotides: Nasopharyngeal samples: September 18th, 2020	May 5th, 2021	June 21st, 2021	July 5th, 2021	August 12th, 2021	October 1st, 2021	November 24th, 2021	Majority nucleotides: Culture supernatants: May 5th, 2021	June 14th, 2021	August 13th, 2021	November 24th, 2021	December 2nd, 2021	Nucleotide diversity (%): Nasopharyngeal samples: September 18th, 2020	May 5th, 2021	June 21th, 2021	July 5th, 2021	August 12th, 2021	October 1st, 2021	November 24th, 2021	Nucleotide diversity (%): Culture supernatants: May 5th, 2021	June 14th, 2021	August 13th, 2021	November 24th, 2021	December 2nd, 2021
241	ORF1a	C	U	U	U	U	U	U	U	U	U	U	U	U	U	U	0,0	0,0	0,0	0,1	10,6	0,0	0,1	0,0	0,1	0,0	0,0	0,4
629		C	-	-	-	U	U	U	U	U	U	-	-	U	U	U	0,2	45,5	13,2	25,1	28,4	0,1	1,0	6,4	0,1	0,2	0,1	0,3
696		A	-	-	-	-	-	U	U	U	U	-	-	U	U	U	0,0	37,5	29,6	32,0	4,5	0,2	0,1	4,9	0,0	0,1	0,5	0,2
913		C	-	U	-	-	-	U	U	U	U	-	-	U	U	U	0,0	33,9	26,8	35,9	10,0	0,1	0,1	9,0	0,1	0,1	0,1	0,3
981		C	-	-	-	-	-	-	-	-	U	-	-	-	-	-	0,0	0,0	0,1	0,0	4,0	0,0	14,0	0,0	0,2	0,1	0,4	0,4
1192		A	-	-	-	-	-	-	-	G	-	-	-	-	-	-	0,2	0,0	0,1	0,2	2,6	33,3	0,0	0,0	0,0	0,1	0,1	0,1
1677		A	-	-	-	-	-	-	-	-	-	-	-	-	C	-	0,0	0,0	0,0	0,0	2,3	0,0	15,0	0,1	0,1	0,1	0,3	0,0
1683		U	-	-	-	-	-	-	-	-	C	-	-	-	-	C	0,0	0,0	0,0	0,0	1,4	0,0	18,5	0,1	0,1	0,0	0,1	0,1
1895		G	-	-	-	-	-	-	-	A	-	-	-	-	-	-	0,5	0,0	0,0	0,0	6,4	0,0	0,0	0,0	0,0	0,5	0,0	0,3
1978		U	-	-	-	-	-	-	-	C	-	-	-	-	-	-	0,1	0,0	0,0	0,0	1,0	12,7	0,2	0,1	0,2	0,0	0,2	0,3
2509		C	-	-	-	-	-	-	-	-	-	-	-	-	U	-	0,5	0,0	0,0	0,0	3,4	0,0	7,4	0,0	0,0	0,0	0,8	0,0
2594		C	-	-	-	-	-	-	-	-	U	-	-	-	U	U	0,0	0,0	0,0	0,0	2,2	0,0	0,0	0,1	0,0	0,0	0,1	0,2
2706		C	-	-	-	U	U	-	-	-	U	U	U	-	U	U	0,1	43,6	25,0	28,1	5,4	0,0	0,1	4,3	0,1	0,0	0,2	0,0
3037		C	U	U	U	U	U	U	U	U	U	U	U	U	U	U	0,3	0,3	1,3	0,0	20,8	0,0	0,1	0,0	0,1	0,0	0,4	0,1
3193		A	-	-	-	-	-	G	G	G	-	-	-	G	-	-	0,3	0,0	25,5	33,4	21,2	0,9	0,0	0,0	0,0	2,6	0,0	0,6
3224		A	-	-	-	-	-	-	-	G	-	-	-	-	-	-	0,0	0,0	0,0	0,0	6,3	0,1	0,1	0,0	0,0	0,0	0,0	0,0
3267		C	-	U	-	-	-	U	U	U	-	-	-	U	-	-	0,0	28,8	23,7	39,3	6,8	0,0	0,0	2,4	0,1	0,2	0,4	0,1
3340		G	-	-	-	-	-	-	-	-	-	-	-	-	U	-	0,0	16,8	0,0	0,0	2,3	0,1	12,7	0,1	0,1	0,0	0,0	0,3
3393		C	-	-	-	-	U	-	-	-	-	U	U	-	-	-	0,3	28,7	43,4	20,8	4,6	0,0	0,2	2,4	0,0	0,0	0,0	0,1
4543		C	U	-	U	U	U	U	U	U	U	U	U	U	U	U	0,1	0,0	0,0	0,0	7,8	0,0	0,1	0,3	0,2	0,2	0,8	0,1
4872		A	-	-	-	G	G	-	-	-	-	G	G	-	-	-	0,1	0,0	12,5	37,1	2,9	0,0	0,0	0,0	0,1	0,0	0,1	0,1
4878		C	-	-	-	U	U	U	U	U	U	U	U	U	U	U	0,0	0,0	0,0	0,0	6,6	0,0	0,2	0,1	0,1	0,1	0,2	0,2
4893		C	-	-	-	-	-	U	U	U	U	-	-	U	U	U	0,1	0,0	27,3	43,3	5,2	0,6	0,0	0,1	0,0	0,2	0,1	0,2
4904		G	-	-	-	-	A	-	-	-	-	A	A	-	-	-	0,0	10,5	30,0	47,9	1,7	0,0	0,0	4,8	0,0	0,0	0,2	0,1
5278		U	-	-	-	-	-	-	-	-	-	-	-	-	C	-	0,9	0,0	0,0	0,0	2,1	0,0	14,6	0,0	0,0	0,0	8,0	0,0
5629		G	U	-	U	U	U	U	U	U	U	U	U	U	U	U	2,8	0,5	1,6	0,2	11,4	0,6	0,0	0,8	1,9	0,9	5,3	1,5
5648		A	-	-	-	-	-	-	-	-	C	-	-	-	C	C	0,1	0,1	0,0	0,0	1,6	0,0	0,1	0,1	0,0	0,0	0,2	0,1
5986		C	-	U	-	-	-	-	-	-	-	-	-	U	-	-	0,0	0,0	0,0	0,0	10,3	0,0	0,0	0,0	0,0	18,4	0,0	0,1
6629		C	-	-	-	-	U	-	-	-	-	-	U	-	-	-	0,0	0,0	37,7	15,6	5,6	0,0	0,0	0,0	0,3	0,3	0,1	0,1
6633		C	-	-	-	U	U	U	U	U	U	U	U	U	U	U	0,0	0,2	0,0	0,0	11,0	0,1	0,0	0,0	0,1	0,3	0,1	0,2
7113		C	-	-	-	-	-	-	-	-	U	-	-	-	-	U	0,0	0,0	/	0,0	3,4	0,0	15,5	0,0	0,0	0,0	0,3	1,5
7185		U	-	-	-	C	C	C	C	C	C	C	C	C	C	C	0,1	0,0	0,0	0,0	10,9	0,0	0,0	0,1	0,0	0,0	0,4	0,1
7279		C	-	-	-	-	-	-	-	-	U	-	-	-	-	-	0,1	0,0	0,0	2,7	2,5	0,0	31,3	0,2	0,3	0,5	0,3	0,5
8191		U	-	-	-	-	-	-	-	-	-	-	-	A	-	-	0,1	0,0	0,0	0,0	12,0	0,2	0,2	0,1	0,0	16,0	0,2	0,0
8770		C	-	-	-	U	-	-	-	-	-	-	-	-	-	-	0,0	47,2	0,0	0,0	10,6	0,0	0,1	3,1	0,1	0,0	0,7	0,1
8950		C	-	-	-	-	-	-	-	-	-	-	U	-	-	-	0,1	0,0	48,1	14,9	1,6	0,1	0,1	0,2	0,1	0,1	0,2	0,1
9042		C	-	-	-	U	U	U	U	U	U	U	U	U	U	U	0,0	0,1	0,1	0,2	17,7	0,0	0,1	0,1	0,5	0,1	0,1	0,3
9120		C	-	-	-	U	U	U	U	U	U	U	U	U	U	U	0,0	0,0	0,0	0,0	8,6	0,0	0,3	0,0	0,0	0,1	0,2	0,3
9501		C	-	-	-	U	-	U	U	U	U	U	U	U	U	U	0,0	0,0	/	0,0	9,1	0,0	0,0	0,6	2,6	0,4	0,5	2,2
9526		G	U	-	U	U	U	U	U	U	U	U	U	U	U	U	0,1	0,0	0,0	0,0	4,6	0,0	0,8	1,2	0,2	0,2	0,0	0,3
9626		U	-	-	-	C	-	-	-	-	-	-	-	-	-	-	0,0	50,0	/	0,0	15,5	0,0	0,0	0,3	0,1	0,0	0,0	0,5
9783		G	-	-	-	-	-	-	C	-	-	-	-	-	-	-	0,0	0,0	/	0,0	35,0	0,0	0,0	0,0	0,0	0,0	0,0	0,2
9792		A	-	-	-	-	C	-	-	-	-	-	-	-	-	-	0,0	16,7	0,0	0,0	2,2	0,0	0,0	1,2	0,0	0,0	0,0	0,0
9811		C	-	-	-	-	-	-	-	-	U	-	-	-	U	U	0,1	0,0	0,0	0,0	6,1	0,0	2,9	0,0	0,1	0,2	0,7	0,6
9972		A	-	-	-	-	-	-	-	-	G	-	-	-	-	-	0,0	0,0	0,0	0,0	5,7	0,0	21,6	0,0	0,0	0,2	0,0	0,1
11083		G	-	-	-	-	-	-	-	-	-	U	-	-	-	-	1,0	1,1	0,0	0,0	2,2	0,0	0,7	27,1	0,1	0,1	0,7	0,5
11208		C	-	-	-	U	U	U	U	U	U	U	U	U	U	U	0,0	0,0	0,2	0,1	4,6	0,2	0,0	0,2	0,1	0,0	0,2	0,1
11365		G	-	-	-	-	-	-	-	-	U	-	-	-	-	U	0,0	0,0	0,0	0,0	2,3	0,0	19,2	0,1	0,0	0,0	0,1	0,1
11497		C	U	-	U	U	U	U	U	U	U	U	U	U	U	U	0,4	0,0	0,0	0,3	2,5	0,4	0,0	0,1	0,3	0,2	0,1	0,2
12459		C	-	-	-	-	U	-	-	-	-	-	U	-	-	-	0,4	0,0	45,8	20,6	3,1	0,0	0,1	0,3	0,1	0,4	0,2	0,3
12592		U	-	-	-	-	C	-	-	-	-	-	C	-	-	-	0,0	0,2	39,0	23,5	2,1	0,0	0,0	0,1	0,1	0,0	0,1	0,1

Supplementary Figure S2a (continued)

Nucleotide positions	Genes	Wuhan-Hu-1 isolate genome	Signature mutations in B.1.160/Marseille-4 genomes	Signature mutations in Alpha genomes	Majority nucleotides: Nasopharyngeal samples: September 18th, 2020	May 5th, 2021	June 21st, 2021	July 5th, 2021	August 12th, 2021	October 1st, 2021	November 24th, 2021	Majority nucleotides: Culture supernatants: May 5th, 2021	June 14th, 2021	August 13th, 2021	November 24th, 2021	December 2nd, 2021	Nucleotide diversity (%): Nasopharyngeal samples: September 18th, 2020	May 5th, 2021	June 21th, 2021	July 5th, 2021	August 12th, 2021	October 1st, 2021	November 24th, 2021	Nucleotide diversity (%): Culture supernatants: May 5th, 2021	June 14th, 2021	August 13th, 2021	November 24th, 2021	December 2nd, 2021
13506	ORF 1b	C	-	-	-	U	U	U	U	U	U	U	U	U	U	U	0,0	0,0	0,0	0,0	13,9	0,0	0,0	0,0	0,1	0,0	0,1	0,1
13517		C	-	-	-	-	-	-	-	U	-	-	-	-	-	-	0,3	0,0	0,0	0,0	3,6	16,7	0,1	0,4	0,1	0,6	0,1	0,2
13860		C	-	-	-	-	-	U	U	U	-	-	-	U	-	-	0,2	0,1	24,2	35,4	3,9	0,1	0,1	0,1	0,2	0,0	0,1	0,2
13936		G	-	-	-	-	-	-	-	-	-	-	-	-	-	U	0,0	0,0	0,1	0,0	4,3	0,3	0,0	0,1	0,0	0,0	0,1	3,0
13993		G	U	-	U	U	U	U	U	U	U	U	U	U	U	U	0,1	0,1	0,0	0,1	5,4	0,1	0,0	0,0	0,1	0,0	0,3	0,1
14408		C	U	U	U	U	U	U	U	U	U	U	U	U	U	U	0,1	0,1	0,0	0,1	11,0	0,1	0,4	0,1	0,3	0,0	0,3	0,2
15766		G	U	-	U	U	U	U	U	U	U	U	U	U	U	U	0,1	0,3	0,0	0,1	9,2	0,0	0,0	0,0	0,1	0,0	0,1	0,1
16128		A	-	-	-	-	-	-	-	-	-	-	-	-	G	-	0,2	0,0	0,1	0,1	2,3	0,0	13,9	0,0	0,0	0,0	12,1	0,0
16675		C	-	-	-	U	U	U	U	U	U	U	U	U	U	U	0,1	0,2	0,0	0,2	13,2	0,1	0,0	0,1	0,7	0,5	0,2	0,1
16751		C	-	-	-	U	U	U	U	U	U	U	U	U	U	U	7,0	0,0	0,0	0,2	11,2	0,4	0,6	0,3	0,9	0,0	0,6	0,0
16889		A	G	-	G	G	G	G	G	G	G	G	G	G	G	G	0,1	0,0	0,0	0,0	7,7	0,0	0,2	0,0	0,1	0,0	0,3	0,1
17019		G	U	-	U	U	U	U	U	U	U	U	U	U	U	U	0,0	0,0	0,0	0,0	15,6	0,0	0,0	0,3	0,2	0,4	0,1	0,1
17122		G	-	-	-	-	-	-	-	-	U	-	-	-	-	U	0,1	0,0	0,0	0,0	11,9	0,1	16,0	0,0	0,1	0,1	0,1	0,1
17125		C	-	-	-	-	-	U	U	U	-	-	-	U	-	-	0,1	0,0	18,9	43,4	18,4	0,1	0,0	0,1	0,0	0,1	0,2	0,2
17251		C	-	-	-	U	U	U	U	U	U	U	U	U	U	U	0,0	0,3	0,1	0,1	11,1	0,2	0,1	0,2	0,1	0,1	0,5	0,5
17721		G	-	-	-	-	-	-	-	-	-	-	-	-	U	-	0,1	0,0	0,0	0,0	5,0	0,0	11,7	0,1	0,1	0,1	0,2	0,3
17738		C	-	-	-	-	-	-	-	U	-	-	-	-	-	-	0,0	0,0	0,0	0,1	9,3	1,0	0,0	0,1	0,1	0,1	0,2	0,1
18023		A	-	-	-	-	-	-	-	-	G	-	-	-	-	-	0,0	0,2	0,1	0,1	0,9	0,2	22,3	0,2	0,0	0,1	0,5	0,4
18877		C	U	-	U	-	U	U	U	U	-	U	U	U	-	-	0,0	47,0	0,0	0,0	6,8	0,0	0,1	2,7	0,2	0,0	0,2	0,1
19186		C	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0,1	0,1	0,1	0,0	4,5	0,0	0,5	0,3	0,0	0,2	44,5	0,1
19547		C	-	-	-	-	-	-	-	-	-	-	-	U	-	-	0,0	0,0	0,0	0,0	4,9	0,0	0,0	0,0	0,0	15,2	0,1	0,3
19947		G	-	-	-	-	-	-	U	-	-	-	-	-	-	-	0,0	0,0	0,0	0,0	59,4	0,0	0,2	0,0	0,2	0,0	0,4	0,1
20930		C	-	-	-	-	-	-	-	-	-	U	-	-	-	-	0,1	18,5	0,0	0,0	5,8	0,1	0,0	6,2	0,1	0,0	0,2	0,1
21219		C	-	-	-	-	-	-	-	-	U	-	-	-	-	-	0,0	0,0	0,0	0,0	0,0	0,0	30,1	0,1	0,0	0,0	0,2	0,1
21232		G	-	-	-	-	-	-	-	-	-	-	-	-	U	-	0,1	0,0	0,0	0,0	0,0	0,0	19,3	0,0	0,1	0,0	0,3	0,3
21575		C	-	-	-	-	U	-	-	-	-	U	U	-	-	-	1,2	44,4	42,4	14,4	8,2	0,0	0,5	4,4	1,2	5,4	0,1	0,5
21765	SPIKE	U	-	Del	-	-	-	-	Del	Del	Del	-	-	Del	Del	Del	0,1	14,7	11,0	38,3	39,2	22,6	22,9	2,1	0,0	21,8	23,7	21,2
21766		A	-	Del	-	-	-	-	Del	Del	Del	-	-	Del	Del	Del	0,0	16,2	11,8	44,0	10,9	1,6	1,1	2,2	0,0	1,2	1,6	1,0
21767		C	-	Del	-	-	-	-	Del	Del	Del	-	-	Del	Del	Del	0,1	15,4	11,1	42,8	19,6	0,0	0,0	2,1	0,0	0,0	0,1	0,0
21768		A	-	Del	-	-	-	-	Del	Del	Del	-	-	Del	Del	Del	0,0	15,4	11,2	43,1	10,4	0,0	0,0	2,1	0,0	0,0	0,1	0,1
21769		U	-	Del	-	-	-	-	Del	Del	Del	-	-	Del	Del	Del	0,0	15,6	11,5	43,7	13,7	0,2	0,2	2,1	0,1	0,2	0,5	0,5
21770		G	-	Del	-	-	-	-	Del	Del	Del	-	-	Del	Del	Del	0,1	18,2	15,2	47,6	9,6	5,6	5,6	3,6	0,0	7,2	6,4	6,5
21991		U	-	Del	-	-	Del	Del	Del	Del	Del	-	Del	Del	Del	Del	0,0	20,3	17,3	18,6	22,1	1,3	4,3	0,1	2,1	2,5	0,7	1,4
21992		U	-	Del	-	-	Del	Del	Del	Del	Del	-	Del	Del	Del	Del	0,1	20,4	16,9	18,1	20,3	1,3	5,6	0,1	2,1	2,5	2,0	2,1
21993		A	-	Del	-	-	Del	Del	Del	Del	Del	-	Del	Del	Del	Del	0,1	20,6	16,7	17,7	17,8	3,7	18,9	0,1	11,5	7,4	10,7	12,6
22992		G	A	-	A	A	/	-	-	-	-	A	A	-	-	-	0,3	33,3	/	20,6	4,8	0,0	0,0	5,5	0,1	0,0	0,2	0,3
23063		A	-	U	-	-	-	U	U	U	U	-	-	U	U	U	0,0	50,0	33,3	27,1	12,4	0,0	0,0	6,7	0,1	0,0	0,4	0,4
23188		C	-	-	-	-	-	-	-	-	-	U	-	-	-	-	0,0	0,0	0,0	0,0	2,3	0,0	0,1	7,3	0,0	0,1	0,2	0,1
23271		C	-	A	-	-	-	A	A	A	A	-	-	A	A	A	0,2	15,7	20,0	44,2	4,2	0,0	0,1	5,1	0,0	0,3	0,2	0,2
23277		C	-	-	-	-	-	-	-	-	-	-	-	-	-	U	0,0	0,0	0,0	0,0	7,9	0,0	0,2	0,2	0,0	0,2	0,1	0,3
23403		A	G	G	G	G	G	G	G	G	G	G	G	G	G	G	0,2	0,0	0,0	0,0	17,4	0,1	0,0	0,1	0,1	0,0	0,1	0,2
23415		C	-	-	-	-	-	-	-	-	A	-	-	-	-	A	0,1	0,0	0,0	0,0	3,6	1,3	14,0	0,2	0,0	0,1	0,1	0,2
23525		C	-	-	-	-	-	-	-	-	-	-	-	-	-	U	0,0	0,0	0,0	0,0	2,4	0,0	0,0	0,1	0,1	0,0	0,2	0,2
23598		A	-	-	-	-	-	-	-	-	-	G	-	-	-	-	0,1	0,0	0,1	0,0	1,6	0,0	0,1	49,0	0,1	0,8	0,4	0,0
23604		C	-	A	-	-	-	A	A	A	A	A	-	A	A	A	0,1	24,8	25,6	37,8	6,0	0,0	0,1	32,7	0,0	0,9	0,6	0,2
23709		C	-	U	-	-	-	U	U	U	U	-	-	U	U	U	0,1	27,5	28,1	37,8	7,8	0,0	0,1	3,5	0,1	0,1	0,1	0,1
23948		G	-	-	-	C	C	-	-	-	-	C	C	-	-	-	0,0	36,9	37,9	29,0	2,2	0,0	0,2	4,1	0,0	0,0	0,3	0,4
24023		C	-	-	-	U	U	-	-	-	-	U	U	-	-	-	0,0	43,5	43,5	27,4	2,0	0,0	0,4	5,7	0,3	0,2	0,2	0,7
24506		U	-	G	-	-	-	G	G	G	G	-	-	G	G	G	0,0	31,2	25,0	31,9	17,7	0,0	0,1	3,2	0,1	0,0	0,1	0,1
24914		G	-	C	-	-	-	C	C	C	C	-	-	C	C	C	0,0	11,6	24,7	43,9	18,7	0,0	0,1	4,4	0,0	0,0	0,2	0,1
25460	ORF 3a	C	-	-	-	-	-	-	U	-	-	-	-	U	-	-	0,2	0,1	49,6	12,9	26,2	0,0	0,3	0,2	0,3	0,1	0,2	0,1
25515		C	-	-	-	-	-	A	-	A	-	-	-	-	-	-	0,0	0,0	29,4	31,9	37,6	0,1	0,3	0,0	0,0	0,1	0,2	0,2
25517		C	-	-	-	-	-	-	-	-	U	-	-	-	-	-	0,1	0,0	0,0	0,1	9,9	0,0	37,0	0,1	0,0	0,2	0,1	0,1
25521		C	-	-	-	-	-	-	-	-	-	-	-	-	U	-	0,1	0,1	0,0	0,1	3,1	0,1	27,9	0,2	0,1	0,1	0,9	0,0
25563		G	U	-	U	U	U	-	U	-	-	U	U	U	-	-	0,0	46,2	40,2	20,4	27,4	0,0	0,0	5,6	0,1	0,3	0,2	0,1
25710		C	U	-	U	-	U	-	U	-	-	U	U	U	-	-	0,1	42,2	34,4	17,6	26,0	0,0	0,1	7,4	0,1	0,0	0,3	0,0
26204		C	-	-	-	-	-	-	U	-	-	U	-	U	-	-	0,0	23,5	42,9	25,9	28,1	0,0	0,0	3,7	23,0	10,6	0,0	0,0
26681		C	-	-	-	-	-	-	-	U	-	-	-	-	-	-	3,1	0,1	0,0	0,1	6,2	49,1	0,3	0,2	0,1	0,1	0,0	0,2
26735	M	C	U	-	U	-	-	-	-	-	-	U	U	U	U	-	0,0	37,9	43,5	9,1	8,6	0,0	19,6	17,0	0,0	0,1	0,3	0,2
26819		U	-	-	-	-	-	-	G	-	-	-	-	-	-	-	0,1	0,0	0,0	0,0	35,4	0,1	0,0	0,0	0,1	0,0	0,1	0,3
26835		C	-	-	-	-	-	-	U	-	-	-	-	-	-	-	0,0	0,0	0,0	0,1	45,4	0,0	0,0	0,1	0,0	0,0	0,2	0,1
26858		C	-	-	-	U	-	U	-	U	U	-	-	-	-	U	0,0	44,1	48,8	13,5	23,6	0,0	28,6	20,2	0,0	0,0	0,1	0,1
26876		U	C	-	C	-	-	-	C	-	-	C	C	C	C	-	0,0	39,2	46,4	12,6	37,2	0,0	22,5	14,7	0,2	0,2	0,2	0,2
26907		C	-	-	-	-	-	-	U	-	-	U	U	U	-	-	0,0	24,9	48,8	43,7	34,9	0,0	0,0	49,1	21,2	15,8	0,6	0,0
26952		A	-	-	-	-	-	-	-	-	-	-	-	-	G	-	0,0	0,0	0,0	0,0	1,5	0,0	28,3	0,0	0,0	0,0	0,1	0,1
27459		G	-	-	-	-	-	-	-	-	U	-	-	-	-	U	0,0	0,0	0,0	0,1	10,0	0,6	11,1	0,1	0,0	0,0	0,0	2,3
27918		A	-	-	-	-	-	-	-	G	G	-	-	-	G	G	0,3	0,0	0,3	5,1	21,8	0,0	0,0	0,0	0,1	0,0	0,0	0,0
28038		G	-	-	-	-	-	-	-	U	U	-	-	-	U	U	0,0	0,0	1,5	6,4	26,7	0,0	0,0	0,1	0,0	0,1	0,0	0,0
28271		A	-	-	-	U	U	U	U	U	U	U	U	U	U	U	0,1	35,5	0,1	0,0	11,4	0,0	0,1	6,2	0,1	0,3	0,3	0,1
28338		A	-	-	-	-	-	-	-	-	-	-	-	-	-	G	0,1	0,0	0,0	0,0	1,3	0,0	0,0	0,0	0,0	0,0	0,1	0,3
28560	ORF 6 / ORF 7a / ORF 7b / ORF 9b / N	G	-	-	-	U	U	U	U	U	C	U	U	U	U	C	0,1	0,0	0,0	0,1	4,4	0,1	19,2	0,0	0,0	0,0	0,3	0,1
28572		A	-	-	-	-	-	-	-	-	G	-	-	-	-	G	0,1	0,0	0,0	0,1	1,2	0,0	21,0	0,0	0,0	0,0	0,1	0,2
28657		C	-	-	-	-	-	-	-	-	-	-	-	-	-	U	0,2	0,0	0,0	0,0	2,5	0,0	0,6	0,1	0,0	0,0	0,1	0,3
28887		C	-	-	-	-	-	-	-	-	U	-	-	-	-	U	0,2	0,0	0,0	0,0	3,5	0,0	15,5	0,2	0,1	0,1	0,4	0,0
28975		G	C	-	C	-	C	C	C	C	C	C	C	C	C	C	0,4	32,8	0,0	0,0	16,6	0,0	5,5	7,3	0,0	0,1	0,1	0,3
28977		C	-	U	-	U	-	-	-	-	-	-	-	-	-	-	0,0	45,1	2,5	0,0	2,8	0,7	0,0	7,3	0,0	0,0	0,2	0,1
29242		A	-	-	-	U	-	-	-	-	-	-	-	-	-	-	0,1	47,9	0,2	0,1	2,0	0,0	0,1	5,8	0,0	0,2	0,3	0,3
29399		G	A	-	A	A	A	A	A	A	A	A	A	A	A	A	0,0	0,1	0,0	0,0	9,3	0,0	0,0	0,0	0,0	0,1	0,1	0,5
29445		C	-	-	-	-	U	-	U	-	-	-	U	U	-	-	0,0	0,1	31,6	30,4	16,2	0,0	0,0	0,0	0,0	0,2	0,0	0,0
29743		C	-	-	-	-	-	-	-	-	-	-	-	U	-	-	0,1	0,0	0,0	0,0	32,1	0,0	0,0	0,1	0,0	13,5	0,1	0,1

Supplementary Figure S2b.

Nucleotide positions	Nasopharyngeal samples							Culture supernatants				
	September 18th, 2020	May 5th, 2021	June 21st, 2021	July 5th, 2021	August 12th, 2021	October 1st, 2021	November 24th, 2021	May 5th, 2021	June 14th, 2021	August 13th, 2021	November 24th, 2021	December 2nd, 2021
241	2150	352	3	1704	20886	164	1666	590	1037	601	134	826
629	994	573	121	1725	15751	960	793	1639	1204	1004	1333	2036
696	3494	3506	3644	3332	67497	2441	2681	2199	2210	1457	1499	2574
913	2521	2860	3041	2330	51997	1753	2001	1205	1600	828	760	1455
981	3342	3018	3306	2933	111794	2344	2404	2416	1979	1514	1651	2601
1192	837	812	668	1499	59773	1396	736	1963	1374	1402	1612	1952
1677	1055	123	236	943	129526	673	646	1717	1372	1169	1202	1635
1683	1056	121	236	973	129521	680	681	1690	1368	1171	1172	1635
1895	3856	128	229	802	163711	96	155	377	249	209	191	376
1978	2898	266	70	316	38991	102	1250	1469	2128	1160	432	1490
2509	1618	146	326	508	6193	91	122	259	197	197	178	269
2594	663	242	23	391	14494	199	1235	1657	2040	1210	558	2141
2706	665	243	20	442	14618	221	1435	1767	1933	1349	609	2476
3037	694	395	230	569	53	722	854	1515	938	1119	1345	1603
3193	1410	1795	1891	1329	33319	231	582	342	455	189	211	361
3224	736	1925	1985	1550	33479	866	1704	980	1360	681	680	967
3267	735	2002	1946	1312	33223	1073	1840	1024	1520	562	788	1049
3340	734	2127	2112	1698	32456	1510	2003	1122	1705	836	712	1142
3393	731	1696	1624	1333	32139	1136	1586	865	1291	602	590	855
4543	3340	489	7	1275	304941	247	1546	1516	1733	1069	441	953
4872	1522	76	8	143	4024	212	669	2538	1645	1127	1153	3332
4878	1521	72	7	118	4012	179	610	2453	1591	1056	1051	3188
4893	1520	91	11	127	3991	178	602	2458	1649	1023	1088	3348
4904	1520	95	10	117	3986	145	535	2329	1682	926	976	3274
5278	850	828	92	343	718	67	151	169	241	78	71	192
5629	1774	799	189	1765	51104	164	617	389	472	212	42	388
5648	1181	797	189	1850	51155	486	1860	1017	1553	703	374	1190
5986	571	149	224	199	913	383	528	1207	849	940	1075	1349
6629	1811	410	284	699	162979	893	857	1394	747	961	1139	1593
6633	1811	425	302	733	163229	937	902	1457	816	1005	1204	1635
7113	1974	30	0	95	288760	6	194	415	297	313	170	473
7185	1653	8	2	68	284759	15	578	1427	810	980	592	1433
7279	1653	15	6	110	285215	18	770	1630	1162	1312	860	1701
8191	2290	1560	928	1493	177777	1106	1893	1083	1614	807	824	1146
8770	1971	108	2	454	153168	28	1543	1035	1537	900	325	1076
8950	2911	1251	1238	2031	259162	1318	1912	2233	1550	1597	1349	2443
9042	963	840	955	1326	107695	1244	778	1495	884	956	1258	1906
9120	961	1099	1287	1677	106579	1590	983	1786	1287	1187	1539	2316
9501	1692	87	0	322	4013	23	516	358	688	266	52	267
9526	1724	186	1	491	4020	57	2429	1801	1370	591	131	1792
9626	1488	2	0	4	181	4	162	340	1708	319	28	660
9783	1493	6	0	2	1111	1	155	261	1631	236	23	428
9792	2650	6	1	2	22549	1	45	81	504	57	5	104
9811	2735	395	178	151	24543	173	718	812	1867	589	261	987
9972	1264	682	223	251	24799	357	1396	870	2043	904	594	1175
11083	1784	3073	3301	2742	114282	2212	2593	1758	2729	1391	1087	1925
11208	1800	2632	2754	2418	114516	1884	2563	1484	2366	1089	925	1684
11365	3333	2531	2218	2663	167240	1993	3056	3704	3270	2401	2028	3693
11497	1557	285	7	695	54023	509	910	2039	1649	1253	1119	2203
12459	2225	1415	896	2354	63912	1377	2333	3008	1870	1939	1796	3089
12592	916	438	688	1380	23200	981	877	1734	1183	1405	1599	1755

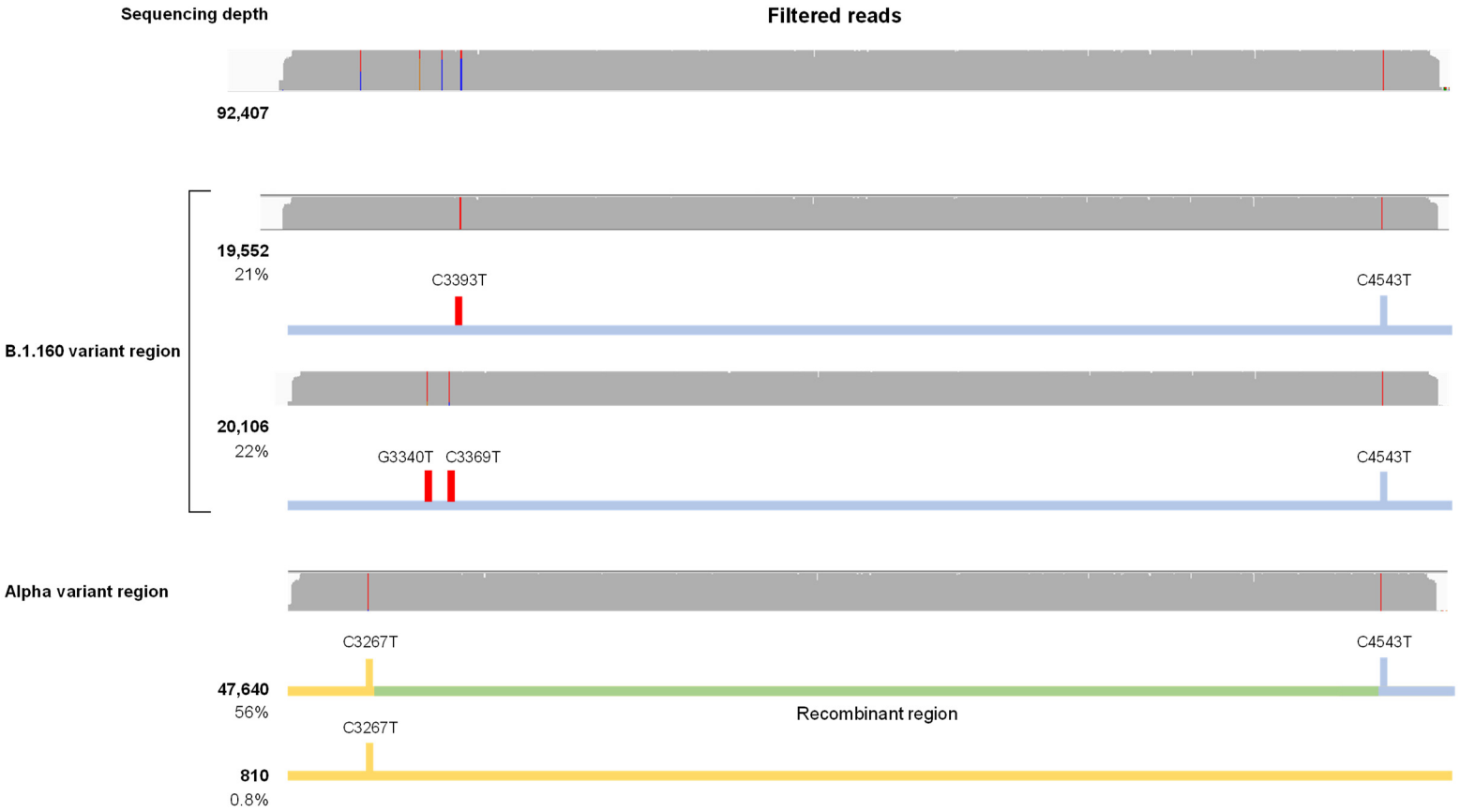
Supplementary Figure S2b (continued)

Nucleotide positions	Nasopharyngeal samples							Culture supernatants				
	September 18th, 2020	May 5th, 2021	June 21st, 2021	July 5th, 2021	August 12th, 2021	October 1st, 2021	November 24th, 2021	May 5th, 2021	June 14th, 2021	August 13th, 2021	November 24th, 2021	December 2nd, 2021
13506	751	19	8	79	7784	6	763	2084	2239	1345	433	2188
13517	751	26	8	98	7798	6	857	2324	2535	1517	479	2448
13860	2406	693	1098	1297	20082	869	913	2767	1799	1537	1948	2481
13936	5135	969	1413	1387	160834	385	430	920	601	570	626	886
13993	2826	1931	1232	1485	143197	1240	2017	2074	1456	743	510	1177
14408	320	876	155	1628	21272	1343	844	1654	1575	1176	1243	1774
15766	1689	1008	1839	1791	47916	1480	840	1480	1395	1162	1463	1717
16128	3156	3403	2881	2744	57062	687	703	501	650	342	153	554
16675	811	818	448	455	31287	727	1562	901	1055	423	386	1129
16751	1830	978	569	546	31093	276	525	297	429	149	154	369
16889	1236	209	12	292	233	509	1206	2272	1850	1715	1421	2741
17019	1237	137	19	175	225	324	791	1510	1304	1017	849	1838
17122	3716	2711	1588	2183	68134	1562	3401	2728	2586	1595	1649	3410
17125	3716	2527	1485	2054	68137	1435	3221	2496	2440	1420	1535	3126
17251	2512	2303	1468	2283	69806	1686	2221	1252	1783	992	1046	1468
17721	2530	1170	552	2074	31757	862	2288	2056	1719	1380	1121	1972
17738	2522	947	408	1750	31777	598	1874	1541	1415	1068	697	1447
18023	2671	1143	1496	2294	109388	1253	2310	2414	1877	1614	1632	2330
18877	1328	1598	2688	2248	62018	2114	1310	1911	1848	1464	1872	2269
19186	1445	1670	1687	1462	37080	1031	1504	926	1445	656	604	956
19547	14	97	78	40	81	124	251	628	396	508	482	861
19947	148	71	19	224	392	58	519	1174	870	1040	648	1478
20930	409	2758	2087	2111	7742	1685	1855	1222	2118	989	831	1493
21219	157	23	127	26	8	93	396	692	1130	873	529	1483
21232	157	25	132	29	7	104	400	690	1091	884	530	1457
21575	221	363	132	132	3101	294	580	271	914	518	288	817
21765	2025	756	1110	1045	3515	702	571	1409	1242	624	800	1091
21766	2025	717	1076	925	3513	552	445	1426	1272	494	630	869
21767	2024	723	1090	934	3513	543	440	1461	1304	488	621	861
21768	2024	721	1075	929	3513	543	440	1454	1267	488	621	861
21769	2024	711	1045	915	3513	544	441	1414	1223	489	623	863
21770	2024	710	1035	933	3513	575	466	1370	1165	526	677	919

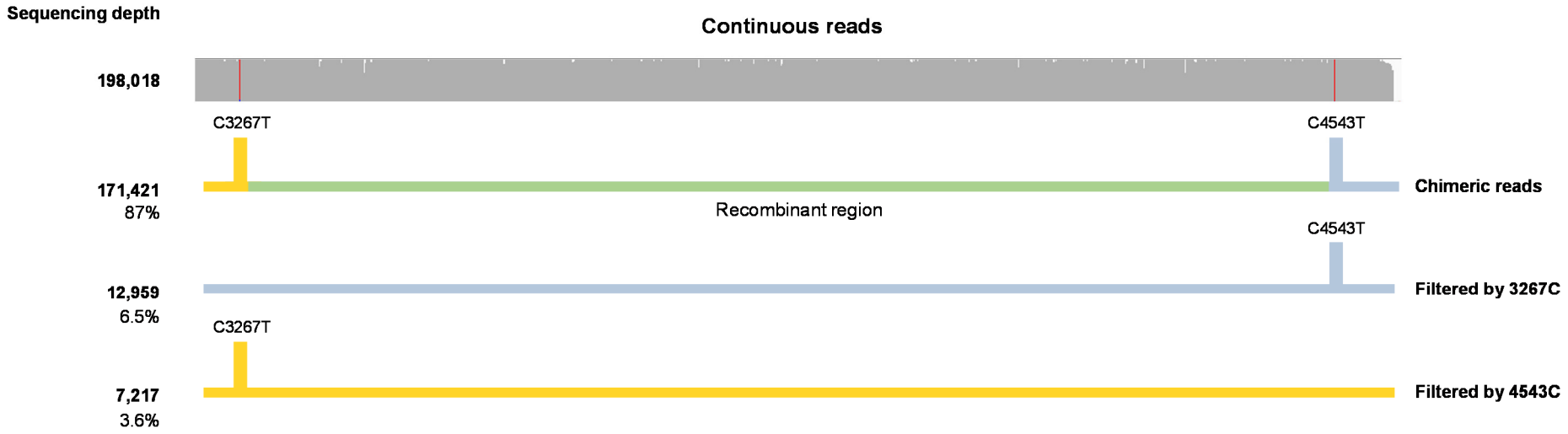
Supplementary Figure S2b (continued)

Nucleotide position	Nasopharyngeal samples							Culture supernatants				
	September 18th, 2020	May 5th, 2021	June 21st, 2021	July 5th, 2021	August 12th, 2021	October 1st, 2021	November 24th, 2021	May 5th, 2021	June 14th, 2021	August 13th, 2021	November 24th, 2021	December 2nd, 2021
21991	3898	1347	1126	1191	11092	80	70	1573	96	80	106	144
21992	3900	1346	1121	1185	11090	80	71	1578	96	80	106	145
21993	3900	1338	1117	1176	11088	82	74	1583	104	81	107	151
22992	1241	9	0	175	151073	12	186	980	1363	997	95	673
23063	1226	16	3	214	150290	7	243	1142	1610	1029	116	776
23188	3006	2370	2725	2115	165844	1184	1572	1338	1947	1026	481	1333
23271	1840	2849	3112	2196	19307	1490	1827	1063	2417	706	601	1437
23277	1840	2957	3275	2345	19312	1631	1980	1151	2603	831	681	1571
23403	1837	2753	3033	2309	20084	1749	1974	1168	2358	921	714	1643
23415	1834	2889	3168	2442	20072	1869	2098	1254	2544	986	767	1754
23525	782	726	1039	1136	1106	956	860	1377	1761	1102	1060	1929
23598	784	787	1096	1200	689	988	977	202	1647	1185	1185	2070
23604	784	828	1135	1228	689	995	976	107	1702	1192	1162	2071
23709	783	703	957	996	688	778	868	1384	1387	862	921	1605
23948	2775	352	243	434	51870	216	998	660	955	397	221	529
24023	2776	469	308	478	51684	280	1282	724	1216	566	304	722
24506	1061	1248	336	993	40365	660	1497	793	1485	562	566	1000
24914	953	1286	2252	2139	56056	1706	1017	1808	1896	1247	1577	2169
25460	942	787	1402	1411	109844	1314	794	2225	1764	1233	1424	1744
25515	942	774	1492	1357	109838	1229	910	2476	1803	1411	1331	1895
25517	942	744	1477	1346	109841	1159	903	2407	1680	1324	1281	1812
25521	942	753	1386	1246	109845	1121	904	2390	1679	1313	1262	1783
25563	942	604	1079	1096	109851	1004	684	2107	1299	899	1122	1489
25710	1647	351	131	493	58375	369	1455	2327	1623	826	565	1447
26204	1014	221	35	239	22478	87	215	652	453	179	391	633
26681	977	1114	950	1537	1767	1408	1001	2401	2245	1384	1753	3572
26735	979	1201	1084	1729	1804	1681	1125	2807	2065	1327	1686	4015
26819	979	1224	1103	1428	5035	1482	1097	2895	2426	1402	1725	4763
26835	991	1301	1184	1407	5549	1238	1063	2756	2116	1243	1538	4249
26858	2548	1491	1290	1574	32691	308	405	1066	555	395	459	1141
26876	2553	1214	1042	1322	32927	1149	1834	4224	2713	1901	1484	3997
26907	2347	233	43	247	32445	10	299	864	631	336	120	707
26952	1594	222	45	282	31569	33	911	3053	2445	1381	409	3308
27459	1380	1048	1563	1661	114777	361	262	761	498	344	431	820
27918	383	618	630	254	133	14	8	1189	1151	771	14	20
28038	383	774	915	362	135	27	22	1690	1456	1007	41	56
28271	1896	2572	3143	2135	92933	1762	1829	1473	1720	760	647	1217
28338	1894	3302	3787	2615	92839	2051	2050	1563	2017	925	876	1423
28560	179	829	1450	1667	12470	1260	798	2017	1535	1200	1277	2103
28572	179	907	1533	1789	12474	1368	1003	2162	1653	1295	1390	2429
28657	179	1106	1816	2053	12462	1579	1057	2834	2155	1644	1663	3113
28887	1241	827	66	212	7521	222	956	2372	2567	1083	555	2506
28975	1241	808	41	139	7558	151	888	2328	1816	754	382	2595
28977	1241	548	40	127	7569	141	841	2227	1672	704	354	2478
29242	1366	771	461	1371	48806	766	838	2431	1909	1475	1265	2651
29399	359	1382	172	329	5301	593	853	1056	1414	687	501	1073
29445	360	1429	133	339	5326	494	749	927	1166	548	425	908
29743	385	450	128	298	4362	483	624	1312	940	1035	765	1412

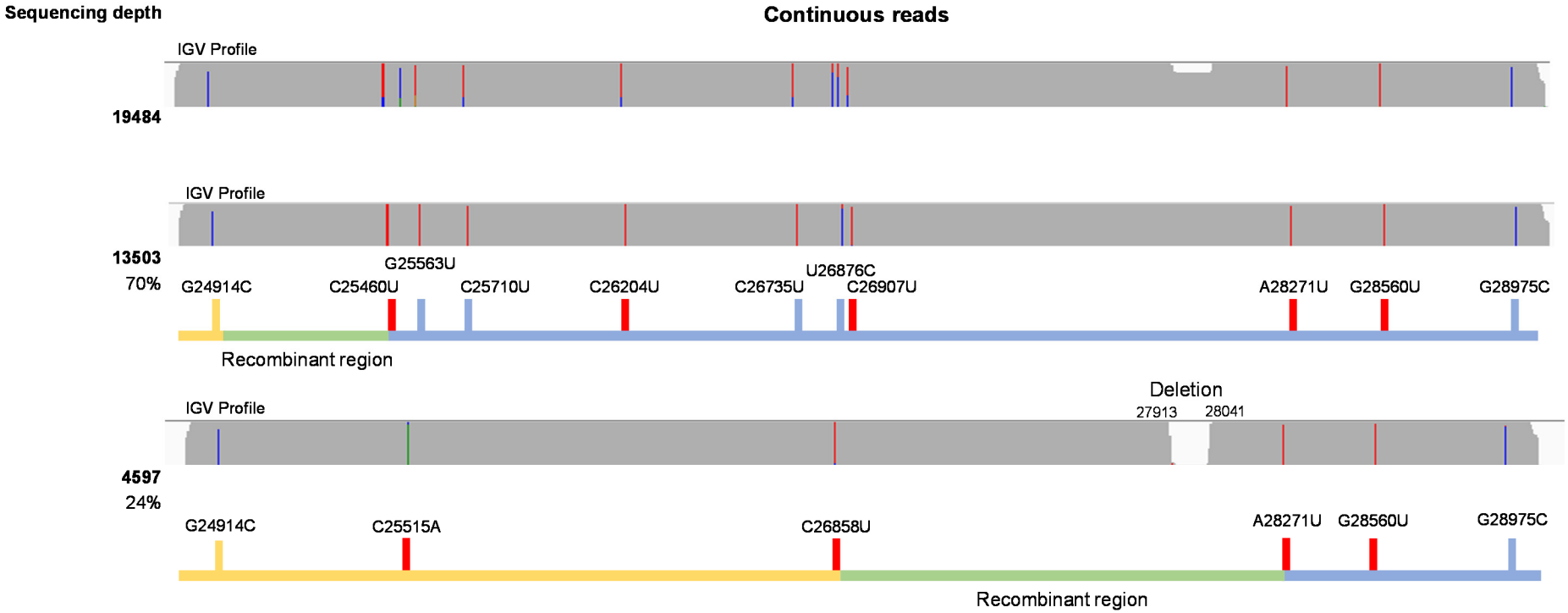
Supplementary Figure S3. Detailed content of amplicons corresponding to positions 3,100-4,570 of the SARS-CoV-2 genome and retrieved from respiratory samples collected on May 5th, 2021. Amplicons were first sorted in order to keep continuous reads for analysis. The number and percentage of each type of reads are indicated. B.1.160 variant regions are indicated by a blue color; Alpha variant regions are indicated by a green color; patient-specific mutations are indicated by a red color. Reads were visualized using the IGV software [3].



Supplementary Figure S4. Detailed content of amplicons corresponding to positions 3,100-4,570 of the SARS-CoV-2 genome and retrieved from respiratory samples collected on August 12th, 2021. Amplicons were first sorted in order to keep continuous reads for analysis. The number and percentage of each type of reads are indicated. B.1.160 variant regions are indicated by a blue color; Alpha variant regions are indicated by a green color; patient-specific mutations are indicated by a red color. Reads were visualized using the IGV software [3].



Supplementary Figure S5. Detailed content of amplicons corresponding to positions 24,880-29,010 of the SARS-CoV-2 genome and retrieved from respiratory samples collected on August 12th, 2021. Amplicons were first sorted in order to keep continuous reads for analysis. The number and percentage of each type of reads are indicated. B.1.160 variant regions are indicated by a blue color; Alpha variant regions are indicated by a green color; patient-specific mutations are indicated by a red color. Reads were visualized using the IGV software [3].



4. SUPPLEMENTARY TABLES

Supplementary Table S1. List of GISAID identifiers for sequences used in the present study.

All genomes were obtained in our laboratory (University hospital institute Méditerranée Infection, Marseille, France). The GISAID sequence database is accessible at: (https://www.gisaid.org/) [4].

Source sample	GISAID identifier
Respiratory samples	
Case-patient's sequences	EPI_ISL_10816731
	EPI_ISL_10816733
	EPI_ISL_10816742
	EPI_ISL_10816743
	EPI_ISL_10816744
	EPI_ISL_11030507
	EPI_ISL_6332079
Most similar sequences from our database	EPI_ISL_10816729
	EPI_ISL_10816740
	EPI_ISL_2757096
	EPI_ISL_3054898
	EPI_ISL_3054899
	EPI_ISL_3055037
	EPI_ISL_3055223
	EPI_ISL_3055225
	EPI_ISL_3055227
	EPI_ISL_3055234
	EPI_ISL_3055235
	EPI_ISL_3830105
	EPI_ISL_3830215
	EPI_ISL_3831874
	EPI_ISL_3933559
	EPI_ISL_3933565
	EPI_ISL_3933569
	EPI_ISL_3933571
	EPI_ISL_3933572
	EPI_ISL_3934656
	EPI_ISL_3934682
	EPI_ISL_3934683
	EPI_ISL_4394365
	EPI_ISL_4944362
	EPI_ISL_4944364
	EPI_ISL_4951784
	EPI_ISL_4951785
	EPI_ISL_6737742
	EPI_ISL_7376142
	EPI_ISL_7376355

Supplementary Table S1 - continued

		EPI_ISL_8321213
		EPI_ISL_8321223
		EPI_ISL_8321228
		EPI_ISL_8321229
		EPI_ISL_8321241
		EPI_ISL_8321242
		EPI_ISL_8321256
		EPI_ISL_8709907
		EPI_ISL_8709910
		EPI_ISL_8711731
		EPI_ISL_9773781
		EPI_ISL_9774389
		EPI_ISL_9774894
		EPI_ISL_9774895
Culture supernatants		
	Case-patient's sequences	EPI_ISL_10816730
		EPI_ISL_10816732
		EPI_ISL_10816734
		EPI_ISL_10816735
		EPI_ISL_10816738

Supplementary Table S2. Sampling time and location for sequences used in the present study.

GISAID identifier	Date	Location
EPI_ISL_3830215	2021-03	Europe / France / Provence-Alpes-Cote d'Azur /Marseille
EPI_ISL_10816729	2020-11	Europe / France / Provence-Alpes-Cote d'Azur / Marseille
EPI_ISL_10816740	2021-03	Europe / France / Provence-Alpes-Cote d'Azur / Marseille
EPI_ISL_10816741	2021-11	Europe / France / Provence-Alpes-Cote d'Azur / Marseille
EPI_ISL_2757096	2020-11-07	Europe / France / Provence-Alpes-Cote d'Azur / Marseille
EPI_ISL_3054898	2020-09-11	Europe / France / Provence-Alpes-Cote d'Azur / Marseille
EPI_ISL_3054899	2020-09-14	Europe / France / Provence-Alpes-Cote d'Azur / Marseille
EPI_ISL_3055037	2020-09-12	Europe / France / Provence-Alpes-Cote d'Azur / Marseille
EPI_ISL_3055223	2020-09-15	Europe / France / Provence-Alpes-Cote d'Azur / Marseille
EPI_ISL_3055225	2020-09-15	Europe / France / Provence-Alpes-Cote d'Azur / Marseille
EPI_ISL_3055227	2020-09-15	Europe / France / Provence-Alpes-Cote d'Azur / Marseille
EPI_ISL_3055234	2020-09-17	Europe / France / Provence-Alpes-Cote d'Azur / Marseille
EPI_ISL_3055235	2020-09-17	Europe / France / Provence-Alpes-Cote d'Azur / Marseille
EPI_ISL_3830105	2021-03	Europe / France / Provence-Alpes-Cote d'Azur / Marseille
EPI_ISL_3831874	2020-08	Europe / France / Provence-Alpes-Cote d'Azur / Marseille
EPI_ISL_3933559	2020-11	Europe / France / Provence-Alpes-Cote d'Azur / Marseille
EPI_ISL_3933565	2020-09	Europe / France / Provence-Alpes-Cote d'Azur / Marseille
EPI_ISL_3933569	2020-08	Europe / France / Provence-Alpes-Cote d'Azur / Marseille
EPI_ISL_3933571	2020-08	Europe / France / Provence-Alpes-Cote d'Azur / Marseille
EPI_ISL_3933572	2020-08	Europe / France / Provence-Alpes-Cote d'Azur / Marseille
EPI_ISL_3934656	2021-04	Europe / France / Provence-Alpes-Cote d'Azur / Marseille
EPI_ISL_3934682	2021-04	Europe / France / Provence-Alpes-Cote d'Azur / Marseille
EPI_ISL_3934683	2021-03	Europe / France / Provence-Alpes-Cote d'Azur / Marseille
EPI_ISL_4394365	2021-04	Europe / France / Provence-Alpes-Cote d'Azur / Marseille
EPI_ISL_4944362	2021-07	Europe / France / Provence-Alpes-Cote d'Azur / Marseille
EPI_ISL_4944364	2021-07	Europe / France / Provence-Alpes-Cote d'Azur / Marseille
EPI_ISL_4951784	2021-04	Europe / France / Provence-Alpes-Cote d'Azur / Marseille
EPI_ISL_4951785	2021-04	Europe / France / Provence-Alpes-Cote d'Azur / Marseille
EPI_ISL_6737742	2021-01	Europe / France / Provence-Alpes-Cote d'Azur / Marseille
EPI_ISL_7376142	2021-01	Europe / France / Provence-Alpes-Cote d'Azur / Marseille
EPI_ISL_7376355	2021-01	Europe / France / Provence-Alpes-Cote d'Azur / Marseille
EPI_ISL_8321213	2020-10	Europe / France / Provence-Alpes-Cote d'Azur / Marseille
EPI_ISL_8321223	2020-10	Europe / France / Provence-Alpes-Cote d'Azur / Marseille
EPI_ISL_8321228	2020-10	Europe / France / Provence-Alpes-Cote d'Azur / Marseille
EPI_ISL_8321229	2020-10	Europe / France / Provence-Alpes-Cote d'Azur / Marseille
EPI_ISL_8321241	2020-10	Europe / France / Provence-Alpes-Cote d'Azur / Marseille
EPI_ISL_8321242	2020-10	Europe / France / Provence-Alpes-Cote d'Azur / Marseille
EPI_ISL_8321256	2020-10	Europe / France / Provence-Alpes-Cote d'Azur / Marseille
EPI_ISL_8709907	2021-11	Europe / France / Provence-Alpes-Cote d'Azur / Marseille
EPI_ISL_8709910	2022-01	Europe / France / Provence-Alpes-Cote d'Azur / Marseille
EPI_ISL_8711731	2021-03	Europe / France / Provence-Alpes-Cote d'Azur / Marseille
EPI_ISL_9773781	2020-03	Europe / France / Provence-Alpes-Cote d'Azur / Marseille
EPI_ISL_9774389	2020-03	Europe / France / Provence-Alpes-Cote d'Azur / Marseille
EPI_ISL_9774894	2021-04	Europe / France / Provence-Alpes-Cote d'Azur / Marseille
EPI_ISL_9774895	2021-04	Europe / France / Provence-Alpes-Cote d'Azur / Marseille

*Supplementary Table S3. Identifiers of r*aw data deposited in the European Bioinformatics Institute (EMBL-EBI) sequence database (https://www.ebi.ac.uk/)

a.

Raw data EMBL-EBI identifier	GISAID identifier	Sampling date	Time from diagnosis (days)	Next-generation sequencing technology, instrument
ERR9671577	EPI_ISL_6332079	September 18th, 2020	0	Illumina, NovaSeq
ERR9671578	EPI_ISL_10816743	May 5th, 2021	229	Illumina, NovaSeq
ERR9671579	EPI_ISL_11030507	June 21st, 2021	276	Illumina, NovaSeq
ERR9671580	EPI_ISL_10816731	July 5th, 2021	290	Illumina, NovaSeq
ERR9692770	EPI_ISL_10816742	August 12th, 2021	328	Nanopore, GridION
ERR9671581	EPI_ISL_10816744	October 1st, 2021	378	Illumina, NovaSeq
ERR9671582	EPI_ISL_10816733	November 24th, 2021	432	Illumina, NovaSeq

b.

Raw data EMBL-EBI identifier	GISAID identifier	Sampling date of the nasopharyngeal sample	Time to cytopathic effect (days)	Next-generation sequencing technology, instrument
ERR9671585	EPI_ISL_10816730	May 5th, 2021	8	Illumina, NovaSeq
ERR9671583	EPI_ISL_10816732	June 14th, 2021	4	Illumina, NovaSeq
ERR9671584	EPI_ISL_10816734	August 13th, 2021	4	Illumina, NovaSeq
ERR9671586	EPI_ISL_10816735	November 24th, 2021	5	Illumina, NovaSeq
ERR9671587	EPI_ISL_10816738	December 2nd, 2021	7	Illumina, NovaSeq

c.

Raw data EMBL-EBI identifier	GISAID identifier / pre-sequencing approach	Sequenced region	Sampling date
ERR9683059	EPI_ISL_10816742 / RT-Klenow	Whole genome	August 12th, 2021
ERR9683060	EPI_ISL_10816743 / H2	3,100-4,570	May 5th, 2021
ERR9683061	EPI_ISL_10816742 / H3	24,813-29,074	August 12th, 2021
ERR9738766	EPI_ISL_10816742 / H2	3,100-4,570	August 12th, 2021

References

1. La Scola, B.; Le Bideau, M.; Andreani, J.; et al. Viral RNA load as determined by cell culture as a management tool for discharge of SARS-CoV-2 patients from infectious disease wards. *Eur J Clin Microbiol Infect Dis.* **2020**, *39*, 1059-1061.
2. Colson, P.; Fournier, PE.; Chaudet, H.; et al. Analysis of SARS-CoV-2 Variants From 24,181 Patients Exemplifies the Role of Globalization and Zoonosis in Pandemics. *Front Microbiol.* **2022**, *12*, 786233. doi: 10.3389/fmicb.2021.786233.
3. Robinson, J.T.; Thorvaldsdottir, H.; Wenger, A.M.; Zehir, A.; Mesirov, J.P. Variant Review with the Integrative Genomics Viewer (IGV). *Cancer Research* **2017**, *77*, 31-34.
4. Alm, E.; Broberg, EK.; Connor, T.; et al. Geographical and temporal distribution of SARS-CoV-2 clades in the WHO European Region, January to June 2020. *Euro Surveill.* **2020**, *25*, 2001410.

Article 3 : **Evolution intra-hôte et adaptation virale à l'hôte**

Avant-propos

Ce troisième article s'inscrit dans la continuité du travail précédent publié dans *Viruses*, qui portait sur l'apparition séquentielle et l'isolement de recombinants SARS-CoV-2 chez un patient immunodéprimé atteint d'un lymphome.

Dans cette étude, en prenant en compte le contexte immunitaire du patient, nous avons étudié les stratégies évolutives du SARS-CoV-2 au niveau des protéines, des gènes et des nucléotides, en intégrant deux mécanismes évolutifs distincts : la recombinaison homologue et l'émergence de mutations au sein de l'hôte. Ce cas rare (440 jours d'infection documentée) a fourni une occasion exceptionnelle d'identifier des convergences évolutives, de révéler des signatures gène par gène de sélection positive, neutre ou purificatrice. Ces signatures ont été analysées en fonction de l'appartenance des mutations aux iSNV (intra-host single nucléotide variation) et aux SNP, mettant en évidence la dynamique évolutive spécifique des gènes. En outre, nous avons analysé l'évolution de la protéine Spike, qui a subi deux trajectoires évolutives successives chez ce patient : une phase initiale d'émergence de mutations au sein de l'hôte, suivie du remplacement par une protéine Spike recombinante, puis à une nouvelle émergence de mutations *de novo,* conduisant à une combinaison optimisée de mutations pour l'adaptation à l'hôte. Ce travail de modélisation a été réalisé en collaboration avec le Pr. Fantini.

Dans cet article, j'ai souhaité explorer l'importance du remodelage adaptatif du génome du SARS-CoV-2 au cours d'une infection prolongée, un travail qui pourrait éventuellement servir à de futures études comparatives. Pour cette étude, 16 prélèvements ont été séquencés avec la même technologie (Illumina). Cette étude souligne également la valeur scientifique des infections chroniques comme environnements d'observation privilégiés de l'évolution virale, tout en pointant la nécessité d'une surveillance attentive chez les patients immunodéprimés.

Cette étude a été soumise dans Virus Evolution le 23 octobre 2024, nous présentons dans cette thèse le preprint publié dans Biorxiv.

OXFORD UNIVERSITY PRESS | Virus Evolution

"Long-Term Intra-Host Evolution of SARS-CoV-2 in an Immunocompromised Patient: Recombination and de novo Mutations Driving Viral Adaptation"

Journal:	*Virus Evolution*
Manuscript ID	VEVOLU-2024-222
Manuscript Type:	Research Article
Date Submitted by the Author:	23-Oct-2024
Complete List of Authors:	Burel, Emilie; Institut Hospitalo-Universitaire Méditerranée Infection, VITROME; Aix-Marseille Université, RITMES Pontarotti, Pierre; IHU Mediterranee Infection, MEPHI; Aix-Marseille Universite, RITMES; CNRS, SNC5039 Fantini, Jacques; Aix-Marseille Universite, INSERM UMR UA 16 Lagier, Jean-Christophe; IHU Mediterranee Infection, Unité 3 IHU ; Assistance Publique - Hôpitaux de Marseille (AP-HM), Pôle de maladies infectieuses CHU Timone; Aix-Marseille Universite, MEPHI Fournier, Pierre Edouard; IHU Mediterranee Infection, VITROME; Assistance Publique - Hôpitaux de Marseille (AP-HM), Pôle de maladies infectieuses; Aix-Marseille Universite, RITMES
Keywords:	Intra-host evolution, SARS-CoV-2, genomics, Spike, ORF8

SCHOLARONE™
Manuscripts

bioRxiv preprint doi: https://doi.org/10.1101/2025.03.25.645210; this version posted March 26, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Long-Term Intra-Host Evolution of SARS-CoV-2 in an Immunocompromised Patient: Recombination and Within-Host Mutations Driving Viral Adaptation

Short title (for the running head):

SARS-CoV2 intra-host evolution led to a more host-optimized virus.

Author list: Emilie Burel[1,2]*, Pierre Pontarotti[1,3,6], Jacques Fantini[4], Jean-Christophe Lagier[1,3,5], Pierre Edouard Fournier[1,2,5]*

Affiliations: [1] IHU Méditerranée Infection, 19-21 boulevard Jean Moulin, 13005 Marseille, France; [2] RITMES, Aix-Marseille Univ., Service de Santé des Armées, IHU Méditerranée Infection, Marseille, France; [3] MEPHI, Aix-Marseille Univ., IHU Méditerranée Infection, Marseille, France ; [4] INSERM UMR UA 16, Aix-Marseille Université, Marseille, France ; [5] Pôle de maladies infectieuses, Assistance Publique - Hôpitaux de Marseille (AP-HM), IHU Méditerranée Infection, Marseille, France ; [6] CNRS SNC5039, Marseille, France.

*** Corresponding author:** Pierre Edouard Fournier, IHU - Méditerranée Infection, 19-21 boulevard Jean Moulin, 13005 Marseille, France. Tel.: +33 413 732 401, Fax: +33 413 732 402; email: pierre-edouard.fournier@univ-amu.fr

Key words (5): Intra-host evolution; SARS-CoV-2, recombination, genomics, iSNV, Spike

Word counts; Abstract, 253 ; text, 7097

Figures: 9; **Tables:** 3; **References: 102**

bioRxiv preprint doi: https://doi.org/10.1101/2025.03.25.645210; this version posted March 26, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

ABSTRACT

An immuno-compromised patient with lymphoma experienced a prolonged SARS-CoV-2 infection lasting 14 months, initially infected with B.1.160, followed by B.1.1.7. This study focused on intra-host single nucleotide variants (iSNVs) and single nucleotide polymorphisms (SNPs), distinguishing their origin as either within-host mutations or parental. The whole genome analysis revealed accelerated evolution, with positive selection detected in key genes such as Spike, N, ORF9b, and nsp13, all involved in viral replication and immune evasion. Of the two evolutionary mechanisms involved, host-driven mutation has played a dominant role in this evolutionary story. C>T transitions emerged as the most widespread mutational signature, consistent with host-driven RNA editing mechanisms. Two host-internal mutations, A28271T in the translation initiation region of the N Kozak gene and C26858T in the M gene, were highly shared among the deposited SARS-CoV-2 sequences. Recombination with parental lineages played a major role particularly in Spike, ORF3a and M genes. In Spike, the B.1.1.7 sequence was selected, wild-type ORF3a and M were restored, suggesting a selective advantage in returning to an ancestral sequence. In addition, the temporary emergence of intra-host convergent mutations in Spike, notably L5F, D796H and T572I, underlines the strong selective pressures exerted on this gene. A 126-nucleotide deletion in ORF8 resulted in a truncated protein, reinforcing its uselessness for viral replication, as observed in circulating variants such as B.1.1.7. The present case highlights the complex interplay between viral recombination and mutations within the host in chronic infections and further underscores the remarkable evolutionary plasticity of SARS-CoV-2 and its potential to generate highly adapted viral strains.

INTRODUCTION

The COVID-19 pandemic has been shaped by the regular emergence of SARS-CoV-2 variants

108

bioRxiv preprint doi: https://doi.org/10.1101/2025.03.25.645210; this version posted March 26, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

with properties that escape natural or vaccine-induced immunity, prompting interest in understanding their origins. Intra-host evolution in long-term infections has emerged as a possible origin[1,2], a theory supported in particular by the emergence of the Alpha (B.1.1.7)[3] and Omicron (B.1.1.529) variants[4], which quickly acquired numerous mutations, in comparison with variants circulating during the previous epidemic waves, conferring immune escape and increased transmissibility. Documenting intra-host evolution that leads to unique patterns is crucial for understanding the mechanisms driving the emergence of host-adapted strains, such as those with increased affinity for the ACE2 Spike receptor[5,6]. Immunocompromised patients, due to AIDS, some cancers or immunosuppressive therapy, for the main cases, are more prone to develop long-term infections due to impaired viral clearance[7,8,9,10,11,12–15]. In addition, weakened immune pressure, combined with treatments such as antiviral therapies, antibodies and convalescent plasma, create favorable conditions for the emergence of new genotypes[16,10,5,17,18] and convergent events[19]. During prolonged infections, co-infections with multiple strains have been documented[20,21] which can lead to recombination through intracellular co-infection and RNA polymerase-induced homologous recombination[22,23]. It has been suggested that this process played a key role in the emergence of SARS-CoV-2[24,25], being particularly impactful on the Spike protein gene[26,27,28,29]. Recombination was extensively described for Omicron variants and subvariants[30,31] and it has recently been shown that SARS-CoV-2 has a higher frequency of RNA recombination than other human CoVs[32]. Moreover, the study of intra-host nucleotide variation (iSNV) is crucial for understanding the early mutagenic mechanisms at the origin of the creation of a pool of quasispecies[33,34,35]. Among the dominant mutations observed during intra-host evolution of SARS-CoV-2, C>T transitions are particularly frequent, and have been linked to the activity of enzymes from the APOBEC (Apolipoprotein B mRNA Editing Catalytic Polypeptide-like) family, including APOBEC3A, APOBEC1, and APOBEC3G, which are cytidine

bioRxiv preprint doi: https://doi.org/10.1101/2025.03.25.645210; this version posted March 26, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

deaminases[36,37]. Combined mechanisms involving host-induced RNA editing and host environmental pressure associated with recombination, may thus contribute to the rapid evolution of viruses.

We previously described and confirmed by culture the emergence of recombinant strains in a patient infected with B.1.160, who was later superinfected by B.1.1.7, resulting in unique strains[38]. Given the exceptional duration of this chronic infection (440 days), this case provided a rare opportunity to investigate intra-host viral evolution over an extended period which combined both mechanisms: recombination and within host emergence of mutations, offering valuable insights into their interplay. To further explore this evolutionary process, we conducted ultra-deep sequencing on 16 nasopharyngeal swabs collected over time using Illumina technology for all samples. Twelve samples offered sufficient coverage in order to make a deep analysis on iSNVs and SNPs patterns.

Our findings revealed extensive intra-host diversity, with positive selection observed in key genes such as N, ORF9b, nsp4, nsp13 and nsp3, highlighting their adaptive potential. Additionally, we identified a convergent event in (A28271T) substitution in a non-coding region, located in the Kozak sequence upstream of the Nucleocapsid gene, potentially optimizing viral gene expression[39]. In the Spike gene, we observed transient convergent mutations: D796H and L5F, which were subsequently eliminated by recombination, while within-host mutations T572I and T618K appeared among the parental B.1.1.7 haplotype. Moreover, recombination led to the restoration of the wild-type (WT) ORF3a and WT-like M sequence, suggesting a selective advantage for reverting to the ancestral sequence. This event emphasizes the role of recombination not only in generating genetic diversity but also in purging mutations. Furthermore, we identified a 126-nt deletion in ORF8, resulting in a truncated protein form, reinforcing its dispensability for viral replication, as previously observed in circulating variants such as B.1.1.7.

bioRxiv preprint doi: https://doi.org/10.1101/2025.03.25.645210; this version posted March 26, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Finally, this case provides a unique example of intra-host evolution of SARS-CoV-2, testifying to its genomic plasticity generating highly host-adapted genotypes, notably an optimized spike, and also evidenced by recombination reversions and the significant accumulation of new within host mutations. The interplay between host-induced mutagenesis, selection pressures and recombination provides a rich framework for study, which could be used in future comparative studies aimed at better understanding host adaptation strategies.

MATERIALS AND METHODS

Ethics statement

This retrospective study received a positive decision from the "Assistance Publique-Hôpitaux de Marseille"(AP-HM) ethics and scientific committee (N° CSE24-4). Access to biological and patient registry data from the hospital information system was approved by the AP-HM data protection committee and registered in the European General Data Protection Regulation register under number RGPD/APHM 2019-73.

Samples

This study was carried out retrospectively on 54 nasopharyngeal swabs collected from September 2020 to December 2021, as part of care of one patient at the IHU Mediterranee Infection Institute (Marseille, France).

SARS-CoV-2 Genome Sequencing

Illumina sequencing was performed only from samples with Ct values <30, as previously described[38]. Viral RNA was extracted from 200 µL of nasopharyngeal swab fluid using the EZ1 Virus Mini kit v2.0 on an EZ1 Advanced XL instrument (Qiagen, Courtaboeuf, France) or using the MagMax Viral/Pathogen Nucleic Acid Isolation kit on the KingFisher Flex system (Thermo Fisher Scientific, Waltham, MA, USA), following the manufacturer's instructions.

bioRxiv preprint doi: https://doi.org/10.1101/2025.03.25.645210; this version posted March 26, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Next-generation sequencing performed on the NovaSeq 6000 instrument followed the Illumina COVIDSeq protocol (Illumina Inc; Illumina Inc., San Diego, CA, USA.), which included first strand cDNA synthesis from extracted viral RNA; cDNA amplification with two COVIDSeq primer pools; fragmentation and tagging of PCR amplicons with adapter sequences; clean up; PCR amplification with index ligation (seven cycles) of tagmented amplicons; pool, clean up, quantification and normalization of libraries; library sequencing on a NovaSeq 6000 sequencing system SP flow cell (Illumina Inc; Illumina Inc., San Diego, CA, USA).

Complete genome analysis

Raw files were basecalled using the Dragen Bcl Convert pipeline (v3.9.3; https://emea.support.
illumina.com/sequencing/sequencing_software/bcl-convert.html; accessed on 18 May 2022), generating fastq files. Raw reads were trimmed using Trimmomatic (V 0.39) (https://doi.org/10.1093/bioinformatics/btu170) and then mapped using bwa-mem2 (Sequence alignment using Burrows-Wheeler Transform) (v2.2.1; https://github.com/bwa-mem2/bwa-mem2; 2019 Intel Corporation, Heng Li) against the Wuhan-Hu-1 isolate reference (GenBank NC_045512.2); primers were cleaned using Samtools ampliconclip (v1.13; https://www.htslib.org/; accessed on 18 May 2022). Variant calling was performed using FreeBayes (v1.3.5; https://github.com/freebayes/freebayes; accessed on 18 May 2022, Erik Garrison, Gabor Marth, University of Tennessee and Utah) and consensus genomes were obtained with Bcftools (v1.13; https://samtools.github.io/bcftools/bcftools.html; accessed on 18 May 2022, 2012–2021 Genome Research Ltd.). The consensus threshold of a SNP (Single Nucleotide Polymorphism) has been set at 70% with a minimum coverage of 10 reads.

To check for the presence of large Indels, given the length of the reads (50 bp), the cleaned reads were also assembled with Unicycler (v0.4.4)

bioRxiv preprint doi: https://doi.org/10.1101/2025.03.25.645210; this version posted March 26, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

(https://doi.org/10.1371/journal.pcbi.1005595), three almost complete genomes were obtained in one contig, the resulting assemblies were deposited in the GISAID database (https://gisaid.org/) (EPI_ISL_18896020, EPI_ISL_18897181, EPI_ISL_18897180).

RNA sequences in this study are represented using thymine (T) instead of uracil (U), in order to facilitate sequence comparisons and computational processing. It should obviously be understood that T represents U in its biological context.

Substitutions detected after a deletion (ex : G28038T) were excluded from the analysis because of artifactual origin.

In depth analysis of intra host diversity

To investigate intra-host SARS-CoV-2 evolution and diversity, we classified mutations based on their origin and allele frequency (AF). Within-host mutations were defined as novel mutations absent from both parental lineages (B.1.160 and B.1.1.7) and were obtained by excluding all signature mutations from B.1.160 and B.1.1.7. Although full reconstruction of the original B.1.1.7 haplotype was not possible due to lack of pre-D229 samples, a detailed analysis of contemporaneous B.1.1.7 strains in Marseille found no evidence of externally introduced mutations (Table S1). Therefore, all mutations except those defining B.1.160 or B.1.1.7 were considered to be internal host mutations, and additional external genomic contributions negligible.

To ensure high-confidence variant detection, we set a minimum allele frequency (MAF) threshold of 5% and a minimum sequencing depth of 50x, following recommendations from previous studies[40]. To further refine variant calling, only iSNVs covered by at least 5 reads were retained, ensuring a practical MAF threshold of ≥10% at 50x depth. Additionally, a genome-wide coverage threshold of >80% was applied to ensure the inclusion of high-quality samples in the analysis at 50x coverage.

iSNV analysis was performed on BAM files obtained from the previously described mapping

113

bioRxiv preprint doi: https://doi.org/10.1101/2025.03.25.645210; this version posted March 26, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

process. Variant calling was conducted using iVar (v1.3.1, https://github.com/andersen-lab/ivar) in combination with Samtools (https://github.com/samtools/samtools), with the precited parameters and a minimum Phred quality score of 30.

VCF files generated by iVar and TSV files from the mapping step were processed using custom Python scripts (https://github.com/EMbBureli/Intra-host-SARS-CoV-2-evolution-study-in-an-immunocompromised-patient) to generate filtered mutation files.

Phylogenetic analysis

The phylogenetic analysis was obtained with the Nexstrain pipeline[41] using the Maximum-likelihood method[42]. The intra-host evolutionary rate was calculated and compared to inter-host rates from 700 locally circulating SARS-CoV-2 sequences (GISAID dataset, September 2020–December 2021). Plots were visualized using the Auspice online tool (https://auspice.us/).

Entropy calculation based at gene level

Shannon entropy (H) was used to measure the mutational diversity of allele frequencies at gene levels within each pool and was computed using the following formula:

$$H = -\sum pi \log_2 (pi)$$

where pi represents the frequency of alternative alleles for a given mutation. To minimize bias due to variable gene coverage, the raw entropy values were divided by the percentage of gene coverage. The normalized entropy values per gene were summed to obtain an overall measure of mutation diversity, and pairwise statistical significance was applied using the Mann-Whitney U test, from scipy.stats (https://docs.scipy.org/doc/scipy/) Python library. Graphs were plotted using the Matplotlib Python library (https://matplotlib.org/stable/plot_types/).

pN/pS calculation

The pN/pS ratio was computed following the Nei-Gojobori method[43], without the Jukes-

bioRxiv preprint doi: https://doi.org/10.1101/2025.03.25.645210; this version posted March 26, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Cantor correction for multiple substitutions, given the relatively short evolutionary time within the host, with $pS = \frac{S+0.1}{ES}$ and $pN = \frac{NS}{EN}$, where S and NS represent the number of observed synonymous and non-synonymous mutations respectively, within a given gene. The corresponding synonymous (ES) and non-synonymous (EN) sites were calculated using a Python script based on the reference SARS-CoV-2 genome (NC_045512.2). For cases where pS = 0 and pN =0, the impossible ratio was annotated as "undetermined", and "infinite" when NS>0. The computed pN/pS ratio, number of infinite and undetermined cases were plotted using the python Matplotlib library (https://matplotlib.org/stable/plot_types/).

Molecular modeling of mutant SARS-CoV-2 Spike proteins

Molecular modeling of mutant spike proteins was performed as described previously[44]. A complete structure of the original SARS-CoV-2 spike protein encompassing the amino acid residues 14-1200 (thus lacking the signal sequence) was generated and characterized as described[45]. Minimized structures of the spike protein of each variant were obtained by sequentially introducing appropriate mutations and/or deletions. Energy minimizations of the variants were performed with the Polak-Ribière conjugate gradient algorithm using a maximum of 3 × 105 steps and a root mean square (RMS) gradient of 0.01 kcal/mol.Å as the convergence condition[46]. The electrostatic potential illustrated in this study is the sum of the Coulomb potentials for each atom of the considered molecule, with a distance-dependent dielectric constant[47].

Within-Host Mutations and Global Convergence

To assess the prevalence of SNPs detected within-host (SNP_WH) in globally circulating SARS-CoV-2 strains, we queried the GISAID (Global Initiative on Sharing Avian Influenza Data, https://www.gisaid.org) database on November 30, 2023, analyzing a dataset of 16,269,440 SARS-CoV-2 genomes. The visualization of SNP_WH proportions in the global dataset was generated using the ggplot2 library in R (https://ggplot2.tidyverse.org/).

bioRxiv preprint doi: https://doi.org/10.1101/2025.03.25.645210; this version posted March 26, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Mapping visualization

The ORF8 deletion was directly visualized from PCR products realized on D378 sample targeting region 24,813-29,074. Reads were mapped against the SARS-CoV-2 reference genome. Sam files were compressed into .bam files with Samtools, (v1.13; https://www.htslib.org/), indexed and visualized using the IGV software (https://www.igv.org/)[48], the 126-nt deletion was visible thanks to continuous reads.

All scripts are available at :(https://github.com/EMbBureli/Intra-host-SARS-CoV-2-evolution-study-in-an-immunocompromised-patient).

RESULTS

An immunocompromised patient with a chronic SARS-CoV-2 infection.

A 56-year-old immunocompromised patient with a severe pneumonia was diagnosed with SARS-CoV-2 infection in September 2020, resulting in admission to intensive care. Despite the improvement in his condition, he remained SARS-CoV-2-positive for 14 months (Figure 1a-b). In addition, in May 2021, a progressive multifocal leukoencephalitis was diagnosed. In his medical history, the patient was previously diagnosed with mixed Hodgkin and follicular lymphoma in 2017. The patient was still under therapy for the follicular lymphoma in 2021. In terms of infecting genotype, the patient first contracted the SARS-CoV-2 variant B.1.160 in August 2020 and then B.1.1.7, which was partially detected from May 2021 (Figure 1 a-d), based on 19 lineage-defining mutations. The evolutionary process involved multiple recombination events and the appearance of within host mutations. Different recombinant regions from the B.1.1.7 variant were identified during the infection (Figure 1d) and objectified in the previous article[38], including the 5' end, comprising C913T and C3267T, positions 17,019 to 26,876, including the Spike gene (21,563-25,384) at consensus level, and positions 27,972 to 28,977 at sub consensus level. Notably, the sample corresponding to day

bioRxiv preprint doi: https://doi.org/10.1101/2025.03.25.645210; this version posted March 26, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

229 contained 11 additional B.1.1.7 lineage-defining mutations with allele frequencies (AF) ranging between 13.3% and 36.4%, reflecting early stages of recombination before these mutations were eventually eliminated (Figure S2). Treatment regimens received by the patient included two convalescent plasma treatments as well as Lenalidomide, Rituximab and Pembrolizumab chemotherapies. The serological data for May and October 2021 are presented in Table 1. The patient died in December 2021 from complications of his hematological and neurological diseases.

Genomic evolution of SARS-CoV-2 over time, a global overview

As part of the longitudinal monitoring of the virus' evolution, we first focused on the analysis of 16 consensus genomes, which are artificial reconstructions based on the dominant nucleotide at each position, obtained from nasopharyngeal swabs taken between September 2020 and December 2021. These genomes were sequenced from samples with Ct values below 30, all sequences were classified as B.1.160 variants according to the PangoLIN (Phylogenetic Assignment of Named Global Outbreak LINeages) nomenclature (https://cov-lineages.org/). The GISAID accession numbers for these sequences are presented in Table 2. Phylogenetic reconstruction based on the 16 consensus genomes (≥10× coverage) revealed the emergence of three distinct genotypes over time (Figure 2a): Genotype 1 was dominant at days 229, 259, and 276; Genotype 2 was detected at days 277, 290, 329, 341 and 378; and Genotype 3 emerged at days 285, 330, 431, 432, 425, 434, and 440.

The temporal overlap between these genotypes can be explained by the coexistence of viral subpopulations within the host, potentially replicating in different anatomical compartments, as observed in long-term SARS-CoV-2 infections[49,6,50]. At coverage 10x, the intra-host evolutionary rate was estimated at 1.39 × 10□³ substitutions per site per year, 1.54 times higher than the estimated inter-host rate of 9.02 × 10□ □ substitutions per site per year (Figure 2b, S1). This value was derived by comparing our dataset with 700 randomly selected SARS-

bioRxiv preprint doi: https://doi.org/10.1101/2025.03.25.645210; this version posted March 26, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

CoV-2 genomes from locally circulating variants (September 2020–December 2021) deposited in the GISAID database (https://gisaid.org/) (Figure S1). These findings align with previous reports demonstrating accelerated viral evolution in chronic infections[5,16,51], especially as in this particular case, the evolutionary dynamic of SARS-CoV-2 was driven by recombinant parental mutations associated with the emergence of *de novo* mutations.

Expanding host diversity analysis with iSNVs and SNPs.

We determined the optimal sequencing depth for analyzing intra-host SARS-CoV-2 diversity by comparing genome-wide and gene-level coverage at 50x and 200x (Figure S3a-b). While 200x coverage, representing 10 times the reciprocal of the MAF (5%), initially seemed preferable[52], it resulted in a loss of SNP information (30 fewer substitutions in SNP_TOT, Figure S3c) and a drop of ≈11% in genome-wide coverage (80.82% ± 10.90 at 200x vs. 91.47% ± 7.24 at 50x, Figure S3a). 50x was selected as the optimal threshold, ensuring both SNP retention and reliable iSNV detection (≥5% MAF, ≥5 reads support) (see Material and Methods). Samples D259 and D440 were excluded due to less than 80% genome coverage (Figure S3c).

We sought to capture the full spectrum of viral evolution within the host, combining the contribution of exogenous mutations with assimilation by recombination and evolution by *de novo* emergence process. Hence, we first categorized mutations based on their frequency, to distinguish transitional mutations from fixed mutations, using a 50% allele frequency (AF) threshold: intra-host single nucleotide variants (iSNVs; AF <50%) and single nucleotide polymorphisms (SNPs; AF ≥50%). Additionally, mutations were grouped into two pools: “Within-host” (WH): mutations arising exclusively within the host, and “Total” (TOT): including all mutations, independently of their origin.

At day 229, a notable increase in iSNV richness was notable, reaching its maximum in the iSNV_TOT pool with 45 iSNV detected, representing the active phase of recombination,

bioRxiv preprint doi: https://doi.org/10.1101/2025.03.25.645210; this version posted March 26, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

where the introduction of recombinant genomic segments temporarily increased intra-host diversity (Figure 3b). After reaching this peak, iSNV richness declined in both pools, showing fluctuations over time, which may be attributed to sampling bias or subpopulation circulation. From day 290, the iSNV_TOT and iSNV_WH curves are nearly indistinguishable, which means a minimal contribution of recombination-derived parental mutations to the iSNVs pools beyond this time point. The accumulation of SNP_TOT ranged from 20 substitutions at the beginning of the infection to a maximum of 56 SNPs at later stages. The number of SNPs originating from recombination remained stable throughout the infection, with eight lineage-defining B.1.1.7 substitutions acquired (C913T, C3267T, A23063T, C23271A, C23604A, C23709T, T24506G, and G24914C), while six lineage-defining B.1.160 substitutions were lost (C18877T, G22992A, G25563T, C25710T, C26735T, and T26876C) (Figure S2). The SNP_WH and SNP_TOT pools exhibited a simultaneous increase starting from day 330, although a decline in the number of SNPs (TOT and WH) was observed between days 329 and 330. This drop can be attributed to uncovered genomic positions as well as recombined segments, which involve both exogenous mutations from parental genomes and mutations that have emerged within the host. Notably, a decrease in iSNV richness is observed at days 329 and 378 compared to other time points. This drop could reflect variations of stochastic effects in viral population dynamics, emphasizing the need for careful interpretation of temporal fluctuations in intra-host diversity.

The allele frequency distribution (Figure 3a) shows contrast in evenness between iSNVs and SNPs. SNPs (WH and TOT) predominantly clustered within the 90-100% frequency range, indicating low evenness, exceptions were observed at day 229 (D229), D276 and D290, where SNPs exhibit a broader distribution across the 50-100% frequency intervals. In contrast, iSNVs show a more dispersed frequency distribution, spanning a wider range of frequency bins than SNPs, particularly in samples D229, D277, D290, D330, and D431. The

bioRxiv preprint doi: https://doi.org/10.1101/2025.03.25.645210; this version posted March 26, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

fluctuations in iSNV frequency distribution can be attributed to sampling bias or circulation of viral subpopulations. In contrast, the tighter clustering of SNPs (TOT and WH) within high-frequency bins reflects the stabilization and strong fixation over time.

We then examined the proportion of synonymous and non-synonymous mutations in the four pools (Figure 3c-S4). Non-synonymous mutations were significantly more frequent in the SNP pools (SNP_WH and SNP_TOT), while they were more frequent without reaching statistical significance in the iSNV pools (iSNV_WH and iSNV_TOT) (Figure S4). This more balanced distribution of non-synonymous and synonymous mutations may reflect the influence of genetic drift on transient mutations, where mutations arise stochastically and are not necessarily subject to immediate selective pressures.

Next, we analyzed the mutational spectrum of iSNVs in search of special signatures associated with pools. Across all pools, 11 different types of mutation were observed. A massive enrichment in C>T transitions was evident in all pools (Figure 3d), consistent with APOBEC-mediated RNA editing, a major contributor to SARS-CoV-2 evolution[53,36,37]. Additionally, the G>T transversion and A>G transition emerged as the second and third most frequently observed mutations in SNP_TOT and iSNVs, respectively. The high prevalence of G>T mutations in SNP_TOT is particularly noteworthy, as previous studies have highlighted its frequent occurrence and its association with oxidative damage-induced mutagenesis[54], a process that can influence SARS-CoV-2 evolution within individual hosts. This is further supported by the second most prevalent G>T mutations in iSNVs, suggesting its active role in intra-host viral diversity. Furthermore, the A>G transition, which is also prominent in iSNV pools, has been linked to ADAR deaminase activity, another host-driven mechanism affecting viral RNA editing[55]. The mutation types T>C and G>A were slightly over-represented in iSNVs, and may result from ADAR and APOBEC activity respectively, acting on the complement strand on double-stranded RNA[56].Given that SARS-CoV-2 RNA is mostly

120

bioRxiv preprint doi: https://doi.org/10.1101/2025.03.25.645210; this version posted March 26, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

present in its positive-strand form, the C>T transition dominates, while G>A, which reflect APOBEC activity on the negative strand, is less prevalent[57]. Collectively, these three mutation types, C>T, G>T, and A>G, along with the more sporadic yet notable presence of T>C and G>A in iSNVs constitute the dominant mutational signature, underscoring the host antiviral machinery as a driving force in within host evolutionary landscape[58,59].

We next examined the distribution of mutations across codon positions (Figure 3d-e). In iSNV pools, the distribution appeared relatively homogeneous, with no codon position standing out statistically, but a slight dominance of 3rd position, known for tolerating synonymous substitutions, with regard to the redundancy of the genetic code. In SNP pools, mutations preferentially affected the second codon position, leading to a significantly heterogeneous distribution ($p < 0.05$; Chi-square test). This preference was most pronounced in SNP_WH. The more balanced distribution observed in iSNVs is consistent with genetic drift, reflecting a more stochastic process where mutations accumulate as a reservoir of diversity without strong selective constraints. In contrast, the significant enrichment of mutations in the second codon position in SNPs suggests that selection is actively shaping mutation retention. This observation aligns with the fact that second-position mutations are more likely to cause amino acid changes, driving non-synonymous substitutions and potentially influencing viral adaptation.

When analyzing mutation type relative to codon position (Figure 3f), we similarly observed a more even distribution of mutations across positions for iSNV pools, particularly for C>T, A>G, G>T, G>C, and A>C mutations. In contrast, SNP pools exhibited a more localized distribution: for instance, G>A, T>G, and A>C mutations predominantly occurred at the first codon position, while C>T mutations were heavily biased toward the second position in SNPs. In iSNV pools, C>T mutations were more dispersed, occurring across all three codon positions. This visualization highlights fundamental differences in mutational preferences

121

bioRxiv preprint doi: https://doi.org/10.1101/2025.03.25.645210; this version posted March 26, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

between transient within-host variants (iSNVs) and fixed mutations (SNPs). The preferential localization of SNPs in specific codon positions, particularly at the second position, strongly suggests that selection is actively shaping the fixation of certain mutations, while iSNVs act as a more neutral reservoir of genetic diversity.

Across all four mutation pools, a broad distribution of C>T mutations was observed across different nucleotide contexts (Figure 4,S4). In SNP_TOT and SNP_WH, C>T occurred in 16 and 17 different contexts, respectively, with the most represented being ACA, ACT, GCT in the two pools with TTC and CCT also more prevalent in the SNP_TOT pool. In iSNV_TOT and iSNV_WH, C>T is present in 21 and 20 contexts, primarily ACA, ACT, TTC. The selection process may shape the representation of C>T contexts in SNPs, potentially reflecting the mutational signature of APOBEC enzymes, which preferentially target single-stranded RNA in UC/AC motifs[57].

In addition, A>G mutations, associated with ADAR activity, exhibit distinct context preferences. In iSNV_WH, the ATC, TAT, and GAA motifs were the most represented, consistent with ADAR's tendency to target adenosines flanked by uridines or enriched in adenines, or eventually sensitive to GAA motifs[60]. In SNP_WH, A>G mutations predominantly occurred in AAA and GAA. This further supports the role of host-driven RNA editing mechanisms in shaping SARS-CoV-2 intra-host diversity.

Quantitative diversity

As a precursor to the pN/pS analysis, we examined Shannon entropy along the genome, which quantifies mutational diversity. The whole genome entropy was adjusted according to gene coverage (normalized entropy), cumulative values by mutation group are shown in Figure 5c. Entropy was significantly higher in iSNVs from both WH and TOT pools, reinforcing the transient nature of intra-host mutations. Notably, no significant differences were observed between iSNV_WH and iSNV_TOT or between SNP_WH and SNP_TOT (Mann-Whitney U

122

bioRxiv preprint doi: https://doi.org/10.1101/2025.03.25.645210; this version posted March 26, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

test), suggesting that within-host emergence of mutation dominates the mutational landscape, while the contribution of exogenous mutations inherited from parental genomes remains more limited at the whole-genome scale.

We assessed further the impact of evolutionary dynamics at the gene level. Entropy was consistently higher in iSNVs than in SNPs (Figure 5a-b). This pattern reflects the greater transient diversity within the host before fixation.

Notably, the Spike (S) gene exhibited a distinct mutational pattern (Figure 5b). The lowest entropy was observed in iSNV_WH, suggesting strong purifying selection limiting intra-host diversification. In contrast, iSNV_TOT displayed the highest entropy, likely driven by recombination-mediated diversity. Additionally, SNP_WH exhibited low entropy with minimal variation, indicating that few iSNV_WH mutations reached fixation. In contrast, SNP_TOT showed higher entropy than SNP_WH, reinforcing the key role of recombination in shaping the Spike gene's evolution. This suggests an initial phase of mutational influx in Spike (iSNV_TOT), followed by selection acting on beneficial variants (SNP_TOT).

Some genes exhibited low entropy across all pools, indicating strong fixation or evolutionary conservation. These included nsp5, nsp8, nsp9, E, nsp6, ORF6, nsp15, and nsp14.

The nucleocapsid (N) , nsp13, nsp1 and nsp12 gene exhibited very low median entropy in SNPs; however, the observed variability (interquartile range) reflects the heterogeneity across samples (Figure S7). The high outlier in the iSNV-TOT pool for N and ORF8 corresponds to the active recombination phase, during which B.1.1.7 mutations were still present (at AF ≈15%) before being eliminated (Figure S7).

Highly conserved genes

Highly conserved genes throughout the infection course were identified based on low pN/pS ratios, tolerance for high proportion of indeterminate pN/pS values (i.e., cases where both non-synonymous and synonymous mutations were absent, leading to an undefined ratio), and

bioRxiv preprint doi: https://doi.org/10.1101/2025.03.25.645210; this version posted March 26, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

a very limited number of cases where pS = 0 and N > 3 (Non-synonymous), resulting in an infinite ratio (Figure 6 a-b,S6). Across all pools, several genes displayed low pN/pS ratios (M and nsp12) or no fixation of non-synonymous mutations (E, ORF6, ORF8, nsp16, nsp5, nsp9) with high undetermined ratio and no infinity value reported. This highlights a transient emergence of mutations, followed by their elimination, preserving the functional integrity of the proteins. NSP12, NSP7 and NSP8 form the RNA-dependent RNA polymerase (RdRp) holoenzyme[61]. These genes play critical roles in viral replication[61]. During this intra-host evolution, NSP12 retained its non-synonymous mutations inherited from the B.1.160 lineage, including G13993T, C14408T (ancestral), and G15766T (B.1.160). However, among the mutations that emerged within the host, only the synonymous mutation C13506T became fixed, while C13860T was only transiently fixed. This pattern suggests that the optimized changes in NSP12 from the B.1.160 lineage were preserved, while additional non-synonymous intra-host mutations did not persist, reinforcing the high conservation of NSP12 throughout this evolutionary trajectory, given its fundamental role in the SARS-CoV-2 life cycle.

Moreover, nsp14 , nsp15, and nsp5 consistently displayed undetermined pN/pS values (pN = 0, pS = 0) in all pools, and sporadic low pN/pS values (Table 3, Figure 6a-S6). These genes are likely under functional constraints, where mutations that do emerge are rapidly eliminated by purifying selection. These arguments for nsp14 and nsp15 conservation reinforce their critical role in RNA proofreading and frequency of recombination[32] respectively, which is relevant here.

Similarly, nsp6, for which an indeterminate pN/pS ratio has been found on numerous samples in iSNVs (Figure S6), exhibited strong negative pN/pS ratio in SNPs. Consistent with the literature, in SARS-CoV-2 and other coronaviruses, nsp6 has been shown to be highly conserved both in terms of nucleotides and protein sequence[62].

bioRxiv preprint doi: https://doi.org/10.1101/2025.03.25.645210; this version posted March 26, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

To notice, the coverage of some genes was poor, including ORF7a, nsp11 and ORF8 in particular, for which we were unable to obtain valuable information on their selection.

Positive selection and adaptive evolution

We characterized genes under positive selection by a pN/pS ratio >1 and/or the presence of infinite pN/pS values associated with NS>3, to account for the accumulation of non-synonymous likely providing adaptive advantage.

While most SARS-CoV-2 genes were constrained by purifying selection, several exhibited a positive selection, particularly in SNP_WH and SNP_TOT pools, as expected.

One of the strongest signals of positive selection was observed in SNPs from nsp13 (helicase), a gene critical for viral replication and immune evasion. The nsp13 gene displayed variable selective pressures in iSNVs, with most samples exhibiting a pN/pS ratio <1, indicative of purifying selection, and two samples showing an infinite pN/pS ratio (S=0 and N>3) (Table S2, Figure S7). As the infection progressed, positively selected within-host mutations progressively reached fixation in SNP_WH, driving a pN/pS>1 (Figure 6b). This transition from negative selection and neutral drift in iSNVs stages to strong positive selection underscores nsp13 as a key hotspot for adaptive mutations in immunocompromised patients[63,64].

Similarly, nsp4 (replication complex) and ORF9b (an interferon antagonist) displayed strong positive selection, indicating their possible involvement in immune evasion and viral replication optimization. ORF9b, in particular, exhibited infinite pN/pS values in SNP pools, with three non-synonymous mutations detected at D434. Notably, the mutations that appeared in ORF9b were of within-host origin and not of the parental genomes.

While initially under negative selection in iSNVs, the Nucleocapsid gene underwent strong positive selection in both SNP_WH and SNP_TOT pools, particularly in the late stages of infection (Figure 6a-b). This was evidenced by a high number of infinite pN/pS (N>3) values.

125

bioRxiv preprint doi: https://doi.org/10.1101/2025.03.25.645210; this version posted March 26, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

The accumulation of 8 non-synonymous mutations in iSNV_TOT visible on day 229 came from parental genome and recombination process eliminated these B.1.1.7 mutations. These results are consistent with the fact that N is considered a vital hotspot for mutations[65].

Initially, nsp2 showed negative selection or had many samples with undetermined values in iSNVs. Over time, it shifted to positive selection in SNP_WH, with three samples carrying four or more non-synonymous mutations and no synonymous mutations. Similarly, three SNP_TOT

samples showed a pN/pS ratio >1, indicating positive selection emerging late in the infection (from day 431 onward), possibly linked to immune evasion.

At the iSNV_WH stage, only the non-synonymous mutation C2706T was detected, while all other mutations in SNP_WH appeared directly at >50% allele frequency (AF). This suggests a rapid evolutionary process, though it may also result from sampling limitations or transient mutations not being captured.

The case of nsp3: selection analyzed on two levels

During its 440 days of evolution, nsp3 retained three synonymous mutations inherited from its parental genomes: S-C3037T (common ancestor), S-C4543T (B.1.160 lineage), and S-G5629T (B.1.160 lineage). As a result, purifying selection dominated in SNP_TOT (pN/pS <1), largely due to the fixation of these synonymous mutations. However, when focusing exclusively on SNP_WH and iSNVs, positive selection was detected (Table 3, Figure 6a-b). This suggests a fast evolutionary process, where non-synonymous mutations are arising and selected within the host, probably providing an adaptive advantage and consequently are reaching fixation. Hence, nsp3 exemplifies the case in which positive selection favors specific mutations within the host, increasing their allele frequency, while purifying selection prevails still after fixation due to the high number of inherited synonymous mutations.

Optimized selection in the Spike protein

126

bioRxiv preprint doi: https://doi.org/10.1101/2025.03.25.645210; this version posted March 26, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Spike exhibited strong positive selection in SNP_TOT pool, accumulating 6 to 8 non-synonymous mutations over the course of infection while no synonymous mutations were detected, leading to an infinite pN/pS ratio—except for the within host L821L, observed at D229 and D276 (Figure 6b). This suggests that positive selection favored amino acid changes leading to their persistence over time. The evolutionary trajectory of S indicates that mutations inherited from the parental genome B.1.1.7 were strongly selected and retained, while within-host mutations were less contributive. The iSNV_WH pool exhibited variable pN/pS, sporadically >1, and signs of positive selection (Figure 6b). This positive selection seemed weak on promoting the transition from iSNV_WH to SNP_WH; less sign of positive selection was observed for SNP_WH, highlighting a strong purifying filter that eliminated non-optimal variations.

This pattern suggests that recombination-introduced mutations had already undergone prior filtering, leading to a highly optimized combination in SNP_TOT. Spike transitioned from genetic drift in iSNVs (negative selection with occasional positive selection in iSNV_WH) to strong positive selection in SNP_TOT, reinforcing the idea that only the most functionally advantageous mutations were fixed, shaping a highly refined adaptive profile.

Gene-scale optimized Spike configuration

During this evolution time period, five within host SNPs have arisen in the Spike (L5F, D796H, L821L, T618K, T572I), and some were evolutive convergence events (Figure 7a). L5F, a potentially CTL escape mutation, is located at a site recognized by Human Leukocyte Antigen (HLA) molecules, which are key players in presenting viral peptides to T cells for immune recognition[66]. L5F has been found in numerous VOCs such as B.1.177 or B.1.526 and has been shown to increase viral infectivity[67]. This mutation arose in parallel with D796H in the S2 Spike subunit at the first stage of infection, before being replaced by the Spike B.1.1.7 gene during the recombination process. D796H, a signature mutation of the variant of

127

bioRxiv preprint doi: https://doi.org/10.1101/2025.03.25.645210; this version posted March 26, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

concern B.1.3.118, appeared within the host in the first described immunocompromised long term shedding patient[5], following convalescent plasma therapy. As the patient received convalescent plasma therapy before the detection of this mutation (Figure 1a), it is possible that we faced a similar situation, promoted by exogenous antibodies. After recombination of the original Spike gene, the mutations T618K and T572I arose gradually (Figure 7a). T572I is a convergent mutation also present in the AY.32 variant while T618K is a rare substitution. The consensus total mutations observed in the Spike protein are illustrated in Figure 7b.

Molecular-scale optimized Spike configuration

As part of our analysis, we also studied the surface potential to gain a clear understanding of the protein-level impact and cell interaction resulting from the recombination and within host mutations observed (Figure 7c). Regarding within-host SNPs, an increase of electropositive surface is associated with T618K (upper panel) and D796H (lower panel), no additional significant changes have been observed regarding within host mutations. A first recombination event is visible at D259, in the N-terminal domain (NTD), where deletions 144 and 69-70 (from lineage B.1.1.7) appear, decreasing the NTD surface area facing the host cell plasma membrane (Figure 7c lower panel). This optimizes the virus's initial interaction with lipid rafts[68]. The NTD condensation induced by deletions 144 and 69-70 is clearly visible in the lower panel of figure 6. At D285, the B.1.1.7 Spike variant is fully recombined, and the N501Y mutation is observed in the receptor-binding domain (RBD), permanently replacing the S477N mutation (present in B.1.160), thus increasing affinity for ACE2[69]. The rearrangement of the RBD surface is visible in the upper panel of figure 7. Other mutations include D614G, common to both variants B.1.160 and B.1.1.7, which facilitates the conformational unmasking of the RBD in the trimeric Spike[70], S982A, which stabilizes this unmasked "up" conformation of the RBD, A570D and P681H, which may alter the Furin cleavage site. The D1118H mutation has a stabilizing effect on SARS-CoV-2 virions[71]. The

128

bioRxiv preprint doi: https://doi.org/10.1101/2025.03.25.645210; this version posted March 26, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

NTD condensation induced by deletions 144 and 69-70 is clearly visible in the lower panel. The RBD surface rearrangement induced by the N501Y mutation is visible in the upper panel. Recombination has thus fixed a new assembly of upstream-optimized mutations, leading to subtle adjustments to the electrostatic surface potential. The deletions at positions 144 and 69-70 reduce the surface area of the NTD facing the host plasma membrane, optimizing the virus initial interaction with lipid rafts[47]. The absence of mutations linked to immune escape (e.g., K147 or R246) suggests that this patient did not produce neutralizing antibodies targeting the NTD, consistent with the patient's immunocompromised status[72] with IgG serology was negative as of May 25, 2021 (Table 1). The evolutionary significance of mutations in the Spike protein is closely related to their structural localization. While the NTD reshaping enhances its electrostatic binding to negatively charged membrane components, the RBD mutations (S477N and N501Y) increase Spike's affinity for the ACE2 receptor, with N501Y providing a stronger advantage over S477N[69]). This recombination-driven selection likely optimizes virus-cell interactions, independent of immune escape processes, demonstrating the evolutionary benefits of such mutations.

Convergent detection

Given the potential for SARS-CoV-2 long-term infection to generate new variants[51], we investigated if within host SNPs selected for having reached or exceeded a 70% allele frequency, indicating for a strong tendency toward fixation, could be found in the general population and could therefore represent evolutionary convergences. We searched the GISAID database for their prevalence worldwide (Figure 8, Table S3), while it is also important to bear in mind that these results are subject to limitations due to sampling bias and inconsistent submission of data to GISAID. 20.31% of the selected within host SNPs were found in VOC (Variant of Concern) signature mutations, and 2 substitutions, C26858T, a synonymous substitution present in the Membrane gene and A28271T, a non-coding region substitution, were found re-

bioRxiv preprint doi: https://doi.org/10.1101/2025.03.25.645210; this version posted March 26, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

spectively 22% and 52% of all sequences. The massive proportion of the C26858T synonymous substitution could be the result of optimized codon usage. Although UUC is more optimized in terms of codon usage, with codon frequencies of 20.3 for UUC and 17.6 for UUU[73], the widespread replacement by UUU may confer advantages related to RNA structure or stability, suggesting an adaptation during within-host viral evolution. This change might enhance viral fitness through mechanisms beyond simple codon optimization, such as improving RNA folding or influencing the regulation of translation. The mutation A28271T is in the N gene Kozak translation initiation box[74] and the A to T substitution is responsible for a reduced protein expression of the ORF9b gene, and therefore increases the production of the N protein relative to the ORF9b protein. A similar effect was observed with the B.1.1.7 deletion A28270- [39]. The parental strain B.1.1.7 that infected the patient following the initial infection carried the A28270- deletion, within the Kozak sequence. However, this deletion was not maintained during the intra host evolutionary process, as it was detected at an allele frequency of 15% on day 229 (Figure S2) before being completely eliminated. The substitution A28271T, emerged at the same position, potentially restoring the functional advantage previously provided by the deletion.

The wild type-like restoration of ORF3a and M

From D0 to D329, the two parental mutations remained detectable but were subsequently eliminated (Figure S7). While this could result from a reversion event, it is more likely due to recombination, given that five mutations disappeared completely and did not reappear in later samples: NS-25460T (WH), S-G25563T (B.1.160), S-C25710T (B.1.160), NS-C26204T (WH). Following this elimination, only two additional mutations—one synonymous (S) and one non-synonymous (NS), both within-host (WH) transiently reached the fixation threshold (50%) before disappearing: NS-C25517T, which peaked at 61.9% (D432) before being lost

130

bioRxiv preprint doi: https://doi.org/10.1101/2025.03.25.645210; this version posted March 26, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

and S-C25515A, which briefly reached 83.2% (D330) before vanishing. . This case represents a scenario where recombination facilitated a reversion to the WT ORF3a .

For the Membrane gene, we observed a similar situation: the parental B.1.160 S-C26735T and S-T26876C mutations were eventually eliminated at day 434 and the S-C26858T (WH) reached fixation.

Impact of intra-host evolution on the ORF8 gene

The accessory protein ORF8 plays a crucial role in the pathogenesis of SARS-CoV-2 due to its secretory nature and high plasticity, which allow it to interact with a wide array of host proteins, making it a key player in both viral replication and immune evasion[75,76,77]. In sample D378, a PCR analysis (Figure 9a) of the genomic region spanning 24,813-29,074 revealed a 126-nucleotides deletion from positions 27,915 to 28,040 (see supplementary Material and Methods). This within host deletion, which gradually became the dominant variant after 378 days of infection resulted in a loss of 42 amino acids, accounting for 34.43% of the ORF8 protein. Similar truncated forms of ORF8 have been observed in variants such as B.1.1.7, where a stop codon was introduced[76]. This 126-nt deletion in ORF8 was fortuitously discovered following a targeted PCR covering this region (Figure 9a) but was initially identified as an uncovered region in consensus genomes (Figure 9b). The deletion was obtained by the read assembly (Figure 9b), which was able to circumvent the analysis bias caused by the use of 50 bp read lengths in the mapping strategy. Indeed, in mapping approaches, reads are directly aligned to a reference genome, serving as a backbone, and deletions are inferred from gaps or reduced coverage. If a deletion is larger than the length of the reads (for example, a 126-nt deletion compared with 50 bp reads), the reads covering the newly deleted region may not align correctly and may be rejected as ambiguous, leading to an uncovered region in the consensus genome. In contrast, *de novo* assembly reconstructs the genome without relying on a reference sequence, enabling the detection of novel structural

bioRxiv preprint doi: https://doi.org/10.1101/2025.03.25.645210; this version posted March 26, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

variations such as large deletions. This issue mirrors the challenges identified by Brandt et al.[78] who showed that a large deletion (168 nt) in ORF8 was not detected by variant calling pipelines and standard sequencing technologies.

DISCUSSION

This case study illustrates the dynamic evolution of SARS-CoV-2 within an immunocompromised patient, initially infected with the B.1.160 variant and later with B.1.1.7. The combination of immunodeficiency and the treatments received, such as convalescent plasma, created a favorable environment for accelerated viral evolution. The intra-host evolution rate observed in this case (1.39 x 10^-3 substitutions per site per year) was significantly higher than the inter-host rate (9.02 x 10^-4 substitutions per site per year), aligning with previous findings[5,16,51,79] and relies only partially on recombination from parental genomes. The temporal overlap and coexistence of distinct genotypes within the host suggest the circulation of viral subpopulations across different anatomical sites[80,49,6,50]. Although this study is based on a single case, its significance is reinforced by the exceptionally long duration of the reported infection (440 days), making it one of the longest documented cases of SARS-CoV-2 persistence in an immunocompromised patient. The robustness of the analysis is further supported by the availability of high-quality samples (12 whole genomes sequences with more than 80% coverage at 50x depth).

However, our study has several limitations, the first being that as a retrospective study, we were not able to plan sequencing data points or analyzing other anatomical sites of interest such the lower respiratory system, as sample collection was dependent of previously banked specimens. Consequently, some key aspects of viral evolution, particularly the very initial stages of recombination could not be detailed. Reconstructing multiple recombination events through haplotype analysis, including within-host mutations, may have provided valuable

bioRxiv preprint doi: https://doi.org/10.1101/2025.03.25.645210; this version posted March 26, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

insights and could represent an interesting working perspective, giving the challenge due to the short read length (50 bp), which restricts the ability to confidently link adjacent mutations within the same haplotype.

Another limitation concerns the restricted analysis of the immune response. Only three serological data points were available, which limited our ability to assess how immune pressure may have shaped viral evolution. A more comprehensive approach, including HLA typing and T-cell response characterization, could have provided deeper insights into the mechanisms driving the evolutionary processes at play. Since serology was negative from the end of May, evolutionary immune-pressure came essentially from innate and T-cell immunity since June. Once again, the retrospective nature of the study prevented us to explore these aspects in greater depth.

Patients with prolonged infections have been shown to accumulate more intra-host single nucleotide variants (iSNVs) than patients with shorter infections[63], and should not be neglected as critical insights into the mechanisms driving intra-host viral evolution and the emergence of new variants[81,34,82,83]. Given the potential bias introduced by sequencing errors, we focused on iSNVs with allele frequencies between 5% and 50%, used iVar, a highly accurate tool for detecting intra-host single-nucleotide variants[84] and applied optimal depth and quality thresholds to ensure reliable results.

In our study, some genes showed very few iSNV-WH but high SNP_WH, as they may have reached fixation too rapidly for detection, likely due to sampling limitations. Additionally, setting a MAF threshold at 5% may have restricted the ability to capture early-stage evolutionary dynamics, though this choice was necessary to ensure data reliability while minimizing false positives.

Stanevich et al[79]. demonstrated the emergence of mutations that facilitated immune evasion, including S:del141-144, associated with reduced antigen presentation by both HLA class I and

bioRxiv preprint doi: https://doi.org/10.1101/2025.03.25.645210; this version posted March 26, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

673 II molecules, S:R273S, nsp3:T1456I and N:P6T, previously identified as immunogenic
674 peptides targeted by T cells, which induced the loss of their ability to elicit T-cell responses.
675 Comparing our findings with subsequent studies would help assess whether similar immune
676 escape mechanisms could be at play in this prolonged infection.

677 The Envelope (E) and Membrane (M) remained highly conserved throughout the 440-day
678 infection, suggesting strong functional constraints preserving their roles in viral assembly and
679 exit. Similarly, nsp12 displayed high iSNV entropy but remained strongly constrained at the
680 SNP level, aligning with its essential function in replication, where excessive mutations are
681 often deleterious. Interestingly, E was previously reported to evolve within-host in another
682 long-term infection case[79] whereas in our study, it remained mostly conserved. This
683 discrepancy could stem from selective pressures and HLA diversity in distinct host
684 environments, warranting further investigation.

685 Several genes exhibited positive selection, including Spike, N (nucleocapsid), ORF9b, nsp13,
686 nsp4, nsp2, and nsp3, highlighting their roles in immune evasion, and adaptation to the host
687 environment.

688 N and ORF9b, known to interfere with type I and III interferon (IFN) responses, showed
689 strong positive selection, reinforcing their role in modulating innate immunity[85,86] . Moreover,
690 nsp2 underwent positive selection too, likely driven by interactions with IFN-β[87] in the same
691 way. Nsp13 (helicase) has been shown to accumulate mutation in immunocompromised
692 patients[63], possibly due to its role in host-pathogen interactions and immune modulation.
693 Nsp3, which plays a role in innate immune modulation, displayed accumulated non-
694 synonymous mutations, similar to findings by Stanevich et al.[79], where mutations in this
695 region led to altered recognition by T lymphocytes.

696 ORF3a is an abundantly accumulated accessory protein facilitating virus exit via lysosome
697 deacidification[88,89]. Interestingly, recombination led to a reversion to a wild-type-like

134

bioRxiv preprint doi: https://doi.org/10.1101/2025.03.25.645210; this version posted March 26, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

sequence, suggesting a selective advantage in retaining its ancestral functional state. The loss of multiple mutations in a coordinated manner suggests a genomic replacement event, likely driven by recombination, which may help restore a more genetically stable ORF3a[90] configuration which needs further investigation giving that the ORF3a has been implicated in modulating the host immune response, notably by interacting with key proteins such as IKKβ and NEMO in the NF-κB signaling pathway[89]. This finding raises the possibility that some mutations associated with reduced pathogenicity, such as those seen in Omicron lineages[91], could be reversible through recombination. Interestingly, the Membrane gene showed the same return to wild-type sequence by recombination. Additionally, since the E gene did not exhibit mutations from either parental strain, it is possible that it was not subject to recombination

The massively predominant C>T mutations encountered both, in iSNVs and SNPs (from within host and parental genomes origin) is known to be induced by the host APOBEC-mediated RNA editing family and has been extensively documented[92,54,93]in the global population and immunocompromised long carrier patients. APOBEC targets cytosine for deamination resulting in G>A mutations on the complementary strand during viral replication[94]. This mechanism, which involves the conversion of cytidines to uridines, likely contributes to increased viral plasticity, aiding the virus adaptation to the host[58,37]A>G and G>T were the two types the most represented among within host iSNVs. G>T mutations are frequently observed globally[95,96] and are linked with oxidative stress caused by reactive oxygen species[59]. In addition, the A>G transition mediated by the ADAR protein has also contributed to the emergence of within host mutations, further emphasizing the importance of host-pathogen interactions in SARS-CoV-2's intra-host evolution[55,97]. The dominance of A>G, C > T, and G > T mutations during intra-host evolution has been observed in cohorts across the world, suggesting that these molecular mechanisms are consistently at play in shaping

135

bioRxiv preprint doi: https://doi.org/10.1101/2025.03.25.645210; this version posted March 26, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

SARS-CoV-2's evolution [81] .

Large deletions in ORF8 have also been reported in other lineages[98]. It is known that the presence of a full-length ORF8 is not essential for viral replication, and the loss of this gene may even confer a selective advantage by facilitating immune evasion or enhancing viral persistence. This phenomenon of "adaptive gene loss" has been described in SARS-CoV-2, where the loss of ORF8 can provide benefits to the virus, likely by reducing the immune system's ability to detect and respond to infection[99]. This highlights the evolutionary flexibility of SARS-CoV-2, with ORF8 deletions serving as a potential adaptive mechanism during prolonged intra-host viral evolution.

Finally, this study emphasizes the need for comparative genomic analyses of SARS-CoV-2 in long-term carriers to deepen our understanding of the virus's adaptive ability within the host. Gaining insights into host-virus interactions is crucial for elucidating the mechanisms that enable prolonged persistence, even in the absence of severe immunosuppression, as suggested in cases of long COVID[100,101,102] , a persistent global public health concern[101]. Understanding how the virus adapts at genomic, proteomic and structural levels will be instrumental in identifying novel therapeutic targets, ultimately contributing to the development of more effective treatment strategies as genomic diversity accumulation could act as a potential mechanism for viral persistence.

Acknowledgments

We are grateful to the IHU NGS platform, Claudia Andrieu, Sofiane Bakhour, Amira Doudou, Idir Kacel and Vincent Bossi. We are also grateful to Jeremy Delerce for providing data on circulating variants of B.1.1.7. We gratefully acknowledge all the contributors of the sequences deposited in GISAID used on the global population SARS-CoV-2 tree.

bioRxiv preprint doi: https://doi.org/10.1101/2025.03.25.645210; this version posted March 26, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Author contributions

Conceptualization: EB, PP, PEF ; Contributed materials/analysis tools: EB, JF ; Analyzed the data: EB, JF ; Writing: EB, JF; review & editing: PP, PEF Supervision: PEF

All authors have read and agreed to the published version of the manuscript.

Funding

This work was supported by the French Government under the "Investments for the Future" program managed by the National Agency for Research (ANR), Mediterranean-Infection 10-IAHU-03 and was also supported by Region "Provence Alpes Côte d'Azur" and European funding FEDER PRIMMI (Fonds Européen de Développement Régional-Plateformes de Recherche et d'Innovation Mutualisées Méditerranée Infection), FEDER PA 0000320 PRIMMI.

Conflicts of interest

The authors have no conflicts of interest to declare. Funding sources had no role in the design and conduct of the study; collection, management, analysis, and interpretation of the data; and preparation, review, or approval of the manuscript.

REFERENCES

1. Moelling, K. Within-Host and Between-Host Evolution in SARS-CoV-2—New Variant's Source. *Viruses* **13**, 751 (2021).
2. Burki, T. The origin of SARS-CoV-2 variants of concern. *Lancet Infect. Dis.* **22**, 174–175 (2022).
3. Peacock, T. P., Penrice-Randal, R., Hiscox, J. A. & Barclay, W. S. SARS-CoV-2 one year on: evidence for ongoing viral adaptation. *J. Gen. Virol.* **102**, (2021).
4. Li, P., de Vries, A. C., Kamar, N., Peppelenbosch, M. P. & Pan, Q. Monitoring and managing SARS-CoV-2 evolution in immunocompromised populations. *Lancet Microbe* **3**, e325–e326 (2022).
5. Kemp, S. A. *et al.* SARS-CoV-2 evolution during treatment of chronic infection. *Nature* **592**, 277–282

137

bioRxiv preprint doi: https://doi.org/10.1101/2025.03.25.645210; this version posted March 26, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

(2021).

6. Chaguza, C. *et al.* Accelerated SARS-CoV-2 intrahost evolution leading to distinct genotypes during chronic infection. *Cell Rep. Med.* **4**, 100943 (2023).
7. Fung, M. & Babik, J. M. COVID-19 in Immunocompromised Hosts: What We Know so Far. *Clin. Infect. Dis.* **72**, 340–350 (2021).
8. Lythgoe, K. A. *et al.* SARS-CoV-2 within-host diversity and transmission. *Science* **372**, eabg0821 (2021).
9. Valesano, A. L. *et al.* Temporal dynamics of SARS-CoV-2 mutation accumulation within and across infected hosts. *PLoS Pathog.* **17**, (2021).
10. Avanzato, V. A. *et al.* Case Study: Prolonged Infectious SARS-CoV-2 Shedding from an Asymptomatic Immunocompromised Individual with Cancer. *Cell* **183**, 1901-1912.e9 (2020).
11. Scherer, E. M. *et al.* SARS-CoV-2 Evolution and Immune Escape in Immunocompromised Patients. *N. Engl. J. Med.* **386**, 2436–2438 (2022).
12. Marques, A. D. *et al.* SARS-CoV-2 evolution during prolonged infection in immunocompromised patients. *mBio* **15**, e00110-24 (2024).
13. Raglow, Z. *et al.* SARS-CoV-2 shedding and evolution in patients who were immunocompromised during the omicron period: a multicentre, prospective analysis. *Lancet Microbe* **5**, e235–e246 (2024).
14. Lynch, M. *et al.* Genomic Evolution of SARS-CoV-2 Virus in Immunocompromised Patient, Ireland. *Emerg. Infect. Dis.* **27**, 2499–2501 (2021).
15. Quaranta, E. G. *et al.* SARS-CoV-2 intra-host evolution during prolonged infection in an immunocompromised patient. *Int. J. Infect. Dis.* (2022) doi:10.1016/j.ijid.2022.06.023.
16. Choi, B. *et al.* Persistence and Evolution of SARS-CoV-2 in an Immunocompromised Host. *N. Engl. J. Med.* **383**, 2291–2293 (2020).
17. Borges, V. *et al.* Long-Term Evolution of SARS-CoV-2 in an Immunocompromised Patient with Non-Hodgkin Lymphoma. *mSphere* **6**, (2021).
18. Heyer, A. *et al.* Remdesivir-induced emergence of SARS-CoV2 variants in patients with prolonged infection. *Cell Rep. Med.* **3**, 100735 (2022).
19. Focosi, D., Quiroga, R., McConnell, S., Johnson, M. C. & Casadevall, A. Convergent Evolution in SARS-CoV-2 Spike Creates a Variant Soup from Which New COVID-19 Waves Emerge. *Int. J. Mol. Sci.* **24**, 2264 (2023).
20. Pedro, N. *et al.* Dynamics of a Dual SARS-CoV-2 Lineage Co-Infection on a Prolonged Viral Shedding

bioRxiv preprint doi: https://doi.org/10.1101/2025.03.25.645210; this version posted March 26, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

805 COVID-19 Case: Insights into Clinical Severity and Disease Duration. *Microorganisms* **9**, 300 (2021).

806 21. Combes, P. *et al.* Evidence of co-infections during Delta and Omicron SARS-CoV-2 variants co-circulation
807 through prospective screening and sequencing. *Clin. Microbiol. Infect.* **28**, 1503.e5-1503.e8 (2022).

808 22. Francisco Junior, R. D. S. *et al.* Emergence of Within-Host SARS-CoV-2 Recombinant Genome After
809 Coinfection by Gamma and Delta Variants: A Case Report. *Front. Public Health* **10**, 849978 (2022).

810 23. Lai, M. M. *et al.* Recombination between nonsegmented RNA genomes of murine coronaviruses. *J. Virol.*
811 **56**, 449–456 (1985).

812 24. Gribble, J. *et al.* The coronavirus proofreading exoribonuclease mediates extensive viral recombination.
813 *PLOS Pathog.* **17**, e1009226 (2021).

814 25. Wells, H. L. *et al.* The coronavirus recombination pathway. *Cell Host Microbe* **31**, 874–889 (2023).

815 26. Turakhia, Y. *et al.* Pandemic-scale phylogenomics reveals the SARS-CoV-2 recombination landscape.
816 *Nature* **609**, 994–997 (2022).

817 27. Sekizuka, T. *et al.* Genome Recombination between the Delta and Alpha Variants of Severe Acute
818 Respiratory Syndrome Coronavirus 2 (SARS-CoV-2). *Jpn. J. Infect. Dis.* **75**, 415–418 (2022).

819 28. Wertheim, J. O. *et al.* Detection of SARS-CoV-2 intra-host recombination during superinfection with Alpha
820 and Epsilon variants in New York City. *Nat. Commun.* **13**, 3645 (2022).

821 29. Lacek, K. A. *et al.* SARS-CoV-2 Delta–Omicron Recombinant Viruses, United States. *Emerg. Infect. Dis.*
822 **28**, 1442–1445 (2022).

823 30. Perez-Florido, J. *et al.* Detection of High Level of Co-Infection and the Emergence of Novel SARS CoV-2
824 Delta-Omicron and Omicron-Omicron Recombinants in the Epidemiological Surveillance of Andalusia. *Int.*
825 *J. Mol. Sci.* **24**, 2419 (2023).

826 31. Singh, P., Sharma, K., Shaw, D., Bhargava, A. & Negi, S. S. Mosaic Recombination Inflicted Various
827 SARS-CoV-2 Lineages to Emerge into Novel Virus Variants: a Review Update. *Indian J. Clin. Biochem.*
828 **38**, 418–425 (2023).

829 32. Zhou, Y. *et al.* SARS-CoV-2 EndoU-ribonuclease regulates RNA recombination and impacts viral fitness.
830 Preprint at https://doi.org/10.1101/2024.11.11.622995 (2024).

831 33. Karamitros, T. *et al.* SARS-CoV-2 exhibits intra-host genomic plasticity and low-frequency polymorphic
832 quasispecies. *J. Clin. Virol.* **131**, (2020).

833 34. Sun, F. *et al.* SARS-CoV-2 Quasispecies Provides an Advantage Mutation Pool for the Epidemic Variants.
834 *Microbiol. Spectr.* **9**, e00261-21 (2021).

139

bioRxiv preprint doi: https://doi.org/10.1101/2025.03.25.645210; this version posted March 26, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

35. Jary, A. *et al.* Evolution of viral quasispecies during SARS-CoV-2 infection. *Clin. Microbiol. Infect.* **26**, 1560.e1-1560.e4 (2020).

36. Kim, K. *et al.* The roles of APOBEC-mediated RNA editing in SARS-CoV-2 mutations, replication and fitness. *Sci. Rep.* **12**, 14972 (2022).

37. Nakata, Y. *et al.* Cellular APOBEC3A deaminase drives mutations in the SARS-CoV-2 genome. *Nucleic Acids Res.* **51**, 783–795 (2023).

38. Burel, E. *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. *Viruses* **14**, 1266–1266 (2022).

39. Yang, J. *et al.* A Noncoding A-to-U Kozak Site Change Related to the High Transmissibility of Alpha, Delta, and Omicron VOCs. *Mol. Biol. Evol.* **40**, msad142 (2023).

40. Naderi, S., Sagan, S. M. & Shapiro, B. J. Within-host genetic diversity of SARS-CoV-2 across animal species. *Virus Evol.* **11**, veae117 (2025).

41. Hadfield, J. *et al.* Nextstrain: real-time tracking of pathogen evolution. *Bioinformatics* **34**, 4121–4123 (2018).

42. Sagulenko, P., Puller, V. & Neher, R. A. TreeTime: Maximum-likelihood phylodynamic analysis. *Virus Evol.* **4**, (2018).

43. Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. *Mol. Biol. Evol.* (1986) doi:10.1093/oxfordjournals.molbev.a040410.

44. Fantini, J., Yahi, N., Azzaz, F. & Chahinian, H. Structural dynamics of SARS-CoV-2 variants: A health monitoring strategy for anticipating Covid-19 outbreaks. *J. Infect.* **83**, 197–206 (2021).

45. Colson, P. *et al.* Culture and identification of a “Deltamicron” SARS□CoV□2 in a three cases cluster in southern France. *J. Med. Virol.* **94**, 3739–3749 (2022).

46. Di Scala, J. In Silico/In Vitro Approaches for the Identification of Functional Cholesterol-Binding Domains in Membrane Proteins. (2017).

47. Fantini, J., Azzaz, F., Chahinian, H. & Yahi, N. Electrostatic Surface Potential as a Key Parameter in Virus Transmission and Evolution: How to Manage Future Virus Pandemics in the Post-COVID-19 Era. *Viruses* **15**, 284 (2023).

48. Robinson, J. T., Thorvaldsdóttir, H., Wenger, A. M., Zehir, A. & Mesirov, J. P. Variant Review with the Integrative Genomics Viewer. *Cancer Res.* **77**, e31–e34 (2017).

bioRxiv preprint doi: https://doi.org/10.1101/2025.03.25.645210; this version posted March 26, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

865 49. Wang, Y. *et al.* Intra-host variation and evolutionary dynamics of SARS-CoV-2 populations in COVID-19
866 patients. *Genome Med.* **13**, (2021).
867 50. Farjo, M. *et al.* Within-host evolutionary dynamics and tissue compartmentalization during acute SARS-
868 CoV-2 infection. *J. Virol.* **98**, e01618-23 (2024).
869 51. Harari, S. *et al.* Drivers of adaptive evolution during chronic SARS-CoV-2 infections. *Nat. Med.* **28**, 1501–
870 1508 (2022).
871 52. Lauring, A. S. Within-Host Viral Diversity: A Window into Viral Evolution. *Annu. Rev. Virol.* **7**, 63–81
872 (2020).
873 53. Li, Y. *et al.* C-to-U RNA deamination is the driving force accelerating SARS-CoV-2 evolution. *Life Sci.*
874 *Alliance* **6**, e202201688 (2023).
875 54. Simmonds, P. Rampant C→U Hypermutation in the Genomes of SARS-CoV-2 and Other Coronaviruses:
876 Causes and Consequences for Their Short- and Long-Term Evolutionary Trajectories. *mSphere* **5**, e00408-
877 20 (2020).
878 55. Ringlander, J. *et al.* Impact of ADAR-induced editing of minor viral RNA populations on replication and
879 transmission of SARS-CoV-2. *Proc. Natl. Acad. Sci.* **119**, e2112663119 (2022).
880 56. Mourier, T. *et al.* Host-directed editing of the SARS-CoV-2 genome. *Biochem. Biophys. Res. Commun.* **538**,
881 35–39 (2021).
882 57. Azgari, C., Kilinc, Z., Turhan, B., Circi, D. & Adebali, O. The Mutation Profile of SARS-CoV-2 Is
883 Primarily Shaped by the Host Antiviral Defense. *Viruses* **13**, 394 (2021).
884 58. Lamb, K. D. *et al.* Mutational signature dynamics indicate SARS-CoV-2's evolutionary capacity is driven
885 by host antiviral molecules. *PLOS Comput. Biol.* **20**, e1011795 (2024).
886 59. Graudenzi, A., Maspero, D., Angaroni, F., Piazza, R. & Ramazzotti, D. Mutational signatures and
887 heterogeneous host response revealed via large-scale characterization of SARS-CoV-2 genomic diversity.
888 *iScience* **24**, 102116 (2021).
889 60. Doherty, E. E. *et al.* ADAR activation by inducing a *syn* conformation at guanosine adjacent to an editing
890 site. *Nucleic Acids Res.* **50**, 10857–10868 (2022).
891 61. Singh, D. *et al.* Redefining NSP12 activity in SARS-CoV-2 and its regulation by NSP8 and NSP7. *Mol.*
892 *Ther. Nucleic Acids* **36**, 102452 (2025).
893 62. Zhang, C. *et al.* SARS-CoV-2 NSP6 reduces autophagosome size and affects viral replication via sigma-1
894 receptor. *J. Virol.* **98**, e00754-24 (2024).

141

bioRxiv preprint doi: https://doi.org/10.1101/2025.03.25.645210; this version posted March 26, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

63. Voloch, C. M. *et al.* Intra-host evolution during SARS-CoV-2 prolonged infection. *Virus Evol.* **7**, (2021).

64. Jia, Z. *et al.* Delicate structural coordination of the Severe Acute Respiratory Syndrome coronavirus Nsp13 upon ATP hydrolysis. *Nucleic Acids Res.* **47**, 6538–6550 (2019).

65. Ye, Q., West, A. M. V., Silletti, S. & Corbett, K. D. Architecture and self□assembly of the SARS□CoV □2 nucleocapsid protein. *Protein Sci.* **29**, 1890–1901 (2020).

66. Ahmad, W., Ahmad, S. & Basha, R. Analysis of the mutation dynamics of SARS-CoV-2 genome in the samples from Georgia State of the United States. *Gene* **841**, 146774 (2022).

67. Li, Q. *et al.* The Impact of Mutations in SARS-CoV-2 Spike on Viral Infectivity and Antigenicity. *Cell* **182**, 1284-1294.e9 (2020).

68. Fantini, J., Chahinian, H. & Yahi, N. Convergent Evolution Dynamics of SARS-CoV-2 and HIV Surface Envelope Glycoproteins Driven by Host Cell Surface Receptors and Lipid Rafts: Lessons for the Future. *Int. J. Mol. Sci.* **24**, 1923 (2023).

69. Barton, M. I. *et al.* Effects of common mutations in the SARS-CoV-2 Spike RBD and its ligand, the human ACE2 receptor on binding affinity and kinetics. *eLife* **10**, e70658 (2021).

70. Guérin, P. *et al.* Structural Dynamics of the SARS-CoV-2 Spike Protein: A 2-Year Retrospective Analysis of SARS-CoV-2 Variants (from Alpha to Omicron) Reveals an Early Divergence between Conserved and Variable Epitopes. *Molecules* **27**, 3851 (2022).

71. Magazine, N. *et al.* Mutations and Evolution of the SARS-CoV-2 Spike Protein. *Viruses* **14**, 640 (2022).

72. Jaafar, R. *et al.* High Individual Heterogeneity of Neutralizing Activities against the Original Strain and Nine Different Variants of SARS-CoV-2. *Viruses* **13**, 2177 (2021).

73. Roy, A. *et al.* Base Composition and Host Adaptation of the SARS-CoV-2: Insight From the Codon Usage Perspective. *Front. Microbiol.* **12**, 548275 (2021).

74. Kozak, M. Possible role of flanking nucleotides in recognition of the AUG initiator codon by eukaryotic rnbosomes. *Nucleic Acids Res.*

75. Arduini, A., Laprise, F. & Liang, C. SARS-CoV-2 ORF8: A Rapidly Evolving Immune and Viral Modulator in COVID-19. *Viruses* **15**, 871 (2023).

76. Cecchetto, R. *et al.* Detection of SARS-CoV-2 Δ426 ORF8 Deletion Mutant Cluster in NGS Screening. *Microorganisms* **11**, 2378 (2023).

77. Móvio, M. I. *et al.* SARS-CoV-2 ORF8 as a Modulator of Cytokine Induction: Evidence and Search for Molecular Mechanisms. *Viruses* **16**, 161 (2024).

bioRxiv preprint doi: https://doi.org/10.1101/2025.03.25.645210; this version posted March 26, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

925 78. Brandt, D. *et al*. Multiple Occurrences of a 168-Nucleotide Deletion in SARS-CoV-2 ORF8, Unnoticed by
926 Standard Amplicon Sequencing and Variant Calling Pipelines. *Viruses* **13**, 1870 (2021).
927 79. Stanevich, O. V. *et al*. SARS-CoV-2 escape from cytotoxic T cells during long-term COVID-19. *Nat.*
928 *Commun.* **14**, 149 (2023).
929 80. Rueca, M. *et al*. Compartmentalized Replication of SARS-Cov-2 in Upper vs. Lower Respiratory Tract
930 Assessed by Whole Genome Quasispecies Analysis. *Microorganisms* **8**, 1302 (2020).
931 81. Armero, A., Berthet, N. & Avarre, J.-C. Intra-Host Diversity of SARS-Cov-2 Should Not Be Neglected:
932 Case of the State of Victoria, Australia. *Viruses* **13**, 133 (2021).
933 82. Kimbrel, J. *et al*. Multiple Mutations Associated with Emergent Variants Can Be Detected as Low-
934 Frequency Mutations in Early SARS-CoV-2 Pandemic Clinical Samples. *Viruses* **14**, 2775 (2022).
935 83. Domingo, E. *et al.* A general and biomedical perspective of viral quasispecies. *RNA* **31**, 429–443 (2025).
936 84. Grubaugh, N. D. *et al*. An amplicon-based sequencing framework for accurately measuring intrahost virus
937 diversity using PrimalSeq and iVar. *Genome Biol.* **20**, 8 (2019).
938 85. Han, L. *et al.* SARS□CoV□2 ORF9b antagonizes type I and III interferons by targeting multiple
939 components of the RIG□I/MDA□5–MAVS, TLR3–TRIF, and cGAS–STING signaling pathways. *J. Med.*
940 *Virol.* **93**, 5376–5389 (2021).
941 86. Chen, K. *et al*. SARS-CoV-2 Nucleocapsid Protein Interacts with RIG-I and Represses RIG-Mediated IFN-
942 β Production. *Viruses* **13**, 47 (2020).
943 87. Thorne, L. G. *et al*. Evolution of enhanced innate immune evasion by SARS-CoV-2. *Nature* **602**, 487–495
944 (2022).
945 88. Zhang, J. *et al*. SARS-CoV-2 ORF3a Protein as a Therapeutic Target against COVID-19 and Long-Term
946 Post-Infection Effects. *Pathogens* **13**, 75 (2024).
947 89. Azad, G. K. & Khan, P. K. Variations in Orf3a protein of SARS-CoV-2 alter its structure and function.
948 *Biochem. Biophys. Rep.* **26**, 100933 (2021).
949 90. Patiño-Galindo, J. Á. *et al*. Recombination and lineage-specific mutations linked to the emergence of
950 SARS-CoV-2. *Genome Med.* **13**, 124 (2021).
951 91. Bálint, G., Vörös-Horváth, B. & Széchenyi, A. Omicron: increased transmissibility and decreased
952 pathogenicity. *Signal Transduct. Target. Ther.* **7**, 151 (2022).
953 92. Matyášek, R. & Kovařík, A. Mutation Patterns of Human SARS-CoV-2 and Bat RaTG13 Coronavirus
954 Genomes Are Strongly Biased Towards C>U Transitions, Indicating Rapid Evolution in Their Hosts. *Genes*

143

bioRxiv preprint doi: https://doi.org/10.1101/2025.03.25.645210; this version posted March 26, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

11, 761 (2020).

93. Wang, D. *et al.* Population Bottlenecks and Intra-host Evolution During Human-to-Human Transmission of SARS-CoV-2. *Front. Med.* **8**, 585358 (2021).
94. Zhao, L. *et al.* The mutational spectrum of SARS-CoV-2 genomic and antigenomic RNA. *Proc. R. Soc. B Biol. Sci.* **289**, 20221747 (2022).
95. Mercatelli, D. & Giorgi, F. M. Geographic and Genomic Distribution of SARS-CoV-2 Mutations. *Front. Microbiol.* **11**, 1800 (2020).
96. De Maio, N. *et al.* Mutation Rates and Selection on Synonymous Mutations in SARS-CoV-2. *Genome Biol. Evol.* **13**, evab087 (2021).
97. Peng, X. *et al.* RNA editing increases the nucleotide diversity of SARS-CoV-2 in human host cells. *PLOS Genet.* **18**, e1010130 (2022).
98. DeRonde, S., Deuling, H., Parker, J. & Chen, J. Identification of a novel SARS-CoV-2 variant with a truncated protein in ORF8 gene by next generation sequencing. *Sci. Rep.* **12**, 4631 (2022).
99. Colson, P. *et al.* The return of the "Mistigri" (virus adaptative gain by gene loss) through the SARS□CoV□2 XBB.1.5 chimera that predominated in 2023. *J. Med. Virol.* **95**, e29146 (2023).
100. Davis, H. E., McCorkell, L., Vogel, J. M. & Topol, E. J. Long COVID: major findings, mechanisms and recommendations. *Nat. Rev. Microbiol.* **21**, 133–146 (2023).
101. Proal, Amy D et al. Targeting the SARS-CoV-2 reservoir in long COVID. *Lancet Infect. Dis.* **Volume 0, Issue 0**,.
102. Zuo, W. *et al.* The persistence of SARS-CoV-2 in tissues and its association with long COVID symptoms: a cross-sectional cohort study in China. *Lancet Infect. Dis.* **24**, 845–855 (2024).

LEGENDS

Figure 1: Timeline of the patient care and SARS-CoV-2 variant evolution undergoing recombination and *de novo* emergence of mutations.

a) Timeline of the patient's treatment regimen, including chemotherapy, immunotherapy, and antiviral interventions, in relation to the detection of SARS-CoV-2 variants. b) The sequenced nasopharyngeal samples are marked on the timeline with their corresponding qPCR Ct values. c) Superimposition of the epidemiological curves of the two parental genomes, B.1.160 and B.1.1.7 lineages, that infected the patient, in Marseille, the patient's city of residence. d)

144

bioRxiv preprint doi: https://doi.org/10.1101/2025.03.25.645210; this version posted March 26, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Genomic structure of the detected recombinant consensus-level strains. The parental B.1.160 genome is represented in blue, while segments originating from the B.1.1.7 variant are shown in yellow, illustrating the progressive acquisition of recombination-derived genomic regions.

Figure 2: Within host evolution at consensus level

a) Maximum likelihood phylogenetic tree constructed with Nextstrain, based on the 16 whole SARS-CoV-2 genomes obtained from the immunocompromised patient between September 2020 and December 2021. Branch lengths represent the extent of genetic divergence. The phylogenetic tree reveals the placement of the sequences into three distinct genotypic groups, highlighted by dotted lines. b) Maximum likelihood phylogenetic tree constructed using Nextstrain, based on the 16 SARS-CoV-2 genome sequences taken from the patient. The y-axis reflects genetic divergence, the tree shows the rate of evolution of the sequences over time.

Figure 3: Distribution of iSNVs and SNPs: abundance, diversity, proportion of synonymous and non-synonymous mutations, and classification of mutation type

a)Abundance of iSNVs and SNPs illustrated by the distribution of mutations according to allelic frequency intervals (x-axis), membership of a mutation pool being indicated by a color code: blue for ‘within host’ iSNVs, yellow for ‘total’ iSNVs, green for ‘within host’ SNPs and red for ‘total’ SNPs, samples are indicated next to the y- axis. b) Richness represented by the number of mutations in the different pools as a function of time. c) Temporal accumulation of synonymous and non-synonymous mutations, classified as iSNV_WH (within host), iSNV_TOT (total), SNP_WH (within host) and SNP_TOT (SNP_TOT). d) Distribution of substitution classes in the four pools, displaying cumulative counts and normalised frequencies representing the proportion of each mutation class to the total number of mutations in each pool. e) Stacked bar charts showing the distribution of mutations at

bioRxiv preprint doi: https://doi.org/10.1101/2025.03.25.645210; this version posted March 26, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

codon positions (1st, 2nd and 3rd) for each substitution type in the four pools. f) Number of mutations per codon position for each pool, with χ^2 test results and p-value.

Figure 4) Trinucleotide context associated with mutation type across the four mutation pools: distribution and principal component analysis.

Stacked bar plots showing the distribution of number of mutation types across different trinucleotide contexts within each of the four mutation pools: iSNV_WH, iSNV_TOT, SNP_WH, and SNP_TOT. Each bar represents the number of a given trinucleotide context, color-coded by mutation type.

Figure 5) Gene level and wide genome normalized entropy

a)Scatter plots depicting the distribution of Shannon entropy along the viral genome with position (x-axis) within the four pools. The background colors highlight different genomic regions, including structural (S, N, E, M) and non-structural (ORFs, nsp) proteins. b)Boxplots show the normalized sequence entropy across different genomic regions for each mutation pool. c)Boxplots compare overall entropy distributions across the four mutation pools, with statistical significance denoted by p-values from Mann-Whitney U test.

Figure 6: pN/pS ratio across SARS-CoV-2 genes in the four mutation pools

a)pN/S ratio calculated across different genomic regions of SARS-CoV-2, categorized into iSNV_WH, iSNV_TOT, SNP_WH, and SNP_TOT. Each panel corresponds to one of the four mutation pools b) Histogram representing the number of cases with pN/pS = ∞, referring to cases where pS=0 and N≥1, stars indicate pools where samples with pS=0 and N≥3 are found. c) Number of synonymous and non-synonymous mutations in Spike gene per sample according to the different pools.

bioRxiv preprint doi: https://doi.org/10.1101/2025.03.25.645210; this version posted March 26, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Figure 7) Evolution of allele frequencies of *Spike* within host substitutions, localization of total consensus mutations in the SARS-CoV-2 Spike protein and electrostatic potential of SARS-CoV-2 Spike proteins.

a)Allele frequencies (%) of various within host Spike substitutions in the viral population across different time points (days). Each colored line represents a different mutation, with transient emerging SNPs (L5F, D796H, L821L, T618K, T572I) labeled. These mutations show distinct temporal patterns of allele frequency, reflecting their rise and fall in the viral quasispecies population over the course of the infection. b) The molecular structure of each mutant protein (from Day 0 to Day 440) is represented in grey ribbons and the mutations in atomic spheres (blue, nitrogen; grey, carbon; white, hydrogen; red, oxygen). Deletions are highlighted in orange (Del 144) and blue (Del 69-70) disks, taking as reference the last amino acid before the deletion (respectively V143 and I68). The signal sequence (amino acid residues 1-13) does not appear in the structure. NTD, N-terminal domain; RBD, receptor binding domain. c) The electrostatic potential of each mutant protein (from Day 0 to Day 440) is represented on the molecular surface (blue, positive; red, negative; white, neutral) in two opposite views. The upper panel corresponds to the same orientation of the Spike proteins as in Figure JF1. The lower panel shows the rear view obtained after a 180° rotation. Several regions are highlighted by orange circles to better visualize the changes associated with mutations.

Figure 8) Global proportion of high fixed within host SNPs based on GISAID sequences.

Graph showing the proportion (%) of within host SNPs, selected for having reached or exceeded a 70% allele frequency, observed in the patient among the general population, at different positions in the viral genome.

147

bioRxiv preprint doi: https://doi.org/10.1101/2025.03.25.645210; this version posted March 26, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Figure 9: Emergence of ORF8 126-nt deletion since D378

Date	Nature	Technical	Results : IgG anti SARS-CoV-2	
2021-05-03	Serum	Immunoluminometric technique Binding	Ti Index 16	Positive
2021-05-25	Serum	Immunoluminometric technique Binding	Ti Index 10	Negative
2021-08-12	Serum	Immunoluminometric technique Binding	Ti Index <4	Negative
2021-11-23	Serum	Immunoluminometric technique Binding	Ti Index <5	Negative

a)Visualization of a PCR product sequencing on the region made on the sample D378. Mapped and sorted reads are visualized by IGV, a large nucleotide deletion is visible. b) Visualisation of the ORF-8 deletion on the consensus genomes obtained by assembly (top panel) and visualisation of the deletion detected as a missing region, obtained by mapping (bottom panel).

TABLES

Table 1: Serological data obtained on serum

GISAID Identifier	Strain name (Time for Diagnosis)	Sampling date	PangoLIN lineage	Coverage	NGS Technology, Instrument
EPI_ISL_6332079	D0	2020-09-18	B.1.160	99.30 %	Illumina, NovaSeq
EPI_ISL_10816743	D229	2021-05-05	B.1.160	97.10 %	Illumina, NovaSeq
EPI_ISL_19195977	D259	2021-06-04	B.1.160	91.80 %	Illumina, NovaSeq
EPI_ISL_11030507	D276	2021-06-21	B.1.160	90.20 %	Illumina, NovaSeq
EPI_ISL_19195978	D277	2021-06-22	B.1.160	96.80 %	Illumina, NovaSeq
EPI_ISL_19195979	D285	2021-06-30	B.1.160	97.90 %	Illumina, NovaSeq
EPI_ISL_10816731	D290	2021-07-05	B.1.160	97.70 %	Illumina, NovaSeq
EPI_ISL_19195980	D329	2021-08-13	B.1.160	99.20 %	Illumina, NovaSeq
EPI_ISL_19195981	D330	2021-08-14	B.1.160	97.40 %	Illumina, NovaSeq
EPI_ISL_19195982	D341	2021-08-25	B.1.160	83.60 %	Illumina, NovaSeq
EPI_ISL_10816744	D378	2021-10-01	B.1.160	95.80 %	Illumina, NovaSeq
EPI_ISL_19196018	D425	2021-11-17	B.1.160	82.60 %	Illumina, NovaSeq
EPI_ISL_19196019	D431	2021-11-23	B.1.160	98.70 %	Illumina, NovaSeq
EPI_ISL_10816733	D432	2021-11-24	B.1.160	98.80 %	Illumina, NovaSeq
EPI_ISL_19196020	D434	2021-11-26	B.1.160	98.00 %	Illumina, NovaSeq
EPI_ISL_19196021	D440	2021-12-02	B.1.160	93.20 %	Illumina, NovaSeq

Table 2: Summary of SARS-CoV-2 genomes used in this study

bioRxiv preprint doi: https://doi.org/10.1101/2025.03.25.645210; this version posted March 26, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

1079 Table showing sequencing information for the 16 SARS-CoV-2 consensus genomes obtained with a 10x
1080 coverage. Displayed are each sample's GISAID identifier, strain name (associated with sampling date) and
1081 PangoLIN lineage.

1082

Gene	pN/pS	Infinite (N>3)	Indeterminate	Behavior across pools
nsp1	None	None	Present	Indeterminate
nsp2	Variable: <1 in iSNV pools and SNP_WH, >1 in SNP_TOT in later samples	3 samples in SNP_WH	Present (iSNV_WH and SNP_WH)	Few synonymous mutations in iSNV_WH/SNP_WH, positive selection in SNP_TOT at the final stage (global evolution process)
nsp3	Variable: pN/pS<1 in SNP_TOT, some samples >1 in SNP_WH and iSNVs	Present in iSNVs and SNP_WH	None in SNP_TOT	Strong evolutionary constraint in SNP_TOT, sporadic positive selection in iSNVs/SNP_WH, occasional infinite values
nsp4	Variable: low in iSNVs, >1 in some SNP_WH and SNP_TOT samples	Only in SNPs	Present in iSNVs	Positive selection in SNPs, accumulation of beneficial non-synonymous mutations
nsp5	Very low	None (<3 non-synonymous mutation)	High	Conserved protein, possible purifying selection
nsp6	Very low (pN/pS<1 in 6 SNP_TOT samples, sporadic <1 ratio in SNP_WH/iSNVs)	None (<3 non-synonymous mutation)	High in iSNVs	Highly conserved, purifying selection in SNPs and iSNVs (linked to autophagy interactions)
nsp7	None	None	None	Indeterminate
nsp8	Low, sporadic in iSNVs and SNPs	None	High	Indeterminate, signs of purifying selection
nsp9	Low in iSNVs	None (<3 non-synonymous mutation)	High	Indeterminate, signs of purifying selection
nsp10	None	None	None	None
nsp11	Not covered	Not covered	Not covered	Not analyzable
nsp12	Very low in all pools	Sporadic only in SNP_TOT	None in SNP_TOT, few in other pools	Highly conserved, negative selection since iSNV stage
nsp13	Low in iSNVs, moderate in SNP_WH, high in SNP_TOT (>1 in 8 samples)	Sporadic in iSNVs only	None in SNP_TOT, few in other pools	Positive selection in SNPs (accumulation of non-synonymous mutations), negative selection in iSNVs
nsp14	Very low in SNP_TOT	None	High in all pools (moderate in SNP_TOT)	Indeterminate, signs of purifying selection
nsp15	Very low, sporadic in iSNVs (1 sample with ratio <1)	None	High	Indeterminate, signs of purifying selection
nsp16	Very low, sporadic in iSNVs (2 samples <1 in iSNVs, 1 sample <1 in SNPs)	None (<3 non-synonymous mutation)	High	Indeterminate, signs of purifying selection
S (Spike)	Variable: 3 samples with ratio >1 in iSNV_TOT, 4 samples with ratio <1 in iSNVs_WH, 2 samples	Strongly present in SNP_TOT	Few in SNP_WH	Signs of positive selection in SNP_TOT: Impact of recombination-driven diversity. Minor contribution
M	Very low in all pools	None (<3 non-synonymous mutation)	Few	Highly conserved, purifying selection in SNPs and iSNVs
E	None	None (<3 non-synonymous mutation)	High	Indeterminate, signs of conserved protein
ORF3a	Very low in all pools	Few in iSNV_WH, iSNV_TOT and SNP_WH	Few	Highly conserved in SNP_TOT, but signs of positive evolution within the host.
ORF6	None	None	High	Indeterminate, seems preserved
ORF7a	Not covered	Not covered	Not covered	Not analyzable
ORF8	Low, sproadic in iSNV_WH and iSNV_TOT (1 sample with ratio <1)	None (<3 non-synonymous mutation)	Few	Indeterminate, signs of conserved protein
ORF9b	None	High in SNPs	Few	Signs of positive selection in SNPs
N	Variable : sporadic ratio >1 for 1 sample in SNPs and iSNV_TOT, 2 samples with ratio <1 iniSNVs	High in SNPs	Moderate in iSNVs	Signs of positive selection in SNPs

1083
1084

149

bioRxiv preprint doi: https://doi.org/10.1101/2025.03.25.645210; this version posted March 26, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Table 3: Evolutionary constraints and selection dynamics of SARS-CoV-2 genes across mutation pool

bioRxiv preprint doi: https://doi.org/10.1101/2025.03.25.645210; this version posted March 26, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

FIGURES

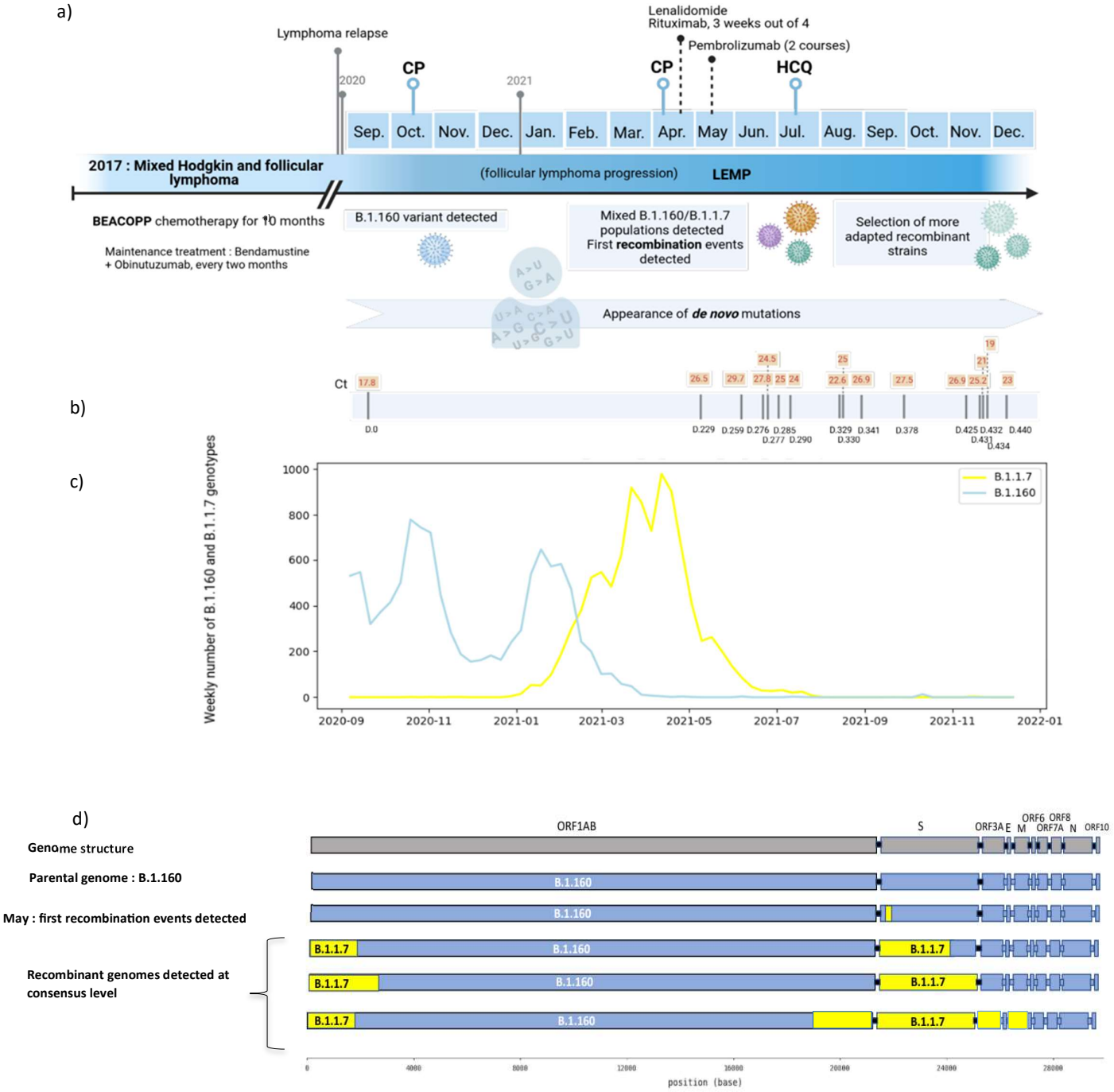
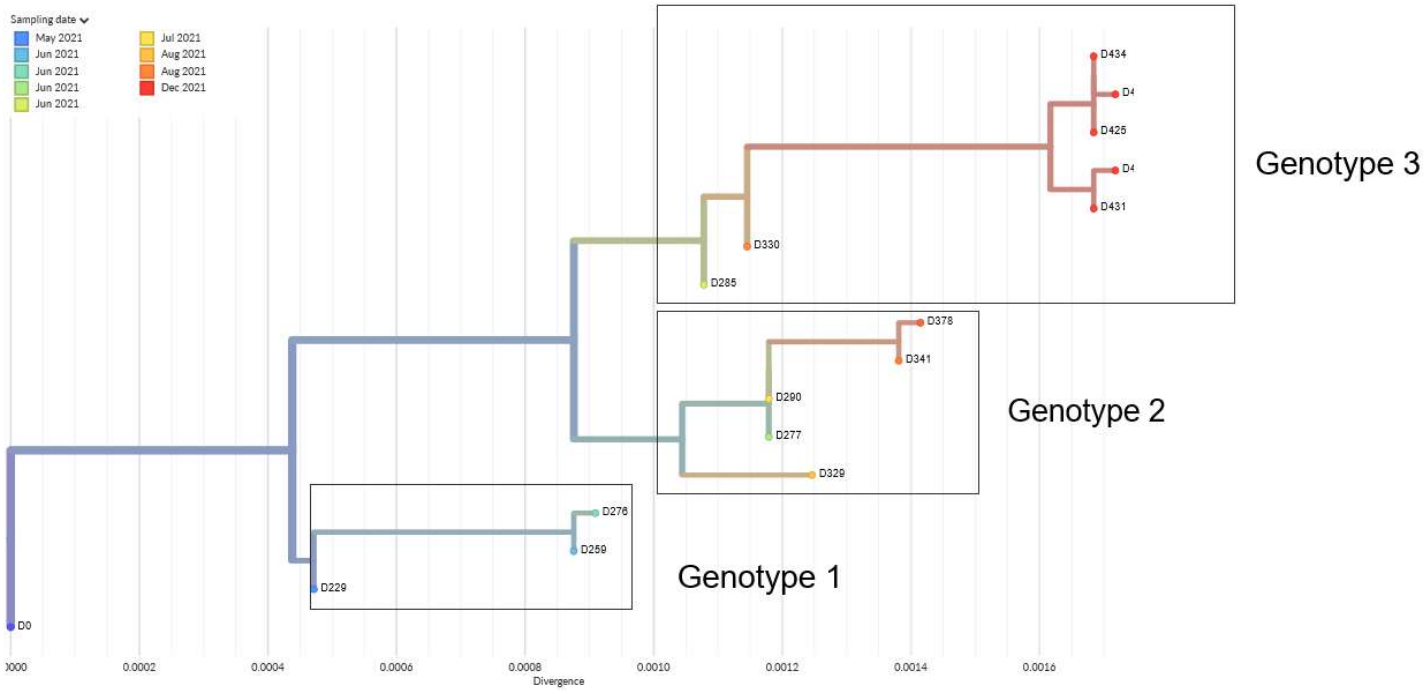


Figure 1: Timeline of the patient care and SARS-CoV-2 variant evolution undergoing recombination and *de novo* emergence of mutations.

bioRxiv preprint doi: https://doi.org/10.1101/2025.03.25.645210; this version posted March 26, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

a



b

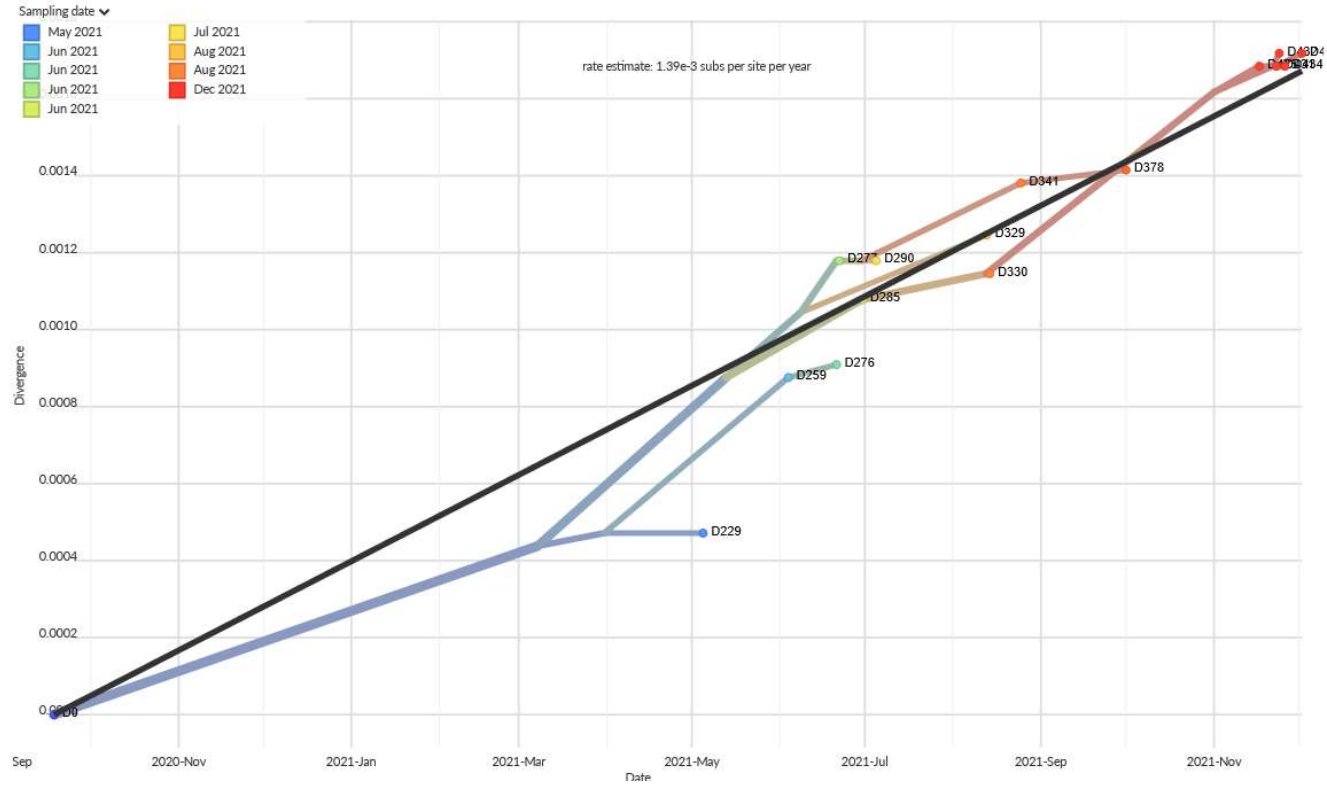


Figure 2: SARS-CoV-2 evolution at consensus level (10x): three genotypes identified and evolution rate

bioRxiv preprint doi: https://doi.org/10.1101/2025.03.25.645210; this version posted March 26, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

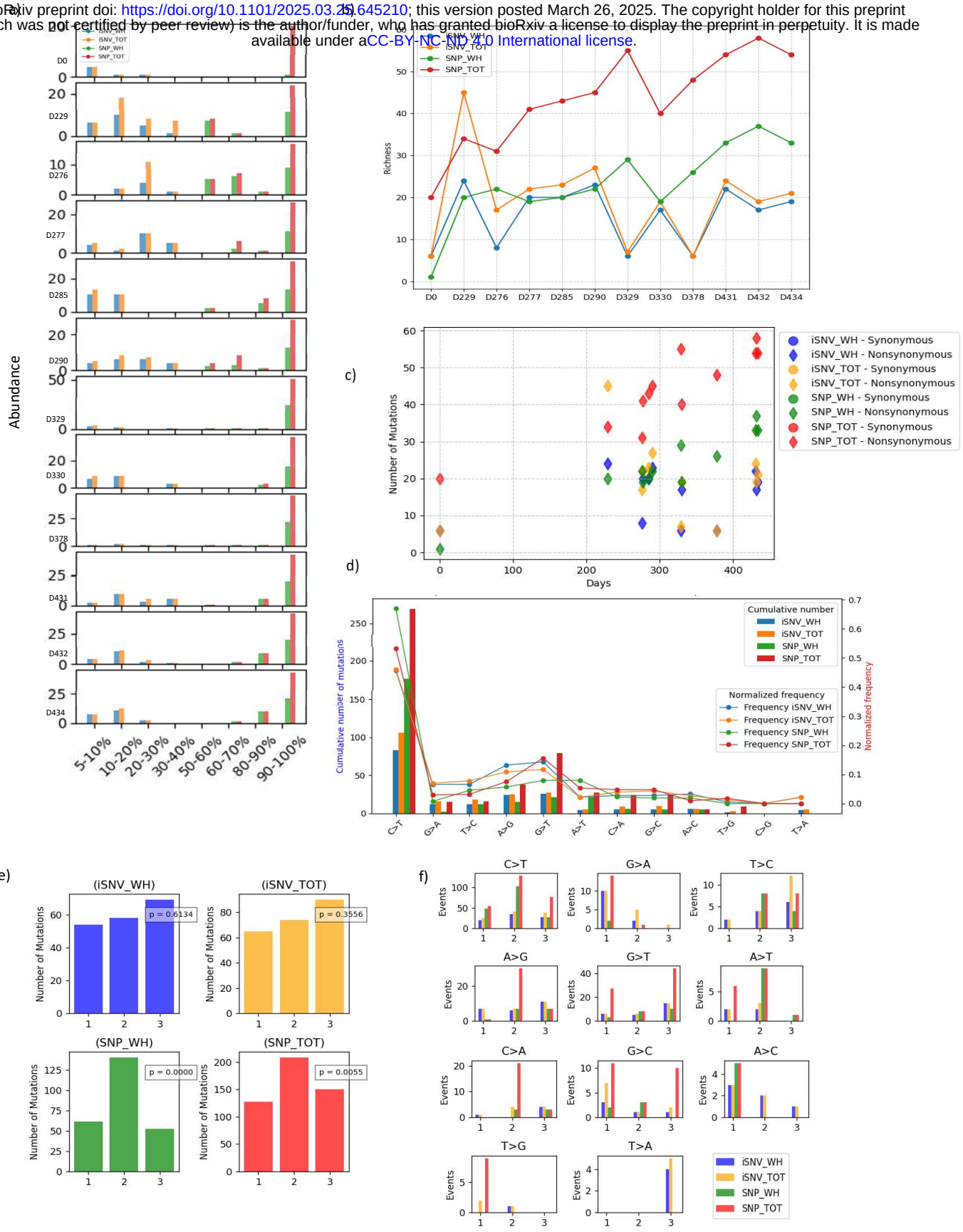


Figure 3: Distribution of iSNVs and SNPs: abundance, diversity, proportion of synonymous and non-synonymous mutations, and classification of mutation types.

bioRxiv preprint doi: https://doi.org/10.1101/2025.03.25.645210; this version posted March 26, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

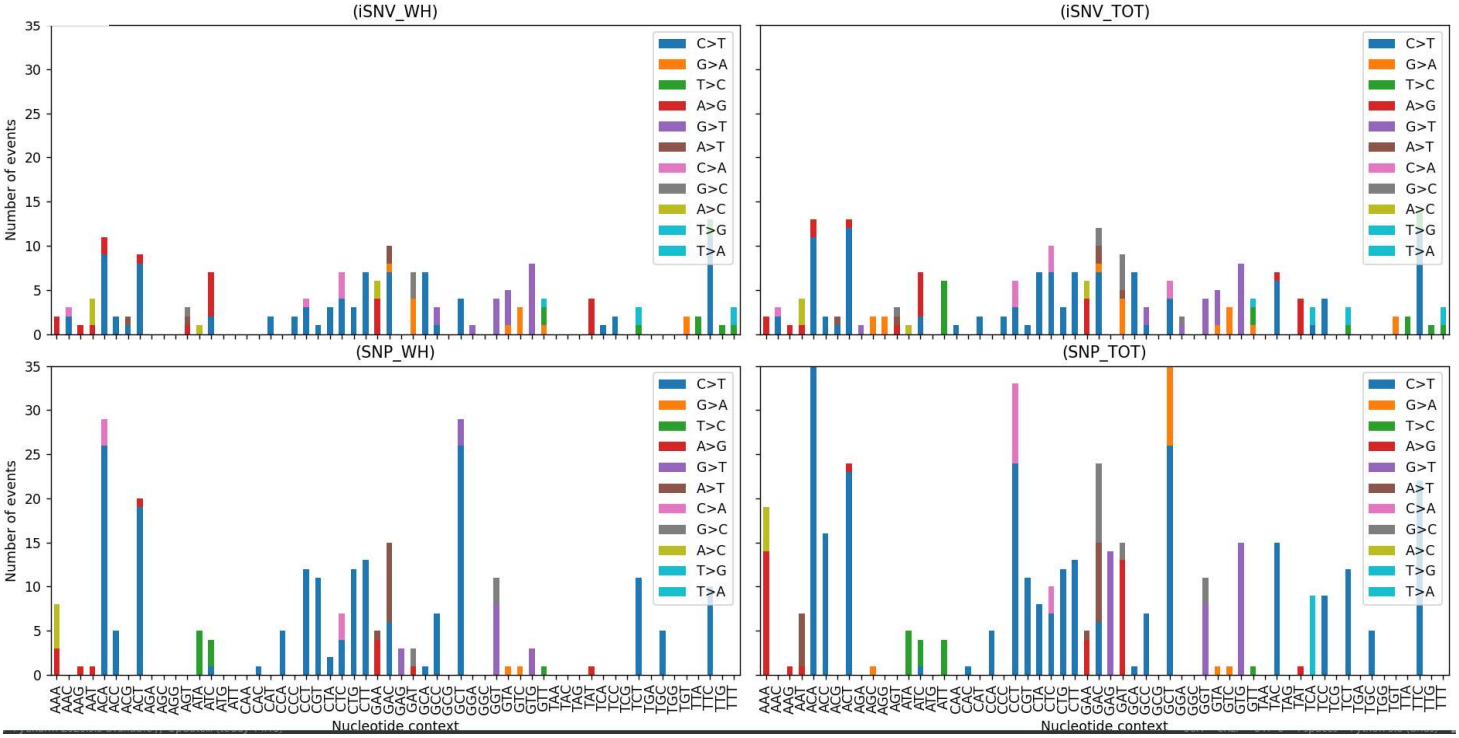


Figure 4: Trinucleotide context associated with mutation type, across the four mutation pools:

bioRxiv preprint doi: https://doi.org/10.1101/2025.03.25.645210; this version posted March 26, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

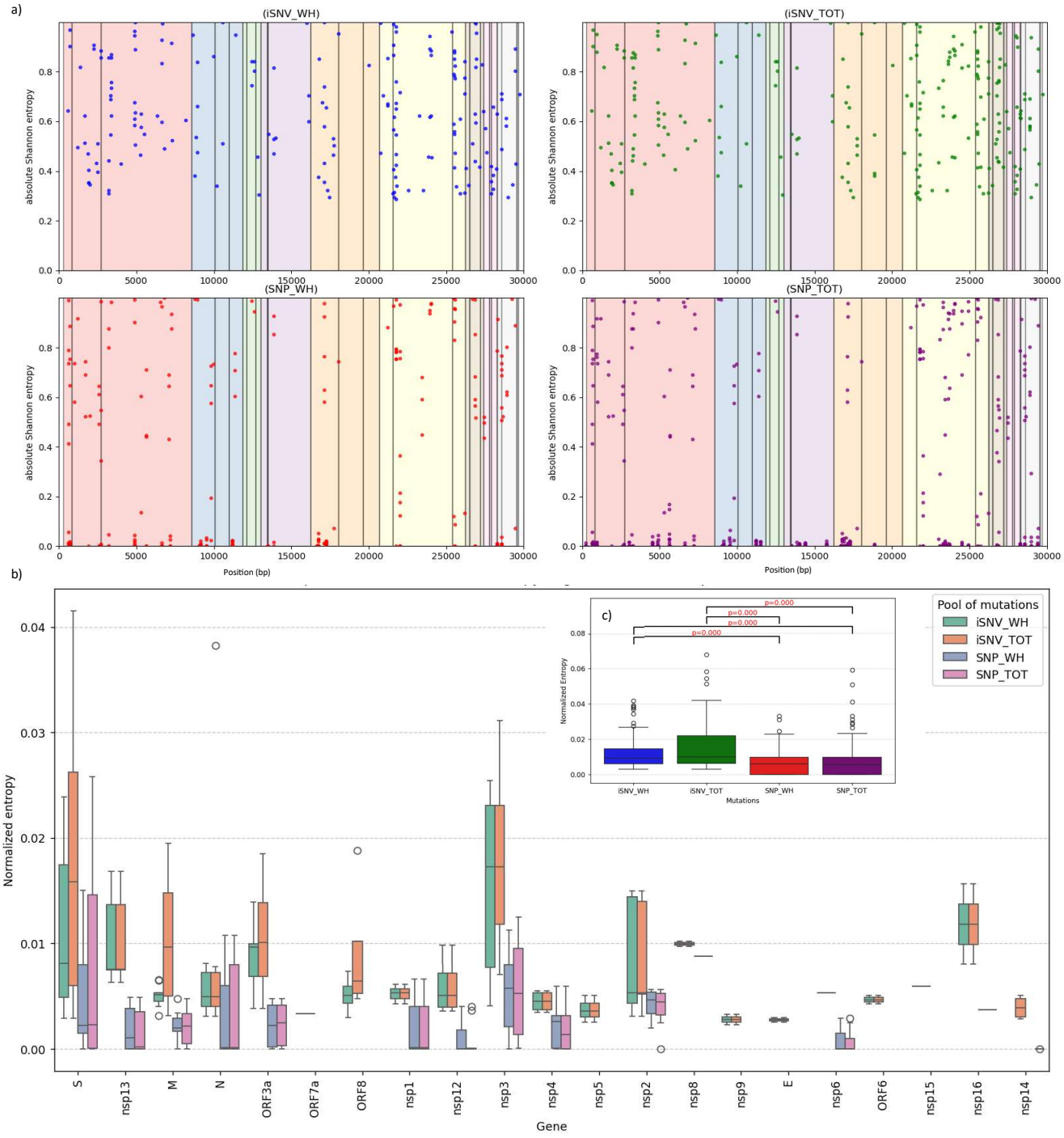


Figure 5: Normalized entropy at gene and whole genome level

bioRxiv preprint doi: https://doi.org/10.1101/2025.03.25.645210; this version posted March 26, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

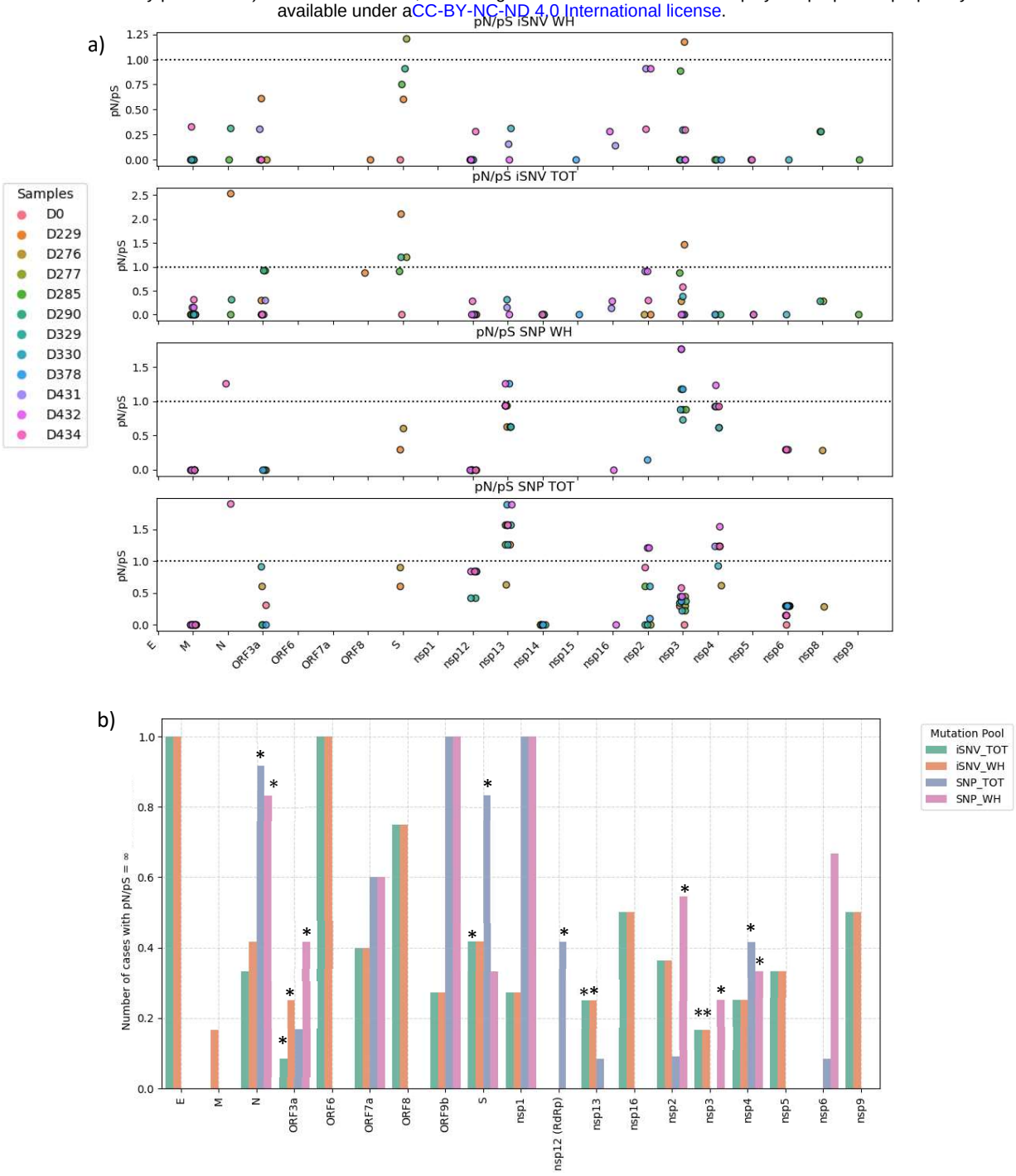


Figure 6: pN/pS ratio across SARS-CoV-2 genes in the four mutation pools

bioRxiv preprint doi: https://doi.org/10.1101/2025.03.25.645210; this version posted March 26, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

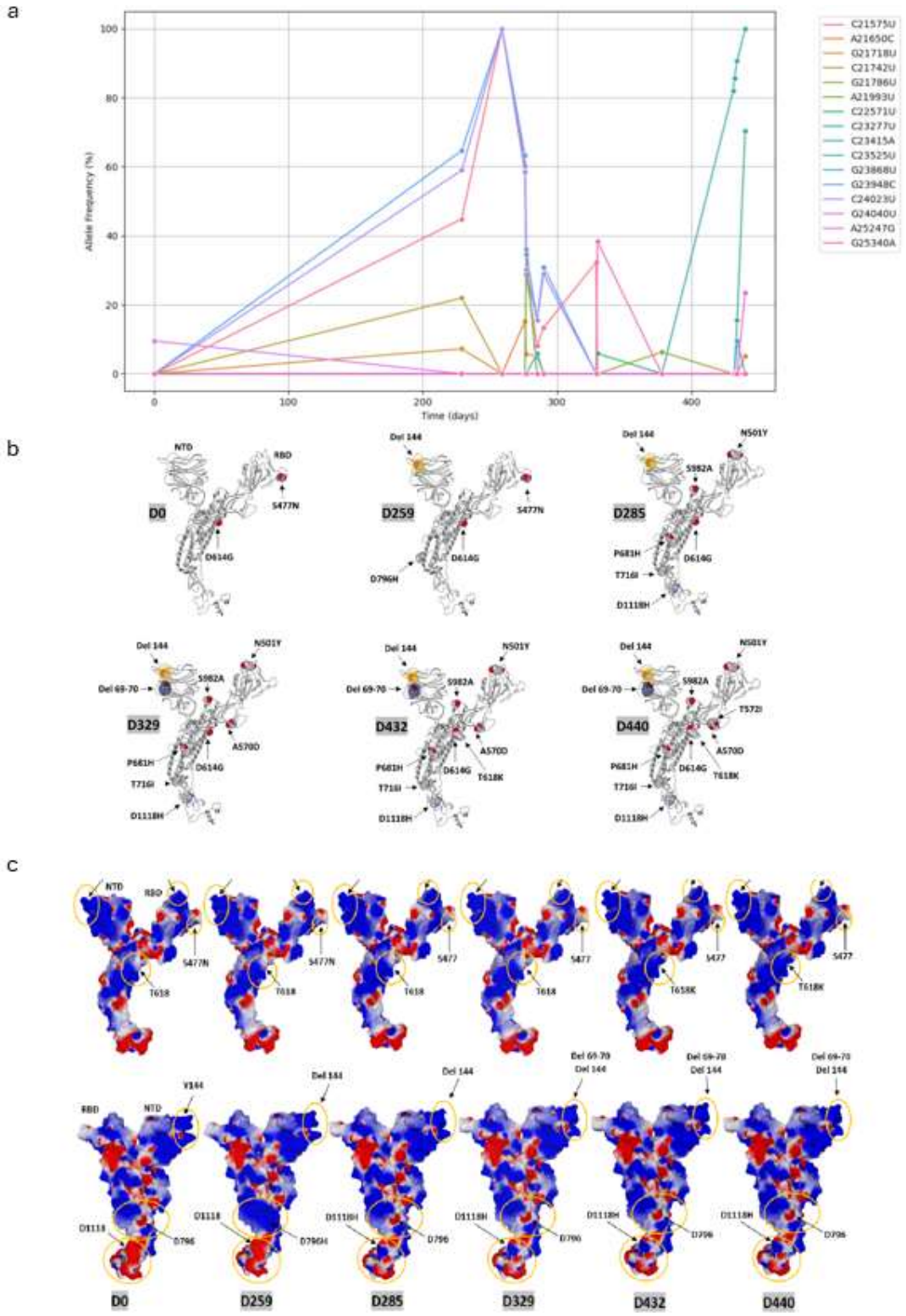


Figure 7: Evolution of allele frequencies of *Spike* within host substitutions, localization of total consensus mutations in the SARS-CoV-2 Spike protein and electrostatic potential of SARS-CoV-2 Spike proteins.

bioRxiv preprint doi: https://doi.org/10.1101/2025.03.25.645210; this version posted March 26, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

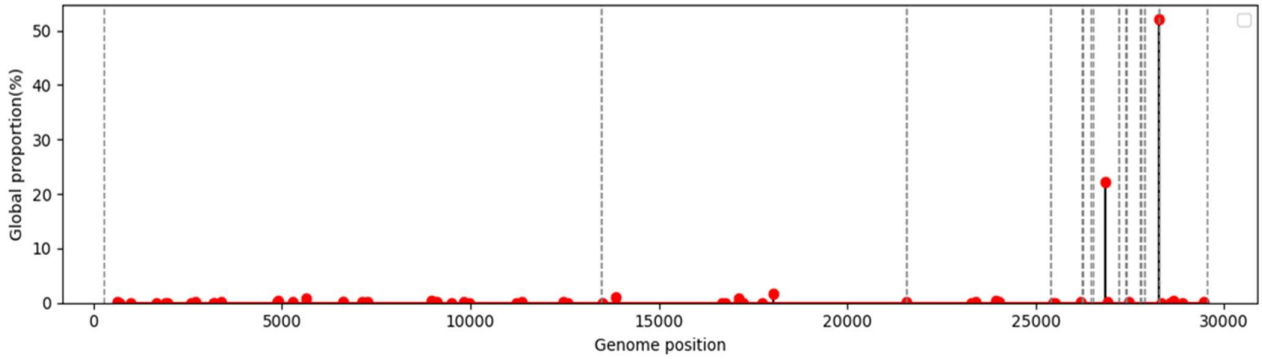


Figure 8: Global proportion of high fixed within host SNPs based on GISAID sequences.

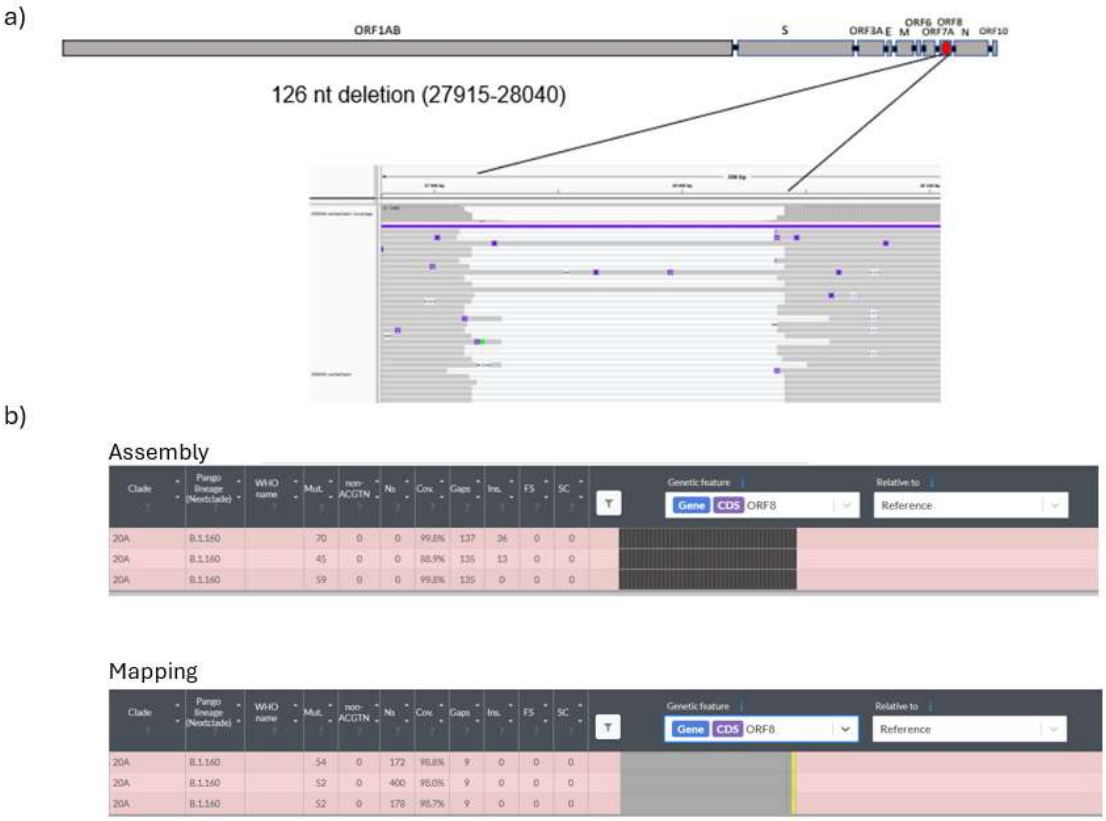


Figure 9: Illustration of the ORF8 126-nt deletion which emerged since day 378

SUPPLEMENTARY

SUPPLEMENTARY METHODS

PCR amplification of [24,813-29,074]

Briefly, RNA samples were extracted using the EZ1 Virus kit with an EZ1 Advanced XL instrument (Qiagen) following the manufacturer's recommendations. PCR amplification of the [24,813-29,074] region was performed in a 25 µL total volume using the SuperScript III One-Step RT-PCR Kit (Invitrogen, Carlsbad, CA, USA), using the forward: ATGGAAAAGCACACTTTCCT; reverse: GCTTTAGTGGCAGTACGTTT primers at concentrations of 200 nM per reaction.

PCR was performed using the following conditions: 50 ◦C for 25 min, 95 ◦C for 2 min, then 40 cycles including 15 s at 95 ◦C, 45 s at 60 ◦C, and 2 min at 68 ◦C. Amplicons were sequenced using the Nanopore technology on a GridION instrument (Oxford Nanopore Technologies Ltd., Oxford, UK), following the manufacturer's instructions. Raw reads were mapped on the SARS-CoV-2 reference genome using Minimap2 (https://github.com/lh3/minimap2), bam files were then filtered using SAMtools (v1.13; https://github.com/samtools/2008-2023 Genome Research Ltd.) combined with an in-house awk script.

TABLES

Nextclade_Lineage	PANGO_Lineage	Mutation_defining_B.1.1.7 lineages circulating in Marseille (from January to April _2021)
20I.1	B.1.1.7	C241T,C3037T,C14408T,A23403G,G28881A,G28882A,G28883C,C913T,C3267T,C5388A,C5986T,T6954C,GTCTGGTTTT11287G,C14676T,C15279T,T16176C,ATACATG21764A,TTTA21990T,A23063T,C23271A,C23604A,C23709T,T24506G,G24914C,C27972T,G28048T,A28111G,TA28270T,G28280C,A28281T,T28282A,C28977T,A17615G
20I.1 [S:681R]	B.1.1.7	C241T,C3037T,C14408T,A23403G,G28881A,G28882A,G28883C,C913T,C3267T,C5388A,C5986T,T6954C,GTCTGGTTTT11287G,C14676T,C15279T,T16176C,ATACATG21764A,TTTA21990T,A23063T,C23271A,C23709T,T24506G,G24914C,C27972T,G28048T,A28111G,TA28270T,G28280C,A28281T,T28282A,C28977T,A17615G,A12162G,A4964G,C7224T,C23604G
20I.2	B.1.1.7	C241T,C3037T,C14408T,A23403G,G28881A,G28882A,G28883C,C913T,C3267T,C5388A,C5986T,T6954C,GTCTGGTTTT11287G,C14676T,C15279T,T16176C,ATACATG21764A,TTTA21990T,A23063T,C23271A,C23604A,C23709T,T24506G,G24914C,C27972T,G28048T,A28111G,TA28270T,G28280C,A28281T,T28282A,C28977T,T15096C,A28095T
20I.2 [S:484K]	B.1.1.7	C241T,C3037T,C14408T,A23403G,G28881A,G28882A,G28883C,C913T,C3267T,C5388A,C5986T,T6954C,GTCTGGTTTT11287G,C14676T,C15279T,T16176C,ATACATG21764A,TTTA21990T,A23063T,C23271A,C23604A,C23709T,T24506G,G24914C,C27972T,G28048T,A28111G,TA28270T,G28280C,A28281T,T28282A,C28977T,T15096C,A28095T,G23012A
20I.3	B.1.1.7	C241T,C3037T,C14408T,A23403G,G28881A,G28882A,G28883C,C913T,C3267T,C5388A,C5986T,T6954C,GTCTGGTTTT11287G,C14676T,C15279T,T16176C,ATACATG21764A,TTTA21990T,A23063T,C23271A,C23604A,C23709T,T24506G,G24914C,C27972T,G28048T,A28111G,TA28270T,G28280C,A28281T,T28282A,C28977T,C21855T
20I.4	B.1.1.7	C241T,C3037T,C14408T,A23403G,G28881A,G28882A,G28883C,C913T,C3267T,C5388A,C5986T,T6954C,GTCTGGTTTT11287G,C14676T,C15279T,T16176C,ATACATG21764A,TTTA21990T,A23063T,C23271A,C23604A,C23709T,T24506G,G24914C,C27972T,G28048T,A28111G,TA28270T,G28280C,A28281T,T28282A,C28977T,C2110T,C14120T,A28095T
20I.5	B.1.1.7	C241T,C3037T,C14408T,A23403G,G28881A,G28882A,G28883C,C913T,C3267T,C5388A,C5986T,T6954C,GTCTGGTTTT11287G,C14676T,C15279T,T16176C,ATACATG21764A,TTTA21990T,A23063T,C23271A,C23604A,C23709T,T24506G,G24914C,C27972T,G28048T,A28111G,TA28270T,G28280C,A28281T,T28282A,C28977T,A5250G,C6706T,A10646G,G17058T,G27261T

Table S1: List of mutations found in sequenced B.1.1.7 variants circulating in Marseille from January 2020 to May 2021

Sample	Region	S_Mutations_iSNV_WH	NS_Mutations_iSNV_WH	S_Mutations_iSNV_TOT	NS_Mutations_iSNV_TOT	S_Mutations_SNP_WH	NS_Mutations_SNP_WH	S_Mutations_SNP_TOT	NS_Mutations_SNP_TOT
D0	E	0	0	0	0	0	0	0	0
D229	E	0	0	0	0	0	0	0	0
D276	E	0	0	0	0	0	0	0	0
D277	E	0	0	0	0	0	0	0	0
D285	E	0	0	0	0	0	0	0	0
D290	E	0	0	0	0	0	0	0	0
D329	E	0	0	0	0	0	0	0	0
D330	E	0	1	0	1	0	0	0	0

D378	E	0	0	0	0	0	0	0	0
D431	E	0	0	0	0	0	0	0	0
D432	E	0	0	0	0	0	0	0	0
D434	E	0	1	0	1	0	0	0	0
D0	M	0	0	0	0	0	0	2	0
D229	M	1	0	3	0	1	0	1	0
D276	M	1	0	3	0	0	0	0	0
D277	M	0	0	2	0	1	0	1	0
D285	M	1	0	1	0	0	0	2	0
D290	M	1	0	3	0	1	0	1	0
D329	M	1	0	1	0	1	0	3	0
D330	M	1	0	1	0	0	0	2	0
D378	M	0	0	0	0	2	0	2	0
D431	M	0	1	2	1	1	0	1	0
D432	M	0	1	2	1	1	0	1	0
D434	M	1	1	1	1	1	0	1	0
D0	N	0	0	0	0	0	0	0	2
D229	N	0	2	1	8	0	2	0	4
D276	N	0	0	0	0	0	2	0	3
D277	N	0	1	0	1	0	1	0	3
D285	N	1	0	1	0	0	1	0	3
D290	N	1	1	1	1	0	1	0	3
D329	N	0	0	0	0	0	2	0	4
D330	N	0	0	0	0	0	1	0	3
D378	N	0	0	0	0	0	1	0	3
D431	N	0	1	0	1	0	3	0	5
D432	N	0	2	0	2	0	3	0	5
D434	N	0	1	0	1	1	4	1	6
D229	nsp1	0	1	0	1	0	1	0	1
D276	nsp1	0	1	0	1	0	1	0	1
D277	nsp1	0	0	0	0	0	2	0	2
D285	nsp1	0	0	0	0	0	2	0	2
D290	nsp1	0	0	0	0	0	2	0	2
D329	nsp1	0	0	0	0	0	2	0	2

D330	**nsp1**	0	0	0	0	0	2	0	2
D378	**nsp1**	0	0	0	0	0	2	0	2
D431	**nsp1**	0	1	0	1	0	2	0	2
D432	**nsp1**	0	0	0	0	0	2	0	2
D434	**nsp1**	0	0	0	0	0	2	0	2
D229	**nsp11**	NC	NC	NC	NC	NC	NC	NC	NC
D276	**nsp11**	NC	NC	NC	NC	NC	NC	NC	NC
D277	**nsp11**	NC	NC	NC	NC	NC	NC	NC	NC
D285	**nsp11**	NC	NC	NC	NC	NC	NC	NC	NC
D330	**nsp11**	NC	NC	NC	NC	NC	NC	NC	NC
D378	**nsp11**	NC	NC	NC	NC	NC	NC	NC	NC
D0	**nsp12**	0	0	0	0	0	0	0	3
D229	**nsp12**	1	0	1	0	0	0	0	3
D276	**nsp12**	1	0	1	0	0	0	0	3
D277	**nsp12**	0	0	0	0	1	0	1	3
D285	**nsp12**	0	0	0	0	0	0	0	3
D290	**nsp12**	0	0	0	0	2	0	2	3
D329	**nsp12**	0	0	0	0	2	0	2	3
D330	**nsp12**	1	0	1	0	0	0	0	3
D378	**nsp12**	0	0	0	0	1	0	1	3
D431	**nsp12**	2	0	2	0	1	0	1	3
D432	**nsp12**	1	0	1	0	1	0	1	3
D434	**nsp12**	1	1	1	1	1	0	1	3
D0	**nsp13**	0	5	0	5	0	0	0	2
D229	**nsp13**	0	0	0	0	1	2	1	4
D276	**nsp13**	0	1	0	1	1	2	1	2
D277	**nsp13**	0	0	0	0	1	3	1	5
D285	**nsp13**	0	0	0	0	1	2	1	4
D290	**nsp13**	0	0	0	0	1	3	1	5
D329	**nsp13**	0	0	0	0	1	3	1	5
D330	**nsp13**	1	1	1	1	1	2	1	4
D378	**nsp13**	0	0	0	0	1	4	1	6
D431	**nsp13**	2	1	2	1	1	3	1	5
D432	**nsp13**	2	0	2	0	1	4	1	6

D434	**nsp13**	0	2	0	2	1	3	1	5
D0	**nsp14**	0	0	0	0	0	0	1	0
D229	**nsp14**	0	0	1	0	0	0	0	0
D276	**nsp14**	0	0	0	0	0	0	1	0
D277	**nsp14**	0	0	0	0	0	0	1	0
D285	**nsp14**	0	0	1	0	0	0	0	0
D290	**nsp14**	0	0	0	0	0	0	1	0
D329	**nsp14**	0	0	0	0	0	0	1	0
D330	**nsp14**	0	0	1	0	0	0	0	0
D378	**nsp14**	0	0	0	0	0	0	1	0
D434	**nsp14**	0	0	1	0	0	0	0	0
D378	**nsp15**	1	0	1	0	0	0	0	0
D229	**nsp16**	0	1	0	1	0	0	0	0
D285	**nsp16**	0	1	0	1	0	0	0	0
D431	**nsp16**	2	1	2	1	0	0	0	0
D432	**nsp16**	1	1	1	1	1	0	1	0
D229	**nsp2**	0	0	1	0	0	1	0	1
D276	**nsp2**	0	0	1	0	0	0	0	0
D277	**nsp2**	0	1	0	1	0	0	1	0
D285	**nsp2**	0	0	0	0	0	2	1	2
D290	**nsp2**	0	1	0	1	0	0	1	0
D329	**nsp2**	0	1	0	1	0	0	1	0
D330	**nsp2**	0	1	0	1	0	2	1	2
D378	**nsp2**	0	0	0	0	2	1	3	1
D431	**nsp2**	1	3	1	3	0	4	1	4
D432	**nsp2**	1	3	1	3	0	4	1	4
D434	**nsp2**	2	2	2	2	0	3	1	3
D0	**nsp3**	0	0	0	0	0	0	3	0
D229	**nsp3**	1	4	1	5	0	3	3	3
D276	**nsp3**	1	0	1	1	0	3	2	3
D277	**nsp3**	0	4	0	4	1	3	4	4
D285	**nsp3**	1	3	1	3	1	3	4	3
D290	**nsp3**	0	5	0	5	1	4	4	5
D329	**nsp3**	1	0	2	0	2	5	5	6

D330	**nsp3**	3	3	3	4	1	3	4	3
D378	**nsp3**	0	0	0	0	1	4	4	5
D431	**nsp3**	4	0	4	0	1	6	4	6
D432	**nsp3**	3	0	3	0	1	6	4	6
D434	**nsp3**	1	1	1	2	0	6	3	6
D0	**nsp4**	0	0	0	0	0	0	0	1
D229	**nsp4**	0	1	0	1	1	3	1	4
D276	**nsp4**	NC	NC	NC	NC	NC	NC	NC	NC
D277	**nsp4**	1	0	1	0	0	2	0	3
D285	**nsp4**	0	0	0	0	1	3	1	4
D290	**nsp4**	1	0	1	0	0	3	0	4
D329	**nsp4**	0	0	0	0	0	3	0	4
D330	**nsp4**	0	0	0	0	1	2	1	3
D378	**nsp4**	2	0	2	0	0	2	0	3
D431	**nsp4**	0	1	0	1	1	3	1	4
D432	**nsp4**	0	1	0	1	1	4	1	5
D434	**nsp4**	0	0	0	0	1	3	1	4
D229	**nsp5**	0	1	0	1	0	0	1	4
D378	**nsp5**	1	0	1	0	0	0	0	0
D434	**nsp5**	1	0	1	0	0	0	0	0
D276	**nsp5**	0	0	0	0	0	0	1	2
D290	**nsp5**	1	0	1	0	0	0	0	0
D285	**nsp5**	0	0	0	0	0	0	0	0
D431	**nsp5**	0	0	0	1	0	0	0	0
D432	**nsp5**	0	1	0	0	0	0	0	0
D277	**nsp5**	1	0	1	0	0	0	0	3
D0	**nsp5**	0	0	0	0	0	0	0	1
D0	**nsp6**	0	0	0	0	0	0	1	0
D229	**nsp6**	0	0	0	0	0	1	1	1
D276	**nsp6**	0	0	0	0	0	1	0	1
D277	**nsp6**	0	0	0	0	0	1	1	1
D285	**nsp6**	0	0	0	0	0	1	1	1
D290	**nsp6**	0	0	0	0	0	1	1	1
D329	**nsp6**	0	0	0	0	0	1	1	1

D330	**nsp6**	1	0	1	0	0	1	1	1
D378	**nsp6**	0	0	0	0	0	1	1	1
D431	**nsp6**	0	0	0	0	1	1	2	1
D432	**nsp6**	0	0	0	0	1	1	2	1
D434	**nsp6**	0	0	0	0	1	1	2	1
D276	**nsp8**	0	0	0	0	1	1	1	1
D277	**nsp8**	1	1	1	1	0	0	0	0
D290	**nsp8**	1	1	1	1	0	0	0	0
D285	**nsp9**	1	0	1	0	0	0	0	0
D329	**nsp9**	0	1	0	1	0	0	0	0
D0	**ORF3a**	0	0	0	0	0	1	1	1
D229	**ORF3a**	1	2	2	2	0	1	0	1
D276	**ORF3a**	1	0	1	0	0	2	1	2
D277	**ORF3a**	0	3	0	3	1	0	1	0
D285	**ORF3a**	0	3	1	3	0	0	0	0
D290	**ORF3a**	0	3	1	3	1	0	1	0
D329	**ORF3a**	0	0	0	0	0	3	1	3
D330	**ORF3a**	2	0	2	0	0	0	0	0
D378	**ORF3a**	0	0	0	0	1	0	1	0
D431	**ORF3a**	1	1	1	1	0	0	0	0
D432	**ORF3a**	1	0	1	0	0	1	0	1
D434	**ORF3a**	1	0	1	0	0	0	0	0
D378	**ORF6**	0	1	0	1	0	0	0	0
D434	**ORF6**	0	1	0	1	0	0	0	0
D229	**ORF7a**	0	1	0	1	0	0	0	0
D276	**ORF7a**	NC	NC	NC	NC	NC	NC	NC	NC
D277	**ORF7a**	NC	NC	NC	NC	NC	NC	NC	NC
D285	**ORF7a**	NC	NC	NC	NC	NC	NC	NC	NC
D290	**ORF7a**	NC	NC	NC	NC	NC	NC	NC	NC
D330	**ORF7a**	NC	NC	NC	NC	NC	NC	NC	NC
D378	**ORF7a**	NC	NC	NC	NC	NC	NC	NC	NC
D431	**ORF7a**	NC	NC	NC	NC	NC	NC	NC	NC
D432	**ORF7a**	0	0	0	0	0	1	0	1
D434	**ORF7a**	NC	NC	NC	NC	NC	NC	NC	NC

D229	ORF8	1	0	1	3	0	0	0	0
D277	ORF8	0	2	0	2	0	0	0	0
D285	ORF8	0	2	0	2	0	0	0	0
D290	ORF8	0	2	0	2	0	0	0	0
D229	ORF9b	0	0	0	0	0	1	0	1
D276	ORF9b	0	0	0	0	0	1	0	1
D277	ORF9b	0	0	0	0	0	1	0	1
D285	ORF9b	0	0	0	0	0	1	0	1
D290	ORF9b	0	0	0	0	0	1	0	1
D329	ORF9b	0	0	0	0	0	1	0	1
D330	ORF9b	0	0	0	0	0	1	0	1
D378	ORF9b	0	0	0	0	0	1	0	1
D431	ORF9b	0	1	0	1	0	2	0	2
D432	ORF9b	0	1	0	1	0	2	0	2
D434	ORF9b	0	1	0	1	0	3	0	3
D0	S	1	0	1	0	0	0	0	2
D229	S	1	2	1	7	1	1	1	2
D276	S	0	1	0	6	1	2	1	3
D277	S	1	4	1	4	0	0	0	6
D285	S	2	5	2	6	0	0	0	7
D290	S	1	3	1	4	0	0	0	7
D329	S	0	1	0	1	0	1	0	8
D330	S	0	2	0	2	0	0	0	6
D378	S	0	1	0	1	0	0	0	6
D431	S	0	0	0	0	0	1	0	8
D432	S	0	0	0	0	0	1	0	8
D434	S	0	2	0	2	0	1	0	8

Table S2

nt	aa	Signature mutation of VOC (PANGOLin lineage)	Number of sequences (GISAID database)	Overall proportion (%) (GISAID total sequence)
C629U	ORF1a L122F		4207	0.03
C696U	ORF1a D144V		0	0.00
C981U	ORF1a A239V		1297	0.01
U1683C	ORF1a I473T		3265	0.02
G1895A	ORF1a V544I		2041	0.01
U1978C	ORF1a D571D		1437	0.01
C2594U	ORF1a P777S		2753	0.02
C2706U	ORF1a T814I	B.1.416.1 / AY.4.1	33735	0.21
A3193G	ORF1a E976E		1451	0.01
A3224G	ORF1a T987A		351	0.00
C3393U	ORF1a A1043V		7465	0.05
A4872G	ORF1a Y1536C		1229	0.01
C4878U	ORF1a T1538I	B.1.621	22054	0.14
C4893U	ORF1a T1543I		60657	0.37
G4904A	ORF1a D1547N		1335	0.01
U5278C	ORF1a D1671D		4237	0.03
A5648C	ORF1a K1795Q	P.1	137052	0.84
C6629U	ORF1a L2122F		13261	0.08
C6633U	ORF1a A2123V	B.1.619	26389	0.16
C7113U	ORF1a T2283I		19351	0.12
U7185C	ORF1a I2307T		1607	0.01
C7279U	ORF1a F2338F		19138	0.12
C8950U	ORF1a I2895I	AY.28	49388	0.30
C9042U	ORF1a S2926F		5781	0.04
C9120U	ORF1a T2952I		10286	0.06
C9501U	ORF1a A3079V		2324	0.01
C9811U	ORF1a T3182T		8466	0.05
A9972G	ORF1a K3236R		937	0.01
C11208U	ORF1a A3648V		3804	0.02
G11365U	ORF1a V3700V	AY.7	29093	0.18
C12459U	ORF1a T4065I		18057	0.11
U12592C	ORF1a V4109V		516	0.00
C13506U	ORF1b C4414C		3136	0.02
C13860U	ORF1b D4532D	P.1	156831	0.96
C16675U	ORF1b L5471L		2705	0.02
C16751U	ORF1b P5496L		335	0.00
G17122U	ORF1b A5620S		116599	0.72
C17125U	ORF1b L5621F		8382	0.05
C17251U	ORF1b R5663C		279	0.00
C17738U	ORF1b T5825I		3727	0.02
A18023G	ORF1b N5920S		729	0.00
C21575U	S L5F	B.1.177/B.1.1.241/B.1.526	272544	1.68
C23277U	S T572I	AY.32	35360	0.22
C23415A	S T618K		272	0.00
G23948C	S D796H	B.1.1.318	11108	0.07
C24023U	S L821L		42377	0.26
C25460U	ORF3a A23V		21298	0.13
C25515A	ORF3a L41L		210	0.00
C26204U	ORF3aT271I		3501	0.02
C26858U	M F112F	B.1.1.529 / BA4 / BA.2.75 / BA.2	3598172	22.12
C26907U	M L129L		21240	0.13
G27459U	ORF7a E22D	B.1.619	9731	0.06
A28271U	NC	B.1.620 / C.37 / B.1.1.529 / BA.2/ BA4 / BA.5 / BA.2.75	8492736	52.20
A28338G	N D22G		2528	0.02
G28560U	N G96V		990	0.01
	ORF9b V93L		990	0.01
G28560C	N G96A		25	0.00
	ORF9b V93L		25	0.00
A28572G	N K100R		337	0.00
	ORF9b K97E		337	0.00
C28657U	N D128D		67991	0.42
G28887C	N T205I		0	0.00
	ORF14 L52L		0	0.00
C29445U	N T391I		26184	0.16

Table S3

FIGURES

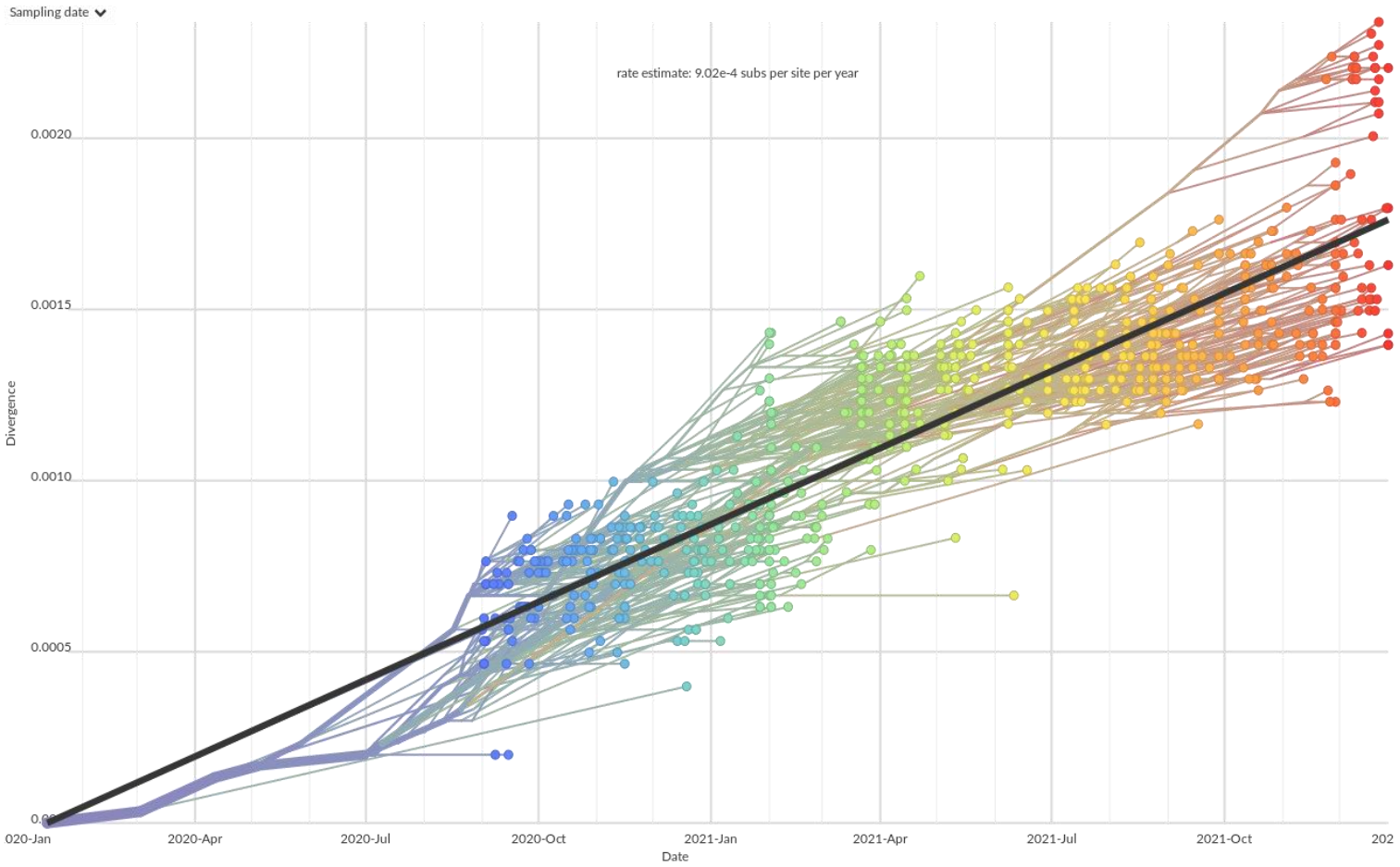


Figure S1

Sample	C913T	C3267T	C4543T	G5629T	C5986T	G9526T	C11497T	G13993T	G15766T	A16889G	G17019T	C18877T	G22992A	A23063T	C23271A	C23604A	C23709T	T24506G	G24
D0	0	0	99,8783	97,5069	0	99,843	99,8179	100	99,9296	100	100	100	99,6933	0	0	0	0	0	
D229	36,982	29,6631	100	99,7419	0	100	100	99,9487	99,7961	100	100	47,622	NaN	53,3333	17,0116	26,2315	28,3046	32,539	12,
D259	0	0	100	100	0	100	100	100	100	100	100	99,9328	100	0	0	0	0	NaN	
D276	29,77	25,3313	NaN	97,8723	0	NaN	NaN	99,9192	100	100	100	99,9636	NaN	NaN	22,2606	27,193	29,8633	25,3776	25,
D277	77,1664	70,4444	99,5918	100	0	100	100	100	100	100	100	99,9141	39,2857	75,7576	61,3372	67,144	70,3436	63,37	63,
D285	77,9201	0	99,8925	99,8733	0	100	100	99,9162	100	100	100	7,44017	8,95522	94,8718	70,6897	85,636	83,905	81,4677	77,
D290	66,748	64,4728	99,9238	99,718	0	100	100	100	100	100	100	100	20,5882	73,5849	57,4591	64,4569	64,513	68,7373	57,
D329	99,9154	100	99,8366	99,7101	8,14577	100	100	100	100	100	100	100	0	100	100	100	100	99,8377	1
D330	100	9,92	100	100	0	100	100	99,9087	100	99,8106	100	7,71375	0	100	99,8891	100	99,9136	100	99,
D378	99,9493	100	100	99,4475	0	100	99,8158	100	100	100	100	100	0	NaN	99,9404	100	100	100	1
D431	99,8578	0	100	100	0	100	100	100	99,8902	100	100	0	0	100	100	100	100	99,9032	1
D432	99,8346	0	100	100	0	99,2245	100	100	100	100	100	0	0	100	99,9425	100	100	100	1
D434	100	10,7015	100	100	0	100	99,9012	100	99,9241	99,6732	99,6499	12,6469	0	100	99,9403	99,8905	100	100	1
D440	100	0	100	100	0	100	99,8462	100	100	100	100	0	0	100	100	100	100	100	1

Sample	G25563T	C25710T	C26735T	T26876C	C27972T	G28048T	A28111G	28270-A	G28280C	A28281T	T28282A	G28881A	G28882A	G28883C	G28975C	C28977T	G29399A
D0	100	100	99,9403	100	0	0	0	0	0	0	0	0	0	0	99,8195	0	99,8843
D229	54,7541	42,2287	40,1305	41,4708	36,3718	35,4142	31,5013	15,8006	15,1655	15,068	16,3377	13,7097	13,3768	18,4733	35,4497	52,8158	99,9242
D259	100	100	100	100	0	0	0	0	0	0	0	0	0	0	100	0	100
D276	62,2476	66,1538	46,7206	48,8889	0	0	0	0	0	0	0	0	0	0	100	0	100
D277	25,7294	35,4839	11,9715	9,03095	0	0	0	0	0	0	0	0	0	0	100	0	100
D285	10,1367	8,43373	100	99,3448	0	0	0	0	0	0	0	0	0	0	100	0	100
D290	22,6364	18,9076	9,5826	13,079	0	0	0	0	0	0	0	0	0	0	100	0	100
D329	98,8889	96,6507	92,6973	93,6055	0	0	0	0	0	0	0	0	0	0	99,6124	0	100
D330	0	0	90,036	87,2886	NaN	0	0	0	0	0	0	0	0	0	100	0	99,6212
D378	0	0	0	0	NaN	0	0	0	0	0	0	0	0	0	100	0	100
D431	0	0	21,7319	22,0487	NaN	0	0	0	0	0	0	0	0	0	100	0	100
D432	0	0	19,1053	22,3001	NaN	0	0	0	0	0	0	0	0	0	94,8276	0	100
D434	0	0	0	0	NaN	0	0	0	0	0	0	0	0	0	100	0	100
D440	0	0	0	0	NaN	0	0	0	0	0	0	0	0	0	100	0	100

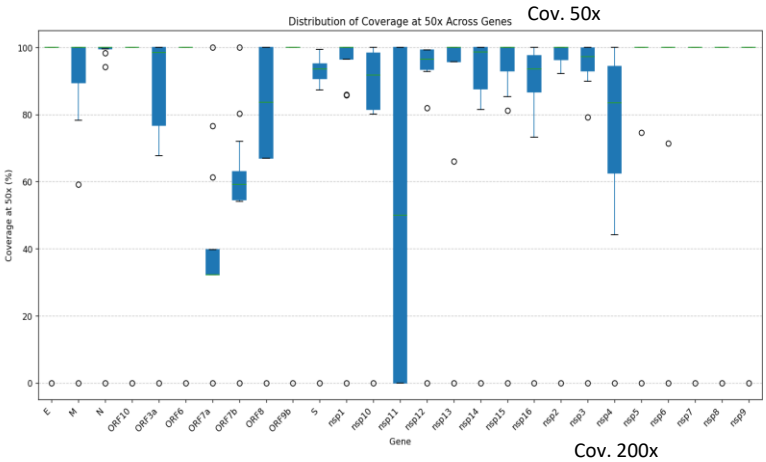
Figure S2

a)

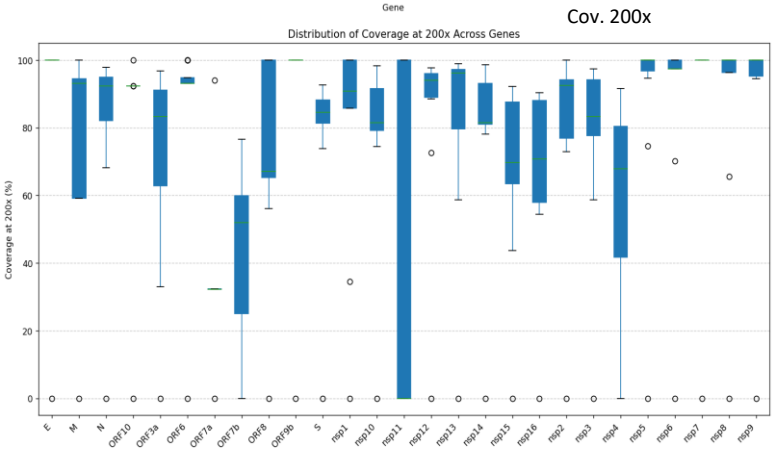
	cov. 50x	cov. 200x
D.0	99.52125879	89.88282558
D.229	93.2038835	81.12152662
D259	77.27150988	59.07599598
D276	80.78674255	65.97924339
D277	91.00770003	76.40107131
D285	93.98393036	85.04184801
D290	95.22597924	87.08068296
D329	98.52695012	91.46970204
D330	91.46300636	78.04486106
D378	89.9832608	78.22564446
D431	96.73585537	90.95748242
D432	97.97455641	92.45731503
D434	95.29293606	90.35487111
D440	79.58218949	65.44693673
Average	91.46855421	80.82428619
SD	7.244008148	10.90205006

b)

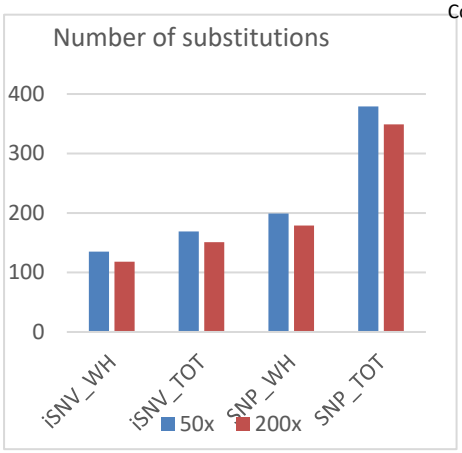
Cov. 50x



Cov. 200x

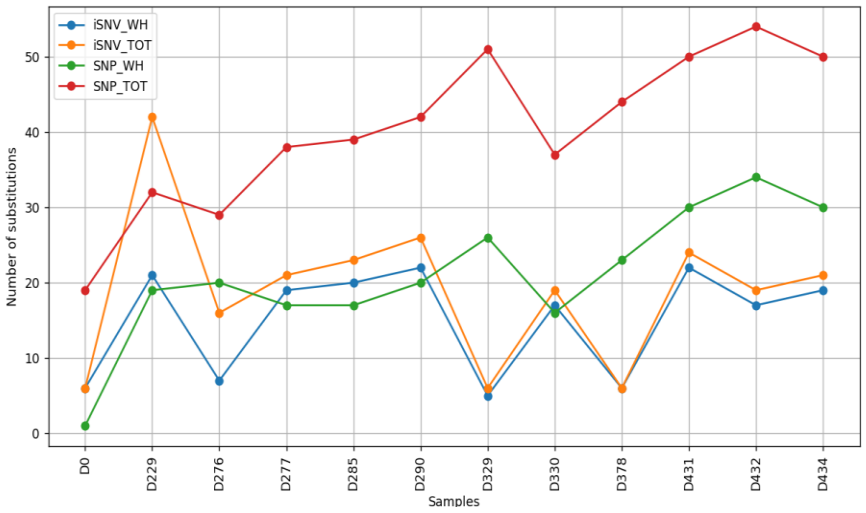


c)



d)

Cov. 50x



Cov. 200x

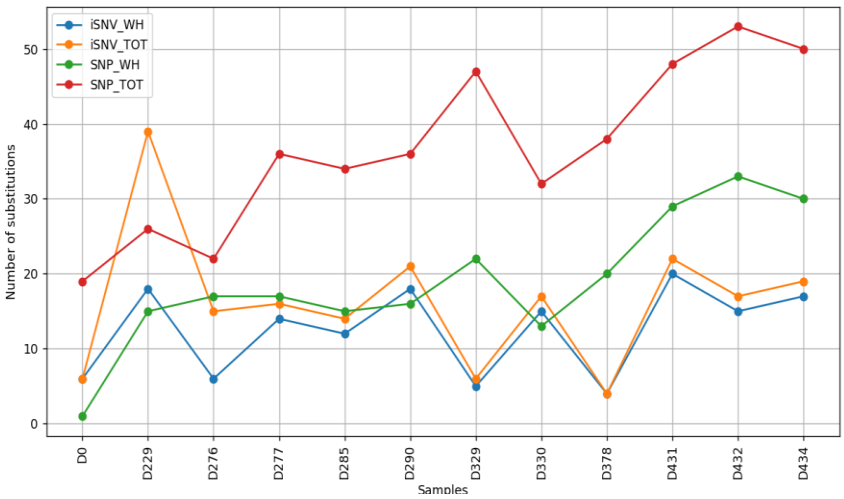


Figure S3

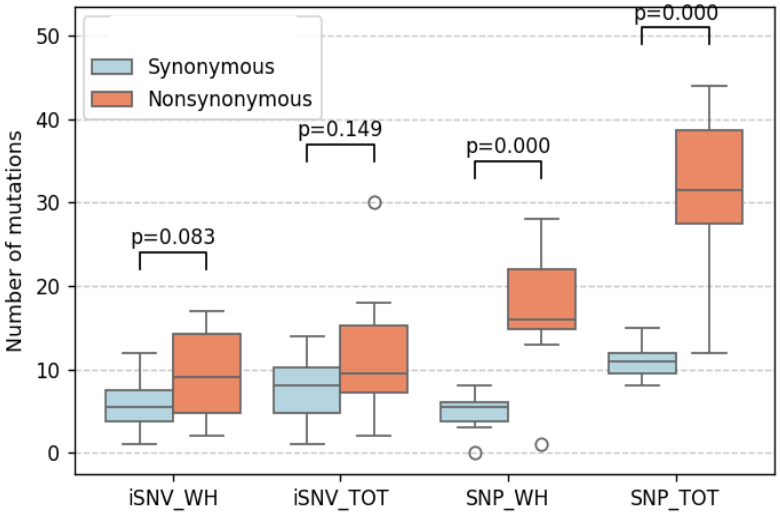


Figure S4

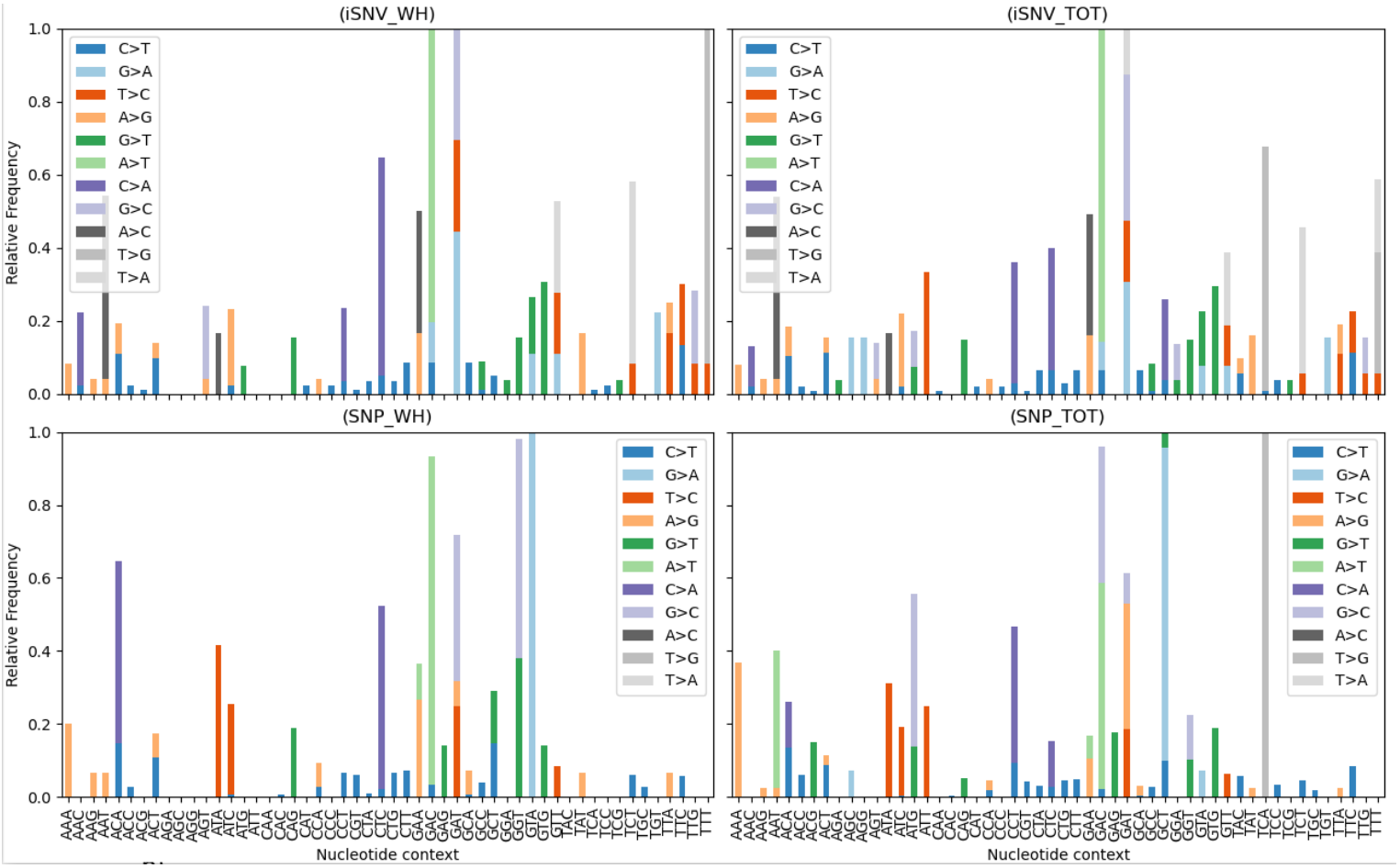


Figure S5

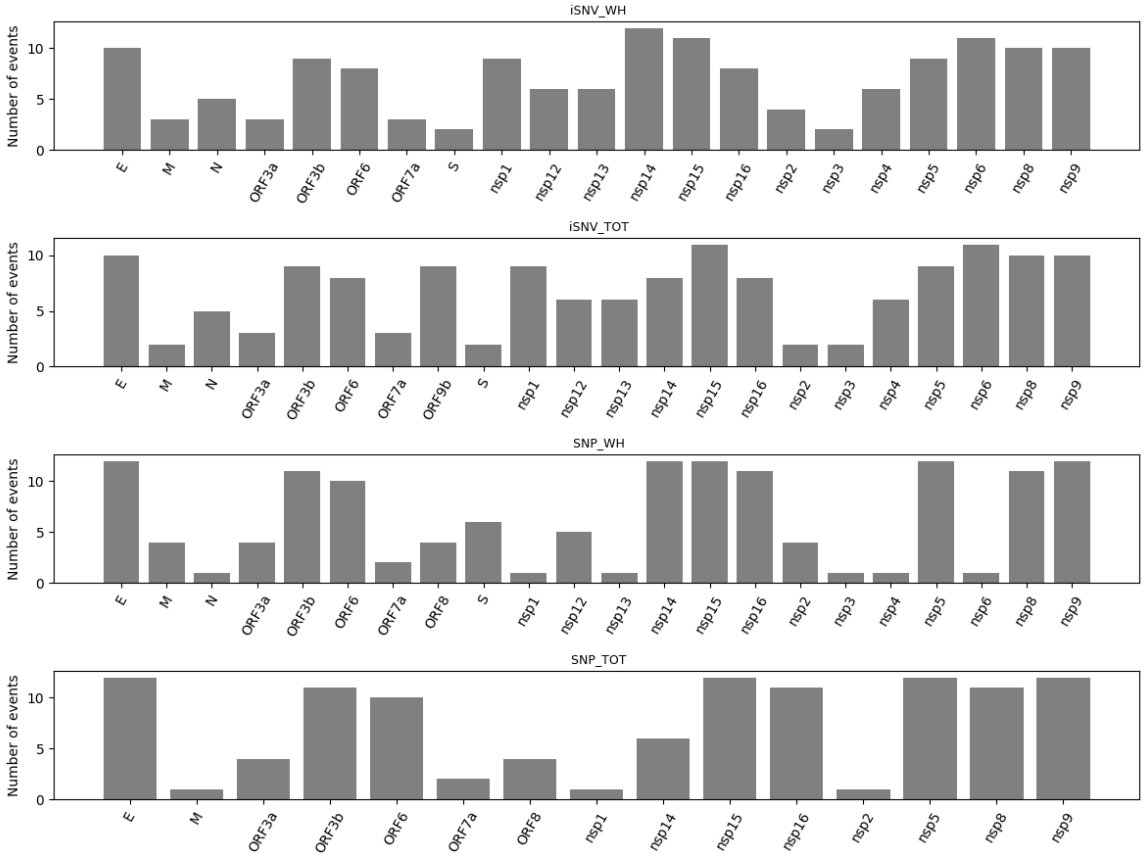


Figure S6 : Indeterminate value : pN=0 and pS=0

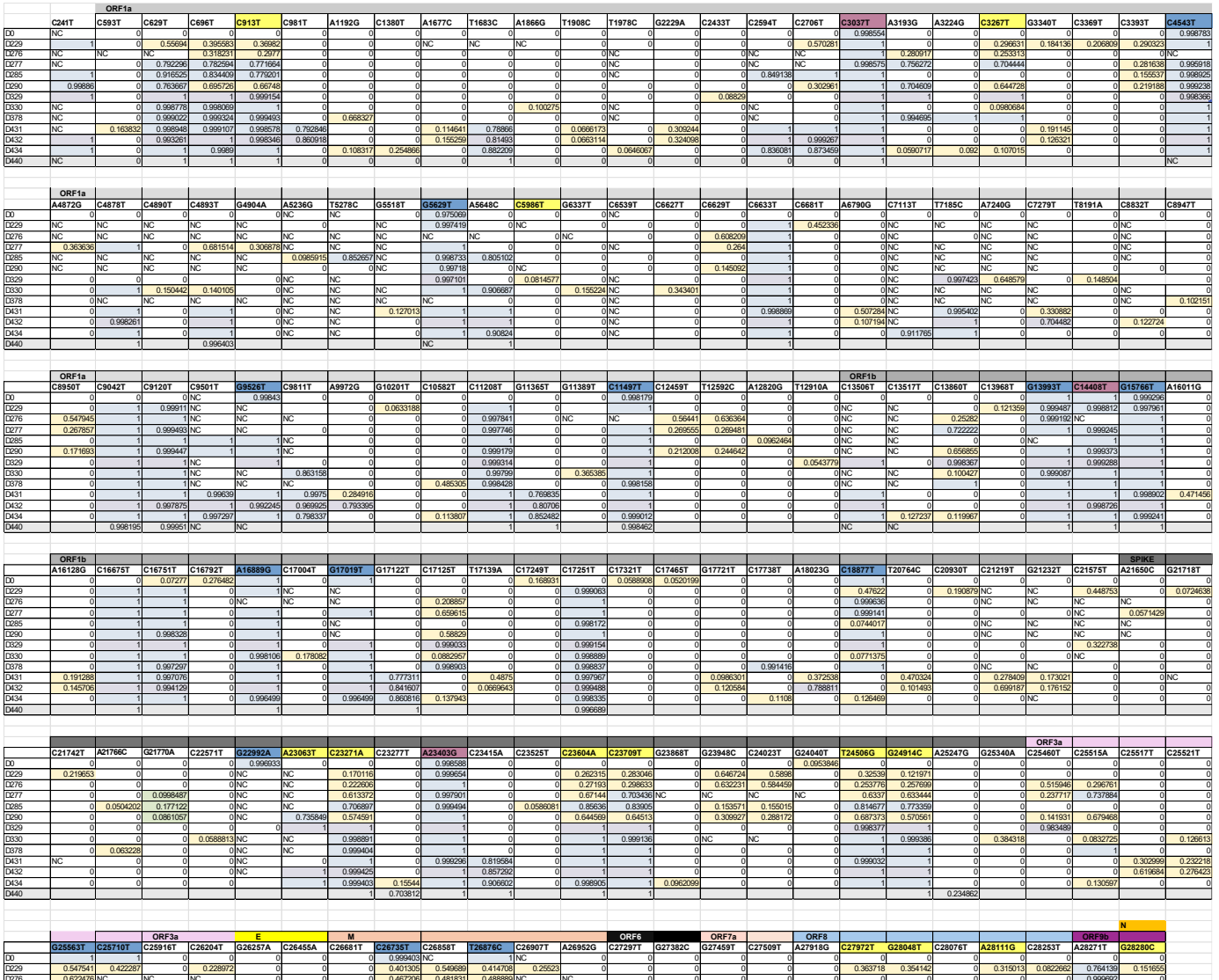


Figure S7 Mutations associated with allele frequency

LEGEND OF SUPPLEMENTARY

Table S1

Nextclade nomenclature (column 1) showing the different genotypes of B.1.1.7 circulating in Marseille from January 2020 to May 2021. We noticed the circulation of seven different B.1.1.7 genotypes. No classified as “Within host” mutation was found between them.

Table S2

Table of non-synonymous and synonymous mutation count within each pool of mutations. Genes are indicated in column 2.

Figure S1

Time-scaled maximum-likelihood phylogenetic tree constructed using Nextstrain©, comprising 700 sequences randomly selected locally in the GISAID database to cover the period from January 2020 to December 2021. Time is represented on x-axis and divergence on the y-axis.

Figure S2

Figure Legend: Table representing the intra-host sequential frequencies of mutations resulting exclusively from recombination events between parental genomes B.1.160 and B.1.1.7. Sampling dates are indicated by D0, D229, D259, etc., representing different time points during the intra-host evolution. Mutations highlighted in blue correspond to those associated with the B.1.160 variant, while mutations highlighted in yellow represent those from the B.1.1.7 variant. The frequencies (%) of each mutation at each time point are shown in the cells. Cells marked as NaN indicate regions not covered by sequencing or missing data.

Figure S3

a) Table indicating coverage (% of genome covered) at 50 and 200 reads depth. In red, genomes with < 80% coverage. b) Percentage of genes covered at 50 (top panel) and 200 (bottom panel) reads depth. c) Number of mutations counted in each pool for the 12 samples selected, at 50 and 200 read depth.

Figure S4

Boxplot comparing the number of synonymous and non-synonymous mutations in the four pools, with statistical significance indicated by p-values (Wilcoxon test).

Figure S5

Relative frequency of the different mutation types associated with their nucleotide contexts (x-axis), meaning the percentage of mutation type within a trinucleotide context.

Figure S6

Histogram showing the number of samples for which the ratio could not be calculated with pN=0 and pS=0, leading to indeterminate selection.

Figure S7

The parental genomes are highlighted with blue (B.1.160), yellow (B.1.1.7),purple for ancestral mutation, NC meaning “Not covered”,

Conclusion et perspectives

Cette thèse s'inscrit dans le contexte d'une pandémie mondiale ayant donné lieu à une mobilisation scientifique sans précédent, notamment en génomique. Le séquençage massif de génomes viraux, initié dès les débuts de la pandémie de COVID-19, a permis de suivre en temps réel l'évolution du SARS-CoV-2 à l'échelle mondiale. Toutefois, une dimension encore insuffisamment explorée demeure celle de l'évolution virale intra-hôte, en particulier dans des contextes spécifiques tels que les infections persistantes chez les patients immunodéprimés.

À travers une approche combinant données virologiques (isolement viral en culture cellulaire), génomiques (séquençage profond, PCR), bio-informatiques (analyse de SNPs, iSNVs, mesures de diversité, pN/pS) et cliniques, cette thèse a mis en évidence la richesse évolutive du SARS-CoV-2 au sein d'un même hôte sur une période prolongée. Elle illustre la complémentarité de deux mécanismes évolutifs majeurs : la recombinaison et l'émergence de mutations *de novo*. Dans un premier temps, nous avons documenté la survenue *in vivo* d'une recombinaison entre deux variants majeurs (B.1.160 et Alpha) chez un patient immunodéprimé atteint d'un lymphome. Cette observation, soutenue par des analyses PCR ciblées, des cultures virales, et le séquençage complet de génomes, constitue l'une des premières preuves expérimentales de recombinaison du SARS-CoV-2 chez l'humain avec isolement du virus. Elle démontre que les infections chroniques peuvent servir de réservoirs de biodiversité, favorisant l'émergence de génotypes mosaïques dotés de propriétés potentiellement nouvelles.

Dans un second temps, cette thèse s'est attachée à explorer de manière approfondie l'évolution intra-hôte sur la durée, grâce à un séquençage profond réalisé sur 16 prélèvements successifs issus du même patient. Cette étude a révélé un taux d'évolution intra-hôte 1,5 fois supérieur au taux inter-hôte ainsi que des dynamiques évolutives complexes. L'émergence de mutations *de*

novo, leur fixation ou élimination progressive, et les signaux de sélection positive détectés dans plusieurs gènes clés (Spike, N, ORF9b, nsp13), montrent une adaptation virale, notamment dans un contexte de déplétion lymphocytaire B. Un enrichissement marqué en mutations C>T, signature des mécanismes d'édition de l'ARN par les enzymes APOBEC, a également été observé, renforçant le rôle du microenvironnement cellulaire et la place de l'interaction hôte-pathogène dans la plasticité génomique virale. La délétion de 126 nucléotides dans ORF8, un gène pour lequel des codons stop prématurés sont apparus dans certains lignages comme le variant Alpha, souligne quant à elle la dispensabilité de ce gène pour la réplication virale.

En conclusion, ces résultats apportent une illustration unique de la dynamique intra-hôte du SARS-CoV-2 dans un contexte de portage chronique, documentant l'interaction entre mutations émergentes, recombinaisons, et pressions sélectives. Ils plaident pour une surveillance génomique ciblée des patients immunodéprimés long porteur, qui représentent de véritables incubateurs de variants évolués, porteurs potentiels d'échappement immunitaire ou de gain de fonction.

À l'issue de ces travaux, plusieurs pistes d'approfondissement peuvent être envisagées :

Une approche immunologique complémentaire aurait été pertinente, l'ajout de tests de séroneutralisation ou d'analyses du profil immunitaire (anticorps neutralisants, réponses des cellules T/B) aurait permis d'évaluer les conséquences fonctionnelles des mutations observées. Ensuite, étendre l'analyse à d'autres compartiments anatomiques (tractus digestif, voies respiratoires inférieures) aurait pu permettre de détecter des sous-populations virales distinctes, et d'évaluer un éventuel tropisme tissulaire évolutif, cet aspect serait à prendre en compte dans les prochaines études sur l'évolution intra-hôte du SARS-CoV-2. D'un point de vue technique, le recours aux sondes UMI (Unique Molecular Identifiers), déjà utilisées en virologie du VIH[39], permettrait une reconstitution plus fiable des haplotypes et l'élimination des artefacts techniques. Cette méthode favoriserait une détection plus précise des variants minoritaires, il

serait envisageable de développer cette technique au laboratoire pour améliorer la détection des iSNVs et permettre une reconstruction fiable et aisée des haplotypes.

D'autre part, à l'heure actuelle, les bases de données internationales ne permettent pas encore un dépôt fluide des données brutes de type iSNVs. Il serait opportun de développer des formats de dépôt standardisés pour intégrer de manière systématique ces données dites sub-consensuelles dans la surveillance mondiale.

Sur le plan structurel, poursuivre la modélisation moléculaire des autres protéines affectées par l'évolution, en particulier celles codées par l'ORF1ab, permettrait d'établir un lien entre les mutations observées et leurs conséquences sur la structure et potentiellement sur la fonction protéique.

Enfin, la limite de cette étude repose sur le nombre n=1 de cas analysé. Une comparaison inter-hôtes en étendant cette démarche à d'autres cas similaires permettrait d'étudier plus en détail les mécanismes de convergence évolutif.

Annexe 1 Etude d'une épidémie à Trichosporon dans un hôpital de Marseille

Avant-propos

Au cours de ma thèse, j'ai eu l'opportunité de participer à un projet collaboratif mené avec les services de parasitologie, néphrologie et cardiologie de l'hôpital de Marseille, à la suite de la détection de plusieurs cas de trichosporonose sur une période de deux ans.

Quatre cas d'infection invasive causée par *Trichosporon austroamericanum*, un pathogène fongique émergent[40] initialement identifié comme *T. inkin* par MALDI-TOF, ont été signalés dans un même service, suivis de deux autres cas dans des services distincts.

Une analyse phylogénétique, basée sur les génomes nucléaires et mitochondriaux de ces isolats cliniques, a permis de révéler une épidémie polyclonale, probablement d'origine environnementale.

Ma contribution a porté sur la définition et la mise en œuvre de la stratégie analytique la plus adaptée pour évaluer la clonalité de l'épidémie, en exploitant des données issues de séquençages mixtes Illumina et Nanopore.

La stratégie d'assemblage *de novo* a permis de fournir le génome mitochondrial en un seul contig et de comparer ces génomes MT en fonction des isolats ; une collaboration avec l'Institut Pasteur nous a permis d'obtenir d'avantage d'isolats et de proposer ainsi la séquence mitochondriale comme séquence potentiellement discriminante intra espèce. L'annotation du génome complet de *Trichosporon austroamericanum* a constitué un défi, s'agissant des premiers génomes complets décrits. Le génome de *Neoccoccus neoformans* a ici été utilisé comme génome annoté de référence, par sa proximité phylogénétique. Cette annotation a permis par la suite de construire la phylogénie du core-génome de ces isolats, révélant une épidémie polyclonale, renforçant les données obtenues à partir des génomes MT. Ce travail a été soumis dans Emerging Infectious Diseases.

EMERGING INFECTIOUS DISEASES®

A genome-based investigation of emerging *Trichosporon austroamericanum* infections at the University Hospital of Marseille

Journal:	*Emerging Infectious Diseases*
Manuscript ID	EID-25-0503
Manuscript Type:	Research
Date Submitted by the Author:	04-Apr-2025
Complete List of Authors:	Burel, Emilie; Aix-Marseille Universite Sartor, Catherine; AP-HM Moal, Valerie; AP-HM, Service de Néphrologie Bossi, Vincent; IHU Méditerranée Infection, Microbes Evolution Phylogeny and Infections (MEPHI) Sevestre, Jacques; Assistance Publique Hopitaux de Marseille, IHU Mediterranee Infection Solignac, Justine; AP-HM Charrel, Rémi; AP-HM Desnos-ollivier, Marie; Pasteur Institute, Natl. Ref. Center for Invasive Mycoses and Antifungals Ranque, Stéphane; Aix-Marseille Universite, Laboratory of Parasitology and Mycology; AP-HM, Laboratory of Parasitology and Mycology MENU, Estelle; AP-HM, Parasitology-Mycology; Aix-Marseille Universite
Keywords:	Trichosporon austroamericanum, Emerging disease, Trichosporonosis, yeast, opportunistic agents, outbreak, genomic
Abstract:	Invasive fungal infections were investigated at Marseille University Hospital, involving six immuno-compromised patients during two years. Isolates were initially misidentified as *Trichosporon inkin* by MALDI-TOF. Phylogenetic analysis of IGS1 and whole-genome sequencing revealed a genetically distinct species, *Trichosporon austroamericanum*. Identified for the first time in 2024, but potentially underestimated, this species is now being reported in grouped cases in French hospital setting. We present the first complete genome assemblies of seven *T. austroamericanum* strains, including a hospital environmental isolate. Phylogenetic analysis based on core-genome and mitogenome revealed significant genetic diversity among strains, indicating a polyclonal outbreak. Furthermore, the mitochondrial genome emerged as a potential marker for intra-species differentiation, with potential utility in epidemiological investigations. This research highlights the need for accurate fungal identification in hospital settings and suggests re-evaluating potentially misidentified *T. inkin* cases. Our findings emphasize the significance of *T. austroamericanum* as an emerging pathogen.

Target journal and type of article: Emerging Infectious Disease, Research Article

Full-length title:

A genome-based investigation of emerging *Trichosporon austroamericanum* infections at the University Hospital of Marseille

Short title (for the running head):

***Trichosporon austroamericanum* infections: A genomic study**

Author list:

Emilie Burel[1,5], Catherine Sartor[2], Valérie Moal[3,4], Vincent Bossi[5], Jacques Sevestre[6,7], Justine Solignac[3], Rémi Charrel[2,8,9], Marie Desnos-Ollivier[10], Stéphane Ranque[6,7], Estelle Menu[6,7,*]

[1]Aix-Marseille Université

[2] Equipe Opérationnelle d'Hygiène Hospitalière, Hôpital de la Conception, 13385 Marseille, France.

[3]Aix Marseille Université, Assistance Publique Hôpitaux de Marseille, Hôpital Conception, Centre de Néphrologie et Transplantation Rénale, 147 Boulevard Baille, 13005, Marseille, France.

[4]Aix Marseille Université, Institut de Recherche et Développement, Microbes Evolution Phylogeny and Infections, 27 Boulevard Jean Moulin, 13005, Marseille, France.

[5]Institut Méditerranée-Infection, Marseille, France

[6]UMR D257 RITMES (Aix-Marseille University, Public Assistance-Marseille Hospitals, French Military Health Service), Marseille, France.

Peer Review

[7]Assistance publique-hôpitaux de Marseille (AP-HM), Service de Parasitologie-Mycologie, Marseille, France.

[8]Unité des Virus Émergents (UVE: Aix-Marseille Univ, Università di Corsica, IRD 190, Inserm 1207, IRBA), Marseille, France.

[9]Assistance publique-hôpitaux de Marseille (AP-HM), Service de virologie aiguë et tropicale, Marseille, France.

[10]Natl. Ref. Center for Invasive Mycoses and Antifungals, Institut Pasteur, 25 Rue du Dr Roux, 75015 Paris, France.

First author biography sentence : Emilie is a PhD student in Genomics and Bioinformatics at the University of Aix Marseille, France. ORCID: 0009-0000-6068-7658

*** Corresponding author:** Estelle Menu, Assistance publique-hôpitaux de Marseille (AP-HM), Service de Parasitologie-Mycologie, 19-21 boulevard Jean Moulin, 13005 Marseille, France. Tel.: +33 413 732 401, Fax: +33 413 732 402; email: estelle.menu@ap-hm.fr

ORCID: 0000-0002-0681-8431

Key words (6): *Trichosporon austroamericanum*, Emerging disease, Trichosporonosis, yeast, opportunistic agents, outbreak

Figures: 4; **Tables:2** ; **References: 26**

Peer Review

ABSTRACT

Invasive fungal infections were investigated at Marseille University Hospital, involving six immuno-compromised patients during two years. Isolates were initially misidentified as *Trichosporon inkin* by MALDI-TOF. Phylogenetic analysis of IGS1 and whole-genome sequencing revealed a genetically distinct species, *Trichosporon austroamericanum*. Identified for the first time in 2024, but potentially underestimated, this species is now being reported in grouped cases in French hospital setting. We present the first complete genome assemblies of seven *T. austroamericanum* strains, including a hospital environmental isolate. Phylogenetic analysis based on core-genome and mitogenome revealed significant genetic diversity among strains, indicating a polyclonal outbreak. Furthermore, the mitochondrial genome emerged as a potential marker for intra-species differentiation, with potential utility in epidemiological investigations. This research highlights the need for accurate fungal identification in hospital settings and suggests re-evaluating potentially misidentified *T. inkin* cases. Our findings emphasize the significance of *T. austroamericanum* as an emerging pathogen.

INTRODUCTION

Fungi belonging to various species of the *Trichosporon* genus species can cause a wide range of infections from superficial skin damage to serious systemic infections in immunocompromised individuals with high mortality rate (1). This genus, which belongs to the Basidyomycota, includes approximately twenty species that are pathogenic to humans (2). Although *Cutaneotrichosporon cutaneum* and *Trichosporon asahii* are largely predominant species, new species are increasingly being described in human pathology in recent years. In 2024, a new species named *Trichosporon austroamericanum* has been identified from a urine sample of a Brazilian kidney transplant recipient. It has also been detected in several

Peer Review

clinical samples in France, South America and Asia (3). Closely related to *Trichosporon inkin*, both species have been confounded in the past, in the absence of a molecular tool analysis to distinguish them, notably sequence analysis of the IGS1 region.

Ubiquitous in the environment, some species of *Trichosporon* were isolated from soil, leaf mould and decayed wood (4). The environmental reservoir of *T. inkin* and *T. austroamericanum* remains unknown. Moreover, the emergence of hospital cases of trichosporonosis, caused by *T. asahii,* is described in the literature, highlighting hospital sources during epidemiological and environmental investigations (5). These pathogens should not be overlooked in the hospital environment.

From July 2022 to June 2024, 4 cases of invasive infection caused by *T. austroamericanum* were recorded in the kidney transplant department in La Conception university Hospital in Marseille, France. Due to the rarity of the species a putative common source was suspected. Moreover 2 other cases were subsequently diagnosed in patients admitted in cardiology and digestive surgery in 2 others hospitals of our institution. In an attempt to explain this sudden upsurge of trichosporonosis cases, complete genome sequences were obtained by WGS and compared with sequences from public databases.

MATERIALS AND METHODS

Patients were identified prospectively over a 2-year period during routine patient care at Marseille University Hospital, France. All patients with *T. austroamericanum* isolated from sterile sites with associated signs and symptoms of infection were included.

Patients' data

Retrospective medical data were collected, when available, using the Axigate DPI (version 5.11.3P10.8.3) for each patient and processed in an anonymized MS Excel 2013 software. These data included demographic data, exposure factors, underlying diseases, symptoms. Biological Data including fungal culture on Sabouraud dextrose agar supplemented with

Peer Review

gentamicin and chloramphenicol (Bio-Rad, Marnes-la-Coquette, France), cryptococcal antigen in serum using CrAg lateral flow assay (Immy, Norman, USA) and Bio-Rad Platelia *Aspergillus* Antigen (Bio-Rad, Marnes-la-Coquette, France) in serum were collected.

Mycological investigation

An environmental investigation was initiated in 2023 by the Infection Control Team (ICT) and pursued in 2024 when the last infected patient was identified. It focused on the rooms and units where the patients had been cared, from the date of their admission for kidney transplantation to their discharge, including the radiology unit, the intensive care unit, in the urological surgery operating theater and the dialysis unit. The investigation was also performed in the nephrology daycare, where patients had been weekly hospitalized after the kidney transplantation. Air samples were collected with a microbiological air sampler (SAMPL'AIR™ Lite AESAP1075, AES Laboratoire, France), collecting 330L on a Sabouraud Dextrose Agar. Water samples were collected in a 250mL bottle containing sodium thiosulfate. Surface sample were collected using swab with liquid amies preservation medium (eSwab, Copan Italia SpA, Italy) and cultured on Sabouraud Dextrose Agar.

Next generation sequencing

DNA Extraction

DNA extraction was performed on pure subculture from Sabouraud dextrose agar supplemented with gentamicin and chloramphenicol (Bio-Rad, Marnes-la-Coquette, France), using the EZ1™ DNA Tissue Kit (Qiagen, France) with the 10,677,990 V1.0 DNA bacteria card in an EZ1 Advanced XL extractor. Colonies from the plate agar were pre-treated in a tube containing ceramic beads (Lysing matrix D tube, MP Biomedicals Germany GmbH, Eschwege, Germany) with 500 µL of NucliSENS™ easyMAG™ lysis buffer (BioMérieux, Marcy-l'Etoile, France) and then disrupted in a FastPrep BIO 101 apparatus (Qbiogene, Strasbourg, France) at maximum power for 40 s. The tubes were then centrifuged at 10,000×

Peer Review

g for one minute. The procedure was then performed on 200 µL of supernatant.

Illumina Novaseq protocol

Genomic DNA was sequenced on the NovaSeq 6000 Technology (Illumina Inc, San Diego, CA, USA) with the paired end strategy and was barcoded in order to be mixed respectively with many other genomic projects prepared with the IHU adaptation of the covidseq test sample prep kit (Illumina). The « tagmentation » step fragmented and tagged the DNA. Then limited cycle PCR amplification (12 cycles) completed the tag adapters and introduced dual-index barcodes. After purification on ITB beads (Illumina Inc, San Diego, CA, USA,) the libraries have been normalized to the same molarity. Normalized libraries were pooled into a single library for sequencing on the NovaSeq. The pooled single strand library was loaded onto the reagent cartridge and then onto the instrument along with the flow cell. Automated cluster generation and paired end sequencing with dual index reads were performed in a single 25-hours run in 2x150-bp.

Nanopore Protocol (GridION + PromethION)

Genomic DNA was sequenced on the P2Solo (Oxford Nanopore Technologies, Oxford, Royaume-Uni) individually using the LSK109 ligation kit on a FLO-PRO002 flowcell. The “end-prep” step allows the fixation of specific nucleotides which will be used for the ligation of the adapters. After ligation of the adapters, purification of the DNA is carried out on CleanNA beads (Coenecoop 75, 2741 PH Waddinxveen, Pays-Bas). The flowcell is activated by the addition of a Flush Buffer and a Tether. The libraries are then loaded onto the flowcell for a 72-hour run.

Bioinformatics

Illumina raw bcl files were converted into fastq file using bclconverter v4.2.4 © 2024 Illumina, fastq files were then trimmed and adapters were removed using Trimmomatic v0.39 (6), the minimum base quality Phred score was set at 33 and the minimum read length at 36.

Peer Review

Nanopore raw reads were trimmed using ProwlerTrimmer (7), with a Phred score ranging from 20 to 35, and a minimum read length of 1000.

Complete genomes obtained by mixed de novo assembly: Illumina and Nanopore reads were assembled using Unicycler v0.4.4 (8), the resulting assemblies were cleaned by removing the few contigs < 1000 bp, GC content and contamination evaluation were obtained using the tool available at https://www.ezbiocloud.net/tools/contest16s. Assembly quality was assessed using BUSCO v.5.7.1 (9) with fungi_odb10 as the reference database.

Mitochondrial genomes: Mitochondrial genomes were isolated into a single contig from the assemblies, identified as the only recognized circular contig of around 40kb in size. Contigs were aligned using mauve v2.4.0 (10).The identified mitochondrial contigs were manually adjusted to ensure they started from the COX1 start codon on the forward strand. When necessary, reverse complements were generated using the tool available at https://reverse-complement.com/terms.html. Genomes were annotated using https://chlorobox.mpimp-golm.mpg.de/geseq.html and sequences deposited in Genbank database (https://www.ncbi.nlm.nih.gov/genbank/).

Phylogenetic analysis: Trimmed illumina reads were mapped against the reference IGS1 CBS17435 sequence using Burrows-Wheeler Aligner (11), consensus sequences were obtained from .bam files using sam2consensus (https://github.com/edgardomortiz/sam2consensus). Consensus sequences were aligned using mafft (https://mafft.cbrc.jp/alignment/server/index.html) and subsequent phylogenetic analysis was made using IQ-TREE (bstp 1000) (12). Tree was visualized and edited on the online platform iToL (https://itol.embl.de/upload.cgi).

Maximum likelihood tree based on the core-genome

The core-genome tree was constructed using the Galaxy platform (usegalaxy.eu). Repeats were masked with RepeatMasker (Galaxy Version 4.1.5+galaxy0), and the assemblies were

Peer Review

subsequently annotated using the Maker annotation pipeline (13) with the Augustus predefined prediction model *Cryptococcus neoformans gattii* (Galaxy Version 2.31.11+galaxy2). The alignment core-genome was derived from the resulting annotation GFF files with the tool Roary (Galaxy Version 3.13.0+galaxy3), then the SNPs were collected using the tool Find SNP sites (Galaxy Version 2.5.1+galaxy0) and used to build the maximum likelihood tree with IQ TREE (2.3.6 built Aug 4 2024) (12), under the model ModelFinder (14). The resulting tree, in Newick format, was visualized on the iTOL platform (https://itol.embl.de/upload.cgi).

Nucleotide sequence accession numbers

RESULTS

Patient's data

By June 2024, a total of six patients with *T. austroamericanum* infection had been identified (Table 1). The median age was 65.5 years (range 55–83 years); 3 (50%) were male. Four patients were kidney transplant recipients who presented clinical signs of an infection between 2 and 4 months after transplantation. All four presented with scar disunion and subcutaneous abscesses, suggesting that the scar was the portal of entry. The two other patients were immuno-competent. Patient 4 was a pacemaker holder with a history of *Listeria monocytogenes* endocarditis, not having undergone recent surgery, initially hospitalized for heart failure decompensation, with a complicated progression in the department. The patient P4 was once referred to a nephrology consultation for an assessment regarding renal failure. Patient 5 developed a shock and acute respiratory distress syndrome in the context of acute post-cholangio-pancreatography retrograde endoscopic pancreatitis. Of the 25 samples testing positive for *T. austroamericanum*, 16 (64%) were blood cultures in two patients, three scar swab in one patient, three abscess aspiration in three patients, three respiratory samples (two

Peer Review

bronchial aspirate and on broncho-alveolar lavage fluid) in one patient.

Mycological investigation

An environmental survey was carried out from April 20th 2023 to July 19th 2024 in the Conception university Hospital. A total of 145 surface samples, 33 air samples and 17 water samples were collected in the urological surgery operating theater and 6 units : the nephrology intensive care unit, the kidney transplantation unit, the nephrology daycare, the radiology unit, and the dialysis unit. Samples were mainly carried out in the room of the patients, the reserve of medical devices and the decontamination room. All the 195 samples were negative. A sample taken as part of a routine assessment in a pediatric intensive care unit at La Timone Hospital isolated a strain of *T. austroamericanum* in an air sample.

Molecular investigation with Next generation sequencing

Multi-Locus Sequence Typing (MLST) approach

Whole-genome sequencing (WGS) was initially performed on the first three reported cases, which had been identified as *T. inkin* using MALDI-TOF. To investigate genetic variations, a Multi-Locus Sequence Typing (MLST) approach was applied (Supplementary Figure S1). Sequences of *cytb*, *rpb*1/2, *tef*1 genes and the SSU and D1/D2 regions were mapped against the reference genome of *T. inkin* (strain MT801082). Although the sequences were identical among the isolates, significant differences from *T. inkin* emerged, except for the D1/D2 and SSU markers which were identical except for one nucleotide variation. Additionally, high divergence was observed between the strains and *T. inkin* reference mitochondrial genome and IGS1 marker, characterized by numerous point mutations, deletions, and insertions (Supplementary Figure S2,3). The consistent mutation patterns across the strains, along with their divergence from *T. inkin*, strongly supported the classification of these isolates as a distinct taxonomic entity. This preliminary analysis served as the starting point for further investigation of the six patients-derived isolates into the IGS1 region. The IGS1 sequences

Peer Review

from the six patients and hospital environment, mapped against the newly described *T. austroamericanum*, were fully covered. Phylogenetic analysis, which included three reference sequences provided by the National Reference Centre for Invasive Mycoses & Antifungals (NRCMA), confirmed that the 7 strains belonged to *T. austroamericanum* species and not to the *T. inkin* species (Figure 1).

WGST (Whole-Genome Sequencing Typing)

We conducted Whole-Genome Sequencing Typing (WGST) to assess the clonal nature of the *T. austroamericanum* outbreak. The genomes were assembled using a hybrid approach combining Illumina and Nanopore technologies. Out of the seven genomes, six demonstrated excellent assembly quality, with a very low number of contigs relative to the genome size of approximately 21 Mb. Specifically, the number of contigs ranged from 12 to 31, with L0221 having 14 contigs, L0385 with 12, L0399 with 16, L0419 with 13, L0453 with 31, and L0458 with 24 contigs. These assemblies showed consistent genome lengths between 20.8 and 21 Mb, and GC content close to 61.31%. The BUSCO scores, which reflect genome completeness, were all above 96%, indicating highly complete and accurate assemblies (ranging from 96.50% to 97.40%). However, the assembly for strain L0445 stood out with a much higher number of contigs (1,051), a larger genome size of 26 Mb, and a higher GC content (62.59%). This assembly also showed contamination with bacterial *Achromobacter xylosoxidans* 16S sequence and was excluded from further analyses.

The core-genome consisting of 11,957 core genes, was obtained using the *Cryptococcus neoformans* Augustus predefined prediction model and was subsequently aligned, allowing for the construction of a phylogenetic tree based on the extracted SNPs, forming six sequences with 542 nucleotide sites and 196 parsimony-informative positions. This analysis revealed distinct genetic differences among the strains (Figure 2A). Out of the 542 sites, 129 distinct site patterns were observed, indicating a significant level of sequence variation across

Peer Review

the strains. Strains L0453 and L0221, separated by only 85 SNPs (Figure 2B), are the closest in the tree compared to the other strains. L0453, from the ICU air-duct, also appears near L0458, with a difference of 97 SNPs between them. Despite these proximities, the genetic differences remain substantial, until 434 SNPs, indicating that none of these strains form a homogeneous clade or emerged from a recent clonal expansion and that we are facing a *T. austroamericanum* polyclonal epidemic. In addition, 7,668 shell genes were identified, representing genes shared by a subset of strains, in favor of genetic diversity beyond the core genome.

Mitochondrial genome

As a complementary approach, we analyzed the mitochondrial genome to assess its potential for discriminating between strains. The mitochondrial genomes with a size of approximately 44 kb of three *T. austroamericanum* reference strains from CNRMA (highlighted in dark red on the tree) and the L0445 strain were obtained through mapping, while the remaining six were generated via *de novo* assembly (Figure 3). Surprisingly, synteny was identical between the mitogenomes assembled *de novo*.

The alignment of these mitochondrial genomes against the *de novo* assembled L0221 sequence revealed four distinct mutational profiles including three among Marseille's strains (Figure 4). It was by three key mutated positions in the mitochondrial genome located in the *cox3*, *apocytochrome b*, and *trnL* (tRNA-Leu) genes (positions 22,240, 31,837, and 39,897). These mutations provide clear genetic distinctions among the strains. For instance, strain CNRMA15 exhibits a unique mutation (C at position 22,240), whereas other strains, like L0221, have a G at this position. This analysis underscores the value of mitochondrial genome sequencing for identifying and distinguishing between *T. austroamericanum* strains involved in this outbreak and beyond. The phylogenetic placement showed four distinct clades, L0221_MT and CNRMA15 form a separate branch due to their unique SNP profile,

Peer Review

L0458_MT and L0453_MT are closely related, as seen in the core-genome analysis, while L0385_MT, L0419_MT, and L0399_MT, L0445_MT, CNRMA22 and CNRMA16 cluster together.

Discussion

With approximately twenty pathogenic species, *Trichosporon* are responsible for the second leading cause of Basidiomycota infection in humans (2). Although rare, *Trichosporon* infections in solid organ transplant patients have previously been reported and should be considered in cases of breakthrough infection, or therapeutic failure with echinocandins (15). In this case series, 67% of patients with invasive *T. austroamericanum* infection were kidney transplant recipients. For all kidney transplant recipients, the starting point was the graft scar and infection occurred within 2-4 months of transplantation, suggesting a common source of contamination at the time of surgery. Interestingly, two patients in two different hospitals also presented with invasive *T. austroamericanum* infections thus questioning the "common source" assumption.

T. austroamericanum was first described in May 2024 by Francisco et al. (3). As our first case appeared in July 2022, *Trichosporon* species were initially misidentified as *T. inkin* with MALDI TOF, the closely linked species, as our database did not yet contain references for *T. austroamericanum*. First, in the absence of a satisfactory reference genome we decided to focus on the mitochondrial genome, the only one available and well described in the literature (16). Considerable differences were observed between the genomes of strains isolated from our patients #1 to #6, and the reference mitochondrial genome of *T. inkin* available, in term of nucleotidic sequence (Supplementary Figure 3). The work of Francisco et al. has reinforced our analyses, with the description of a new species (3). The intergenic spacer 1 regions is the region of interest for discriminating between different *Trichosporon* species, which other widely used as the internal transcribed spacer region or D1/D2 domains of ribosomal DNA,

Peer Review

fail to do (3,17,18). In fact, comparing the rRNA IGS1 region nucleotide sequence between our strains with the *T. austroamericanum* strains supplied by the National Reference Centre for Invasive Mycoses & Antifungals, enabled us to confirm the result of mitochondrial genome analysis. We were indeed facing an emergence of *T. austroamericanum* cases in our hospital. This highlights the fact that mitochondrial genome analysis of *Trichosporon* species would enable us to distinguish between species.

Moreover, the mitochondrial genome showed three different profiles within the *T. austroamericanum* strains isolated in Marseille, with three key mutated positions in the mitochondrial genome found. The appearance of spontaneous mutations has been described in the mitochondrial genome of certain phytopathogenic species, probably due to replication errors and the presence of mobile genetic elements (19). Genomic analysis then enabled us to answer the question of a potential common origin (20). We have produced the first complete genomes of *T. austroamericanum* by de novo whole genome sequencing. Whole genome sequencing and single nucleotide polymorphism (SNP) analysis has already been used to identify clusters of *Trichosporon asahii* and *Rhodotorula* infection, another Basidiomycete (21,22). Regarding SNPs, there is no consensus among fungal species given the need for a large number of strains to study mutation rates, which is not always possible in the case of rarer species. Thus, the cut-off point seems to be set at <13 SNPs for *Candida auris* to determine whether strains are related, and approach 100 SNPs for others like *Rhodotorula mucilaginosa* (21,23). The rDNA usually used in fungal epidemic investigation based on SNPs does not allow to discriminate strains belonging to the *Trichosporon* genus at the species level (18). Moreover, the rRNA IGS1 region nucleotide is limited to identification at the species level. Drawing inspiration from epidemic investigations of bacteria, we chose to target conserved sequences for a core-genome SNP-typing approach (24). These types of approaches have also been used in the investigation of previous fungal epidemics (25).

Peer Review

During our investigation, *T. austroamericanum* strains ranged from 85 to 434 SNPs, restricted to the core-genome, suggesting absence of sustained transmission and polyclonal pseudo-outbreak. Interestingly, phylogenetic analysis of the mitochondrial genome and the core genome showed similar results, highlighting it interest among epidemiological fungal analysis.

One limitation of this study is the relatively low number of genes identified in the core-genome, with only 11,957 genes, which is small considering the genome size of 21 Mb and the diversity of the strains is probably underestimated, but it has nevertheless been highlighted by our approach. We used the predefined prediction model from Augustus for *Cryptococcus neoformans*, another Basidiomycota being the closest available model, although it remains distantly related to the *Trichosporon* genus. This highlights the importance of performing a more specific functional annotation of the genome, particularly with the support of RNA-seq data. RNA-seq has been shown to enhance gene annotation and uncover species-specific characteristics. Grandaubert et al. used RNA-seq to identify novel orphan genes and species-specific transposable elements in fungal pathogens (26). Applying RNA-seq-based annotation in *Trichosporon* could potentially reveal strain-specific features, such as antifungal resistance mechanisms or environmental adaptations that might not be evident using a distant reference model like *C. neoformans.*

An environmental survey was also carried out to identify a source of contamination. Investigation of the cases in the kidney transplant and nephrology departments did not reveal any common source of contamination. Other *Trichosporon* involved in outbreaks have already been isolated from the hospital environment. Fanfair et al. identified 53 patients with positive *T. asahii* cultures in a Jamaican hospital, four of whom hospitalized in an intensive care unit developed invasive *T. asahii* infection (5). Following an environmental investigation, ten surface swabs from the patients' bedrooms showed *T. asahii*, including

Peer Review

drawers, bed rails, faucets and sinks (5). In our investigation, *T. austroamericanum* (L0453) was isolated from an air sample of a pediatric intensive care unit at La Timone Hospital while no infection or colonization with this microorganism was noted among the patients. There was no epidemiological link between this unit and the reported clinical cases, who were treated in other hospitals. This suggests that this species is part of the hospital environment. The analyses regarding the studies showed between 97 and 434 SNPs of differences between the clinical strains and the environmental strain (L0453). However, this does not exclude the possibility of multiple strains circulating, acquired within hospital environmental 'reservoir' containing clonally unrelated isolates.

The results of this study demonstrate that the core-genome analysis can effectively differentiate strains of *Trichosporon austroamericanum* showing a polymorphic species. Furthermore, the mitochondrial genome shows strong potential as an excellent marker for intra-species differentiation. This approach is particularly valuable in the absence of available annotated genomes for *Trichosporon austroamericanum.* By providing the first annotated mitochondrial genome obtained by *de novo* assembly, this study contributes to initiating the knowledge field on the *T. austroamericanum* mitogenome.

Conclusion

This study underscores the necessity of vigilance regarding *Trichosporon austroamericanum* infections, including the potential for nosocomial involvement. It also encourages a re-evaluation of certain deposited sequences previously identified as *T. inkin*, to account for the emergence of this newly described species.

Author contributions

Conceptualization: EM, SR, EB; Contributed materials/analysis tools: EM, EB, CS, VM, JS, JSo, MDO, RC; Analyzed the data: EM, EB; Writing: EM, EB; review & editing: EM, EB, CS, VM, JSo, RC, MDO, SR; Supervision: EM.

Peer Review

All authors have read and agreed to the published version of the manuscript.

Funding

This work was supported by the French Government under the "Investments for the Future" program managed by the National Agency for Research (ANR), Mediterranean-Infection 10-IAHU-03 and was also supported by Region "Provence Alpes Côte d'Azur" and European funding FEDER PRIMMI (Fonds Européens de Développement Régional-Plateformes de Recherche et d'Innovation Mutualisées Méditerranée Infection), FEDER PA 0000320 PRIMMI.

Conflicts of interest

The authors have no conflicts of interest to declare. Funding sources had no role in the design and conduct of the study; collection, management, analysis, and interpretation of the data; and preparation, review, or approval of the manuscript.

Ethics

A declaration of the database to the Assistance-Publique des Hôpitaux de Marseille Health Data Access Portal (PADS24-275) and ethical comity (CSE24-49) has been made.

REFERENCES

1. Miceli MH, Díaz JA, Lee SA. Emerging opportunistic yeast infections. Lancet Infect Dis. 2011 Feb;11(2):142–51.
2. Menu E, Filori Q, Dufour JC, Ranque S, L'Ollivier C. A Repertoire of the Less Common Clinical Yeasts. J Fungi (Basel). 2023 Nov 11;9(11):1099.

Peer Review

3. Francisco EC, Desnos-Ollivier M, Dieleman C, Boekhout T, Santos DW de CL, Medina-Pestana JO, et al. Unveiling Trichosporon austroamericanum sp. nov.: A Novel Emerging Opportunistic Basidiomycetous Yeast Species. Mycopathologia. 2024 May 6;189(3):43.
4. Sugita T, Nishikawa A, Ichikawa T, Ikeda R, Shinoda T. Isolation of Trichosporon asahii from environmental materials. Med Mycol. 2000 Feb;38(1):27–30.
5. Fanfair RN, Heslop O, Etienne K, Rainford L, Roy M, Gade L, et al. Trichosporon asahii among intensive care unit patients at a medical center in Jamaica. Infect Control Hosp Epidemiol. 2013 Jun;34(6):638–41.
6. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics. 2014 Aug 1;30(15):2114–20.
7. Lee S, Nguyen LT, Hayes BJ, Ross EM. Prowler: a novel trimming algorithm for Oxford Nanopore sequence data. Bioinformatics. 2021 Nov 5;37(21):3936–7.
8. Wick RR, Judd LM, Gorrie CL, Holt KE. Unicycler: Resolving bacterial genome assemblies from short and long sequencing reads. PLoS Comput Biol. 2017 Jun;13(6):e1005595.
9. Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics. 2015 Oct 1;31(19):3210–2.
10. Darling ACE, Mau B, Blattner FR, Perna NT. Mauve: multiple alignment of conserved genomic sequence with rearrangements. Genome Res. 2004 Jul;14(7):1394–403.
11. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 2009 Jul 15;25(14):1754–60.
12. Minh BQ, Schmidt HA, Chernomor O, Schrempf D, Woodhams MD, von Haeseler A, et al. IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. Mol Biol Evol. 2020 May 1;37(5):1530–4.

Peer Review

13. Cantarel BL, Korf I, Robb SMC, Parra G, Ross E, Moore B, et al. MAKER: an easy-to-use annotation pipeline designed for emerging model organism genomes. Genome Res. 2008 Jan;18(1):188–96.
14. Kalyaanamoorthy S, Minh BQ, Wong TKF, von Haeseler A, Jermiin LS. ModelFinder: fast model selection for accurate phylogenetic estimates. Nat Methods. 2017 Jun;14(6):587–9.
15. Almeida Júnior JN, Song ATW, Campos SV, Strabelli TMV, Del Negro GM, Figueiredo DSY, et al. Invasive Trichosporon infection in solid organ transplant patients: a report of two cases identified using IGS1 ribosomal DNA sequencing and a review of the literature. Transpl Infect Dis. 2014 Feb;16(1):135–40.
16. Liu Q, Wang X. Characterization and phylogenetic analysis of the complete mitochondrial genome of pathogen Trichosporon inkin (Trichosporonales: Trichosporonaceae). Mitochondrial DNA B Resour. 2021 Mar 11;6(3):803–5.
17. Araujo Ribeiro M, Alastruey-Izquierdo A, Gomez-Lopez A, Rodriguez-Tudela JL, Cuenca-Estrella M. Molecular identification and susceptibility testing of Trichosporon isolates from a Brazilian hospital. Rev Iberoam Micol. 2008 Dec 31;25(4):221–5.
18. Sugita T, Nakajima M, Ikeda R, Matsushima T, Shinoda T. Sequence analysis of the ribosomal DNA intergenic spacer 1 regions of Trichosporon species. J Clin Microbiol. 2002 May;40(5):1826–30.
19. Mendoza H, Perlin MH, Schirawski J. Mitochondrial Inheritance in Phytopathogenic Fungi-Everything Is Known, or Is It? Int J Mol Sci. 2020 May 29;21(11):3883.
20. Douglas AP, Stewart AG, Halliday CL, Chen SCA. Outbreaks of Fungal Infections in Hospitals: Epidemiology, Detection, and Management. J Fungi (Basel). 2023 Oct 29;9(11):1059.
21. Huang JJ, Chen XF, Tsui CKM, Pang CJ, Hu ZD, Shi Y, et al. Persistence of an epidemic

Peer Review

cluster of Rhodotorula mucilaginosa in multiple geographic regions in China and the emergence of a 5-flucytosine resistant clone. Emerg Microbes Infect. 2022 Dec;11(1):1079–89.

22. Desnos-Ollivier M, Maufrais C, Pihet M, Aznar C, Dromer F, French Mycoses Study Group. Epidemiological investigation for grouped cases of Trichosporon asahii using whole genome and IGS1 sequencing. Mycoses. 2020 Sep;63(9):942–51.
23. Mitchell BI, Kling K, Bolon MK, Rathod SN, Malczynski M, Ruiz J, et al. Identifying Candida auris transmission in a hospital outbreak investigation using whole-genome sequencing and SNP phylogenetic analysis. J Clin Microbiol. 2024 Oct 16;62(10):e0068024.
24. Aggelen H van, Kolde R, Chamarthi H, Loving J, Fan Y, Fallon JT, et al. A core genome approach that enables prospective and dynamic monitoring of infectious outbreaks. Sci Rep. 2019 May 24;9(1):7808.
25. Menu E, Criscuolo A, Desnos-Ollivier M, Cassagne C, D'Incan E, Furst S, et al. Saprochaete clavata Outbreak Infecting Cancer Center through Dishwasher. Emerg Infect Dis. 2020 Sep;26(9):2031–8.
26. Grandaubert J, Bhattacharyya A, Stukenbrock EH. RNA-seq-Based Gene Annotation and Comparative Genomics of Four Fungal Grass Pathogens in the Genus Zymoseptoria Identify Novel Orphan Genes and Species-Specific Invasions of Transposable Elements. G3 (Bethesda). 2015 Apr 27;5(7):1323–33.

LEGENDS

Figure 1: IGS1 phylogenetic tree reveals the affiliation of strains to *T. austroamericanum*. Maximum likelihood phylogenetic tree of *Trichosporon* species based on IGS1 sequences. The tree includes strains isolated from six patients (marked in red) and environmental sample (L0453), mapped against the newly described *T. austroamericanum* and related species.

Peer Review

Reference sequences provided by the National Reference Centre for Invasive Mycoses & Antifungals (NRCMA) are also included. The clustering confirms that the patient and environmental strains belong to *T. austroamericanum*, forming a distinct clade. The tree also shows the relationships between other *Trichosporon* species, such as *T. inkin*, *T. caseorum,* and *T. ovoides*. Bootstrap values are indicated at the nodes, supporting the robustness of the clades.

Figure 2: Phylogenetic relationships and SNP distance matrix of *T. austroamericanum* isolates based on core-genome analysis. (A) Phylogenetic tree based on the core-genome of the strains analyzed. The tree shows the evolutionary relationship between strains (L0385, L0419, L0399, L0453, L0458, L0221) with a distance scale (Tree scale: 0.1). The distance between branches reflects the degree of genetic divergence, based on SNP variations in the core genome. (B) SNP distance matrix based on the core-genome. Each value represents the number of SNP positions that differ between genome pairs.

Figure 3: Mitochondrial genome maps of the seven *T. austroamericanum* strains assembled de novo. Circular representations and annotation of the six mitogenomes obtained by de novo assembly, for strains L0221, L0385, L0399, L0419, L0453 and L0458. The outer ring depicts annotated genes, color-coded according to their function: protein-coding genes (yellow), rRNA genes (green), tRNA genes (pink), and non-coding regions (red). The size of each genome is indicated in the center, ranging between 44,424 to 44,494 base pairs, obtained in one contig from the de novo assemblies. Differences in the arrangement and number of certain SNPs (to be completed in the figure) and genomic features are noted, highlighting potential mitochondrial genotypic diversity among the strains.

Peer Review

Figure 4: SNP map and phylogenetic analysis of mitogenomes from *T. austroamericanum* isolates. SNP map (top) and phylogenetic placement (bottom) of mitochondrial genomes from *T. austroamericanum* isolates. The SNP alignment highlights three key positions in the mitochondrial genome (22,240, 31,837, and 39,897 located in the cox3, apocytochrome b, and trnL (tRNA-Leu) genes). Clinical isolates are shown in light red, the environmental strain in blue, and CNRMA strains labeled "CNRMA" in dark red.

TABLES

Table 1. Characteristics of patients with a culture positive for *Trichosporon austroamericanum* in Marseille, France, July 2022–June 2024

	#1	#2	#3	#4	#5	#6
Strain number	L0221	L0385	L0399	L0419	L0445	L0458
Characteristic						
Age, y	67	63	55	83	74	64
Sex	M	F	M	F	F	M
Location						
Hospital	C	C	C	T	N	C
Hospitalization ward						
Kidney transplantion unit	Yes	Yes	Yes	No	No	Yes
Cardiac surgery	No	No	No	Yes	No	No
Digestive surgery	No	No	No	No	Yes	No
Clinical signs at the time of positive culture						
Fever	Yes	Yes	No	No	No	No
Asthenia	Yes	Yes	No	Yes	Yes	No

Peer Review

Dyspnea	No	Yes	No	No	Yes	No
Digestive symptoms	No	No	No	No	Yes	No
Scar disunion	Yes	Yes	Yes	NA	NA	Yes
Biology at the time of positive culture						
Creatinine (µmol/L)	271	128	103	177	80,7	216
CRP (mg/L)	86	73	43	22	NP	1,3
Positive culture results						
Date of first positive culture	2022 July, 6	2023 April, 3	2023 June, 29	2023 Dec., 6	2024 April, 28	2024 June, 24
Days after transplantation	71	78	135	NA	NA	115
No. Positive samples	4	11	1	5	3	1
Blood	0	11	0	5	0	0
Scar swab	3	0	0	0	0	0
Abcess aspiration	1	0	1	0	0	1
Respiratory sample	0	0	0	0	3	0
Microbiology						
CrAg	Positive	Positive	Positive	Negative	NP	Negative
Titer	160	5120	20	-	-	-
Aspergillus antigen	Negative	NP	NP	NP	Negative	NP
Index	0,04	-	-	-	0,13	-
Outcome						
Death within 90 d	No	No	No	Yes	Yes	No
Days after first positive culture	-	-	-	30	30	-
Treatment						
Posaconazol	No	Yes	No	No	No	No

Peer Review

Voriconazol	Yes	Yes	Yes	No	No	Yes
Amphotericin B	No	Yes	No	No	No	No
Surgical trimming	Yes	Yes	Yes	NA	No	Yes
IS reduction	Yes	Yes	No	NA	NA	Yes

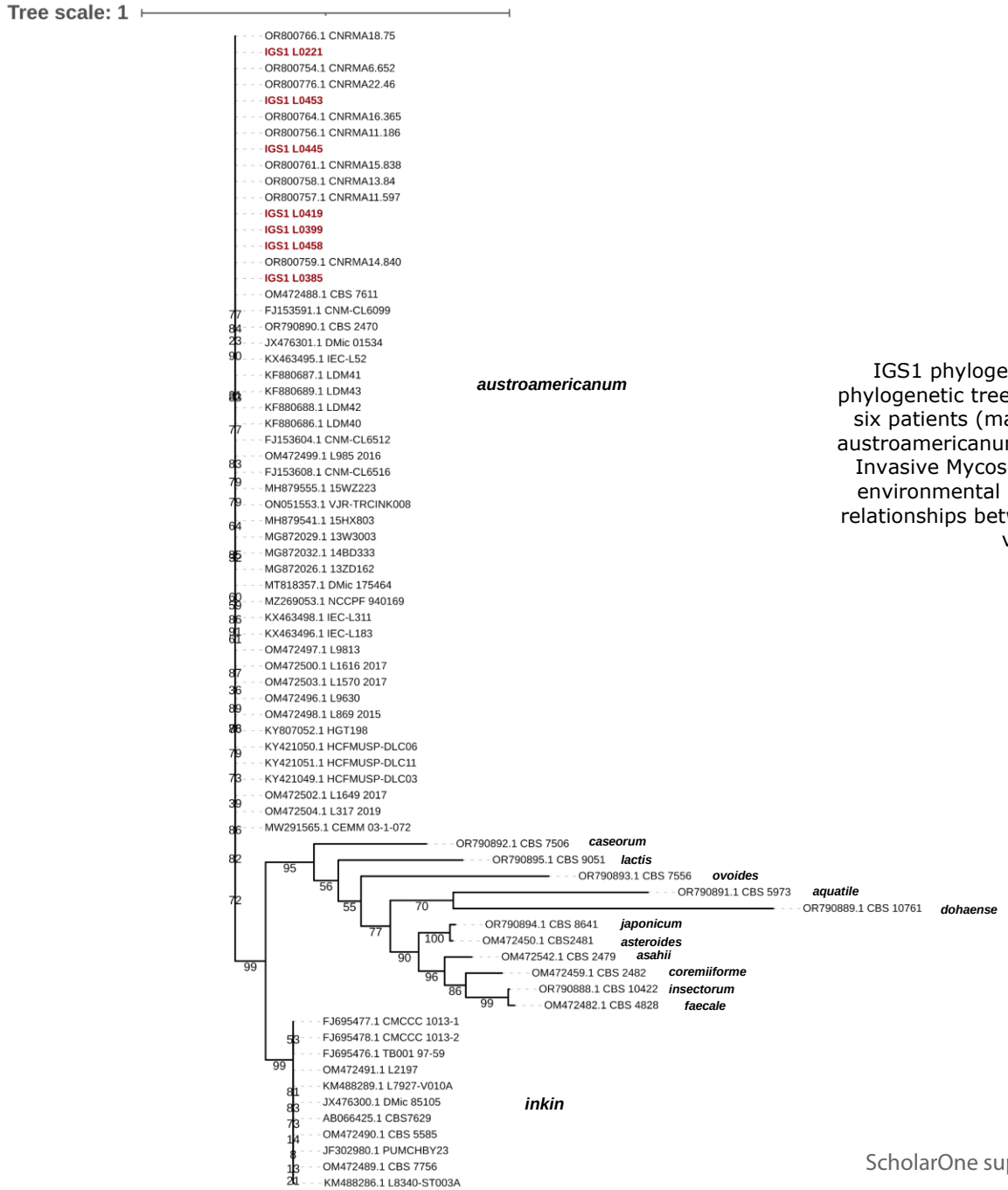
NA: Not applicable; NP: not performed; CRP: C Reactive protein, CrAg: Cryptococcal antigen; y: year; d: day; Dec.: December; IS: immunosuppression

Table 2. Genome assembly results for *Trichosporon austroamericanum* isolates.

Strain number	**L0221**	**L0385**	**L0399**	**L0419**	**L0445**	**L0453**	**L0458**
Contigs	14	12	16	13	1,051	31	24
Length	20,873,347	20,931,201	21,007,957	20,906,681	26,088,532	20,871,584	20,893,304
GC (%)	61.3	61.31	61.31	61.31	62.59	61.31	61.31
Contamination	No	No	No	No	Yes	No	No
BUSCO C-score (%)	96.50	96.90	97.40	96.60	96.10	96.90	96.90

Peer Review

Figure 1



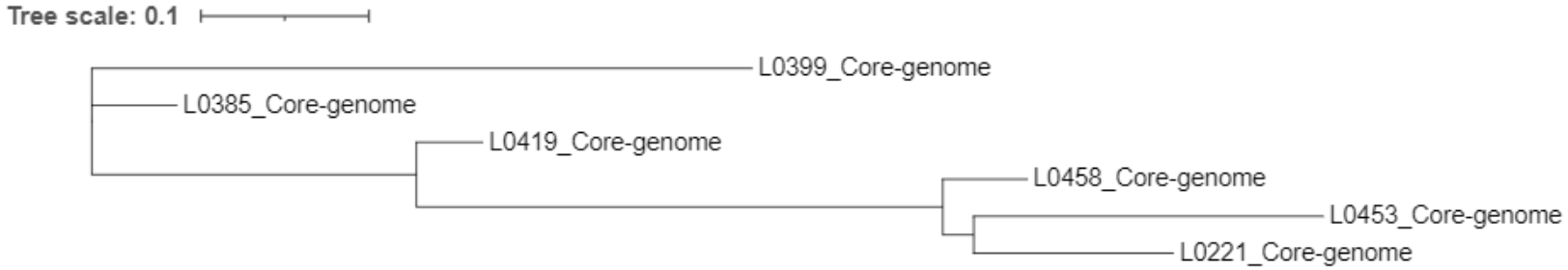
IGS1 phylogenetic tree reveals the affiliation of strains to T. austroamericanum. Maximum likelihood phylogenetic tree of Trichosporon species based on IGS1 sequences. The tree includes strains isolated from six patients (marked in red) and environmental sample (L0453), mapped against the newly described T. austroamericanum and related species. Reference sequences provided by the National Reference Centre for Invasive Mycoses & Antifungals (NRCMA) are also included. The clustering confirms that the patient and environmental strains belong to T. austroamericanum, forming a distinct clade. The tree also shows the relationships between other Trichosporon species, such as T. inkin, T. caseorum, and T. ovoides. Bootstrap values are indicated at the nodes, supporting the robustness of the clades.

338x190mm (300 x 300 DPI)

ScholarOne support: (434) 964-4100

Figure 2

a)



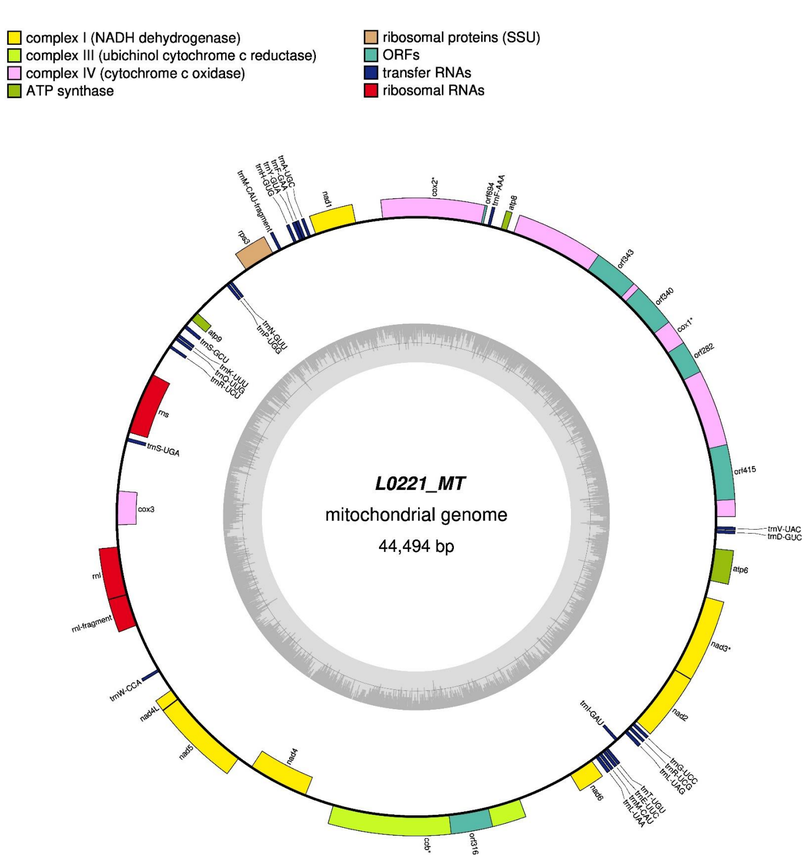
b)

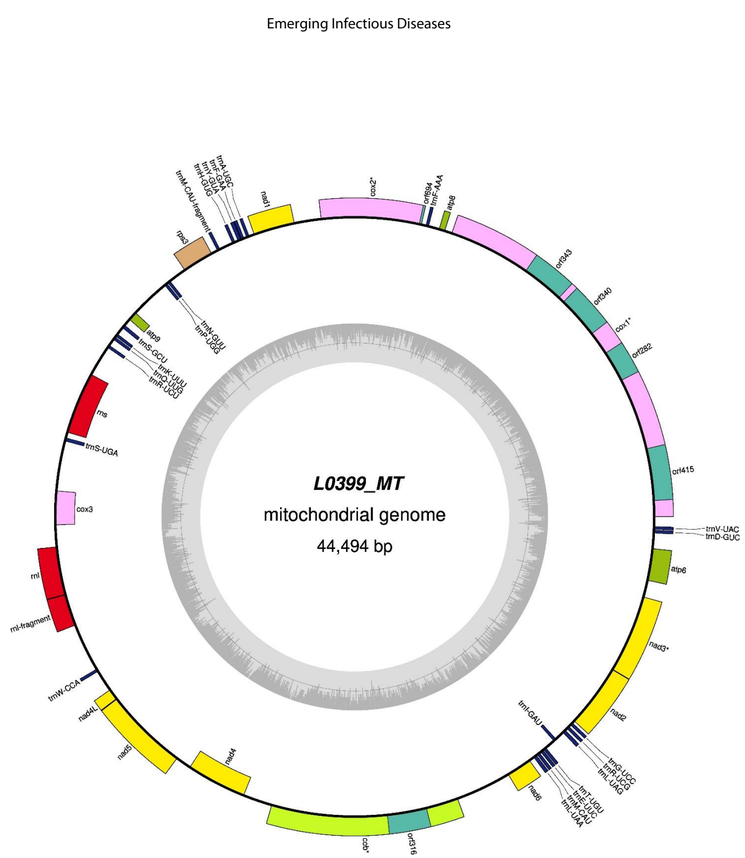
snp-dists 0.8.2	L0385	L0399	L0419	L0221	L0453	L0458
L0385	0	183	124	244	246	187
L0399	183	0	232	353	434	356
L0419	124	232	0	152	226	153
L0221	244	353	152	0	132	85
L0453	246	434	226	132	0	97
L0458	187	356	153	85	97	0

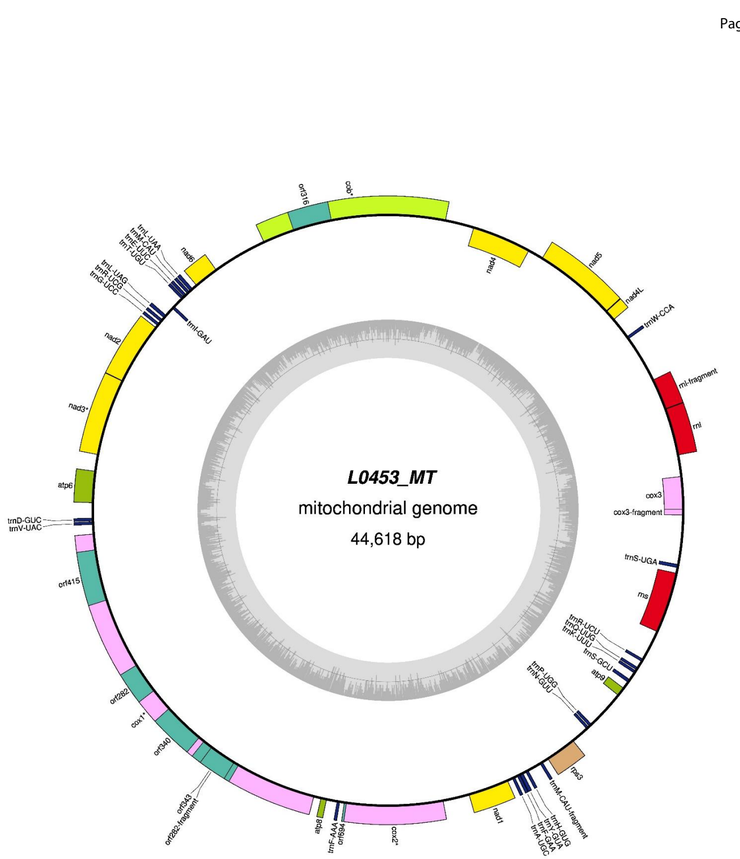
Phylogenetic relationships and SNP distance matrix of T. austroamericanum isolates based on core-genome analysis. (A) Phylogenetic tree based on the core-genome of the strains analyzed. The tree shows the evolutionary relationship between strains (L0385, L0419, L0399, L0453, L0458, L0221) with a distance scale (Tree scale: 0.1). The distance between branches reflects the degree of genetic divergence, based on SNP variations in the core genome. (B) SNP distance matrix based on the core-genome. Each value represents the number of SNP positions that differ between genome pairs.

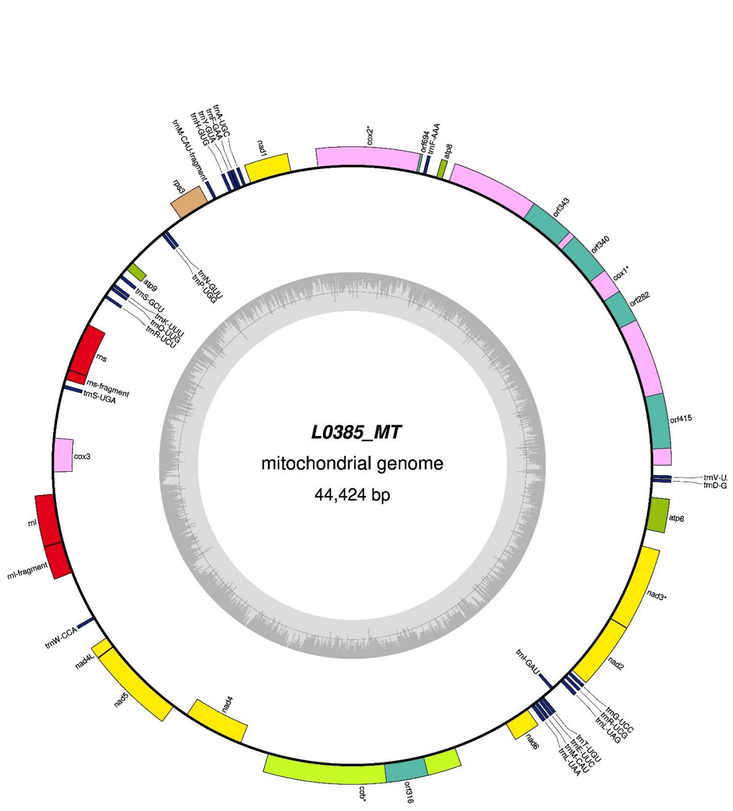
338x190mm (300 x 300 DPI)

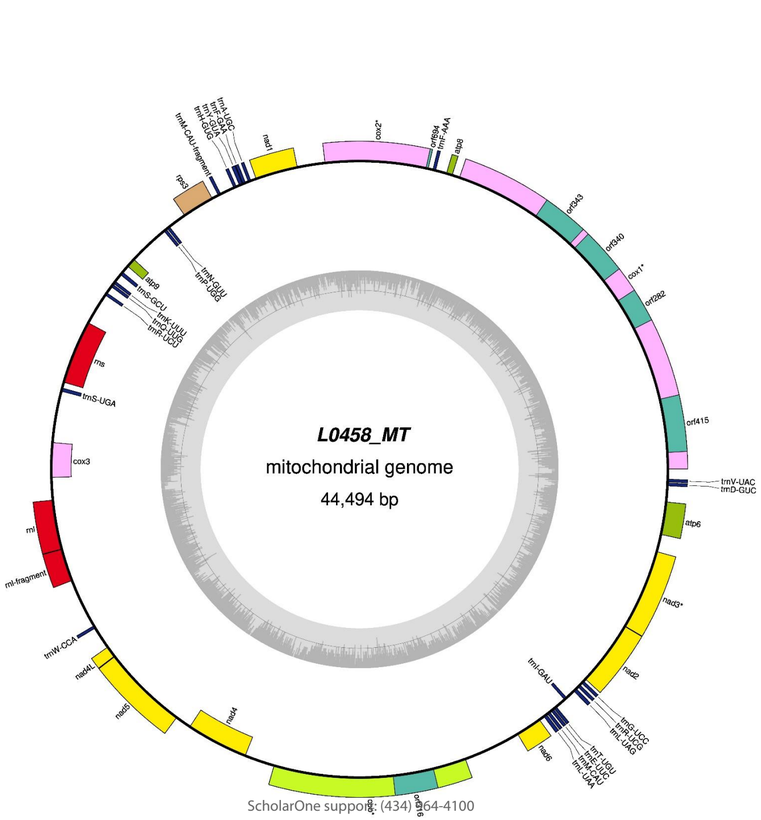
ScholarOne support: (434) 964-4100


complex I (NADH dehydrogenase)
complex III (ubichinol cytochrome c reductase)
complex IV (cytochrome c oxidase)
ATP synthase
ribosomal proteins (SSU)
ORFs
transfer RNAs
ribosomal RNAs
L0221_MT
mitochondrial genome
44,494 bp


L0399_MT
mitochondrial genome
44,494 bp


L0453_MT
mitochondrial genome
44,618 bp


L0385_MT
mitochondrial genome
44,424 bp


L0458_MT
mitochondrial genome
44,494 bp

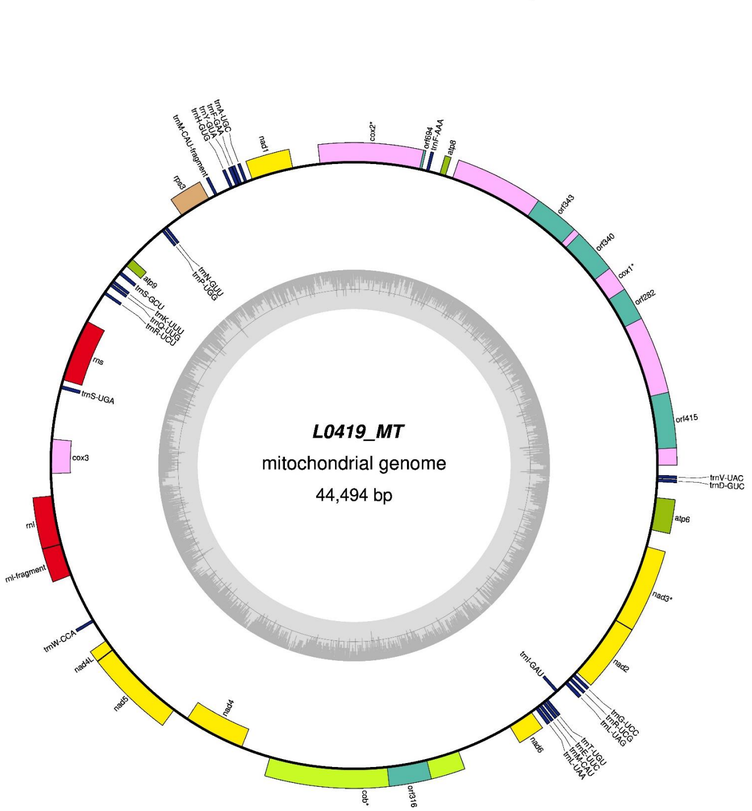
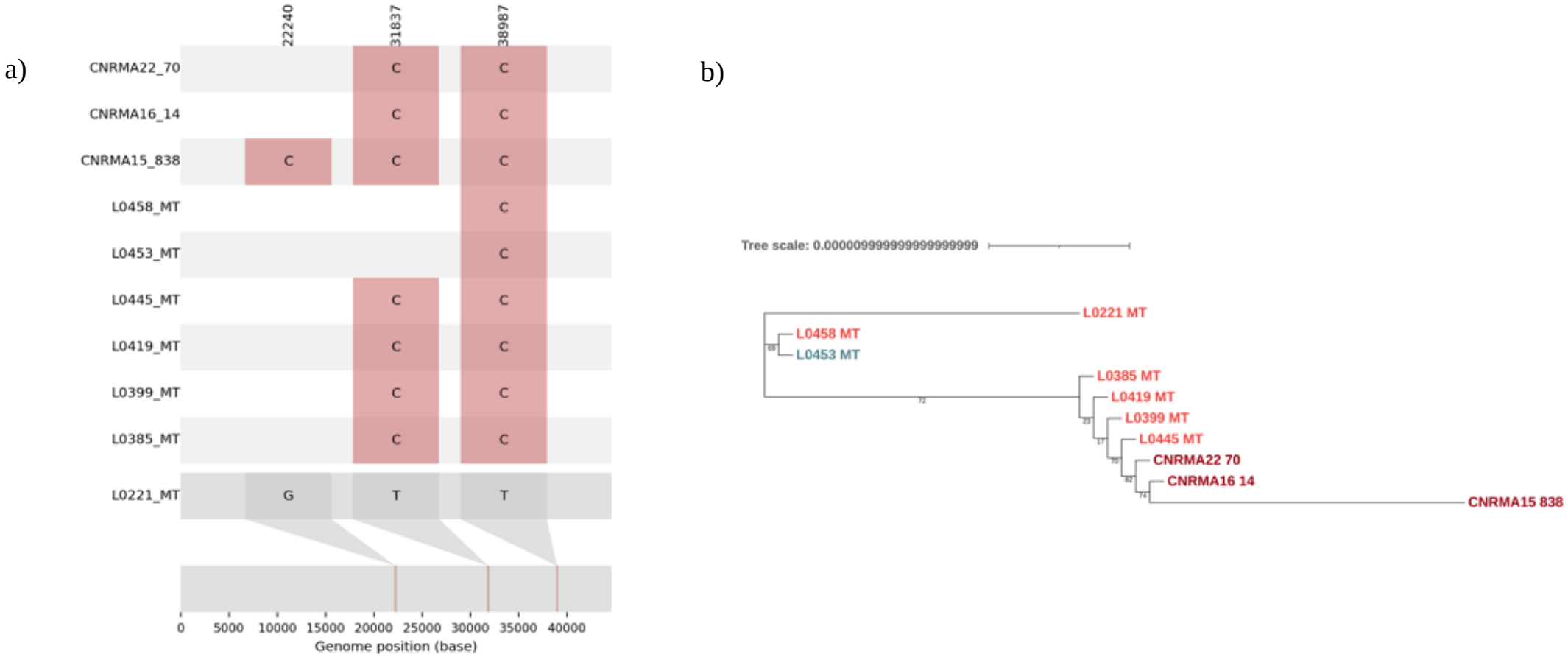

L0419_MT
mitochondrial genome
44,494 bp

Figure 4



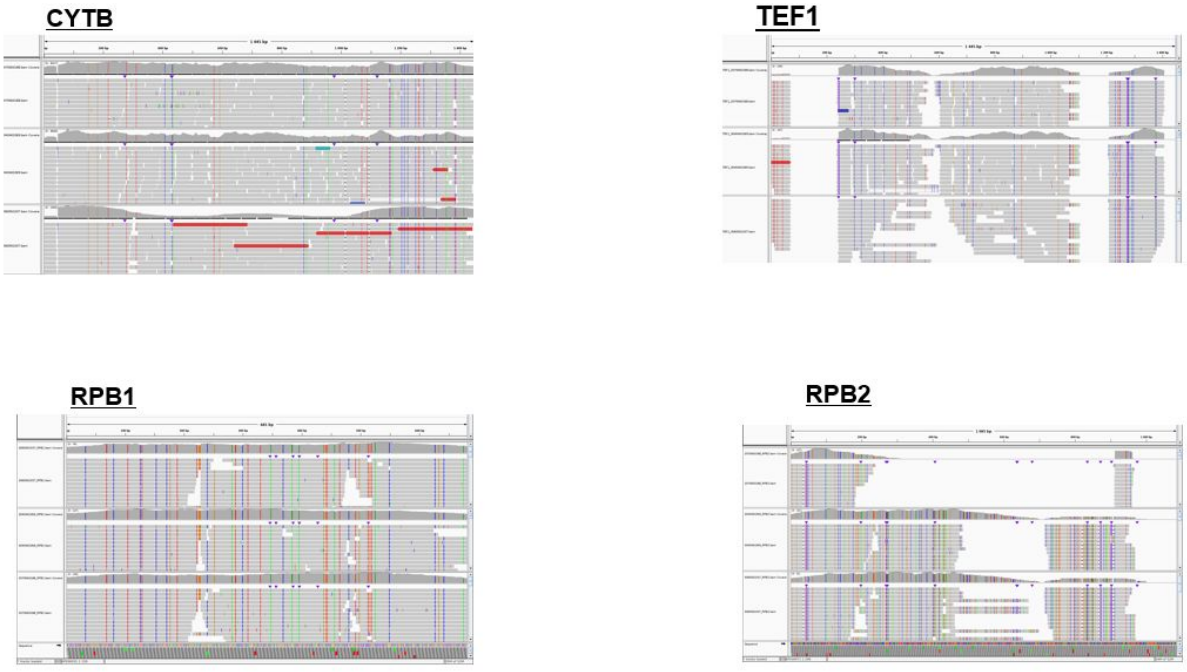
SNP map and phylogenetic analysis of mitogenomes from T. austroamericanum isolates. SNP map (top) and phylogenetic placement (bottom) of mitochondrial genomes from T. austroamericanum isolates. The SNP alignment highlights three key positions in the mitochondrial genome (22,240, 31,837, and 39,897 located in the cox3, apocytochrome b, and trnL (tRNA-Leu) genes). Clinical isolates are shown in light red, the environmental strain in blue, and CNRMA strains labeled "CNRMA" in dark red.

338x190mm (300 x 300 DPI)

ScholarOne support: (434) 964-4100

SUPPLEMENTAL

Supplementary Figure 1: Alignment of CYTB, TEF1, RPB1, and RPB2 genes for the first three cases against *Trichosporon inkin* reference (CBS5585). Multiple sequence alignments for four target genes: CYTB (cytochrome b), TEF1 (elongation factor 1-alpha), RPB1, and RPB2 (large subunits of RNA polymerase II). Each alignment compares sequences from the three first isolated strains, highlighting nucleotide variations. Differences between sequences are color-coded, indicating mutations, insertions, or deletions across the genetic sequences.



Supplementary Figure 2: Alignment of IGS1 Sequences of Trichosporon Strains L0221 and L0385 against *Trichosporon inkin* reference (CBS5585). Alignment of IGS1 sequences from *Trichosporon* strains L0221 and L0385 mapped against the *Trichosporon inkin* reference strain (CBS5585). The horizontal axis represents the length of the IGS1 region, with nucleotide variations displayed as colored bars. Differences between the sequences and the reference are indicated by colored lines, with mismatches, insertions in purple, and deletions highlighted with

ScholarOne support: (434) 964-4100

17 a dash.

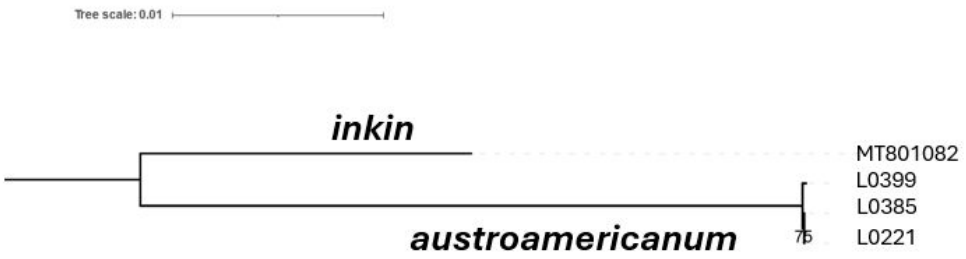
18



19

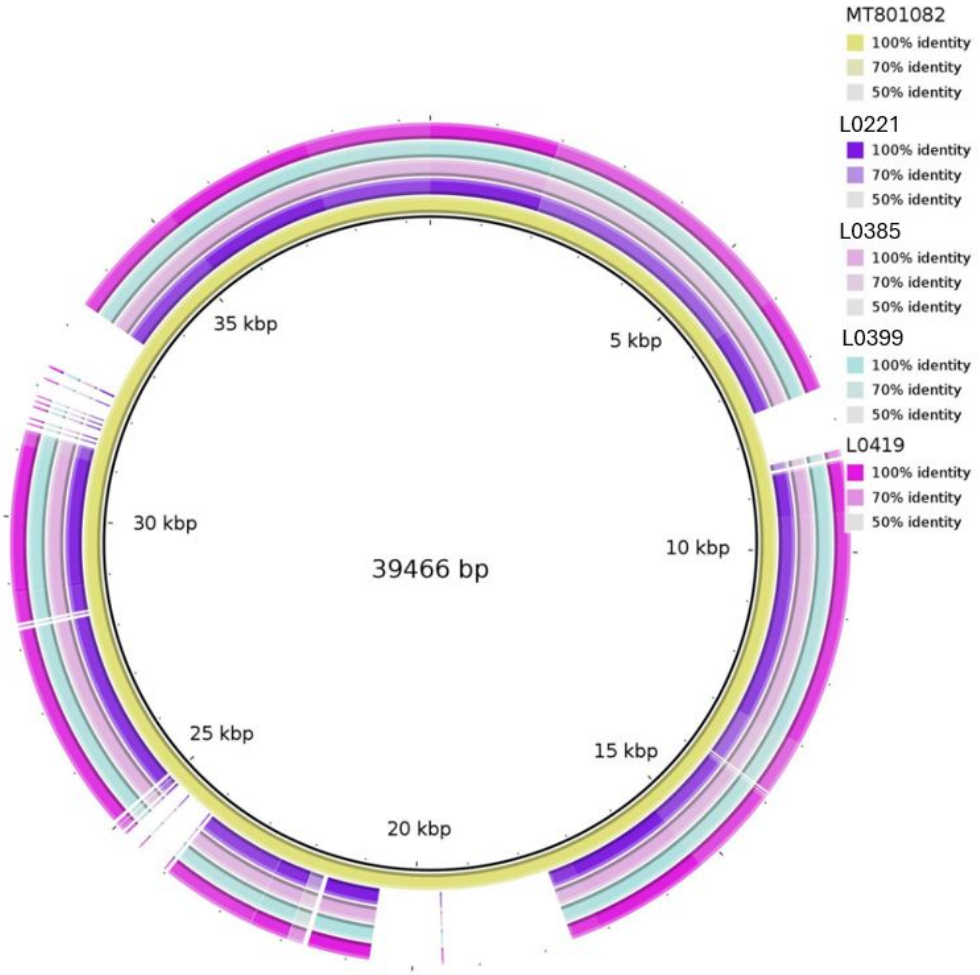
20 Supplementary Figure 3. Maximum likelihood phylogenetic tree based on the mitochondrial
21 genome, showing the relationship between *Trichosporon inkin* (MT801082) and *Trichosporon*
22 *austroamericanum* isolates (L0399, L0385, L0221). Schematic representation of the
23 mitogenome alignment of *Trichosporon austroamericanum* isolates against the mitogenome of
24 *Trichosporon inkin* (MT801082), obtained with BLAST Ring Image Generator.

25



26

27



28
29
30
31
32
33
34
35
36

Annexe 2 : Etude génomique sur la circulation de HCoV-229e dans le sud de la France

Avant-propos

Ce projet a consisté à caractériser les génotypes des coronavirus hCoV-229E, encore faiblement étudiés, ayant circulé dans le sud de la France (Marseille) sur une période donnée. Il a nécessité une approche de biologie moléculaire pour un design précis des amorces génomiques permettant de séquencer le génome entier (séquençage par amplicons), puis d'établir une stratégie analytique afin d'obtenir les génomes consensus puis analyser ces génomes consensus. Dans ce projet, ma contribution a porté sur l'aspect analytique, en procédant par mapping, en combinant les résultats issus des technologies ONT et Illumina, et enfin paramétrer les seuils optimaux d'obtention des génomes consensus.

bioRxiv preprint doi: https://doi.org/10.1101/2025.03.12.639729; this version posted March 14, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

TITLE PAGE

Implementation of a multiplex PCR amplfication system combined with next-generation genome sequencing to decipher the circulation of Human coronavirus 229E lineages in Southern France

Short title (for the running head): HCoV-229E phylogenomics in Southern France

Author list: Houmadi HIKMAT[1,2], Justine PY[1,2], Céline BOSCHI[1,2,3], Emilie BUREL[1], Lorlane LE TARGA[1,2,4], Matthieu MILLION[1,2,3], Lucile LESAGE[3,5], Aurélie MORAND[3,5,6], Bernard LA SCOLA[1,2,3*], Philippe COLSON[1,2,3*]

Affiliations: [1] IHU Méditerranée Infection, 19-21 boulevard Jean Moulin, 13005 Marseille, France; [2] Microbes Evolution Phylogeny and Infections (MEPHI), Aix-Marseille Univ. (AMU), 27 Bd Jean Moulin, 13005 Marseille, France; [3] Assistance Publique-Hôpitaux de Marseille (AP-HM), 264 rue Saint-Pierre 13005 Marseille, France; [4] Biosellal, 27 Chemin des Peupliers, 69570 Lyon, France; [5] Service de Pédiatrie générale, hôpital Timone, AP-HM, 264 rue Saint-Pierre, 13005 Marseille, France; [6] Service d'accueil des Urgences Pédiatriques, hôpital Nord, AP-HM, chemin des Bourrely, 13015 Marseille, France

*** Corresponding authors:** Philippe Colson and Bernard La Scola, IHU Méditerranée Infection, 19-21 boulevard Jean Moulin, 13005 Marseille, France. Tel.: +33 413 732 401, Fax: +33 413 732 402; email: bernard.la-scola@univ-amu.fr, philippe.colson@univ-amu.fr

Word counts: abstract, 249; main text, 2,395

Figures: 5; **Table:** 2; **References:** 38

Keywords: Human Coronavirus 229E, next-generation sequencing, genome, phylogenomics, evolution, Southern France

213

bioRxiv preprint doi: https://doi.org/10.1101/2025.03.12.639729; this version posted March 14, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

ABSTRACT

Coronaviruses are known to evolve rapidly and be prone to new virus emergence. Human coronavirus-229E is one of the seven coronaviruses currently known to cause respiratory symptoms in humans. Genomic data are very scarce for this virus.

Here, we implemented an in-house multiplex PCR strategy to amplify HCoV-229E genomes from nasopharyngeal samples diagnosed as positive for this virus RNA in Southeastern France. Then, next-generation sequencing was performed using Nanopore or Illumina technologies on Gridion or Novaseq 6000 instruments, respectively. HCoV-229E genomes were assembled and analyzed using MAFFT, MEGA, Itol, Nexstrain and Nextclade softwares.

Thirty-one PCR primer pairs were designed to amplify overlapping fragments of the HCoV-229E genome. They allowed obtaining 123 genomes, which were classified in an emerging HCoV-229E lineage first reported in China, with two sublineages being delineated. Relatively to genome NC_002645.1 dating back to 1962, regarding nucleotide mutations, 1,167 substitutions, 72 insertions and 34 deletions were detected in viral genomes obatined here, while regarding amino acid mutations, 415 susbstitutions, 39 deletions and 14 amino acid insertions were detected. The genes with the greatest diversity were S (that encodes the spike protein) then Nsp3. Signature mutations were identified for the two sublineages.

In summary, we almost doubled the set of HCoV-229E genomes available worldwide and provided the first genomes from France. Further studies are needed to strengthen the knowledge about the phylogenomics and evolutionary dynamics of this neglected respiratory virus, and about their specificities, which may also purvey clues to contribute improving knowledge for all other human coronaviruses.

214

bioRxiv preprint doi: https://doi.org/10.1101/2025.03.12.639729; this version posted March 14, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

TEXT

INTRODUCTION

Human coronavirus-229E (HCoV-229E) was the first discovered, in 1966, of the seven coronaviruses currently known to infect and cause respiratory diseases in humans[1,2]. Infections are most often associated with mild clinical symptoms but can be severe and life-threatening in children, elder people and in case of underlying illness[3]. They show a seasonality in temperate countries with the greatest incidence during winter and spring[4,5].

HCoV-229E is classified in genus *Alphacoronavirus*[6]. Its genome is a single-stranded positive sense RNA with an approximate size of 27 kilobases (kb). Its first two-thirds encode non-structural proteins (namely, NSP1 to NSP16) and the remaining third encodes structural proteins including the spike (S), envelope (E), membrane (M) and nucleocapsid (N) proteins. An accessory protein is encoded by a gene located between those encoding the spike and the envelope [7–9]. HCoV-229E is classified into six genotypes named 1 to 6, while an emerging lineage was reported in China in 2023[10,11]. Aminopeptidase-N is the primary cell surface receptor for this virus[12]. Natural and intermediate hosts for HCoV-229E are deemed to be bats and alpaca, respectively[1,13].

The HCoV-229E genome was primarily studied betweeen 1978 and 2001[14–16] and the first complete genome sequence from a clinical isolate was described in 2012[17]. Still, there is currently a huge discrepancy between the million SARS-CoV-2 genomes available in worldwide databases[18] (https://gisaid.org/; https://www.ncbi.nlm.nih.gov/genbank/) and the only 130 genomes (as of 01/01/2024) available in the NCBI Genbank nucleotide sequence database. Besides, none of these genomes originated from France. There is therefore a considerable paucity of data and, consequently of understanding of HCoV-229E genetic diversity and evolution. Hence, here we aimed to sequence and analyze HCoV-229E genomes

bioRxiv preprint doi: https://doi.org/10.1101/2025.03.12.639729; this version posted March 14, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

from respiratory samples that had been diagnosed as HCoV-229E RNA-positive in Marseille, Southeastern France.

MATERIALS AND METHODS

Respiratory samples

Next-generation sequencing (NGS) of HCoV-229E genomes was carried out retrospectively from remains of nasopharyngeal samples that had been sent to our clinical microbiology laboratory at University and Public Hospitals of Marseille, Southeastern France, for the purpose of routine diagnosis of respiratory infections in the setting of clinical routine management, and that were stored at -20°C or -80°C after processing. HCoV-229E RNA testing had been carried out by multiplex real-time reverse-transcription (RT)-PCR (qPCR), as previously described[19].

PCR primer design and PCR amplification of overlaping regions covering the whole genomes

All near complete or complete HCoV-229E genomes available from GenBank (https://www.ncbi.nlm.nih.gov/genbank/)[20] as of 28/02/2022 were retrieved. Recovered genomes were aligned using the MAFFT software[21]. PCR primers targeting the most conserved regions of the genomes were then designed using the Gemi software[22] to implement a PCR amplification primer set that enables generating overlaping amplicons covering the whole genome sequence, following the 'ARTIC' strategy used for instance for SARS-CoV-2 genomes (https://artic.network/ncov-2019/ncov2019-bioinformatics-sop.html). The list of PCR primers and the PCR conditions used for HCoV-229E genome amplification are provided in Table 1.

Next-generation sequencing

For the step of testing designed PCR primers and PCR conditions, NGS used the

bioRxiv preprint doi: https://doi.org/10.1101/2025.03.12.639729; this version posted March 14, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Oxford Nanopore technology (ONT), with the Ligation sequencing kit SQK-LSK109 then library deposit on a SpotON flow cell Mk I, R9.4.1 and a GridION instrument, according to the manufacturer's protocols (Oxford Nanopore Technologies Ltd., Oxford, UK). Thereafter, we performed NGS on RNA extracts obtained using the KingFisher Flex system (Thermo Fisher Scientific, Waltham, MA, USA) from available remains of nasopharyngeal samples that had been diagnosed as HCoV-OC43 RNA-positive in our laboratory. At this step, NGS was carried out using ONT, and Illumina technology on a NovaSeq 6000 instrument, with the CovidSeq protocol (Illumina Inc., San Diego, CA, USA) but with the replacement of the Covid-19 ARTIC PCR primers by the PCR primers we previously designed and according to the PCR conditions we previously set up. Loading procedure on a NovaSeq 6000 SP flow cell followed the NovaSeq-XP workflow and a procedure previously described[23] with a reading of 2x50 nucleotides.

Processing and bioinformatic analyses of NGS reads and viral genomes

Genome sequences were assembled by mapping on the HCoV-229E genome GenBank accession no. LC654445.1 (Fukushima_H829_2020 isolate) with Minimap2[24]. Samtools13 was used for soft clipping of Artic primers and removing sequence duplicates[25]. Consensus genomes were generated using Sam2consensus (https://github.com/edgardomortiz/sam2consensus). A phylogenetic tree was created with the MEGA[26] software (v.11) using the Neighbor-Joining and Maximum composite likelihood parameter methods with 1,000 replicates after sequence alignment with the MAFFT software[21]. All HCoV-229E genomes available from GenBank including those corresponding to genogroups were incorporated in the phylogeny reconstruction. The Itol tool [27] was used to visualize the phylogenetic tree. Nextstrain [28] and Nextclade[29] tools were adapted to enable identifying viral lineages and mutations. Nucleotide and amino acid diversity was obtained relatively to the HCoV-229E reference genome no. NC_002645.1.

bioRxiv preprint doi: https://doi.org/10.1101/2025.03.12.639729; this version posted March 14, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

RESULTS

A total of 524 nasopharyngeal samples had been diagnosed as HCoV-229E RNA-positive in our institution between January 2017 and October 2022, but only remains for 195 of them, which had been collected between December 2022 and March 2022, were available as stored frozen and in sufficient volume (Figure 1). These 195 specimens were tested using a multiplex PCR amplification strategy to amplify the genome per short overlaping fragments for further NGS. After designing and testing individually possible PCR primer pairs, 31 primer pairs generating amplicons covering the entire HCoV-229E genome were selected (Table 1); primer concentration in PCR ranged from 10 to 15.3 µM. These PCR primer pairs were used in two separate pools to prevent hybridization of generated amplicons. They allowed obtaining 123 genomes with a completion corresponding to ≥80% coverage of reference genome NC_002645.1; mean (±standard deviation) coverage was 92.14%±0.05 (range, 80.0-98.0%). qPCR cycle threshold values (Ct) were available for 75 of the samples from which these 123 genomes were obtained. Mean Ct was 21.2±4.0 (15.0-29.0). Regarding the 72 other specimens for which HCoV-229E genomes could not be obtained with previously defined criteria, their mean Ct was 26.3±2.7 (22.0-30.0).

The 123 near-full genomes were obtained from nasopharyngeal samples collected between March 2021 and March 2022. They were classified by the Nextclade tool and the phylogeny reconstruction as belonging to an emerging lineage reported in China in 2023[11], which was designated as a putative genotype 7 in two recent reports[30,31](Figure 2). Based on the phylogenetic analysis, this emerging lineage comprised two sublineages, one of which appeared to match with previously designated sublineages 7b[30,31] whereas there seems to be discrepancies between matches for the second sublineages and previously designated submineage 7a[30,31]. Whatever, 30 of the 123 genomes obtained here belong to a sublineage

bioRxiv preprint doi: https://doi.org/10.1101/2025.03.12.639729; this version posted March 14, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

150 "a" and 93 belong to a sublineage "b" (Figure 2), revealing that these two sublineages co-
151 circulated in our geographical area, with a predominance of viruses of sublineage b (Figure
152 3). Sublineage a was detected since March 2021 while sublineage b was detected since
153 September 2021. For the four months during which the number of genomes obtained from
154 collected specimens were above 10, the proportion of genomes of sublineage a decreased
155 from 32% in November 2021 to 7% in February 2022.

156 The nucleotide and amino acid diversity of the 123 HCoV-229E genomes was
157 determined relatively to genome NC_002645.1 first submitted to Genbank in 2001. Regarding
158 nucleotide mutations, 1,167 substitutions, 72 insertions and 34 deletions were detected in
159 these genomes. Regarding amino acid mutations, 415 substitutions, 39 deletions and 14 amino
160 acid insertions were detected. Of the 415 amino acid substitutions, only 211 were present in
161 ≥5 genomes. These 211 amino acid substitutions were found in several genes, including in the
162 Nsp1, Nsp2, Nsp3, Nsp4, Nsp6, Nsp8, Nsp9, Nsp10, Nsp11, Nsp12, Nsp13, Nsp14, Nsp15,
163 spike, ORF4a, E, M, and N genes. A total of 78 (37%) of these 211 amino acid substitutions
164 were in the spike. They displayed various prevalence (Figure 4, Table 2). Seventy-one were
165 already reported, being mentioned as new substitutions in 13 cases[11]. Spike deletions S:V353-
166 and S:Y354- were observed in all genomes obtained here, while S:A352- was observed in all
167 but four genomes (97%) that harbor four other deletions: S:A355-, S:N356-, S:V357- and
168 S:G358- (Table 2). Two other deletions were present in NSP3 (NSP3:L105-, NSP3:P106-) of
169 all genomes (except one for which these codons were not covered).

170 The genes with the greatest diversity, regardless of the mutation type, were S, which
171 encodes the spike protein, with 67 amino acid mutations per 1,000 amino acids, then Nsp2, N,
172 Nsp3, which encodes a large multi-domain non-structural protein and is an essential
173 component of the viral replication/transcription complex, and E, with between 26 and 28
174 amino acid mutations per 1,000 amino acids. The genes with the lowest diversity were NSP7,

219

bioRxiv preprint doi: https://doi.org/10.1101/2025.03.12.639729; this version posted March 14, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

175 which is part of the RNA-dependent RNA polymerase complex, and NSP16, which encodes a
176 2′-O-methyltransferase. The two sublineages a and b each harbored specific amino acid
177 mutations. Positions 349, 358, 391, 406 and 971 of the spike protein were found to harbor two
178 different amino acid substitutions according to the sublineage. Thus, genomes of sublineage a
179 shared the following mutations: Nsp2:H40R; S:K349Q; S:D391A; S:Y406N; and N:E33D
180 (Figures 4,5). Regarding genomes of sublineage b, they shared the 22 following mutations,
181 which were not carried by genomes of sublineage a: Nsp2:A350V; Nsp3:V423M;
182 Nsp3:K428Q; S:T19I; S:F229V; S:S248A; S:V288E; S:Y305H; S:K309E; S:P310H;
183 S:Q311G; S:S312R; S:F318Y; S:Y321R; S:G324V; S:V325I; S:K349R; S:N365D; S:N377I;
184 S:D391Y; S:Y406G; and S:V739A (Figures 4, 5). Among these mutations, 19 were present in
185 the S gene. Four amino acid mutations located in the S gene that we detected in sublineages a
186 or b genomes were previously reported as newly detected[11].
187
188 **DISCUSSION**
189 We report here a multiplex PCR system to amplify whole HCoV-229E genomes
190 before their sequencing by next-generation technologies. The present study allowed obtaining
191 and analyzing 123 new near full-length HCoV-229E genomes, which was a number almost
192 equal to the global number of HCoV-229E genomes at the beginning of Year 2024. In
193 addition, we provided here the first HCoV-229E genomes from patients sampled in France.
194 Another similar multiplex PCR systems for HCoV-229E whole genome amplification were
195 reported in 2024 by Musaeva et al[31] and McClure et al[30], in Russia and UK, respectively. The
196 first one[31] used twenty-nine PCR primer systems, compared to 31 in the present study, and
197 enabled obtaining 39 genomes with a coverage of a full-length genome >70%, from 50
198 HCoV-229E RNA-positive nasopharyngeal swab samples. The second one[30] used 29–36
199 primer pairs for all four endemic human coronaviruses (HCoV-229E, -OC43, -NL63, and -

220

bioRxiv preprint doi: https://doi.org/10.1101/2025.03.12.639729; this version posted March 14, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

229E), and enabled obtaining 64 HCoV-229E genomes with a coverage of a full-length genome >95%. Overall, as of 01/02/2025, 269 HCoV-229E genomes were available in GenBank, and the 123 genomes obtained here grew the global set to 392.

Most of the HCoV-229E genomes obtained here were retrieved from nasopharyngeal samples collected between 2019 and 2022. This can be explained by the Covid-19 pandemic that led to a massive diagnosis of respiratory virus infections during this period of time. Between July 2020 and June 2021, few HCoV-229E infections were diagnosed in our center, as previously reported, which could be explained at least partly by the implementation of hygiene measures to fight the SARS-CoV-2 pandemic[32]. All the genomes obtained here belong to the emerging lineage previously reported by Ye et al[11], with 43 full-length genomes, who demonstrated that this lineage circulated between 2016 and 2020 in different countries and suggested that it was the majority lineage circulating in the world. We report here that this lineage, which was tentatively reported as genotype 7 in Musaeva et al's and McClure et al's study[30,31], also circulated in France. Two sublineages 7a and 7b had been reported for this genotype[30,31], while we also observed in the present study two sublineages "a" and "b" with an overall predominance of sublineage b. Hence, our findings further support that the emerging lineage initially reported in China, then in Japan, Haiti, the United States, Russia, and UK likely became the predominant lineage worldwide. In addition, our findings together with previous data'[11,30,31] indicate that this emerging lineage are evolving with new mutations whose occurrence may depend on the time and geographical area of HCoV-229E circulation. Sublineages in this emerging lineage should be more precisely delineated in further studies with additional HCoV-229E genomes.

A high amino acid diversity was observed here in the spike and non-structural proteins, which is consistent with previous findings in coronaviruses[11,30,31,33–35]. The HCoV-229E receptor binding domain (RBD) contains three loops, named 1, 2 and 3, which are

bioRxiv preprint doi: https://doi.org/10.1101/2025.03.12.639729; this version posted March 14, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

involved in the binding of the virus to the host aminopeptidase N cellular receptor and are located at amino acid positions 308-325, 352-359 and 404-408, respectively[34]. Some amino acid mutations in these regions were observed here as in three previous studies[11,30,31,35]. These notably involved four amino acid positions. For three of them, two different mutations were observed here (K349R or Q; G358P or A; Y406G or N). For position 391, two types of substitutions were observed here that are signatures of either sublineage a (D391A) or sublineage b (D391Y). Besides, a tryptophan at position 404 of the spike RBD was reported to be very important for loop 3 binding to the cellular receptor[34]. Notwithstanding, tryptophan was replaced here by a leucine in all genomes, indicating that this latter does not preclude viral infection. Mutations Q430K, D444N and K488N also encountered in the present study were already reported and would result in an N-glycosylation site at position 488[35]. Besides, apolar bonds were predicted that involve certain amino acids in the RBD, notably at position 318[34]. A mutation at this position was reported to be associated with an ≈13-fold reduction in the affinity with the cellular receptor[34]. Here, a mutation (F318Y) at this position was found in 76% of the genomes. This same mutation was also reported previously in HCoV-229E genotypes 3, 4, 5, 6 and in the emerging lineage as well[11]. Taken together, these data highlight the broad diversity of spike amino acid patterns and may be useful for the interpretation of structural analyses performed previously and in future studies including to investigate the putative impact of these different mutations on the interaction between HCoV-229E and the host receptor[36–38].

Overall, the present study and two other recent studies[30,31]enrich the set of HCoV-229E genomes available worldwide, with 123 and 103 genomes respectively. Nonetheless, genomic data remain scarce and they cover a limited number of countries with still limited data sets, therfore not necessarily reflecting the circulation of this virus at the global scale. Further studies are therefore needed to gain a more global view of the evolutionary dynamics

bioRxiv preprint doi: https://doi.org/10.1101/2025.03.12.639729; this version posted March 14, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

of HCoV-229E, which will clarify the specificities of this virus and may contribute to a more general understanding of the evolution of human coronaviruses.

Acknowledgments

We are very grateful to the technical team of the next-generation sequencing platform of IHU Méditerranée Infection.

Data Availability

HCoV-229E genomes analysed here have been submitted to the GenBank sequence database[20] (https://www.ncbi.nlm.nih.gov/genbank/) (GenBank Submission ID 2929308).

Author contributions

Conceived and designed the experiments: BLS, PC. Contributed materials, analysis tools: All authors. Analyzed the data: HH, JP, EB, PC. Writing—original draft preparation: HH, JP, PC. Writing—review and editing: All authors. All authors have read and agreed to the published version of the manuscript.

Conflicts of interest

Lorlane Le Targa works for Biosellal, a company located in Dardilly, France. The other authors have no conflicts of interest to declare. Funding sources had no role in the design and conduct of the study; collection, management, analysis, and interpretation of the data; and preparation, review, or approval of the manuscript.

Funding

This work was supported by the French Government under the “Investments for the Future” program managed by the National Agency for Research (ANR), Méditerranée-Infection 10-IAHU-03; and by the French Ministry of Higher Education, Research and Innovation (Ministère de l’Enseignement supérieur, de la Recherche et de l’Innovation) and the French

bioRxiv preprint doi: https://doi.org/10.1101/2025.03.12.639729; this version posted March 14, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Ministry of Solidarity and Health (Ministère des Solidarités et de la Santé).

Ethics

The present study has been registered on the Health Data Access Portal of Marseille public and university hospitals (Assistance Publique-Hôpitaux de Marseille (AP-HM)) with No. PADS24-190 and was approved by the Ethics and Scientific Committee of AP-HM.

bioRxiv preprint doi: https://doi.org/10.1101/2025.03.12.639729; this version posted March 14, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

REFERENCES

1. Tang G, Liu Z, Chen D. Human coronaviruses: Origin, host and receptor. J Clin Virol 2022;155:105246.

2. Hamre D, Procknow JJ. A new virus isolated from the human respiratory tract. Proc Soc Exp Biol Med 1966;121:190-3.

3. Kistler KE, Bedford T. Evidence for adaptive evolution in the receptor-binding domain of seasonal coronaviruses OC43 and 229e. ELife 2021;10:e64509.

4. Killerby ME, Biggs HM, Haynes A, et al. Human coronavirus circulation in the United States 2014–2017. J Clin Virol 2018;101:52-56.

5. Moriyama M, Hugentobler WJ, Iwasaki A. Seasonality of Respiratory Viral Infections. Annu Rev Virol 2020;7:83-101.

6. Corman VM, Muth D, Niemeyer D, Drosten C. Hosts and Sources of Endemic Human Coronaviruses. Adv Virus Res 2018;100:163-188.

7. Uhteg K, Carroll KC, Mostafa HH. Coronavirus Detection in the Clinical Microbiology Laboratory: Are We Ready for Identifying and Diagnosing a Novel Virus? Clin Lab Med 2020;40:459-472.

8. Ma M, Yang Y, Wu L, et al. Conserved protein targets for developing pan-coronavirus drugs based on sequence and 3D structure similarity analyses. Comput Biol Med 2022;145:105455.

9. Zhang R, Wang K, Lv W, et al. The ORF4a protein of human coronavirus 229E functions as a viroporin that regulates viral production. Biochim Biophys Acta Biomembr 2014;1838:1088-1095.

bioRxiv preprint doi: https://doi.org/10.1101/2025.03.12.639729; this version posted March 14, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

10. Lau SKP, Lung DC, Wong EYM, et al. Molecular Evolution of Human Coronavirus 229E in Hong Kong and a Fatal COVID-19 Case Involving Coinfection with a Novel Human Coronavirus 229E Genogroup. mSphere 2021;6:e00819-20.
11. Ye R-Z, Gong C, Cui X-M, et al. Continuous evolution and emerging lineage of seasonal human coronaviruses: A multicenter surveillance study. J Med Virol 2023;95:e28861.
12. Yeager CL, Ashmun RA, Williams RK, et al. Human aminopeptidase N is a receptor for human coronavirus 229E. Nature 1992;357:420-422.
13. Corman VM, Baldwin HJ, Tateno AF, et al. Evidence for an Ancestral Association of Human Coronavirus 229E with Bats. J Virol 2015;89:11858-11870.
14. Macnaughton MR, Madge MH. The genome of human coronavirus strain 229E. J Gen Virol 1978;39:497-504.
15. Herold J, Raabe T, Schelle-Prinz B, Siddell SG. Nucleotide sequence of the human coronavirus 229E RNA polymerase locus. Virology 1993;195:680-691.
16. Thiel V, Herold J, Schelle B, Siddell SG. Infectious RNA transcribed in vitro from a cDNA copy of the human coronavirus genome cloned in vaccinia virus. J Gen Virol 2001;82:1273-1281.
17. Farsani SMJ, Dijkman R, Jebbink MF, et al. The first complete genome sequences of clinical isolates of human coronavirus 229E. Virus Genes 2012;45:433-439.
18. Zhu N, Zhang D, Wang W, et al. A Novel Coronavirus from Patients with Pneumonia in China, 2019. N Engl J Med 2020;382:727-733.
19. Mantelli C, Colson P, Lesage L, et al. Coinfections and iterative detection of respiratory viruses among 17,689 patients between March 2021 and December 2022 in Southern France. J Clin Virol 2024;175:105744.

226

bioRxiv preprint doi: https://doi.org/10.1101/2025.03.12.639729; this version posted March 14, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

20. Sayers EW, Cavanaugh M, Clark K, et al. GenBank 2023 update. Nucleic Acids Res 2023;51:D141-D144.

21. Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol Biol Evol 2013;30:772-780.

22. Sobhy H, Colson P. Gemi: PCR Primers Prediction from Multiple Alignments. Comp Funct Genomics 2012;2012:783138.

23. Papa Mze N, Kacel I, Beye M, et al. High Throughput SARS-CoV-2 Genome Sequencing from 384 Respiratory Samples Using the Illumina COVIDSeq Protocol. Genes 2023;14:681.

24. Li H. Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics 2018;34:3094-3100.

25. Li H, Handsaker B, Wysoker A, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics 2009;25:2078-2079.

26. Kumar S, Tamura K, Nei M. MEGA: Molecular Evolutionary Genetics Analysis software for microcomputers. Bioinformatics 1994;10:189-191.

27. Letunic I, Bork P. Interactive Tree Of Life (iTOL): an online tool for phylogenetic tree display and annotation. Bioinformatics 2007;23:127-128.

28. Hadfield J, Megill C, Bell SM, et al. Nextstrain: real-time tracking of pathogen evolution. Bioinformatics 2018;34:4121-4123.

29. Aksamentov I, Roemer C, Hodcroft EB, Neher RA. Nextclade: clade assignment, mutation calling and quality control for viral genomes. J. Open Source Softw. 6:3773.

30. McClure CP, Tsoleridis T, Holmes N, et al. ‘Vivaldi’: An amplicon-based whole genome sequencing method for the four seasonal human coronaviruses 229E, NL63, OC43 and HKU1, alongside SARS-CoV-2. medRxiv 2024;2024.06.26.24308604.

227

bioRxiv preprint doi: https://doi.org/10.1101/2025.03.12.639729; this version posted March 14, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

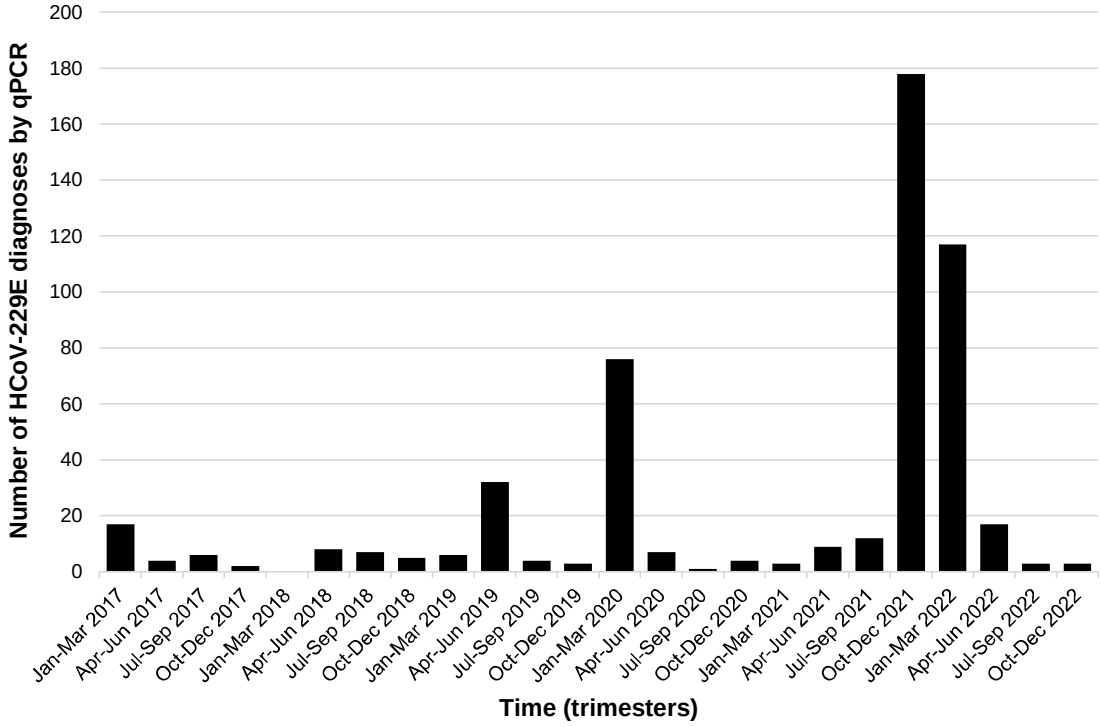
354 31. Musaeva T, Fadeev A, Pisareva M, et al. Development of Primer Panels for Whole-
355 Genome Amplification and Sequencing of Human Seasonal Coronaviruses: hCoV-
356 OC43, hCoV-HKU1, hCoV-229E, and hCoV-NL63. Viruses 2024;17:13.
357 32. Giraud-Gatineau, Kaba L, Boschi C, et al. Control of common viral epidemics but not
358 of SARS-CoV-2 through the application of hygiene and distancing measures. J Clin
359 Virol 2022;150-151.
360 33. Lei J, Kusov Y, Hilgenfeld R. Nsp3 of coronaviruses: Structures and functions of a
361 large multi-domain protein. Antiviral Res 2018;149:58-74.
362 34. Wong AHM, Tomlinson ACA, Zhou D, et al. Receptor-binding loops in
363 alphacoronavirus adaptation and evolution. Nat Commun 2017;8:1735.
364 35. Chibo D, Birch C. Analysis of human coronavirus 229E spike and nucleoprotein genes
365 demonstrates genetic drift between chronologically distinct strains. J Gen Virol
366 2006;87:1203-1208.
367 36. Li Z, Tomlinson AC, Wong AH, et al. The human coronavirus HCoV-229E S-protein
368 structure and receptor binding. eLife 2019;8:e51230.
369 37. Song X, Shi Y, Ding W, et al. Cryo-EM analysis of the HCoV-229E spike glycoprotein
370 reveals dynamic prefusion conformational changes. Nat Commun 2021;12:141.
371 38. Shi J, Shi Y, Xiu R, et al. Identification of a Novel Neutralizing Epitope on the N-
372 Terminal Domain of the Human Coronavirus 229E Spike Protein. J Virol
373 2022;96:e01955-21.
374

228

bioRxiv preprint doi: https://doi.org/10.1101/2025.03.12.639729; this version posted March 14, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

FIGURES

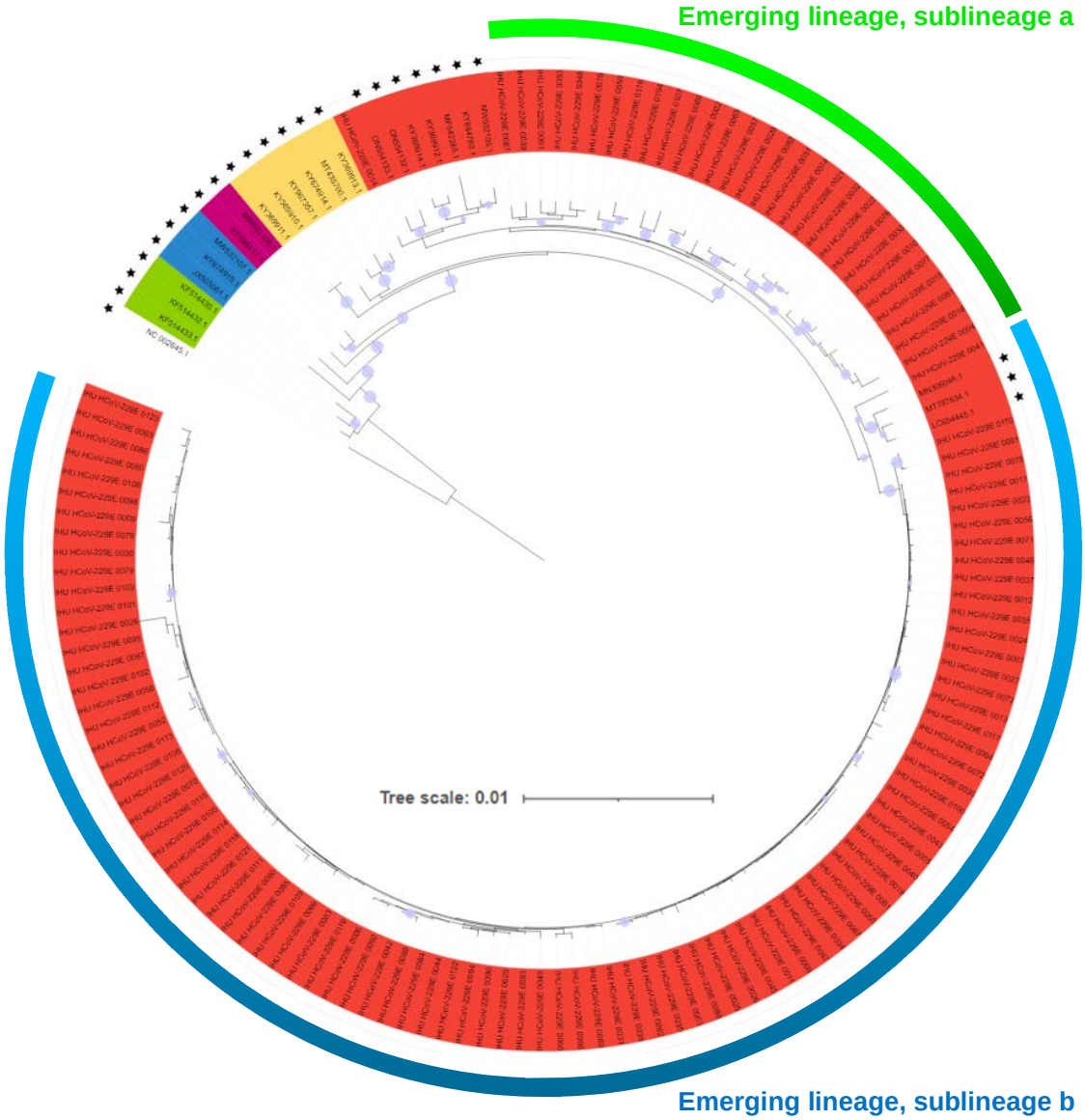
Figure 1. Temporal distribution of HCoV-229E RNA-positive nasopharyngeal samples that had been collected from patients between 2017 and 2022



229

bioRxiv preprint doi: https://doi.org/10.1101/2025.03.12.639729; this version posted March 14, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

383 **Figure 2.** Phylogeny reconstruction based on HCoV-229E genomes recovered in the present
384 study or available from GenBank, including genomes from the different previously delineated
385 lineages or sublineages
386

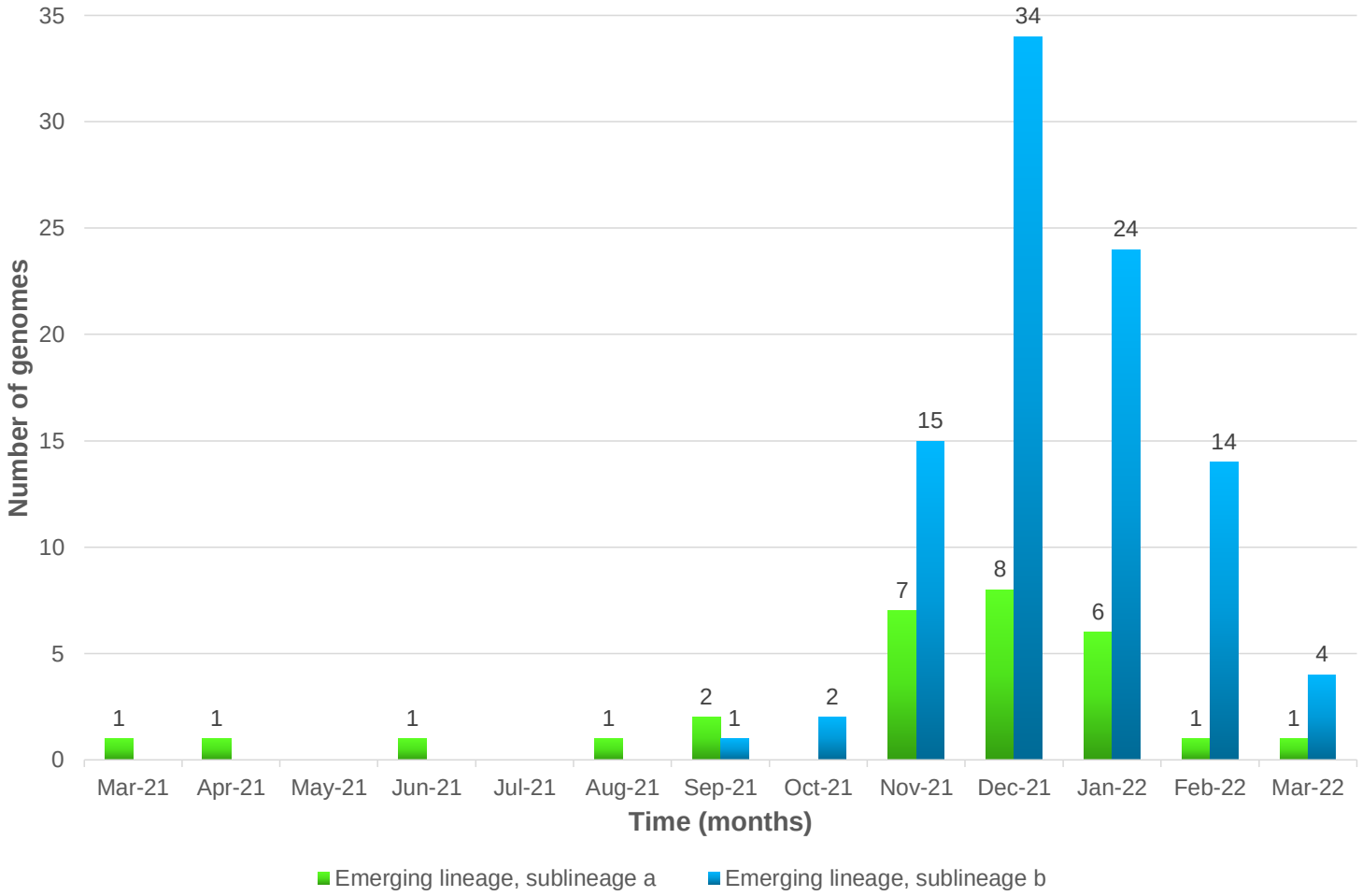


387
388

389 Genomes from GenBank are indicated by a black star.

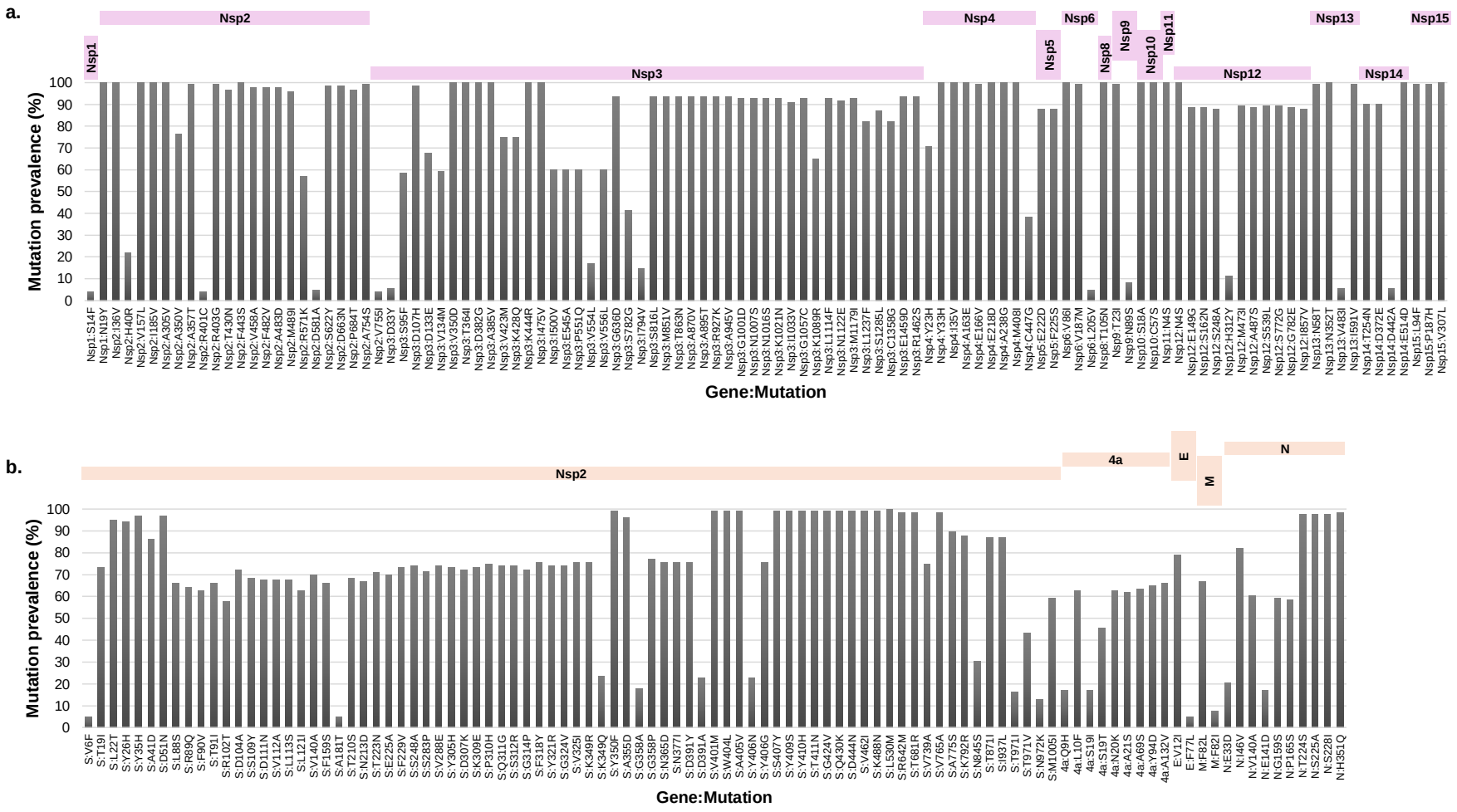
230

Figure 3. Temporal distribution of the two subgenotypes of HCoV-229E detected in the present study



bioRxiv preprint doi: https://doi.org/10.1101/2025.03.12.639729; this version posted March 14, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Figure 4. Amino acid substitution prevalence in non-structural (a) and structural (b) genes



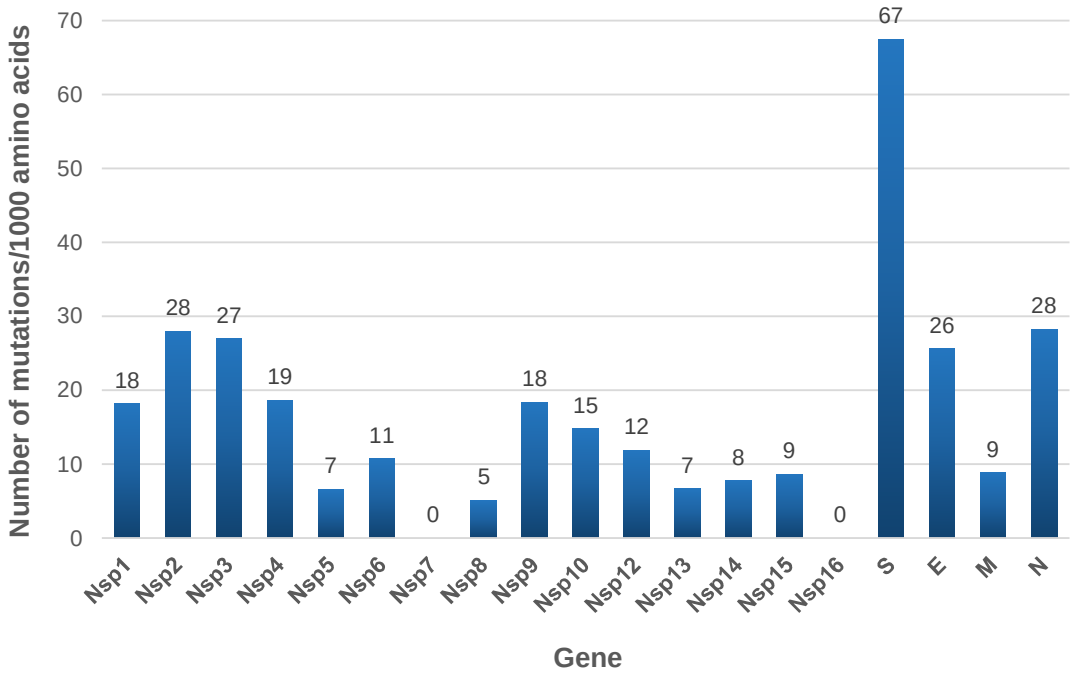
E, envelope; M, Membrane; N, Nucleocapsid; S, Spike.

bioRxiv preprint doi: https://doi.org/10.1101/2025.03.12.639729; this version posted March 14, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

bioRxiv preprint doi: https://doi.org/10.1101/2025.03.12.639729; this version posted March 14, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

396 **Figure 5.** Number of amino acid mutations per 1,000 amino acids

397



398

399

400 Nsp11 was excluded from the analyses as it is only 51 nucleotide-long and overlaps Nsp12.

401 E, envelope; M, Membrane; N, Nucleocapsid; S, Spike.

402

233

bioRxiv preprint doi: https://doi.org/10.1101/2025.03.12.639729; this version posted March 14, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

403 # TABLE

404 **Table 1.** PCR primers and conditions used for the amplification of HCoV-229E genome

405 fragments

406
407

234

bioRxiv preprint doi: https://doi.org/10.1101/2025.03.12.639729; this version posted March 14, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

408

Primer name	Primer sequence (5'-3')	Concentration of primers in pools	Pool
229E_F1	ACGCCTTACGCGAGGTTC	10 pmol/µl	Pool1
229E_R1	ACACCAGCACCAGCTTGC		
229E_F2_alt	CTCAGCAAGCTGGTGCTG	15.3 pmol/µl	Pool2
229E_R2_alt	CATCCTCAAACCAAGACTCG		
229E_F3_alt	CGAGTCTTGGTTTGAGGATG	15.3 pmol/µl	Pool1
229E_R3_alt3	ACATACACAAACCAACATCACC		
229E_F4	ATGGGCCGAGTTGCCAAG	10 pmol/µl	Pool2
229E_R4	AATTGGTGTTAAAGGTGTGCC		
229E_F5_alt	CAAGGCACACCTTTAACACC	15.3 pmol/µl	Pool1
229E_R5_alt	AAGTGTCTTCAGCATCACCC		
229E_F6	GGTGACAAGGGTGATGCTG	10 pmol/µl	Pool2
229E_R6	TGCAAGAACCAATTTGTTTCC		
229E_F7	TGTTGCTCAGATGATAAGCAC	10 pmol/µl	Pool1
229E_R7	CAACTGACAAAGTGGACAAAA		
229E_F8	TCACAGTTGTTGTGTAGGCC	10 pmol/µl	Pool2
229E_R8	CATAACCCTCAAAACTACTCAC		
229E_F9	GTGGTCTTGTGTGTTTGATTC	10 pmol/µl	Pool1
229E_R9	CCAGTACCACAAACRTAACC		
229E_F10	TGTGTTGACTCCAATGCAGG	10 pmol/µl	Pool2
229E_R10	CGGTAATCAGCTTCATTTGCA		
229E_F11	TGCTAATAGCATCTCTCCAGA	10 pmol/µl	Pool1
229E_R11	CTTTCCACACAGACACCAGT		
229E_F12	CACAGGCTAATGGTTTCACA	10 pmol/µl	Pool2
229E_R12	AAGAGTGCATCAAAAGGGCC		
229E_F13_alt	TGATTTCTGTGTTAGTCCAGC	15.3 pmol/µl	Pool1
229E_R13_alt	AATGCCTCCATCACCTTCAC		
229E_F14_alt2	TTGACTTGTGAACGTGTCGTTA	12.6 pmol/µl	Pool2
229E_R14_alt2	ACTGCTCGTGGTCCATAACT		
229E_F15,3	ATGGACCACGAGCAGTCC	15.3 pmol/µl	Pool1
229E_R15,3	GGGTRTTAACATCTTTGTTCC		
229E_F16	TGATGGTGTACCCGTAGTTG	12.6 pmol/µl	Pool2
229E_R16	ACAACCTCGGTCAAAACTTGA		
229E_F17	GCATGTCACATGTTGTACGG	12.6 pmol/µl	Pool1
229E_R17	GGCGCACTTAGTGCACAAC		
229E_F18	TCTTTGTGTAGTATGTGGCTCT	15.3 pmol/µl	Pool2
229E_R18	TCAACACTATAGGMTGTGACAG		
229E_F19	GAGGATCGTGTTCACTGCTT	10 pmol/µl	Pool1
229E_R19	CCACTGTCTGAGTTTGTAAAC		
229E_F20	AYTCCACATGGAGCAAGGC	12.6 pmol/µl	Pool2
229E_R20	AAAATAACGCATAGTGGTCAAC		
229E_F21	TTCTTGCTGGCTCATCTGAT	10 pmol/µl	Pool1
229E_R21	CAGCACCACCAATATTACACT		
229E_F22	GAYGACGGTTCATGTGAGG	10 pmol/µl	Pool2
229E_R22	AAGTCTTCTTCCATTGTGCTKC		
229E_F23	GACTTTTTGCCTCGMAGCA	12.6 pmol/µl	Pool1
229E_R23	GCCAGCACCCAAGTGTAAC		
229E_F24	ACACTTGGGTGCTGGCTC	10 pmol/µl	Pool2
229E_R24	ACCATCTACAACAGATGAGGTA		
229E_F25	TATTTGMTGTTGAGAGTGGTG	12.6 pmol/µl	Pool1
229E_R25	GAAAAAGGGCAATTTCCCGT		
229E_F26	TGTTGAYACATCACACTTCAC	10 pmol/µl	Pool2
229E_R26	GAAACATTACGTGGTTGAACAG		
229E_F27_alt	AGCAAGCTGTTGTTGGTGCT	15.3 pmol/µl	Pool1
229E_R27_alt	TGGTATTGAAGCGGCTGATG		
229E_F28_alt	TGATGCTGAACGWATGGCCA	15.3 pmol/µl	Pool2
229E_R28_alt	ATGGTAGGAATACGTGGTTCA		
229E_F29_alt1	CTCCTGAGGGGCTTGTTTTT	15.3 pmol/µl	Pool1
229E_R29_alt1	AGCATGATCATCCACTAGCTT		
229E_F30	TTGTTTTGTGAACAMCCAGCT	10 pmol/µl	Pool2
229E_R30	GCTTTTGATTGAAGTGTGGTAT		
229E_F31_alt	ATGGCTACAGTCAAATGGGC	15.3 pmol/µl	Pool1
229E_R31_alt	TTGTCAAGCCAAAGCAAGGG		

409

235

bioRxiv preprint doi: https://doi.org/10.1101/2025.03.12.639729; this version posted March 14, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

RNA was extracted by the KingFisher Flex system (Thermo Fisher Scientific, Waltham, MA, USA) according to the manufacter's protocol. Then RNA extracts were amplified by standard RT-PCR using the SuperScript III One-Step RT-PCR Kit with Platinum Taq High Fidelity (Invitrogen, Life Technologies, Carlsbad, CA, USA) and designed PCR primer pairs used in two pools not to have amplicon overlaps. PCR amplification included a reverse transcription step according to the following reaction: 12.5 µL of 2X mix, 0.7 µL of enzyme, 0.5 µL of each PCR pool, 3 µL of nucleic acid extract, and 8.3 µL of water. The PCR protocol included a preliminary reverse transcription step at 50°C for 25 minutes, followed by an initial denaturation at 95°C for 2 minutes, then by 39 PCR cycles including denaturation at 95°C for 15 sec, hybridization at 58°C for 45 sec, and elongation at 70°C for 1.5 min. Finally, a final elongation step at 70°C for 5 min was carried out. The PCR amplicons were purified on a NucleoFast 96-well plate (Macherey-Nagel, Hoerdt, France) with an elution with 40 µL of pure water. Post-PCR amplification, amplicons from the two primer pools were mixed. Degenerated nucleotides are indicated by a red font.

Table 2. Mutations in the spike protein present in ≥5 genomes compared with reference genome NC_002645 dating backing to 1962

Amino acid positions in the spike	6	19	22	26	35	41	51	64	66	88	89	90	91	102	104	109	111	112	113	121	140	159	181	203	210	213	223	225	228	229	248	283	288	305	307	309	310	311	312	314	318	321	324
Amino acid in the reference genome (NC_002645)	V	T	L	Y	Y	A	D	S	V	L	R	F	T	R	D	S	D	V	L	L	V	F	A	I	T	N	T	E	D	F	S	S	V	Y	D	K	P	Q	S	G	F	Y	G
Prevalence (%)	5	73	95	94	97	86	97	5	4	66	64	63	66	58	72	68	67	67	67	63	70	66	5	4	68	67	71	70	72	73	74	72	74	73	72	73	75	74	74	72	76	74	74
Genome (emerging lineage)	F	I	T	H	H	D	N	-	-	S	Q	V	I	T	A	Y	N	A	S	I	A	S	T	-	S	D	N	A	-	V	A	P	E	H	K	E	H	G	R	P	Y	R	V

Amino acid positions in the spike	325	349	349	350	352	353	354	355	358	358	365	377	391	391	401	404	405	406	406	407	409	410	411	424	430	444	462	488	530	642	681	739	765	775	792	845	871	937	971	971	972	1005
Amino acid in the reference genome (NC_002645)	V	K	K	Y	A	V	Y	A	G	G	N	N	D	D	V	W	A	Y	Y	S	Y	Y	T	G	Q	D	V	K	L	R	T	V	V	A	K	N	T	I	T	T	N	M
Prevalence (%)	76	76	24	99	97	100	100	96	18	77	76	76	76	23	99	99	99	23	76	99	99	99	99	99	99	99	99	99	100	98	98	75	98	89	88	30	87	87	16	43	13	59
Genome (emerging lineage)	I	R	Q	F	-	-	-	D	A	P	D	I	Y	A	M	L	V	N	G	Y	S	H	N	V	K	N	I	N	M	M	R	A	A	S	R	S	I	L	I	V	K	I

Prevalence (%)
- <25
- 25-50
- 50-75
- 75-100

bioRxiv preprint doi: https://doi.org/10.1101/2025.03.12.639729; this version posted March 14, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Références

1. Allocution liminaire du Directeur général de l'OMS lors du point presse sur la COVID-19- 11. (2020).
2. Wu, F. *et al.* A new coronavirus associated with human respiratory disease in China. *Nature* **579**, 265–269 (2020).
3. Dockès, P. En sortir, mais dans quel état? De la peste à la covid-19.
4. Edouard Mathieu, Hannah Ritchie, Lucas Rodés-Guirao, Cameron Appel, Daniel Gavrilov, Charlie Giattino, Joe Hasell, Bobbie Macdonald, Saloni Dattani, Diana Beltekian, Esteban Ortiz-Ospina and Max Rose. Excess mortality during the Coronavirus pandemic (COVID-19)" Published online at OurWorldinData.org. Retrieved from: 'https://ourworldindata.org/excess-mortality-covid. (2020).
5. Naseer, S. *et al.* COVID-19 outbreak: Impact on global economy. *Front. Public Health* **10**, 1009393 (2023).
6. Berche, P. The Spanish flu. *La Presse Médicale* **51**, 104127 (2022).
7. Drosten, C. *et al.* Identification of a Novel Coronavirus in Patients with Severe Acute Respiratory Syndrome. *N Engl J Med* **348**, 1967–1976 (2003).
8. Bermingham, A. *et al.* Severe respiratory illness caused by a novel coronavirus, in a patient transferred to the United Kingdom from the Middle East, September 2012. *Eurosurveillance* **17**, (2012).
9. Petersen, E. *et al.* Comparing SARS-CoV-2 with SARS-CoV and influenza pandemics. *The Lancet Infectious Diseases* **20**, e238–e244 (2020).

10. D'Arienzo, M. & Coniglio, A. Assessment of the SARS-CoV-2 basic reproduction number, R0, based on the early phase of COVID-19 outbreak in Italy. *Biosafety and Health* **2**, 57–59 (2020).
11. Liu, Y. & Rocklöv, J. The reproductive number of the Delta variant of SARS-CoV-2 is far higher compared to the ancestral SARS-CoV-2 virus. *Journal of Travel Medicine* **28**, taab124 (2021).
12. Liu, Y. & Rocklöv, J. The effective reproductive number of the Omicron variant of SARS-CoV-2 is several times relative to Delta. *Journal of Travel Medicine* **29**, taac037 (2022).
13. Chu, D. K. *et al.* Physical distancing, face masks, and eye protection to prevent person-to-person transmission of SARS-CoV-2 and COVID-19: a systematic review and meta-analysis. *The Lancet* **395**, 1973–1987 (2020).
14. Di Domenico, L., Pullano, G., Sabbatini, C. E., Boëlle, P.-Y. & Colizza, V. Impact of lockdown on COVID-19 epidemic in Île-de-France and possible exit strategies. *BMC Med* **18**, 240 (2020).
15. Ndwandwe, D. & Wiysonge, C. S. COVID-19 vaccines. *Current Opinion in Immunology* **71**, 111–116 (2021).
16. Young, M., Crook, H., Scott, J. & Edison, P. Covid-19: virology, variants, and vaccines. *bmjmed* **1**, e000040 (2022).
17. Gorbalenya, A. E., Enjuanes, L., Ziebuhr, J. & Snijder, E. J. Nidovirales: Evolving the largest RNA virus genome. *Virus Research* **117**, 17–37 (2006).
18. Fenner F. Classification and nomenclature of viruses. Second report of the International Committee on Taxonomy of Viruses. Intervirology. 7(1-2):1-115 (1976).
19. Woo, P. C. Y. *et al.* ICTV Virus Taxonomy Profile: Coronaviridae 2023: This article is part of the ICTV Virus Taxonomy Profiles collection. *Journal of General Virology* **104**, (2023).

20. Zhu, N. *et al.* A Novel Coronavirus from Patients with Pneumonia in China, 2019. *N Engl J Med* **382**, 727–733 (2020).
21. Arduini, A., Laprise, F. & Liang, C. SARS-CoV-2 ORF8: A Rapidly Evolving Immune and Viral Modulator in COVID-19. *Viruses* **15**, 871 (2023).
22. Konno, Y. *et al.* SARS-CoV-2 ORF3b Is a Potent Interferon Antagonist Whose Activity Is Increased by a Naturally Occurring Elongation Variant. *Cell Reports* **32**, 108185 (2020).
23. Meyerowitz, E. A., Richterman, A., Bogoch, I. I., Low, N. & Cevik, M. Towards an accurate and systematic characterisation of persistently asymptomatic infection with SARS-CoV-2. *The Lancet Infectious Diseases* **21**, e163–e169 (2021).
24. Jones, D. L. *et al.* Shedding of SARS-CoV-2 in feces and urine and its potential role in person-to-person transmission and the environment-based spread of COVID-19. *Science of The Total Environment* **749**, 141364 (2020).
25. Ochani, R. K. *et al.* COVID-19 pandemic: from origins to outcomes. A comprehensive review of viral pathogenesis, clinical manifestations, diagnostic evaluation, and management.
26. Hamming, I. *et al.* Tissue distribution of ACE2 protein, the functional receptor for SARS coronavirus. A first step in understanding SARS pathogenesis. *The Journal of Pathology* **203**, 631–637 (2004).
27. Cevik, M. *et al.* SARS-CoV-2, SARS-CoV, and MERS-CoV viral load dynamics, duration of viral shedding, and infectiousness: a systematic review and meta-analysis. *The Lancet Microbe* **2**, e13–e22 (2021).
28. Shu, Y. & McCauley, J. GISAID: Global initiative on sharing all influenza data – from vision to reality. *Eurosurveillance* **22**, (2017).

29. World Health Organization. Tracking SARS-CoV-2 variants [Internet] [accessed on 2021 October 6]. Available at: https://www.who.int/en/activities/tracking-SARS-CoV-2-variants/.
30. Hadfield, J. *et al.* Nextstrain: real-time tracking of pathogen evolution. *Bioinformatics* **34**, 4121–4123 (2018).
31. Burki, T. The origin of SARS-CoV-2 variants of concern. *The Lancet Infectious Diseases* **22**, 174–175 (2022).
32. Koopmans, M. SARS-CoV-2 and the human-animal interface: outbreaks on mink farms. *The Lancet Infectious Diseases* **21**, 18–19 (2021).
33. Moelling, K. Within-Host and Between-Host Evolution in SARS-CoV-2—New Variant's Source. *Viruses* **13**, 751 (2021).
34. Kemp, S. A. *et al.* SARS-CoV-2 evolution during treatment of chronic infection. *Nature* **592**, 277–282 (2021).
35. Cele, S. *et al.* SARS-CoV-2 prolonged infection during advanced HIV disease evolves extensive immune escape. *Cell Host and Microbe* **30**, 154-162.e5 (2022).
36. Connor, R. *et al.* Recommendations for Uniform Variant Calling of SARS-CoV-2 Genome Sequence across Bioinformatic Workflows. *Viruses* **16**, 430 (2024).
37. Gribble, J. *et al.* The coronavirus proofreading exoribonuclease mediates extensive viral recombination. *PLoS Pathog* **17**, e1009226 (2021).
38. Burel, E. *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. *Viruses* **14**, 1266–1266 (2022).
39. Ko, S. H. *et al.* Rapid Emergence and Evolution of SARS-CoV-2 Variants in Advanced HIV Infection. Preprint at https://doi.org/10.1101/2024.01.05.574420 (2024).

40. Francisco, E. C. *et al.* Unveiling Trichosporon austroamericanum sp. nov.: A Novel Emerging Opportunistic Basidiomycetous Yeast Species. *Mycopathologia* **189**, 43 (2024).