Using viral genomics for the understanding of the epidemiology and evolution of RNA viruses in the context of past and ongoing outbreaks

Fabiana Gámbaro Roglia

Université Paris Cité

Ecole doctorale Bio Sorbonne Paris Cité, ED 562

G5- Génomique évolutive des virus à ARN, Institut Pasteur

Using viral genomics for the understanding of the epidemiology and evolution of RNA viruses in the context of past and ongoing outbreaks

par Fabiana Gámbaro Roglia

Thèse de doctorat de Microbiologie

Dirigée par Jean-Francois Bureau

Présentée et soutenue publiquement le 15 Juin 2022

Devant un jury composé de :
Dr. Sara Moutallier, Directrice de Recherche, ANSES, Rapporteuse
Pr. Nuno Faria, Professeur, Oxford University, Rapporteur
Pr. Diane Descamps, PU-PH, Université Paris Cité, Examinatrice
Dr. Mael Bessaud, Chercheur, Institut Pasteur, Membre invité
Dr. Etienne Simon-Loriere, Chargé de Recherche, Institut Pasteur, Membre invité
Dr. Jean-Francois Bureau, Directeur de Recherche, Université Paris Cité, Directeur de thèse

Université Paris Cité

Doctoral school Bio Sorbonne Paris Cité, ED 562

G5- Evolutionary genomics of RNA viruses, Institut Pasteur

Using viral genomics for the understanding of the epidemiology and evolution of RNA viruses in the context of past and ongoing outbreaks

presented by Fabiana Gámbaro Roglia

Doctoral Thesis in Microbiology

Supervised by Jean-Francois Bureau

Publicly presented and defended on 15th of June 2022

In front of a jury composed by:
Dr. Sara Moutallier, Directrice de Recherche, ANSES, Reviewer
Pr. Nuno Faria, Professeur, Oxford University, Reviewer
Pr. Diane Descamps, PU-PH, Université de Paris, Examinator
Dr. Mael Bessaud, Chercheur, Institut Pasteur, Invited guest
Dr. Etienne Simon-Loriere, Chargé de Recherche, Institut Pasteur, Invited guest
Dr. Jean-Francois Bureau, Directeur de Recherche, Université de Paris, Supervisor

I dedicate this thesis to my dear sister Ximena because she is a fighter, and thanks to that, she will be able to read this.

Le dedico esta tesis a mi querida hermana Ximena porque es una luchadora y gracias a eso podrá leer esto.

RESUME

« Application de la génomique virale pour comprendre l'épidémiologie et l'évolution des virus à ARN dans le context d'épidémies passées et en cours »

Les pandémies qui ont impacté le début du XXIe siècle (la grippe H1N1 en 2009 et la COVID-19 actuellement) soulignent l'importance du défi global que posent les maladies infectieuses virales émergentes et ré-émergentes à nos sociétés. Par exemple, au cours des deux dernières années seulement, le virus chikungunya (CHIKV) a provoqué des épidémies majeures dans plusieurs pays d'Asie du Sud-Est, auxquelles s'ajoute le fardeau croissant du syndrome respiratoire aigu sévère (SARS-CoV-2), avec des conséquences majeures dans le monde entier. Bien que les pandémies ne soient pas un phénomène nouveau, la popularité croissante des technologies de séquençage de nouvelle génération, l'accélération de la génération de données génomiques – qui élargissent les bases de données de génomes viraux – et les progrès des méthodes phylodynamiques ainsi que de l'informatique ont permis d'utiliser les génomes viraux pour répondre à des questions épidémiologiques cruciales, renforçant les réponses de la santé publique aux épidémies.

Ce travail s'est concentré sur l'étude de l'origine, aussi bien spatiale que temporelle, et de la propagation des maladies infectieuses virales d'épidémies passées et en cours. Pour ce faire, nous avons utilisé l'épidémiologie génomique et montré qu'elle peut être un outil très puissant pour enquêter sur les épidémies de maladies infectieuses à différents niveaux. Par exemple, nous avons réussi à mettre en place un protocole de séquençage métagénomique profond qui nous a permis de découvrir que l'agent étiologique responsable d'une série de cas de méningite dans le sud de l'Espagne, au cours de la période 2015-2018, était le virus Toscana (TOSV), un arbovirus responsable d'un nombre croissant d'infections dans les pays bordant la mer Méditerranée. Ensuite, nous avons développé une approche de séquençage basée sur des amplicons qui nous a permis d'obtenir la séquence complète du TOSV à partir d'échantillons de qualité et charge virale très variables. Nous sommes ensuite allés plus loin dans les analyses des épidémies de CHIKV au Cambodge. En effet, l'analyse phylogénétique des génomes que nous avons générés nous a permis d'étudier la diversité génétique du CHIKV et, en ajoutant des données temporelles, d'estimer le moment auquel il a été introduit dans la population. Dans un second temps, une analyse phylogéographique nous a fourni un niveau de détail supplémentaire en mettant en lumière les origines de l'épidémie, les liens avec les épidémies précédentes dans la même région et la dispersion du virus dans le pays. Enfin, à une autre échelle, nous avons pu suivre la dynamique de la population virale du SARS-CoV-2 lors de l'infection à long terme d'un patient immunodéprimé, ce qui a souligné les défis de la prise en charge de ces membres vulnérables de notre société pour lesquels il existe un risque accru de maladie grave.

Pour conclure, ces travaux contribuent à une meilleure compréhension de l'épidémiologie et l'évolution de plusieurs virus à ARN qui représentent encore aujourd'hui des menaces importantes pour la santé humaine, et soulignent les apports de l'épidémiologie génomique contre les épidémies.

Mots clés : Épidémiologie génomique, Phylodynamique, Virus à ARN, Séquençage de nouvelle génération

RESUME LONG

De la pandémie de grippe H1N1 de 2009 à la pandémie de la COVID-19 (due au virus SRAS-CoV-2), les maladies infectieuses émergentes et ré-émergentes constituent l'un des défis mondiaux les plus importants du XXIe siècle. Ils représentent une menace importante pour la santé publique et l'économie mondiale. Cela fait plus de deux ans que les premiers cas d'infection par le SRAS-CoV-2 ont été signalés, puis la déclaration de la COVID-19 comme une pandémie par l'Organisation mondiale de la santé (OMS) le 11 Mars 2020. Cette pandémie a provoqué des conséquences dévastatrices sur l'ensemble de la planète. Plus de 6 millions de morts ont été enregistrés au moment de la rédaction de cette thèse. De plus, la pandémie a causé une crise économique sans précédente, avec des millions de personnes tombées en-dessous du seuil de la pauvreté. Cette perte économique mondiale est estimée à 12 000 milliards de dollars américain jusqu'à la fin de l'année 2021. A cela s'ajoute une crise sociale provoquée par une augmentation des inégalités sociales, ainsi que par l'impact négatif des mesures de distanciation sociale, et de confinement, sur le bien-être psychologique de la population en général, et des enfants en particulier. Bien que les systèmes de surveillance de santé publique se soient améliorés pour faire face aux conséquences de la croissance de la population humaine et de la connectivité mondiale croissante, les effets continus de la pandémie de COVID-19 rappellent le risque de maladies infectieuses émergentes et soulignent l'importance d'avoir un cadre de surveillance et de lutte contre ces maladies.

Néanmoins, l'émergence de maladies infectieuses est un phénomène qui semble se produire d'une façon périodique au cours de l'histoire de l'humanité. Il a été suggéré que l'essor de l'agriculture il y'a près de 11 000 ans était un événement clé pour la propagation des maladies humaines. En effet, l'agriculture a conduit à la sédentarisation des personnes et à l'augmentation de la population humaine. Entre le XVe et le XVIIIe siècle, certaines maladies telles que la variole, la tuberculose et la poliomyélite avaient circulé dans plusieurs régions du monde, favorisées par la colonisation, l'esclavage et la guerre, entraînant ainsi une morbidité et une mortalité importantes. Cependant, au cours des deux dernières décennies, la connectivité mondiale croissante ainsi que les facteurs démographiques et écologiques ont modifié la dynamique et le risque potentiel de maladies infectieuses. Par exemple, en raison du développent et la facilité des moyens de transport, les agents infectieux peuvent maintenant se propager plus rapidement et plus largement, entraînant ainsi l'introduction d'agents pathogènes dans de nouvelles populations hôtes. Une grande partie de ces agents est représentée par des virus à ARN, connus pour leur capacité d'évolution rapide et leur potentiel adaptatif, deux propriétés qui rendent difficile le contrôle de ces virus.

Cependant, l'accessibilité croissante aux technologies de séquençage de nouvelle génération, le rythme accéléré pour la génération des données génomiques, contribuent activement à expandre la diversité des génomes viraux au niveau des data-bases et par conséquent au progrès des analyses phylo-dynamiques. Cela a permis en partie de répondre à des questions d'intérêt crucial en santé publique, notamment lors des épidémies. En effet, cette science qui fait appel à la collecte et à l'analyse des données génomiques sur les agents infectieux circulants est appelée « épidémiologie génomique ».

L'utilisation des génomes des agents pathogènes pour obtenir des informations sur les processus sous-jacents à l'épidémie a conduit à la naissance du domaine de la phylo-dynamique. La phylo-dynamique repose sur le l'hypothèse que les processus épidémiologiques et évolutifs sous-jacents à l'épidémie se produisent à la même échelle temporelle. Cela semble particulièrement vrai pour les virus à ARN, car il s'agit d'agents pathogènes à évolution rapide accumulant des variations génétiques au fur et à mesure que l'épidémie se propage. Par conséquent, les séquences génomiques virales recueillies auprès de quelques individus sont souvent suffisamment divergents pour construire une phylogénie. Cette variation

génétique peut être interprétée comme le résultat d'une sélection liée à l'action d'une combinaison de processus épidémiologiques, évolutifs et immunologiques sur les génomes viraux au fur et à mesure que le virus se propage dans la population. Par conséquent, l'analyse de la phylogénie virale et de la dynamique évolutif d'une épidémie pourrait notamment apporter des réponses sur l'origine de l'épidémie, ainsi que d'autres informations. Ces informations comprennent certains paramètres épidémiologiques nécessaire au control de l'épidémie, comme par exemple, le R0, l'origine spatio-temporelle de l'épidémie, une estimation de la prévalence de la contamination initiale, ainsi que les facteurs environnementaux et sociaux favorisants la propagation virale

Cette thèse explore l'utilisation de l'épidémiologie génomique pour investiguer l'origine et l'échelle temporale, mais aussi la propagation de plusieurs maladies infectieuses virales, contribuant à une meilleure compréhension de l'épidémiologie et de l'évolution de plusieurs virus à ARN représentant une menace importante pour la santé humaine. Parmi ces virus, le virus Toscana (TOSV), Enterovirus (EVs), virus Chikungunya (CHIKV) et SRAS-CoV-2. D'autre part, nous avons démontré que l'épidémiologie génomique est un outil puissant pour enquêter sur les épidémies de maladies infectieuses à différentes étapes de son évolution. Par exemple, lors de la détection d'une maladie, la première question à poser concerne l'agent infectieux responsable.

Dans le chapitre 1, nous avons utilisé le séquençage métagénomique (mNGS) pour identifier l'agent pathogène potentiel à l'origine d'une série de cas de méningite dans le sud de l'Espagne. En effet, notre analyse mNGS a détecté dans plusieurs prélèvements de patients la présence de l'ARN du TOSV, un arbovirus responsable d'un nombre croissant d'infections neuroméningées dans les pays du bassin méditerranéen. La technique mNGS est une approche prometteuse pour le diagnostic des maladies infectieuses car elle permet la détection d'un large spectre d'agents pathogènes en un seul test (viral, bactérien, parasitaire et fongique). Cette technique permet aussi la détection des nouvelles formes recombinantes du virus. En effet, en utilisant cette approche, nous avons identifié une nouvelle forme recombinante E13 parmi les cas de méningite EV-positifs. Toutefois, ce résultat aurait pu potentiellement passer inaperçu avec les méthodes de typage classiques.

Une fois l'agent pathogène responsable de la maladie identifié, la prochaine étape est déterminer si nous disposons des outils de diagnostic des patients positifs (par exemple, tests qPCR ou sérologiques), d'analyses (par exemple, séquençage complet du génome) et de prévention (par exemple, vaccins). Dans le cas contraire, les informations fournies par le mNGS peuvent être utilisées afin de les développer. Dans ce sens, nos découvertes mNGS nous ont permis de concevoir une approche de séquençage basée sur les amplicons. Grâce à cela, nous avons réussi à obtenir plusieurs séquences complètes du génome du TOSV. De plus, ce protocole de séquençage serait d'une potentielle utilité pour les collègues en Espagne mais aussi de la région méditerranéenne, car il permet de détecter et de générer des génomes TOSV complets, en particulier à partir d'échantillons cliniques où l'ARN est souvent en qualité et en quantité limitées.

Le chapitre 2 a également combiné deux méthodes de séquençage (approches métagénomiques et basées sur les amplicons) pour obtenir des séquences complètes de CHIKV à partir de cas détectés au Cambodge lors de deux épidémies différentes : 2011-2013 et 2020. Pour tenter d'obtenir de plus amples informations sur ces épidémies et leur dynamique, nous avons d'abord ajouté des données temporelles à nos données génomiques pour estimer la période où le CHIKV a été introduit dans la population. Ensuite, nous avons inclus des données géographiques et nous avons effectué une analyse phylo-géographique donnant un niveau de détail supplémentaire, mettant ainsi en lumière l'origine potentielle de l'épidémie, les liens avec les épidémies dans la même région et la circulation du virus dans le pays. Dans l'ensemble, notre analyse phylogénétique a montré que les souches de CHIKV circulant au Cambodge de 2011 à 2013 appartenaient

à la lignée IOL, très probablement introduite dans le pays depuis la Thaïlande entre 2009 et 2011. Le CHIKV de cette épidémie abritait la substitution E1: A226V, associée à une augmentation de la dissémination du CHIKV chez les moustiques Aedes albopictus, ainsi que deux autres mutations de la glycoprotéine E2 qui semblaient être caractéristiques des souches appartenant à ce clade. En revanche, les virus signalés au Cambodge en 2020 n'avaient pas la substitution E1:A226V mais abritaient une double substitution dans les protéines de surface, qui ont récemment été associées à une infectiosité et une transmission accrues par Aedes aegypti (E1:K211E et E2:V264A). Notre analyse a également montré que les souches de CHIKV de l'épidémie de 2020 étaient phylogénétiquement plus proches du CHIKV circulant en Asie du Sud-Est. Cela suggère que la récente épidémie n'a pas été déclenchée par le CHIKV circulant auparavant au Cambodge, mais plutôt par l'introduction de virus à partir des pays voisins. En effet, notre analyse phylogénétique a proposé cinq scénarios différents pour expliquer l'introductions du virus, de Thaïlande et de Chine par exemples. Néanmoins, ces résultats doivent être interpréter en prenant notamment en compte la diversité génétique échantillonnée. Nous sommes conscients des limites de notre étude pour les données épidémiologiques et génomiques CHIKV disponibles pour le Cambodge et les pays voisins.

Lorsque des génomes viraux séquencés à partir de la même région au cours de différentes épidémies sont disponibles, la phylo-dynamique peut fournir des informations importantes sur l'évolution du virus au cours de la période inter-épidémique. Ces informations pourraient être utilisées pour répondre aux questions suivantes : le virus a-t-il pu persister dans la population entre les deux épidémies? Ou, le virus a-t-il été introduit dans la population par le biais d'un nouveau passage de la barrière espèce à partir d'un réservoir animal? Ces questions ont déjà été explorées à travers les différentes épidémies précédentes comme celle du virus Ebola (EBOV) par exemple. Dans le deuxième chapitre, nous avons exploré l'évolution inter-épidémique et la propagation du CHIKV au Cambodge. À la lumière des données génomiques et épidémiologiques disponibles pour la région de l'Asie du Sud-Est, notre analyse a suggéré que l'épidémie n'a pas été déclenchée par le CHIKV circulant auparavant au Cambodge, mais plutôt par l'introduction du virus à partir de pays où le CHIKV semble circuler presque continuellement, comme Inde.

La génomique virale peut également être utilisée pour étudier l'évolution intra-hôte. Le NGS permet le séquençage des virus à une profondeur de couverture élevée, caractérisant le répertoire complet des variants de la population virale. L'étude de la diversité de la population virale intra-hôte peut fournir des informations sur la manière dont ces processus pourraient être liés à l'évolution du virus à plus grande échelle. Il s'agit d'un point de discussion majeur sur la pandémie de COVID-19 en cours. En effet, il a été proposé que des infections chroniques conduisant à une accumulation substantielle de modifications nucléotidiques pourraient être responsables de l'émergence de COV tels que les variants Omicron ou Alpha. Le chapitre 3 explore cette thématique en analysant l'évolution du SRAS-CoV-2 au cours d'une infection à long terme chez un patient immunodéprimé. Nous avons constaté que l'évolution à long terme du SRAS-CoV-2 chez un individu immunodéprimé n'est pas toujours associée à une forte accumulation de changements, en particulier dans la protéine de surface. De plus, étant donné le traitement du patient par plasmathérapie convalescente, nous avons étudié l'impact de ce traitement sur l'évolution de la population virale chez le patient. D'une manière intéressante, nous avons noté une modification significative de la diversité virale mais aucune clairance du virus suite au traitement. En effet, le traitement a permet de restaurer le génotype dominant à l'état précoce de l'infection, caractérisé par la variante de la protéine de surface S50 (S:S50) au lieu de L50 (S:L50). Les séquences en acides aminés de la population dominante remplacée et de la nouvelle population dominante ne diffèrent qu'à cette position. Étant donné que la plupart des anticorps neutralisants ciblent la protéine de surface et que le traitement par plasmathérapie a induit un seul changement d'acide aminé dans la population virale, qui était situé dans la protéine de surface, nous proposons les deux scénarios suivants. Dans le premier scénario, la population

émergeait sous sélection positive : les anticorps presents dans le plasma convalescent ciblaient spécifiquement les virus porteurs du variant S:L50, entraînant ainsi le remplacement d'une population virale par un génotype ayant la capacité d'échapper à ce traitement (S:S50). Dans le deuxième scénario, les anticorps ont réussi à éliminer la population dominante dans les voies respiratoires inférieures, quel que soit le génotype de la protéine de pointe (c'est-à-dire S:S50 ou S:L50), qui a ensuite été reconstituée par une population virale moins accessible aux anticorps administrés, éventuellement situés dans un autre compartiment. Ce travail met également en évidence les défis de la prise en charge des personnes immunodéprimées atteintes de la forme chronique de la COVID-19.

En conclusion, ce travail de thèse met le point, en fournissant des exemples concrets, sur l'utilité de l'épidémiologie génomique et de la phylo-dynamique dans la compréhension de l'origine et l'évolution des épidémie causées par les virus à ARN, participant ainsi aux efforts de lutte contre ces virus et la compréhension de leur biologie.

Ce travail fournit des exemples de ce que nous pouvons apprendre des données de séquençage, soulignant l'utilité de l'épidémiologie génomique et de la phylodynamique pour les enquêtes sur les épidémies tout en ajoutant à la compréhension de plusieurs virus à ARN.

ABSTRACT

« Using viral genomics for the understanding of the epidemiology and evolution of RNA viruses in the context of past and ongoing outbreaks »

From 2009 H1N1 influenza pandemic to severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) COVID-19 pandemic, emerging and re-emerging viral infectious diseases have constituted one of the greatest global challenges of the twenty-first century. For instance, in the last two years alone, chikungunya virus (CHIKV) have caused major outbreaks in several countries across Southeast Asia, in addition to the growing burden of SARS-CoV-2, causing major consequences around the globe. While pandemics are not a new phenomenon, the growing popularity of next generation sequencing technologies, the increased pace of genome data generation – which are enlarging viral genome repositories – and advances in both phylodynamic methods and computer power made it possible to use viral genomes to answer crucial epidemiological questions, ultimately strengthening public health response to outbreaks.

This work focused on investigating the origin, timing, and spread of viral infectious diseases of past and ongoing outbreaks. To address these questions we used genomic epidemiology, showing that it can be a very powerful tool to investigate infectious disease outbreaks at various steps. For instance, we succeeded in setting up a metagenomic deep sequencing protocol that allowed us to determine that the probable etiologic agent responsible for a series of meningitis cases in southern Spain during 2015-2018 was Toscana virus (TOSV), an arbovirus responsible for an increasing number of infections in countries enclosing the Mediterranean Sea. Next, using an in-house designed amplicon-based sequencing approach we succeeded to obtain the complete sequence of TOSV from samples of varying viral load and quality. We then went a step further on the analyses of CHIKV outbreaks in Cambodia. In light of the genomes that we generated, phylogenetic analysis allowed us to study the genetic diversity of CHIKV and by adding temporal data, to estimate the time at which it was introduced into the population. Subsequent phylogeographic analysis provided us with an additional level of detail, shedding light on the origins of the outbreak, connections to previous outbreaks in the same region and dispersal of the virus within the country. On a different scale, we were able to track the dynamics of the SARS-CoV-2 viral population during the long-term infection of an immunocompromised patient, highlighting the challenges of treating these vulnerable members of our society for whom there is still significant risk for severe illness.

Together, this work provides a better understanding of the epidemiology and evolution of several RNA viruses representing important threats to human health and simultaneously highlights the importance and main contributions of genomic epidemiology for outbreak investigation.

Key words: Genomic epidemiology, Phylodynamics, RNA viruses, Next generation sequencing

ACKNOWLEGMENTS

I want to start by thanking **Etienne** for accepting me into his laboratory and giving me the opportunity to conduct my Ph.D. Thank you for the trust you have given me in the projects, the freedom to conduct them, and for letting me develop my ideas while challenging me with yours. I am very grateful for our discussions on RNA virus evolution, which I have always found fascinating. I also thank you for insisting that I had a backup plan project(s), which has ultimately become plan A.
I joined your lab a few months after it opened, and now, looking retrospectively, it is great to see how we have all grown along with it. While being the first Ph.D. student from the lab has been a challenge for both of us, in addition to projects that did not go as expected and the COVID-19 pandemic that turned our plans upside down, we managed to make it work, and I take with me great learning from all these years spent in the GEVA lab.

Next, I would like to thank **Jean-François Bureau** for accepting to be the director of this thesis and his support during these years.

I am also very grateful to all members of the **GEVA lab** for making the lab a nice place to work. **Matthieu**, thank you for teaching me all sorts of techniques, from virus infection and titration to library preparation, while coping with my beginner's clumsiness! I also thank you for all your advice on how to bike in Paris and for introducing me to "Massive Attack," which has played in the background during the long hours of writing. **Deborah**, thanks for your support during my first days in Paris, especially for using your MD skills to help me with the French health system! A big thanks go to **Artem**, with whom I enjoyed working on different occasions. You are a great coworker, always being scientifically sharp. I thank you for our interesting discussions about science or not and for your great support during the last period of my Ph.D. To the new members of our team, **Jerome** and **Said,** I thank you for adding to the good atmosphere of the lab, GEVA is in good hands!

I would like to extend my thanks to the **GMFI unit**, mainly to **Isabelle**, **Laurine**, and **Nicholas**. Thank you for putting up with me walking (and sometimes running) around your lab to prepare my gels, launch the sequencing run, and even use all your PCR machines. But above all, I thank you for your kindness and letting me practice my French with you during lunchtime, *merci!*

My sincere thanks to all the members of my Ph.D. committee, **Marco Vignuzzi**, **Etienne Patin**, **Bernard Cazelles**, and **Frederic Lemoine**, for their feedback and advice during our annual meetings.

I want to thank **Molly Ingersoll** for her moral support and advice at key moments during this Ph.D., always meeting me with great humanity and empathy. Molly, your team is very lucky to have you.

To our collaborator from Spain, **Lola**, thank you for trusting me with your project, and very importantly, I thank you for always showing up to our meetings with a big and contagious smile.

To the people back in **Institut Pasteur Montevideo**; First to **Natalia Etcheverria**, **Gonzalo Moratorio** and **Pilar Moreno**. Thank you for opening me to the fascinating world of RNA virus research, encouraging me, and facilitating the opportunity to pursue a Ph.D. in virology. To **Juan Pablo Tosar**, my former mentor. I cannot thank you enough for all the support and teachings of all these years about science and beyond, even when I was no longer your student. I hope someday to match you and inspire a student as you have

inspired me. Thanks to all these people for receiving me in your labs every time I returned home; thank you for keeping me rooted. I hope our paths will cross again.

A big thanks go to my friends here; **Lucia**, **Mara**, **Lena**, **Gaelle**, **Nate**, **Kyrie, Lewis, Bianchi,** and **Guillem**. Paris wouldn't have been the same without you! Thanks for the support, the long night-outs, the beers, the picnics, the trips together. In particular, Mara and Lucia, my rosarinas friends, thank you for all the mates, the talks, and the laughs. Thanks for making Paris a place that looked more like home.

I would also like to thank my two Uruguayan friends (in Paris): **Lu** and **Dani**. I thank you for your guidance and support throughout the whole Ph.D. From the initial paperwork, which I can't imagine having it done without you, to my very last steps. Thank you for making the transition from home much smoother and giving me that sweet little reminder from home.

I thank my two roommates: Ali y Sol; living with you has been one of the highlights of this Ph.D. **Ali**, thanks for your friendship, our scientific and political debates, your contagious laugh and introducing me to Ctangana; I cannot still get over it! **Sol,** doing justice to your name, you have enlightened up my final year of Ph.D. Thanks for supporting me during my thesis writing and giving me a new obsession: Japanese food (veggie, obviously).

Me gustaría agradecer a mi familia por su amor y su apoyo incondicional. Empiezo por mis padres, ***Mónica*** *y* ***Daniel****, gracias por apoyarme en mi decisión de irme del "paisito" en busca de mis sueños. Gracias por siempre creer en mí, incluso cuando yo no lo hacía. Gracias por estimularme a lograr mis objetivos. A mi hermana* ***Xime****, a quien esta tesis va dedicada. Gracias por luchar, por tus ganas de vivir y así, estar hoy acá con nosotros. Tu fuerza me ha inspirado y me dio el empuje que necesitaba para que esta tesis llegara a su final, ¡gracias! A mi hermanito* ***Lucho****, gracias por tus videollamadas que me transportan de forma inmediata junto a ti y me dan un poquito de eso que tanto extraño. Luego, doy las gracias al resto de mi familia, mi* ***madrina****, mi* ***padrino****, mi* ***tia Adri****, mis primas* ***Romi*** *y* ***Vale****, a nuestra pequeña* ***Juli.*** *Gracias por su apoyo y por darme un lugarcito en este mundo donde sé que siempre puedo refugiarme.*

*A la familia que uno elige: mis amigos de casa. A mis amigas del cole, "****Las Pichens****", y a mis amigas y amigos de la facu, "****Las chichis de Fcien****" y "****La fcien y the tox****"; gracias por su amistad y por, sobre todo, por entender mis largos meses de silencio. Quiero destacar a mis* ***"secuaces de la ciencia"****, Caro y Mica, y a* ***"mis amis del coure"****, Silvi, Monchi y Marce, quienes son la prueba viviente que la amistad no se trata de ser inseparables, sino que de estar separadas y que nada cambie. Algo que la distancia me ha enseñado.*

Last but not least, I would like thank **Lucas**, my colleague, my best friend, my R and French teacher, and my partner. *Depuis des jours que j'écris et réécris ces mots, cette page me semble trop petite et triviale pour ce que j'ai à te dire. Mais en gros, merci. Merci de donner de la passion à ma vie, qui s'imprime dans chacun des mots de ce manuscrit. Merci pour ta profonde compréhension de mon être, et grâce à ça, sachant toujours trouver les mots pour m'encourager et me soutenir. Et merci pour ta simplicité et ton bonheur contagieux qui rendent chaque jour un jour heureux. La vida es simplemente mucho mejor y más fácil a tu lado, ¡gracias!*

To conclude, a big thanks to all these people, who are the reminder that none of this really matters: your position, your title, nor your number of publications, but human relationships, how much we impact on people, and the impact they leave on us.

TABLE OF CONTENTS

LIST OF ABBREVIATIONS

AA - amino acid

AIDS - acquired immunodeficiency syndrome

BF - Bayes factor

BSSVS - Bayesian stochastic search variable selection

CHIKV - chikungunya virus

COVID-19 - coronavirus disease 2019

CSF - cerebrospinal fluid

CTMC - continuous-time Markov Chain

DENV - dengue virus

EBOV - Ebola virus

ECSA - East-Central-South-African

ELISA - enzyme-linked immunosorbent assay

EV - Enterovirus

GLM - generalized linear model

HA - hemagglutinin

HCV - hepatitis virus C

IAV - Influenza A virus

ICTV - international committee on taxonomy of viruses

IOL - Indian Ocean lineage

MCC - maximum clade credibility

MCMC - Markov chain Monte Carlo

MERS-CoV - Middle East respiratory syndrome coronavirus

mNGS - metagenomic next-generation sequencing

MuV - Mumps virus

NA - neuraminidase

NGS - next-generation sequencing

NHP - non-human primates

NiV - Nipah virus

NP - nucleoprotein

NS - nonstructural

PA - polymeric acid

PB1 - polymerase basic 1

PB2 - polymerase basic 2

PCR - polymerase chain reaction

PS - path sampling

R_0 - basic reproduction number

RABV - rabies virus

RBD - receptor-binding domain

RRW - relaxed random walk

RT - reverse transcription

Rt - effective reproductive number

RT-qPCR - quantitative reverse transcription PCR

RVDB - reference viral database

SARS-CoV - severe acute respiratory syndrome coronavirus

SARS-CoV-2 - severe acute respiratory syndrome coronavirus 2

SARSr-CoVs - SARS-related coronaviruses

SS - stepping stone

TOSV - Toscana virus

VOC - variants of concern

VOI - variants of interest

VUM - variants under monitoring

WA - West African

WHO - world Health Organization

WNV - West Nile virus

ZIKV - Zika virus

LIST OF TABLES

LIST OF FIGURES

1 INTRODUCTION

From the 2009 H1N1 influenza pandemic to severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) COVID-19 pandemic, emerging and re-emerging infectious diseases have constituted one of the most significant global challenges of the twenty-first century. They have placed a substantial threat to public health and the global economy (1). It has been more than two years since the first cases of SARS-CoV-2 infection were reported (2) and the subsequent declaration of a pandemic by the World Health Organization (WHO) on March 11 2020. Since the start COVID-19 pandemic, we have witnessed devastating consequences: loss of human life (by the time of writing of this thesis, there are more than 6 million deaths globally), a profound economic crisis with millions of people who had fallen into poverty, and an estimated loss of US$12 trillion on the global economy by the end of 2021 (3, 4) and a social crisis provoked by an increase in social inequalities at every scale, along with the negative impact of the social distancing and confinement measures on the psychological well-being of the general population (5), in particular children (6). Although the public health surveillance systems have been improving to face the consequences of the growing human population and increasing world connectivity, the ongoing effects of the COVID-19 pandemic are a reminder of the risk of emerging infectious diseases and highlight the importance of having a framework to survey and curb the ongoing pandemic and any other infectious disease that might arise in the future.

Nonetheless, history has shown us that emerging infectious diseases are not new. It is believed that a key event for the spread of human diseases was the rise of agriculture nearly 11,000 years ago, as it led to the settlement of people and the increase in the human population (7, 8). Between the five-teen and eight-teen centuries, diseases such as smallpox, tuberculosis, and polio circulated in several regions of the world promoted by colonization, slavery, and war, causing substantial morbidity and mortality (9).

However, in the past two decades, the increasing global connectivity together with demographic and ecological factors have changed the dynamics and the potential risk of infectious diseases (9). For instance, the increased international travel and trade have promoted the rapid spread of pathogens over large distances connecting pathogens with new host populations (9, 10).

This brings the following questions: what are the sources and drivers of these emerging infectious diseases? What can we do to control and prevent future outbreaks?

The first part of the introduction briefly describes emerging infectious diseases, i.e., what they are, where they come from, and the components of their emergence. The second part describes the main uses of genomic epidemiology and phylodynamics for studying the pathogen (particularly for this thesis, RNA viruses) underpinning the disease to help answer questions central to disease mitigation and control.

2 INTRODUCTION PART 1: Emerging and re-emerging infectious diseases

An emerging infectious disease is a disease caused by a new and previously unknown pathogen. In some instances, after the number of infections was reduced and the diseases no longer recognized as a public threat, the infectious diseases may reappear or appear in new locations or under a new variant form (e.g., drug-resistant or recombinant form). These are known as re-emerging infectious diseases (1).

Emerging infectious diseases can be caused by different infectious agents, including viruses, bacteria, fungi, protozoa, and helminths (11). This thesis will focus on those caused by viruses.

Depending on the context and its transmissibility among humans, an emerging pathogen can lead to individual or few infections resulting in a local outbreak. It can develop into an epidemic if the number of infections increases and there is geographic expansion. In a worst scenario, the disease can reach several countries or continents, affecting many individuals leading to a pandemic. Also critical to some regions are the so-called neglected diseases. These are diseases that have not received much attention at the national and international level but circulate and affect, in particular, low-income countries in Africa, Asia, and Latin America. Neglected diseases are commonly found in tropical areas and, for this reason, are also known as neglected tropical diseases. It is believed that these diseases continue to persist in communities due to a combination of several factors, including poverty, poor sanitation, and socio-political conflicts. Examples of neglected tropical diseases are dengue fever, rabies, and chagas disease (12).

2.1 What are the sources of emerging infectious diseases?

Most of the emerging infectious diseases have a zoonotic origin meaning that pathogens from wild or domesticated animals cross the species barrier to infect humans, causing disease (13-15). For this reason, these diseases are named zoonosis (or zoonoses in plural), a term that derives from two Greek words, “zoon” and “nosos,” which mean animals and disease, respectively.

The transmission event of a pathogen from animals to humans is also known as jump or spillover and it can occur directly or indirectly (16).

Direct zoonoses are transmitted to humans from infected animals through direct contact with saliva, blood, urine, mucous, feces, or through a mechanical vector (17). An example of this type of zoonosis is rabies, a fatal viral infection caused by the rabies virus (RABV). RABV is maintained in most parts of the world by dogs, foxes, raccoon dogs, raccoons, mongooses, and skunks (18), and it can be transmitted to humans through bites from these infected animals. Dogs constitute the most important host reservoir for RABV, and in fact, dog bites are responsible for the vast majority of human cases (19, 20). Such spillover infections are dead-end infections as further spread to other hosts, e.g., humans, will not occur. Nevertheless, these infections can provoke severe disease in humans, and in the absence of treatment, it can lead to death (18).

Indirect zoonoses are transmitted to humans by being in contact with contaminated objects or surfaces or by being exposed to areas where animals live and roam (16). Additionally, indirect zoonoses are associated with the butchering and consumption of wildlife and domesticated animals (21). For example, the emergence of human immunodeficiency virus (HIV) that causes the Acquired immunodeficiency syndrome (AIDS) is believed to have arisen from hunting non-human primates for food in Africa (22, 23).

Transmission of pathogens from animals to humans can also occur through arthropod vectors such as mosquitoes, ticks, or sandflies, falling under the scope of indirect transmissions of zoonotic diseases (24). Viruses that maintain transmission cycles between vertebrate animal reservoirs as main amplifying hosts and insects as primary vectors are known as arthropod-borne viruses or arboviruses. In particular, when the main vectors are mosquitoes, they can be further named mosquito-borne viruses (24), for example West Nile virus (WNV). WNV is maintained by an enzootic (wild) transmission cycle between wild birds, which act as reservoir hosts, and *Culex* mosquitoes. WNV is then occasionally transmitted to humans and other mammals through the bite of an infected mosquito (25). WNV cannot be directly transmitted from person to person, and humans are dead-end hosts as viremia does not reach a sufficient level to infect new mosquitoes and continue the transmission chain (26, 27). However, while humans do not contribute to the transmission cycle of WNV, other important arboviruses infecting humans, such as dengue virus (DENV), Zika virus (ZIKV), yellow fever virus (YFV), and chikungunya virus (CHIKV), can reach high enough levels of viremia for subsequent transmission to mosquitoes and in this way be transmitted among humans through mosquito bites (28).

2.2 Main components in the emergence of infectious viral diseases

Understanding the processes underpinning the emergence and re-emergence of infectious viral diseases is important to understand their origin and prevent and control current and future epidemics. The main components in the emergence of the infectious viral disease include human, environmental, and viral factors (29, 30).

2.2.1 Human factors

2.2.1.1 Human population growth and encroachment of natural habitats

Although the population growth rate is currently decreasing (around 1.05% per year in 2020, down from 1.08% in 2019 and 1.10% in 2018), in average, the global human population currently increases by 81 million people per year (31).

Such growth in the human population increases the demand for natural resources in several world regions, leading to the fragmentation and degradation of wildlife habitats, with humans and livestock encroaching on natural habitats. Indeed, according to a report from the Intergovernmental Science-Policy Platform on Biodiversity and Ecosystem Services released in 2019, only one-quarter of land areas remain undamaged by human activity. This encroachment of natural habitats and land-use change can alter the size and composition of the animal communities and the overall ecosystem functioning (32), therefore impacting the emergence of infectious diseases. A systematic review identified the most common land-use change

types linked to zoonotic disease transmission: deforestation, agricultural development, irrigation, and urbanization (33).

Habitat destruction, such as deforestation, can increase human exposure to zoonotic diseases due to, for example, the release of host reservoirs (30, 34). A recent study performed a large-scale analysis of the association between deforestation and EBOV outbreaks in Central and West Africa. The authors suggested that EBOV spillovers were associated with recent deforestation in the area, highlighting the importance of preventing the loss of natural forests to reduce the likelihood of future outbreaks (35).

Urbanization – the movement of people from rural to urban areas – is considered a significant driver for the emergence and spread of several arboviral diseases, including dengue (36-38). Arboviruses such as DENV or CHIKV are transmitted by *Aedes* mosquitoes, mainly *Aedes aegypti* and *Aedes albopictus* (39), two mosquitoes species well-adapted to urban settings. It has been proposed that the development of dense and sedentary human societies and the development of irrigation systems played a fundamental role in adapting mosquitoes to breed in human habitats. The reason for this might be the following: female mosquitoes lay their eggs in the water, and humans have developed ways to manipulate and store water like no other species. In the case of a dry season, a female mosquito looking for a place to oviposit would find the stored water in human settlements very appealing (40, 41).

2.2.1.2 Intensification of domestic livestock farming

Population growth has also increased the demand for animal protein, which has led to the intensification and industrialization of animal production (30). Domestic livestock farming, especially without effective disease-control practices, constitutes an important driver for the emergence of diseases. A possible reason for this is how these systems generally work: intensive livestock systems often have large numbers of genetically similar animals kept in close proximity to each other. In such conditions, homogeneous populations may favor the transmission and adaption of pathogens (42). Transmission of pathogens from livestock to humans can occur through direct contact with the animals (the farmworkers being especially at risk) or through the animal waste that is spread on the land and can come into contact with humans, wild animals, or contaminate the water (43). Additionally, intensification might also demand increased movement of vehicles, people, or animals between farms, which can increase the risk of pathogen transmission (44).

A well-known example of how the destruction of natural habitats, land-use change, and livestock farming promoted the emergence of a disease is the case of the Nipah virus (NiV) outbreak in Southeast Asia. NiV is a paramyxovirus that has spilled over from fruit bats to livestock, particularly pigs, and humans, causing disease (45). The first known outbreak of NiV happened in Malaysia in 1998 on a pig farm, causing the death of 108 humans, and in reaction, 1 million pigs had to be euthanized (42). Epidemiological investigations suggested that the emergence of NiV was linked to several related human activities. During the 1970-the 1990s, there has been intense deforestation of Peninsular Malaysia for pulpwood and industrial plantations such as palm oil (46), reducing the fruit bat habitat and their food supply. Consequently, NiV-infected bats went out in search of food and were attracted to orchard fruit trees that had been planted adjacent to the pig farm. Pigs got infected by consuming fruits contaminated with bat saliva or urine. Transmission and spread of NiV among the pigs were facilitated by the farm conditions and the transportation of pigs to other farms, leading to subsequent outbreaks in the South of Malaysia and Singapore (47). Transmission of NiV to humans is thought to have occurred through direct contact with

infectious secretions or excretions of pigs, with no evidence of a direct spillover from bats nor evidence of human-to-human transmission (42). However, the ecology of NiV spillover does not end here. Other routes of human NiV infection in Southeast Asia are associated with human activity. For example, in India and Bangladesh, a common infection mechanism is consuming contaminated date palm sap (48, 49). (Figure 2-1) Date palm sap collection is the main livelihood of the *gachis* (i.e., date palm sap collectors) and consists of shaving the bark of the trees and hanging clay pots in the trunks to collect the sap at night. During the night, large fruit bats, also called flying foxes, are attracted to these settings, and while feeding on these clay pots, they contaminate them with their saliva, urine, or feces (48). Furthermore, in Bangladesh direct contact with infected patients or their secretions consist of a major pathway of human-to-human infection (50). More recently, in 2014, an outbreak in the southern Philippines caused severe illness to humans and horses. The serological and limited genomic data suggested that the etiologic agent was NiV or a closely related virus (*Henipavirus* genus). Furthermore, epidemiologic data suggest that the outbreak was characterized by horse-to-human transmission, probably through direct exposure to infected horses during slaughtering or consumption of the meat of diseased horses and direct human-to-human virus transmission (51).

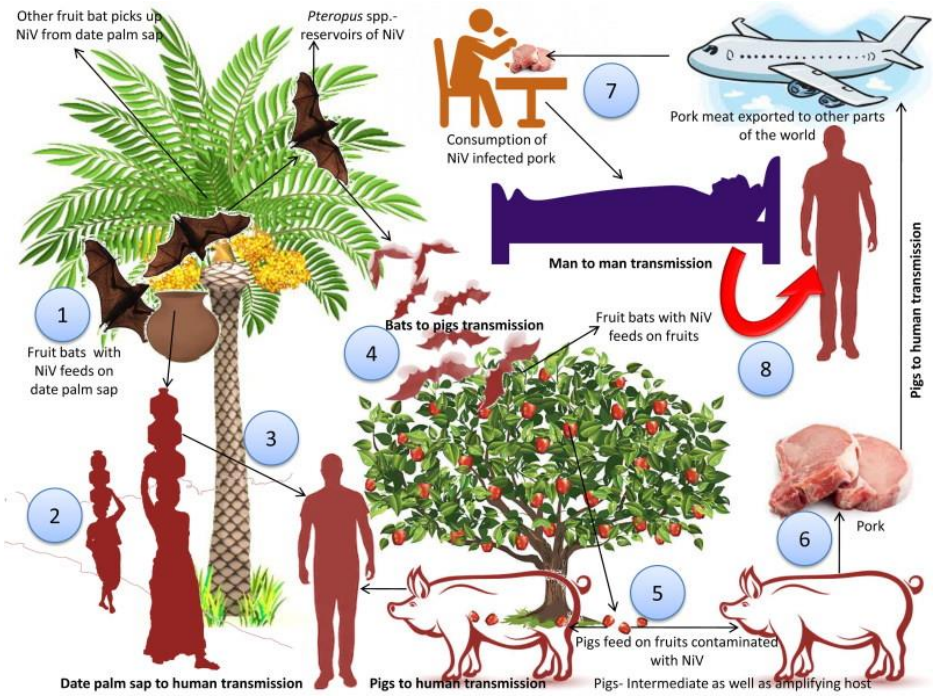


***Figure 2-1**: Schematic representation of NiV ecology. 1) NiV-infected bats feeds on date palm sap. 2), 3) NiV is transmitted to human through the consumption of date palm sap. 4) NiV-infected fruit bats attracted to orchard fruit trees, contaminated the fruits and the farm soil. 5) Pigs get contaminated by the consumption of contaminated fruits. 6) Pork meat infected with NiV are exported to other parts. 7) Human infections can occur through the consumption of contaminated meat. 8) Human-to-human transmission can occur through direct contact with infected patients. Image extracted from Singh RK, et al. 2019*

2.2.1.3 Exploitation of wildlife

Humans have hunted wild animals for consumption since ancient times, and it remains part of some cultures. However, past and recent outbreaks have shown us that hunting wildlife can promote pathogen spillover through consumption or exposure during handling and butchering (52).

Wild meat hunting is still important in several Africa, Asian, and Latin American countries. Considering only Africa and Latin America, estimates indicate a wild meat harvest of nearly 6 million tonnes per year (53). It is believed that the consumption of wild meat in some areas of the world remains a relevant human activity for two main reasons. First, as the human population grows, there is an increased demand for animal food, and low-income countries cannot always afford livestock domestication. In this way, bushmeat is an important source of food in many low-income countries, e.g., West and Central African countries (52, 54). Second, in some regions of the world, the consumption of wild animals is a symbol of status and a luxury good (52, 54).

2.2.1.4 Travel and Transportation

Throughout history, traveling and trade between countries have been accompanied by the spread of diseases. The introduction of smallpox and measles virus in the Americas during the sixteen century are old examples that illustrate this. However, the incredible expansion of global travel, either through air or water, and the development of high-speed railway networks and roads has resulted in unprecedented interconnectivity between human populations. These networks have promoted the rapid movement of people and pathogens over large distances like never before (29). According to the International Civil Aviation Organization, in 2019, before COVID-19, air travel rose to 4.5 billion passengers per year, which is 3.6% higher than in 2018 (55). This explosion of global connectivity fueled by human travel and trade favors viral spread. Indeed, international air travel is believed to have had an important role in the global spread of SARS-CoV-2 (56). Epidemiological and genomic investigations indicate that multiple introductions of SARS-CoV-2 in Brazil (57) and California, United States (58), occurred by air travel. Furthermore, international travel driving the introduction of infectious diseases into new regions has been described for other viruses such as ZIKV during its emergence in the Americas in late 2014 (59) or the DENV outbreak in Madeira, Portugal in 2012 (60). Similarly, although it is hard to demonstrate the exact pathway through which WNV was introduced into North America in the 1990s, either shipping or air travel have been postulated to be the most plausible scenarios (61).

Population mobility at the country level can also play an important role in disease spread; however, this type of data is often limited. Recently it has been proposed to use mobile phone data to assess how people move within communities (62). This data would come from anonymized mobile phone call detail records in which the time and location of each call and text made by each anonymized individual are registered by mobile networks operators. The analysis of such data would allow the inference of the movement of people to trace patterns of local mobility that could help study potential drivers of spatiotemporal spread (62). Owing to the popularity of mobile phones worldwide, obtaining such data could be relatively straightforward, offering an unprecedented source of information on human mobility. Nevertheless, the use of mobile phone data faces several challenges, a significant one is the concern about potential loss of privacy and data protection (63).

2.2.2 Environmental factors: climate change and global warming

Climate change has been described as a significant factor in disease emergence. Indeed, climatic changes can affect key climatic parameters such as global temperatures, soil composition, or rainfall patterns. Changing such environmental conditions can, in turn, influence the survival, reproduction, abundance, and distribution of host reservoirs, vectors, and pathogens (64).

Global temperature is rising, primarily due to the anthropogenic emission of greenhouse gases, with 2019 and 2020 being the warmest years since the pre-industrial era (65). Ectotherms mosquitoes, such as *Aedes aegypti* and *Aedes albopictus*, are sensitive to environmental temperature variations, which directly impact in their internal temperature (66). Consequently, environmental temperature is the main abiotic factor influencing mosquito physiology, ecology and behavior (66). Mosquitoes can only survive and reproduce in specific environments which will depend on the ecological and biological characteristics of each species. For example, for a given activity (e.g., feeding, flying, host-seeking, and reproduction) each species has an optimal temperature range. In the case of *Aedes aegypti,* it can develop (i.e., eggs, larvae and pupae) and survive between 16^{o}C and 34^{o}C while for *Aedes albopictus* the temperature range is wider between 10.4^{o}C and 35^{o}C (66, 67). This allows *Aedes albopictus* to tolerate lower temperatures. While extreme temperatures would kill mosquitoes within their survival range of temperature, it is known that warmer temperature increases mosquito activity, including development, blood-feeding, and reproduction (68). In this way, rising temperatures produce warmer winters, reducing mosquito mortality, and warmer summers, increasing mosquito density and activity. This could extend the mosquito circulation period and lead to a geographic expansion or redistribution of mosquitoes.

Indeed, a recent study used a model of viral transmission by mosquito vectors to predict and map *Aedes aegypti* and *Aedes albopictus* global distribution in the current and future (projected risk for 2050 and 2080) climate (69). Briefly, in agreement with a previous study (70), this work suggests that Aedes mosquitoes' distribution will likely continue to expand. However, while *Aedes albopictus* transmission potential will increase in Europe and North America, it might decline in the tropics (e.g., Southeast Asia, West Africa), where temperatures might become too high for these mosquitoes. In sum, this study suggests climate change might lead to a shift in the geographic distribution of mosquitoes, leading to a net increase and new exposures to Aedes mosquito-borne viruses.

2.2.3 Viral Factors: the role of rapid evolution in RNA viruses

The 2003 SARS-CoV outbreak, the 2009 H1N1 influenza pandemic, the 2012 MERS-CoV outbreak, the 2013–2016 EBOV outbreak in West Africa, the 2015 ZIKV epidemic in the Americas and the SARS-CoV-2 COVID-19 pandemic all have in common that the causative pathogen is an RNA virus.

What is so special about RNA viruses? Their intrinsic propensity for rapid evolution.

RNA viruses can generate high genetic diversity, which allows them to quickly adapt to changing environments, e.g., new host, adaptive immune responses, or antivirals. This constitutes a challenge for their control and possibly, contributes to their emergence risk.

The mechanism of RNA viruses genetic variation include: mutation, recombination, and reassortment; although these mechanisms occur to different extent in different virus families (Figure 2-2) (71, 72). The

fate of mutations or genomic variations will rely on two evolutionary process: natural selection and genetic drift. Selection tends to increase the frequency of beneficial genomic variants in a population while decreasing the frequency of detrimental ones, two processes that are usually referred to as positive and negative selection, respectively. In contrast, genetic drift refers to the stochastic change in the frequencies of genomic variants, which is particularly prominent in small populations (72).

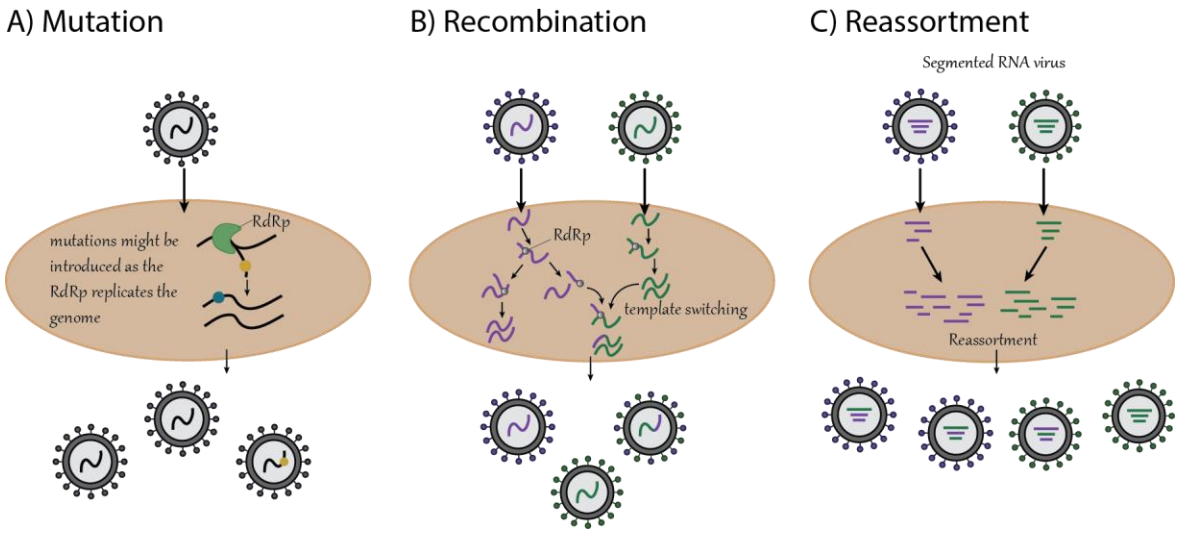


***Figure 2-2**: Mechanisms of RNA virus evolution. A) During replication, the RNA-dependent RNA polymerase (RdRp) might introduce mutations in the genome, which are represented as colored circles. Not all mutations will be present in the progeny (e.g., mutation(s) leading to structural incompatibilities between the capsid proteins during virus assembly). B) Co-infection of a cell by genetically distinct viral strains can result in the generation of recombinant viruses leading to novel mutation combinations, gene amplification or defective genomes (not represented here). C) Co-infection of a cell with different virus genotype can lead to a progeny with a different combination of segments. This progeny is known as reassortants. Panels B and C were adapted from Simon-Loriere and Holmes, 2011.*

2.2.3.1 Mutation

A mutation is a change in the nucleotide sequence. During replication of the genetic information, the DNA or RNA polymerase can incorporate a non-complementary nucleotide, resulting in a point mutation (Figure 2-2A). Mutations that result in a different amino acid are called non-synonymous, while those that do not result in amino acid changes are synonymous (73). In addition, mutations can result in the insertion or the deletion of one or several nucleotides, collectively known as indels. Since in protein-coding nucleotide sequences, three nucleotides (termed a codon) will encode an amino acid, unless the length of the insertion or deletion is a multiple of three, it will alter the reading frame. The result, very likely, will be an entirely different protein (73).

RNA viruses have important features (74):

1. high mutation rates (i.e., the rate at which mutations arise in the genome), with estimates that span between 10^{-4} to 10^{-6} substitutions per nucleotide per generation. Such mutation rates are

largely the consequence of the error-prone RNA-dependent RNA polymerase (RdRp), which, in most cases, lacks proofreading activity, and it can introduce mutations during the replication of the genome

2. short and compact genomes, typically around 10 Kb, with some exceptions, such as the members of the *Coronaviridae* family, whose genomes are roughly 30 Kb in length
3. short generation times and large population sizes.

Each new mutation might have a different effect on the replicative success, or fitness, of the viral population. Mutations that confer a negative effect on viral fitness are known as deleterious mutations, while those with a positive effect are beneficial mutations. In addition, mutations conferring no (or minimal) effect on the fitness of the viral population are called neutral mutations (75). As already mentioned, the fate of every mutation will be determined by natural selection or genetic drift. When the population size is sufficiently large, natural selection is expected to account for the changes in variants frequencies: deleterious mutations will likely be removed through negative or purifying selection, and beneficial mutations may become fixed in the population through positive selection. In this way, natural selection has a deterministic impact on the evolution of the viral population (75). When the size of the viral population is significantly reduced, for example during inter-host transmission (transmission bottlenecks) or strong selective swaps, genetic drift – that is, the stochastic fluctuation of variants frequency in the population – might have an important role. Under such circumstances, deleterious and neutral mutations can be fixed (71, 74).

Experimental studies have shown that most mutations arising during replication are deleterious (76, 77). One possible explanation is their compact genomes, often dense with protein-coding regions and functional RNA structures, and sometimes gene overlapping, leading to structural and functional constraints (78).

Intra-host populations can be very large in RNA viruses. For instance, for HIV-1, it is estimated that nearly 10^{12} virions are produced in a single individual every day (79). As a consequence of such large population sizes and high mutation rates, RNA viruses typically exist within each host as genetically diverse populations or mutant swarms (75).

Although deleterious mutations are commonplace in RNA viruses, the selection of mutations that confer an adaptive advantage can occur, as observed during several outbreaks. For example, it has been proposed that an AA change in the envelope glycoprotein of CHIKV (E1-A226V) augmented the capacity of CHIKV to infect and be transmitted by *Aedes albopictus* (80). Similarly, a single mutation on the glycoprotein of EBOV (GP-A82V) has been proposed to increase infectivity in human cells (81). More recently, throughout the ongoing COVID-19 pandemic, we have observed the emergence of SARS-CoV-2 genetic changes, leading to new SARS-CoV-2 variants. An example of these genetic changes is the D614G mutation in the spike glycoprotein of SARS-CoV-2, the first mutation during the COVID-19 pandemic to gain researchers' attention. It was first detected in March 2020, and it appeared several times independently in the global SARS-CoV-2 population, suggesting convergent evolution and an adaptive benefit of this mutation (82). However, later genomic investigations revealed the presence of this mutation in viruses collected from China in late January, suggesting that it could have resulted from "founder effect" events (83). Later, several studies showed that the D614G substitution confers an advantage for transmissibility (84-86). In early September 2020 in England, a new lineage emerged; it received the name of B.1.1.7 or, more

recently, the Alpha variant (87). The Alpha variant spread rapidly, outcompeting the circulating variants. Seventeen lineage-defining mutations characterized this lineage. Eight of these mutations are located in the spike glycoprotein, including the N501Y substitution in the receptor-binding domain (83). Interestingly, lineages that emerged later in different parts of the world, precisely the lineage B.1.351 and P.1, also known as the Beta and Gamma variants, respectively, harbor the N501Y mutation. Again, convergent evolution suggests an adaptive benefit of this mutation which was further supported by studies showing that the N501Y mutation increases binding affinity to the human cellular receptor Angiotensin-converting enzyme 2 (ACE2) (88) and cell infectivity in mice models (89, 90). A more detailed description of the SARS-CoV-2 variants will be addressed in Chapter 3.

Furthermore, today, next-generation sequencing tools are revolutionizing the study of viral populations. This is because these tools allow the sequencing of viruses at a high depth of genome coverage, resulting in a comprehensive characterization of the full variant repertoire within the viral population, including low frequency variants. Identifying these low frequency variants is relevant for understanding how viral populations evolve within each host and how the immune host response or treatments (e.g., antivirals, convalescent plasma therapy) may impact in the composition of the viral population. Finally, studies of within-host viral population diversity could provide some insights into how within-host processes relate to the evolution of the virus on a larger scale (72).

2.2.3.2 Recombination

Error-prone viral polymerases can also generate recombination during replication. In RNA viruses, the most common recombination mechanism is copy-choice, in which the RdRp releases the template strand (donor strand) while retaining the nascent transcript and then continues RNA synthesis by associating to another template (acceptor strand) or another place on the same template (Figure 2-2B). Viral recombination thus corresponds to the generation of viral RNA from at least two different templates. Hence co-infection is required (71). Several factors can favor this process, the most important one being the sequence homology between the donor and the acceptor strand, and this might result in hot spots of recombination along the genome. This form of recombination can be further classified according to its final product: homologous and non-homologous. In homologous recombination, the process of template switching occurs between regions of sequence homology leading to a complete phylogeny strand. In contrast, in non-homologous recombination, the template jumping occurs in a different genomic region, leading to the synthesis of a progeny with either an additional or missing sequence (74). Additionally, non-replicative recombination has been described (91); however, this mechanism is less often compared to copy-choice in RNA viruses.

Recombination occurs at different rates within RNA viruses (92). It frequently occurs in retroviruses like HIV with an estimated recombination rate of $1.38x10^{-4}$ and $1.4x10^{-5}$ per site per generation. For enteroviruses (EVs), recombination is also a major mechanism of evolution (93, 94). EVs are positive single-stranded RNA viruses which belong to the *Picornaviridae* family. More than 100 types of EVs infect humans and are classified into four genotypes (A-D) (95). In Europe, different genotypes and serotypes of EV co-circulate and replace one another in annual cycles, often leading to the emergence of new genotypes through recombination (96). We will delve into the mechanism of EV recombination in Chapter 1. Recombination is also frequent in viruses from the *Coronaviridae* family For instance, recombination events are reported for SARS-CoV-2. In recent work, Jackson et al. scanned 279 thousand genomes

collected from the UK, assigned to the Alpha variant but did not contain the full set of lineage-defining mutations. By doing so, they identified 16 recombinant sequences (97).

Recombinant variants might have different properties (e.g., pathogenicity) hence the importance of their surveillance. For example, recombination between members of the *Coronaviridae* family led to the emergence of a new coronavirus in turkeys. This recombinant virus resulted from the recombination between infectious bronchitis virus (IBV), a chicken-specific coronavirus, and another coronavirus. Analysis showed that the recombination event led to the replacement of the spike gene of IBV with the sequence of the other coronavirus. This is a remarkable example as it is one of the few documented cases in which recombination seemed to have resulted in cross-species transmission and subsequently viral emergence (98).

2.2.3.3 Reassortment

Reassortment constitutes an important evolutionary mechanism of segmented RNA viruses, these are viruses that maintain their genomes as several RNA molecules (99). In such viruses, when co-infection with different genotypes occurs, the progeny could inherit segments from both parental viruses. We refer to such progeny as reassortants and they can yield new phenotypes (99) (Figure 2-2C).

Reassortment is a mechanism of genetic diversity which can result in numerous combinations of genotypes. For example, a virus with eight segments could potentially generate 256 reassortants. This tremendous genomic novelty could have dramatic consequences for the emergence of infectious diseases. One of the most prominent examples is the antigenic shift in influenza viruses. The influenza virus genome consists of 8 single-stranded RNA segments coding for ten essential proteins and several accessory proteins. The eight gene segments are numbered from one to eight or named after the main protein they encode (Table 1) (100).

***Table 1**: IAV gene segments and main gene products. It is adapted from Gerber et al.,2014.*

Segment	**Main gene products**
1	PB2: Polymerase basic 2
2	PB1: Polymerase basic 1
3	PA Polymeric acid
4	HA: Hemagglutinin
5	NP: Nucleoprotein
6	NA: Neuraminidase
7	M1: Matrix 1, and M2: Matrix 2
8	NS1: Non-structural gene 1, and NS2: Non-structural gene 2

In particular, the influenza A virus (IAV) has been classified into different subtypes based on the two surface proteins, HA and NA. Eighteen subtypes of HA (H1 to H18) and nine NA subtypes (N1 to N9) have

been recognized so far. While a limited number of HA and NA subtypes have been isolated in humans, all have been found in aquatic bird populations, the natural reservoir of IAV (100).

The segmented nature of the IAV genome allows for exchanging segments between different influenza viruses during co-infection in a host. Indeed, in the 20th century, three IAVs caused major pandemics: the 1918 H1N1 virus, the 1957 H2N2 virus, and the 1968 H3N2 virus (101, 102). More recently, the 2009 H1N1 virus (different from the 1918 H1N1 virus) led to another IAV pandemic (103). Briefly, all these pandemics resulted from the introduction and successful spread of a novel HA subtype in the human population from an animal source, leading to antigenic shift. For instance, the 1957 H2N2 virus resulted from the reassortment between circulating in the human population IAV H1N1 and avian IAV H2N2. Similarly, in 1968 H3N2 virus emerged due to the reassortment of human IAV H2N2 with avian IAV H3N2 (101). The 2009 H1N1 virus detected was a reassortant of avian, porcine and human viruses (109).

Two different hypotheses may explain the mechanism of emergence of such IAVs. First, the avian IAV was introduced in the human population, which then reassorted with human IAVs. Second, both the avian and the human IAV infected and reassorted in an unknown mammal (e.g., pigs), and subsequently, this reassortant was transmitted to humans (104).

In sum, given their high mutation rates and recombination and reassortment mechanisms, large population sizes and short generation times, RNA viruses can evolve rapidly, quickly adapting to changing environments. This rapid evolution constitutes a challenge for their control and likely contribute to their emergence risk.

Key points of Introduction part 1: Emerging and re-emerging infectious diseases

- Infectious diseases have occurred periodically throughout history; however, today, the increasing interconnectivity of our world together with demographic and ecological changes have been changing the dynamics and the potential risk of infectious diseases. For example, due to the higher human mobility, pathogens can spread faster and wider, leading to the introduction of pathogens to new host populations.

- Major components in the emergence of infectious disease include: human population growth and expansion, encroachment of natural habitats, climate change and the increased global connectivity.

- Emerging and re-emerging infectious diseases are major threats to public health and a large proportion of them are caused by RNA viruses.

- The rapid evolution of RNA viruses and their adaptive potential likely contribute to their challenging control.

3 INTRODUCTION PART 2: Using viral genomic epidemiology to strengthen public health responses to outbreaks

3.1 Introduction to genomic epidemiology

The emergence of an outbreak brings several questions: what is the causative pathogen? What is its mode of transmission? Where did it come from? How many introductions to the human population have there been? Since when has the pathogen been circulating in the population? What are the drivers of its emergence? Is the outbreak linked to any other outbreak? These (and many more questions) are crucial to better understand the pathogen underpinning the outbreak and develop strategies to mitigate and control the disease (105).

Before the development of next-generation sequencing (NGS) technologies, most of these questions were answered using epidemiological data (i.e., case data) through the inference of key epidemic parameters (105). These epidemiological parameters such as the doubling time (i.e., the time that takes a population to double in size), basic reproductive number R_0 (i.e., the expected number of secondary cases generated in the population where all individuals are susceptible to infection), the effective reproductive number Rt (i.e., the expected number of secondary cases generated in the current state of the population, not necessarily all susceptible to infection) (106) and the incubation period (i.e., the time period from infection to symptom onset of an infected person) (107) are important quantities to inform about the progression of the disease and the impact of control policies. However, the epidemiological data to estimate such parameters might not always be available (105). In addition, when different disease-mitigating interventions are being applied to different populations, a higher resolution may be needed to support interventions, as is the case for the ongoing COVID-19 pandemic (108). Therefore, there are major benefits of integrating epidemiological data with pathogen genomic data and contact-tracing data.

Viral genomic data has been used for epidemiological investigations for decades. However, today, the amount and the speed of genomic data generation have increased, and in parallel, new mathematical and computational tools for their analysis are being developed. As a result, using pathogen genomic data to complement epidemiological analysis increases its popularity (109, 110). Indeed, the science of using genomics and associated analysis to investigate infectious diseases is now referred to as “genomic epidemiology” (105).

3.2 Identification of the viral agent causing the disease

The notification of several patients sharing a pattern of symptoms can reveal the beginning of an outbreak. At this point, the most important task is to identify the etiologic agent. There are several methods for

pathogen identification, including molecular tools such as PCR (polymerase chain reaction), which allows the detection of the genetic material of the pathogen, serology methods such as ELISA (enzyme-linked immunosorbent assay), which allow the detection of viral proteins or antibodies developed against the infectious pathogen on clinical samples (e.g., serum). Although still of great importance, one limitation of these methods is that they require previous knowledge of a given pathogen. An important improvement in NGS technologies has been the advent of metagenomic next-generation sequencing (mNGS) from clinical samples through which all genomic information is sequenced in an untargeted manner (111, 112). In this way, the major advantage of using mNGS is the possibility to identify viral, bacterial, fungal, and any other eukaryotic pathogen directly from the infected individuals without prior knowledge of the etiologic agent. For example, metagenomic sequencing from serum and tissue samples enabled the identification of a novel member of the *Arenaviridae* family, Lujo virus, as the virus responsible for an outbreak in South Africa in 2008 (113). Similarly, unbiased metagenomic sequencing allowed the identification of a novel coronavirus, which today is known as SARS-CoV-2, from samples from patients with pneumonia in China in 2019 (2). In addition, with its untargeted approach, mNGS has the power to detect novel drug-resistant variants form of the pathogen, which is crucial to guide treatments of HIV or Hepatitis C virus (HCV) (111) as well as new recombinant forms of the virus (114). Due to its potential and as they become more affordable and cost-effective, NGS platforms are moving from the research centers to routine use in clinical microbiology laboratories (115).

Nevertheless, there are several challenges when using mNGS. For example, common technical problems during NGS are the low amount or quality of genetic material in the samples, contamination with background RNA/DNA, and cross-contamination of samples. We will expand on the challenges of using mNGS for disease diagnosis in Chapter 1.

The sequencing platforms moved to clinical laboratories and, on some occasions, to the epicenter of disease outbreaks (115). This has been possible as sequencing technologies have become more affordable and mobile, exemplified by MinION (Oxford Nanopore Technologies), introduced in 2014 (116). MinION is a handheld DNA/RNA sequencer that can be easily transported, and unlike most other sequencing platforms, it has the advantage of yielding long-read sequences (116). For example, the MinION has been used in tent laboratories set up at the epicenter of the disease outbreak during the Ebola epidemic (117, 118) and the ZIKV outbreak in Brazil in 2016 (119). It has even been taken to the International Space Center (120). However, in-field metagenomics faces even more challenges than clinical metagenomics. To the technical problems already mentioned, we add an increased likelihood of cross-contamination of samples by not having completely separate environments for sample manipulation and library preparation and, on some occasions, low computer power, poor internet connection, and difficulties in collecting samples metadata (115). Furthermore, despite the possibilities provided by the MinION, it comes at the expense of sequencing accuracy and thus the need for sequencing genomes to high coverage (116). Nonetheless, with the continuous improvement of the sequencing platform, this challenge will likely be overcome shortly.

Once the virus has been identified, this information can be used to answer the following critical questions: do we have diagnostic tools to monitor the outbreak? Are there treatments available to mitigate the disease? The generated genomic data can also be used to reconstruct phylogenetic trees, which can provide an additional level of detail, revealing the origin of the virus and its connection to previous outbreaks, as will be explained in more detail in the following sections.

3.3 Phylodynamics as a framework for outbreak analysis

The term "phylodynamics" was first coined in 2004 by Grenfell et al., referring to the "melding of immunodynamics, epidemiology, and evolutionary biology" to study patterns of viral genetic variation (121). Phylodynamics has become a widely used statistical framework for a better understanding of infectious disease transmission and evolution by extracting the evolutionary and epidemiological information from pathogen genomes (122).

This framework can be applied to study the epidemiology and evolution of pathogens such as RNA viruses because they are rapidly evolving pathogens that accumulate mutations almost in real-time as they spread in the population (105, 115, 122). Phylodynamics relies on two main pillars to recover the evolutionary and epidemiological information from mutations that accumulate in the genomes: (i) phylogenetic inference as the primary analytical tool – that is, the reconstruction of the evolutionary relationships among nucleotide or amino acid sequences, and (ii) the integration of additional data (e.g., sampling time, sampling location) (122).

Several examples illustrate the utility of phylogenetic inference to recover information about the evolutionary and epidemiological processes from mutations that accumulate on viral genomes. For instance, the first phylogenetic analysis of the IAV H1N1 strain during the 2009 pandemic quickly showed that the genomic segments were closely related to the ones already detected in swine suggesting spillover from pigs to humans (103). In a similar way, phylogenetic analyses of Lassa fever virus (LASV), an endemic virus in West Africa, revealed that LASV infections are mostly the results of multiple independent rodent-to-human transmission rather than human-to-human transmission (123). This is in contrast to what was observed in the EBOV outbreak in West Africa. Genomic analyses suggested that EBOV infections during 2014 were the result of a single transmission event from the reservoir followed by human-to-human transmission (124).

The field of phylodynamics includes a wide variety of methods. The following sections briefly describe some of them while providing examples to illustrate their utility.

3.3.1 Phylogenetic reconstruction

A phylogenetic tree or phylogeny is a diagram that shows the evolutionary relationships between different species, organisms, or genes. The resulting branching pattern is often called the topology of the tree, and it indicates how sequences are related to each other. The external or terminal nodes represent the existing taxa (e.g., viral sequences), while the internal nodes represent hypothetical progenitors of the previous ones (125).

There are different methods for phylogenetic reconstruction, and they fall into two categories according to the type of data they use: distance-based and character-based methods (126). Distance methods like neighbor-joining (NJ) calculate the genetic distance between every pair of sequences from a given alignment and use the resulting matrix distance to construct a tree. The most popular character-based

methods are Maximum parsimony (MP), Maximum likelihood (ML), and Bayesian inference. In these methods, each sequence position in the aligned sequences is a "character," and the nucleotides (or amino acids) at that position are the "states" (73). Here is a brief introduction to these methods.

MP follows the principle "plurality should not be posited without necessity" from the philosopher William of Ockham. In short, parsimony methods aim to find the tree topology for a set of aligned sequences that can be explained with the fewest number of character changes (i.e., nucleotide changes). In this way, MP methods will determine the amount of character change for each tree topology. The tree that yields the minimum number of changes is selected as the MP tree. However, as it calculates every possible tree topology, this method might be inefficient for large datasets.

ML is similar to the MP method in that it evaluates different tree topologies and estimates the relative support by summing over all sequence positions. But it differs in that it relies on an explicit model of sequence evolution and the likelihood function (126). The use of an explicit model of evolution allows ML methods to incorporate well-known features of sequence evolution such as different rates between character states (e.g., different rates between transitions and transversions) and different rates of changes among sites (e.g., higher rates at the third codon position). The likelihood function is the conditional probability to observe the data (D) given a hypothesis including a tree topology (*t*) and parameters of the evolutionary model (α), as described by Equation 1:

Equation 1 (73) $$L(t,\theta) = Pr(D|t,\alpha)$$

ML algorithms will search for a tree topology (*t*) and parameters of the evolutionary model (α) that maximizes the probability of observing the data. The α̃ and ẗ are the values of α and *t*, maximizing the likelihood function:

Equation 2 (73) $$\tilde{\alpha},\ddot{t} = arg.\max L(t,\alpha)$$

Based on the search algorithm, the likelihood of each tree is estimated, and the tree topology with the highest likelihood is the ML tree (73, 126).

Bayesian methods also employ an explicit model of sequence evolution and the likelihood function but it differs from ML methods as they do not search for the single best tree. Instead, Bayesian methods compute a distribution over the parameter space (α and *t*), called the posterior probability density function, and they search for a possible set of trees and values parameters for the given data. Bayesian methods require a prior belief or knowledge about α and *t*, which is formalized as a prior probability distribution (73).

The posterior probability can be derived using the Bayes' theorem considering: the likelihood of observing the data D given a tree topology (*t*) and parameters of the evolutionary model (α), f(D|α,t); our prior knowledge or beliefs about α and *t*, f(α,t), that is the prior probability distribution; and, the evidence of the data, *f*(D). The following equation describes the Bayes' theorem:

Equation 3 $$f(\alpha, t|D) = \frac{f(D|\alpha,t) * f(\alpha,t)}{f(D)}$$

Calculating the posterior probability distribution can be very difficult. A solution to this problem would be to estimate it by drawing random samples. However, this is unlikely to work. The reason is that the posterior probability distribution that we are trying to estimate (the one yielding the tree topologies that best explain the data and the prior beliefs) is likely concentrated in a limited part of the vast parameter space. Therefore, the likelihood of sampling the best tree at random is very low (73). This brings the following question: how can one obtain random samples from the actual posterior probability distribution that one wants to estimate? One possible answer is: using Markov chain Monte Carlo (MCMC) sampling (73, 127). MCMC couples two methods: The Monte Carlo and Markov chain. The Monte Carlo approach is a sampling technique that allows estimating a distribution by taking random samples from the distribution. For example, if we would like to estimate the mean height of Ph.D. students at the Institut Pasteur, a Monte Carlo approach would draw a large number of random samples and calculate the sample mean of those instead of directly calculating it from the equations of the distribution. The Markov chain part of the MCMC is the notion that random samples are obtained by a special sequential process. In this way, the Monte Carlo part of the name refers to the sampling purpose, whereas the Markov Chain refers to how we obtain these samples.

Briefly, a Markov chain is a stochastic model describing a sequence of possible events, and is used to estimate the probability distribution among states after a large number of steps. Markov chain has two important properties: (i) the probability of transitioning to any state depends only on the current state (hence the chain), and (ii) regardless of the starting point, the chain will converge towards an equilibrium state ("stationary state").

Therefore, the idea behind the MCMC method is to construct a Markov chain that contains in each state all the parameters of the model, and that has as stationary distribution the posterior probability distribution of the parameters that we want to calculate. This can be achieved using different methods such as the Metropolis–Hastings method (73, 127).

The MCMC initiates by simulating a random set of parameters values (α and *t*). In the next step of the chain, it proposes a "new state" by making small changes to the values of the parameters. In each step, the outcome is evaluated, and the parameter values are either accepted or rejected. If accepted, it becomes the next sample in the MCMC chain. If rejected, the next sample in the chain is a copy of the previous sample. After n number of states is hoped that the chain has converged such that the stochastic algorithm samples from the posterior probability distribution. Those early states in which the chain has not converged are discarded as they do not follow the posterior distribution, and they are usually referred to as "burn-in" (73, 127).

For the tree reconstruction, once we have sampled tree topologies from the posterior distribution using the MCMC method, the sampled trees can be summarized according to one of the several existing methods (128). One of the most widely used approaches is the maximum clade credibility (MCC) tree. Using this approach, each clade within the tree is given a score based on the number of times it appears in the sampled posterior trees. The product of all these clades' scores is taken as the tree score. The tree with the maximum product of the posterior clade probabilities is therefore the MCC tree (128).

3.4 Dating the phylogenetic history of the outbreak

Extracting information from genetic data through phylogenetic inference is the bedrock of phylodynamics, and this method can be further exploited by the integration of additional data. Bayesian inference is particularly attractive for phylodynamics as it can easily integrate different types such as temporal or geographic data. Mr. Bayes (129), BEAST (130) and BEAST2 (131) are three important and widely used software packages for Bayesian phylogenetic analysis. This introduction will be focusing on Bayesian analysis implemented in BEAST.

Incorporating sampling time (i.e., when genomes were sampled) into phylogenetic analysis is a common and important approach (122). It allows calibration of phylogenies in calendar time. By doing this the phylogenetic tree is transformed into a time-structured phylogeny resulting in the following changes: (i) branch lengths are converted from divergence units to time units, (ii) the positions of terminal nodes (or branch tips) correspond to sampling times, and (iii) internal nodes are placed at the most likely time of divergence.

We use molecular clock methods to obtain a time-calibrated tree. In short, molecular clock methods enable calculating the time of divergence between two sequences. Importantly, this calculation depends on the molecular clock hypothesis assumed (122). Indeed, there are different clock models that can be used. The first one to have been developed was the strict clock which assumes nucleotide substitutions (or amino acid substitution when using protein sequences) occur at a constant rate among different sequences (132). In other words, that every branch in a phylogenetic tree evolves according to the same evolutionary rate. However, this model can be too restrictive in some scenarios (133), stimulating the development of more relaxed models. One that has gained popularity is the uncorrelated relaxed clock. In short, the uncorrelated relaxed clock allows each branch of a phylogenetic tree to have its own evolutionary rate with no correlation between neighboring branches (134). The different clock models are reviewed in (135).

The molecular clock can be calibrated using the collection date of the samples. In the same way as we use distance and time to calculate the speed of a car, sequence divergence and time can be used to calculate the evolutionary rate. Using the molecular clock to calculate the evolutionary rate serves two primary purposes. First, to describe the evolutionary process and compare it with previous outbreaks, and second, to date the phylogenetic history (122). Regarding the latter, estimating the evolutionary rate gives an estimation date for an individual branching event on the tree or for the last common ancestor at the root of the tree, which may help resolve the origins of an outbreak. Several examples have illustrated this. For example, the first cases of ZIKV in the Americas were initially detected in Brazil in May 2015; however, phylodynamics analysis suggested that the virus was introduced between May and December 2013, long before the first documented cases in Brazil (59).

Before estimating the evolutionary rate, it is crucial to study the "*clockliness*" of the data. It means to study if the rate at which mutations are accumulated in the genomes is relatively constant, such that it follows a molecular clock of evolutionary change (136). Checking this is important because if the data was collected over a short period or early in an outbreak, we might not have captured a measurable amount of evolutionary change, yielding a poor correlation between sequence divergence and time (137). Additionally, the presence of recombination events could also invalidate the assumptions of the molecular clock (138). If any of these situations is the case, estimates from the molecular clock would be inadequate,

and it is advisable to use the evolutionary rate from a previous outbreak or a related virus in these cases (105, 137).

3.5 Phylodynamics approaches to better understand viral transmission dynamics

At an early stage of a viral outbreak, an important piece of information is the rate at which the virus spreads in the population. As noted in the introduction of this section, R_0 and R_t are two epidemiological parameters used to measure the transmissibility of a given pathogen. These parameters are usually estimated from cumulative incidence data through epidemiological analyses. However, knowing the number of cases during an epidemic might not always be the case, particularly during the initial stages of an outbreak when R_0 might be measured most accurately. Due to the relevance of this information in understanding and controlling the outbreak, a central objective in phylodynamics is to quantify dynamic population processes using genome information (122). Indeed, Bayesian phylogenetic methods are commonly applied to viruses to infer epidemiological processes from genetic data offering an alternative solution. This is possible because, as mentioned before, RNA viruses accumulate genetic changes as the outbreak unfolds. Therefore, specific processes such as population bottlenecks, the rapid expansion of the viral population, or selective sweeps might leave a footprint on the genetic structure of the viral population. Such footprint will be reflected in the topology of the resulting phylogeny, and it could be used to estimate changes in the population size (i.e., number of infected individuals) (139). For instance, Figure 3-1 illustrates idealized scenarios of phylogenies showing the effect of changes in the viral population size. Viruses spreading rapidly in the population will probably produce a phylogenetic tree with external branches relatively longer than the internal ones (Figure 3-1A). This is because quickly expanding viruses are more likely to share a common ancestor when the population is small. On the contrary, a virus population that stays constant in size over time will result in a phylogeny with external branches shorter than the internal ones (Figure 3-1B) (139).

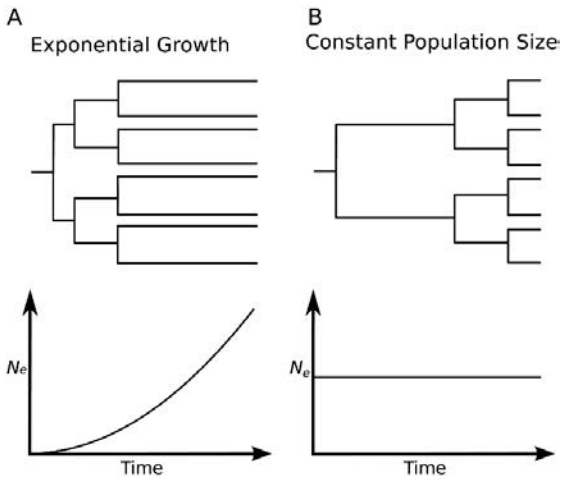


Figure 3-1: *Representation of phylogenies showing the effect of changes in viral population size in the topology of the tree. Extracted from Volz et al. (2013).*

The interest to investigate how these transmission dynamics and population history impact on viral genetic variation led to the development of models that relate patterns of evolution with population genetic approaches. These are known as "tree-generative" models and can be implemented as tree prior in Bayesian inference (122). Example of tree prior models are the exponential growth and constant population size models (represented in Figure 3-1) and the widely used Bayesian Skyline or Bayesian Skygrid. A complete description about these models and their implementation can be found at (https://beast.community/tree_priors). It is important to note that all these models primarily differ in the prior that they put on the effective population size. Due to the joint inference of the phylogeny and the model for the population dynamics, the choice of the tree prior will not only influence the estimated population size but also the phylogeny, in particular the tMRCA. Hence, the resulting phylogeny will strongly depend on the accuracy of the tree prior (140).

Birth-death models are attractive "tree-generative" models as they allow the estimation of R_0. Such calculation is achieved by linking the shape of the tree to the rate at which lineages are added to the tree (birth rate , λ) and the rate at which lineages are removed from the tree (death rate, δ) (141). In this way, R_0, can be estimated using the following the equation:

Equation 4

$$R_0 = \frac{\lambda}{\delta}$$

In the context of epidemiology, λ and δ would correspond to the transmission rate (rate at which infected individuals infect susceptible individuals) and becoming non-infectious rate (rate at which infected individuals recover or die) respectively (141).

There are other methods based on classic susceptible-infected-recovered (SIR) compartmental models which also allow the estimation of R_0 directly from viral sequence data (142).

As an example of application of these methods, the birth-death model was used to investigate the dynamics of the HCV outbreak in Egypt (143). Egypt is the country most affected by hepatitis C virus (HCV) infection with an overall prevalence of 10%, one-fold more than the world population. It is believed this major spread of HCV was driven by campaigns of parenteral antischistosomal therapy (PAT), where millions of people received intravenous injections to fight schistosomiasis (144). Stadler et al. analyzed HCV genomes collected from 63 patients in 1993 and observed an increase in R_t, being larger than one around the 1920s, coinciding with the start of the PAT campaign, peaking in 1960 and then decreasing around the 1970s when PAT was changed to oral therapy (144).

Similarly, this methods have been also used during the early spread of EBOV particularly in Sierra Leone in 2014. Interestingly, the authors used a range of phylodynamic approaches based on birth-death models to estimate the R_0. Overall the authors managed to estimate an R_0 based on sequencing data alone ranging from 1.65 to 2.18, showing that despite efforts to curb the outbreak, it was not enough to reduce the R_0 of the virus (145).

3.6 Reconstructing dispersal history and dynamics of the outbreak

Incorporating spatial data into phylogenetic inference has significant benefits as it allows for a better understanding of viral spread. The approach is based on applying an idea mentioned in the previous section: viruses accumulate genetic variation almost in real-time as the outbreak unfolds. Therefore, as the outbreak progresses, viruses will spread, and this spatial separation may lead to genomic divergences, leaving a genetic footprint on the viral population. Thus, viral genomes constitute an important source of information about the underlying processes shaping the spread of the outbreak. Figure 3-2 illustrates the concept of how spatial spread and therefore spatial heterogeneity might be reflected in the genetic structure. It shows two different phylogenies: panel A represents viruses with strong spatial structure, in contrast to panel B. Viruses that circulate in similar hosts (e.g., humans), or within the same region (e.g., continent or country) are expected to be more closely genetically related as transmission will occur more often between them (139). However, this might not be the case for all viruses which is a reflection of the global interconnectivity.

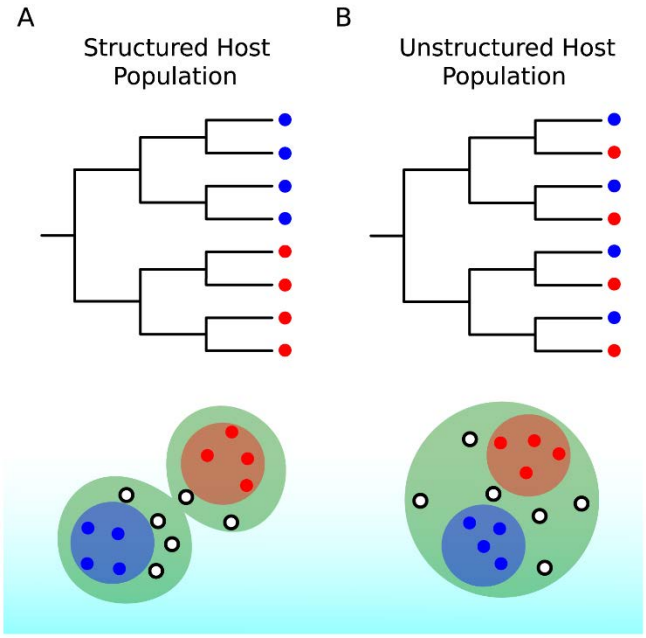


Figure 3-2: *Representation of phylogenies showing the effect of population structure in the topology of the tree. Blue and red circles represent geographic locations from which the samples were collected. White circles represent inferred geographic locations for the internal nodes. Image was extracted from Volz et al. (2013).*

This interaction between genome sequence evolution and geographical dispersal can be studied using phylogeographic approaches, which have developed into a research field referred to as phylogeography (146). These methods rely on assigning a geographical location to unsampled ancestral viruses through a process called ancestral state reconstruction. In this way, each branch of the tree is an independent trajectory of viral movement with a start location, end location, and duration, and the phylogenetic tree, a collection of viral movements which are phylogenetically related. Therefore, if time and location

sampling is known for all the genomes, spatial dynamics of the epidemic can be drawn from these movements (147).

Through the integration of geographic data, phylogeographic methods have two major purposes. First, to reconstruct the complete spatial history and patterns of virus spread (146), and second, to evaluate the impact of external factors on the dispersal history and dynamics of viral spread, which could be used to control and prevent the spread of the ongoing and future outbreaks (148).

3.6.1 Discrete and continuous diffusion models for reconstructing viral spatial spread

The reconstruction of the spatial spread of the virus from genetic data, treat sampling location as a discrete or continuous trait and as an intrinsic property of the collected virus. The information about how the virus spreads is represented by changes in the trait along the branches of the tree. This "trait evolution" approach has the advantage of inferring the location of common ancestors based on the observed locations of the sampled viruses (10, 146). Unlike in other methods, sequences and ancestral states are simultaneously inferred in Bayesian inference frameworks, implying that both genomic and spatial data will impact the phylogeny (147).

Treating locations as discrete or continuous traits in a phylogeographic diffusion model primarily relies on the sampling pattern and the biological question. Discrete diffusion may be preferred when viruses are sampled from different countries or within a space that can be subject to discretization under realistic criteria. The continuous model might be the best choice when working with viruses sampled over a continuous space, e.g., province or district level, or where administrative borders do not confer a reason for discretization. If the sampling space can be considered a discrete or continuous space, the choice should refer to the biological question (Figure 3-3) (146). Concerning the latter, it is necessary to understand the nature of the two models, which determines the information that we can obtain from their implementation.

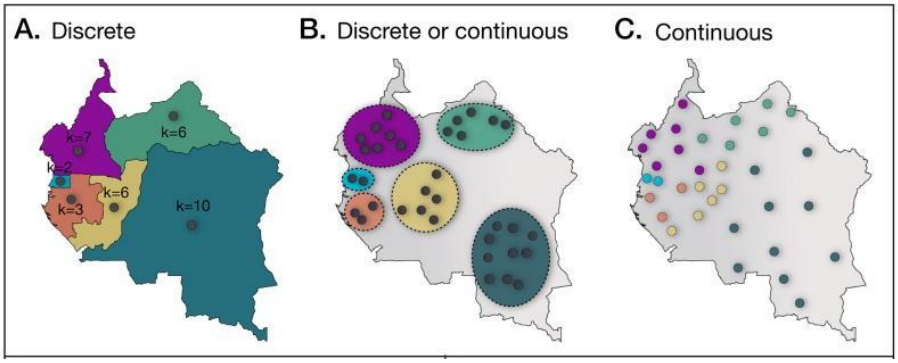


Figure 3-3: *Discrete and continuous phylogeographic approaches. A) Represents the scenario where the sampling scheme is amenable to discretization. B) Represents an intermediate scenario where the sampling scheme is amenable to discretization or not, depending on the biological question. C) Represents the scenario where the sampling has done over a continuous space, where borders do not confer a reason for discretization. Image extracted from Faria et al., 2011.*

In discrete Bayesian phylogeographic approaches, transitions of the virus from one location to another are modeled using continuous-time Markov Chain (CTMCs). In short, considering the transitions between three different states: A, B, C representing changes of three sampling location along a tree (Figure 3-4 A), all the possible transitions will be represented by transition rate matrix Λ, describing the rate at which transitions between states occur (Figure 3-4 B). The dimension of the matrix will be K x K, with K being the total number of states (or sampling locations). For more details about CTMC please refer to (149).

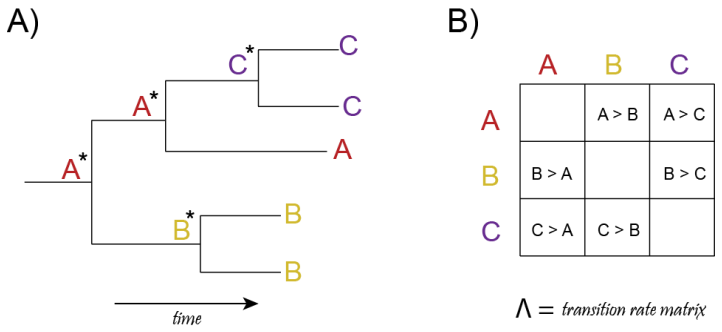


Figure 3-4: *Virus transition location along the phylogeny are modeled as CTMCs in discrete diffusion model. A) Represents a three-state CTMC path. B) Represents the transition rate matrix Λ describing all the possible transitions of the three-state CTMC. C) Based on the sampled locations, the CTMC will infer the location for the internal nodes (here highlighted with an asterisk).*

The transitions among locations can be further described as symmetrical or asymmetrical transitions. In contrast to the symmetrical model, the asymmetrical allows jumps from A to B and B to A to have different rates providing a more realistic scenario. When studying the spatial diffusion process in a "real" context, the space state-trait is not usually reduced to three states as in the example given above but is rather large. Not all transitions are expected to be informative; indeed, it is expected that most transitions will rarely, if ever, occur. Hence, many transitions rates are likely to be zero or close to zero. For this reason, Bayesian discrete phylogeographic frameworks are extended with a stochastic search variable selection (BSSVS) procedure that allows the use of a Bayes Factor (BF) test to identify non-zero transition rates. In this procedure, each transition (*i*) is modeled with an associated coefficient β_i, estimating its contribution, and a binary indicator δ_i indicating if the transition is included in the model (150).

Subsequently, the BF test will assess the strength of the support for a particular transition by comparing the posterior to the prior odds that the transitions are non-zero given by the following formula (150):

Equation 5

$$BF_{(i)} = \frac{posterior\ odds_{(i)}}{prior\ odds_{(i)}} = \frac{\frac{p_i}{1-p_i}}{\frac{q_i}{1-q_i}}$$

where, q_i and p_i define the prior and the posterior probability that the rate *i* is non-zero, respectively. The value of p_i can be extracted from the transition rate matrix obtained after running the analysis.

The discrete model has two main disadvantages. First, the locations of all ancestors are restricted to the sampled locations. Second, discretization of sampling locations may imply an arbitrary or unrealistic grouping of sampling locations leading to an oversimplified abstraction or unrealistic space subdivision.

Bayesian continuous phylogeographic approaches reconstruct viral dispersal and infer ancestral locations of internal nodes using a relaxed random walk (RRW) diffusion process that adapts to the standard Brownian diffusion process (151). In the RRW model, the space is explored in two directions (i.e., latitude and longitude coordinates), allowing ancestral locations to be located at any point in the continuous space. Furthermore, the spatial movement is not assumed to be homogenous over the entire phylogeny; it allows each branch to have its own diffusion rate. In addition to reconstructing the spatiotemporal dispersion of the virus, continuous diffusion models enable the calculation of summary statistics of spatial spread such as dispersal velocity, diffusion coefficients, or evolution of the maximal wavefront distance. All those statistics help describe an outbreak's dynamics further and compare the mode and rate of spatial spread among different outbreaks (152).

In conclusion, the main difference between the two models is in the inference of the likely ancestral location of the internal nodes. While the discrete model infers ancestral locations from the sampled locations in abstraction from a geographical perspective, the continuous model provides a more realistic representation of the diffusion process by allowing the virus to exist in any location within the continuous space, restricted to the assumptions of the model (i.e., diffusion rates) (151).

For this reason, in general, discrete phylogeographic models are preferred for hypothesis testing about an epidemic origin or epidemiological linkage between locations. For instance, during the CHIKV epidemic in Italy, the index case was reported to have hosted a relative from India, information which was concomitantly validated by phylogeographic analysis (153). Similarly, the phylogeographic analysis indicated that the ZIKV responsible for the 2015-2016 outbreak in Cape Verde was likely introduced from northeast Brazil between June 2014 and August 2015 (154). However, although the implementation of CTMCs to discrete models results in parameter estimates, the realization of the process remains unobserved. For this reason, complementary approaches have been developed that allow the calculation of the total number of transitions along the phylogenetic branches (jumps) (155). Recently, this approach has been applied to assess location-transition histories of SARS-CoV-2 from 44 different locations incorporating travel history data (156).

Unlike its discrete counterpart, continuous phylogeographic models are mostly used to characterize the dispersal history and estimate the spatial spread statistics. For example, it was used to characterize the dispersal history and dynamics of YFV during the outbreak in Brazil in 2017, which revealed an initial circulation of the virus among non-human primates and viral dissemination toward densely populated YFV-free areas followed by spillover to humans and an increase in the number of human infections (157). More recently, a continuous phylogeographic model was used to address the spatiotemporal spread of SARS-CoV-2 in Brazil, which allowed the authors to show that during the first phase of the epidemic the virus spread was mostly local and was reduced by the implementation of schools and stores' closure. This was followed by a second phase of long-distance movement events (57).

It is important to note that although phylogeographic reconstructions have proven to be very informative, estimated ancestral locations can be highly uncertain, in particular when the sampling is limited. For this reason, these approaches succeed to be more informative when associated with a thorough sampling process (122, 137, 158). It is also very important to take into account possible sampling bias. An unbalanced sampling between regions will bias the results towards overrepresented locations in the data. This sampling bias issue has been shown for discrete and continuous methods. For instance, in discrete analyses, the overrepresentation of a country in the dataset will probably lead to a higher inference of this location in the internal nodes. Therefore, using this data to infer the origin of a viral lineage could lead to the wrong conclusion that the overrepresented country is seeding more locations than in reality. The reverse might also be true: the underrepresentation of countries will lead to a lower inference of internal nodes for these locations. Finally, sampling bias can affect conclusions when comparing the importance of repeated introductions versus the local spread. If the country of interest is oversampled compared to the rest of the potential source locations, this could lead to estimate smaller number of introductions. In continuous analyses, sampling bias can also incorrectly estimate the "true" origin location of the root and the diffusion rate of the virus (122, 158).

3.6.2 Predicting social and environmental factors driving the outbreak

As we mentioned at the beginning of this section, apart from the spatiotemporal reconstruction of the viral dispersal history, phylogeographic approaches yield yet another important goal: to evaluate social and environmental factors driving the outbreak spread, which could ultimately contribute to better surveillance and control measures (148). This is often achieved by evaluating the virus genetic diversity with geographical and additional data such as human mobility, climate, or vector abundance. As is the case for location, many of these additional data can be represented or treated as a finite number of discrete values. This can be implemented in Bayesian discrete diffusion models, such as the one implemented in BEAST that incorporates additional data (i.e., vector abundance) as covariates or predictors of the transition rates, allowing to evaluate the impact of such variables in the diffusion process (159). This can be done by implementing a generalized linear model (GLM), in which rates are a linear log function of various potential predictors. Descriptions and mathematical details on the GLM model can be found in Text S1 from Lemey et al., 2014 (160).

GLM approach has been applied, for example, to reconstruct the spatiotemporal spread of the EBOV during the 2013-2016 outbreak in West Africa (161). This study consisted of a phylodynamic analysis using more than 1500 genomes collected from 56 different administrative regions with a GLM fitted to the discrete trait model to evaluate the association of 25 predictors with the viral movement among the administrative regions. The study revealed a positive effect of the origin and destination population sizes and an inverse effect of the geographic distance on the spread of the virus. Additionally, this study showed that EBOV dispersal occurred mainly within each country rather than international dispersal, suggesting that the international borders limited virus spread (161).

Also, there are post hoc analyses for continuous phylogeography that investigate the impact of environmental factors on the dispersal of the virus. Such analyses can be implemented using the R package

Seraphim (162). For example, it was used to investigate the impact of environmental factors on the dispersal of WNV in North America from 1999 to 2016 (163). This study found that among the environmental factors tested, temperature was the main predictor of viral dispersal with a tendency to disperse faster in areas with higher temperatures.

In sum, at the core, phylodynamic methods rely on the fact that epidemiological and evolutionary processes underlying viral outbreaks occur on the same temporal scale. Consequently, genomic viral sequences collected from individuals (often) differ enough to reconstruct their phylogeny. This genetic variation can be read as the trace that the combination of epidemiological, evolutionary, and immunological processes leave in their genomes as the outbreak unfolds and viruses spread in the population. Therefore, analyses that reconstruct viral phylogeny and the dynamics of an epidemic in light of additional data (e.g., epidemiological, spatial data) can uncover important information about the processes driving the outbreak that would otherwise remain hidden. Such information includes relevant epidemiological parameters (e.g., R_0), timing and origin of the outbreak, estimated number of introductions that seeded the outbreak, or environmental and social factors fueling the viral dispersal (e.g., air travel or vector density).

3.7 Examples of genomic epidemiology studies shaping intervention strategies in response to an outbreak

Viral genomic data coupled with sufficient sampling, metadata, and appropriate statistical and bioinformatics framework can complement traditional surveillance approaches, providing robust information to control and mitigate the outbreak (164). Critically, this is not always the case, mostly because it is not always possible to acquire and analyze genomes in real-time. Nevertheless, there are several examples where we have observed genomic epidemiology in action informing public-health decisions.

For instance, in the 2013–2016 Ebola epidemic in West Africa, whole-genome virus sequencing was used to reconstruct transmission chains. It subsequently contributed to confirm that a person could sexually transmit the disease even a year after becoming infected with EBOV, providing also evidence of persistent infection of EBOV in semen. Consequently, male survivors were recommended to have their semen tested for persistent EBOV infection (165, 166). Thanks to this, several studies later showed viral persistence infection in more than 50% of male survivors with a maximum duration of persistence in semen up to 500-700 days after being discharged from the Ebola treatment units.

Seven years later, from February to June 2021, Guinea faced a new EBOV outbreak (167). Genomic characterization and phylogenetic analysis showed that the collected viruses fell within the EBOV lineage from the 2013–2016 outbreak suggesting that this new outbreak was not the result of a new spillover event from the virus reservoir. This information was crucial because it confirmed that molecular-diagnostic tools (e.g., PCR) and available therapeutics could be immediately applied to mitigate and control the epidemic. Furthermore, it opened up a new perspective on EBOV reemergence mechanisms. Indeed, the low genetic divergence between 2013 and 2016, and 2021 viruses is compatible with continuous slow replication among humans or a long phase of latency where the virus could have persisted at a low level in survivors from the previous outbreak. This led to the hypothesis of different scenarios of EBOV

transmission to the index case: (i) sexual transmission by exposure to contaminated semen, (ii) contact with the body fluids from a survivor who had relapsed, (iii) relapsed of the index case (although index case did not experience any previous symptomatic EBOV infection), and (iv) the index case was not the index case but part of a small previously unknown chain of human-to-human transmission (167). This study emphasized the risk of EBOV resurgence and the need to survey EBOV-infected survivors to monitor possible reactivation or relapse of EBOV infection and concomitant virus spread.

During the 2016-2017 YFV outbreak in Brazil, YFV genomes have been used to show that human infections result from continuous and direct sylvatic spillover (*Haemagogus* mosquitoes– NHP) rather than urban transmission (*Aedes aegypti* – Human). As the sylvatic cycle involves different mosquito species than urban, this represented critical information about vector control strategies (157).

Likewise, ZIKV genomes sequenced from humans and mosquitoes in Florida showed that the outbreak was fueled by multiple introductions of the virus from the Caribbean rather than local vector-borne transmission, highlighting that traveler education and surveillance are important to reduce future outbreaks (168).

Finally, complete genome sequencing allows to monitor genetic changes in the viral population over time, a piece of crucial information for the design of effective diagnostics and therapeutics. Vaccines, for instance, are an essential line of defense against seasonal influenza, and they are playing a major role in our effort to control the COVID-19 pandemic. Genome sequencing coupled with appropriate bioinformatics analyses provides powerful tools to analyze the virus evolution in real-time, allowing the regular update of vaccines as is the case for the influenza virus vaccine (169). Similarly, in the context of the COVID-19 pandemic, real-time sequencing of SARS-CoV-2 has allowed scientists to track the emergence of the new variants throughout the pandemic, with the Omicron variant being the most recent VOC designated by WHO at the time of writing this thesis (170). Preliminary studies have already shown that the vaccine's effectiveness against Omicron is significantly reduced compared to the Delta variant. For this reason, booster vaccine campaigns have become highly recommended around the world to protect the population against the new variant, and companies such as Pfizer–BioNTech are working on a variant-specific vaccine for Omicron, hoping to induce a high level of protection against the Omicron variant and to prolong immune protection compared to the current vaccine. In addition, Pfizer–BioNTech has already started clinical trials with other variant-specific vaccines to protect against Alpha, Beta, Delta, and simultaneously Alpha/Delta variants (170).

These studies represent clear examples of the contribution of genomic epidemiology to developing public health strategies to mitigate and control an outbreak.

Key points of Introduction part 2: Using viral genomic epidemiology to strengthen public health responses to outbreaks

- Genomic epidemiology has a great potential to inform about a disease outbreak and to collaborate with the public health response, particularly if genomic data can be acquired and analyzed in real-time.

- Viruses, in particular RNA viruses, are rapidly evolving pathogens that accumulate mutations as they spread in the population.

- Certain processes such as expansion or bottlenecks of the viral population, selective sweeps or spatial separation can leave footprints in the genetic structure of the viral population. Extracting this information from pathogen genomes through phylogenetic inference is the bedrock of phylodynamics

- Phylogeography combines genomic and geographic data to reconstruct the dispersal history of the virus.

- Genomic pathogen data can complement "traditional" epidemiological studies to understand better the pathogen underpinning the outbreak and ultimately develop strategies to mitigate and control the disease.

4 OBJECTIVES

Over the last years, large-scale outbreaks caused by RNA viruses such as EBOV, CHIKV, or SARS-CoV-2 have greatly burdened public health and economic stability. Nonetheless, today, thanks to the growing popularity of NGS technologies and the progress in bioinformatics and phylodynamic methods, pathogen genomic information can be used to re-construct outbreak dynamics and contribute to the response to emerging infectious diseases.

This thesis aims to dive into the origin, timing and spread of viral infectious diseases in the context of past and ongoing outbreaks using genomic epidemiology. In particular, it is divided into three different parts addressing specific questions driven by recent outbreaks.

In the first project, in collaboration with the "Instituto de Salud Carlos III" and the "Instituto Maimónides de Investigación Biomédica de Córdoba" from Spain, we aimed at investigating known and unknown meningitis cases from Southern Spain during 2015-2019. More specifically, we used mNGS: (i) to characterize pathogens previously identified via standard laboratory diagnostics, and (ii) to explore the presence of RNA viruses known to cause meningitis in those cases lacking microbiological diagnosis.

The second project aimed to study the emergence and spread of CHIKV in Cambodia during the 2011-2013 period and, after almost a decade of absence, in 2020; two studies in collaboration with Institut Pasteur du Cambodge. In addition, during the 2011-2013 outbreak, multiple cases of encephalitis were detected by our colleagues. Therefore, to gain insights into the difference in disease outcomes upon CHIKV infection, we also aimed to sequence in parallel serum and CSF samples from infected patients with classic chikungunya symptoms or neurological affliction.

In the third project, we addressed different aspects of the epidemiology and evolution of SARS-CoV-2. In particular, we performed phylogenetic studies to investigate the introduction and early spread of SARS-CoV-2 in the northern regions of France. Additionally, we aim at monitoring the long-term intra-host evolution of SARS-CoV-2 in an immunocompromised patient treated with convalescent plasma therapy.

To answer all these questions, we followed the workflow depicted in Figure 4-1.

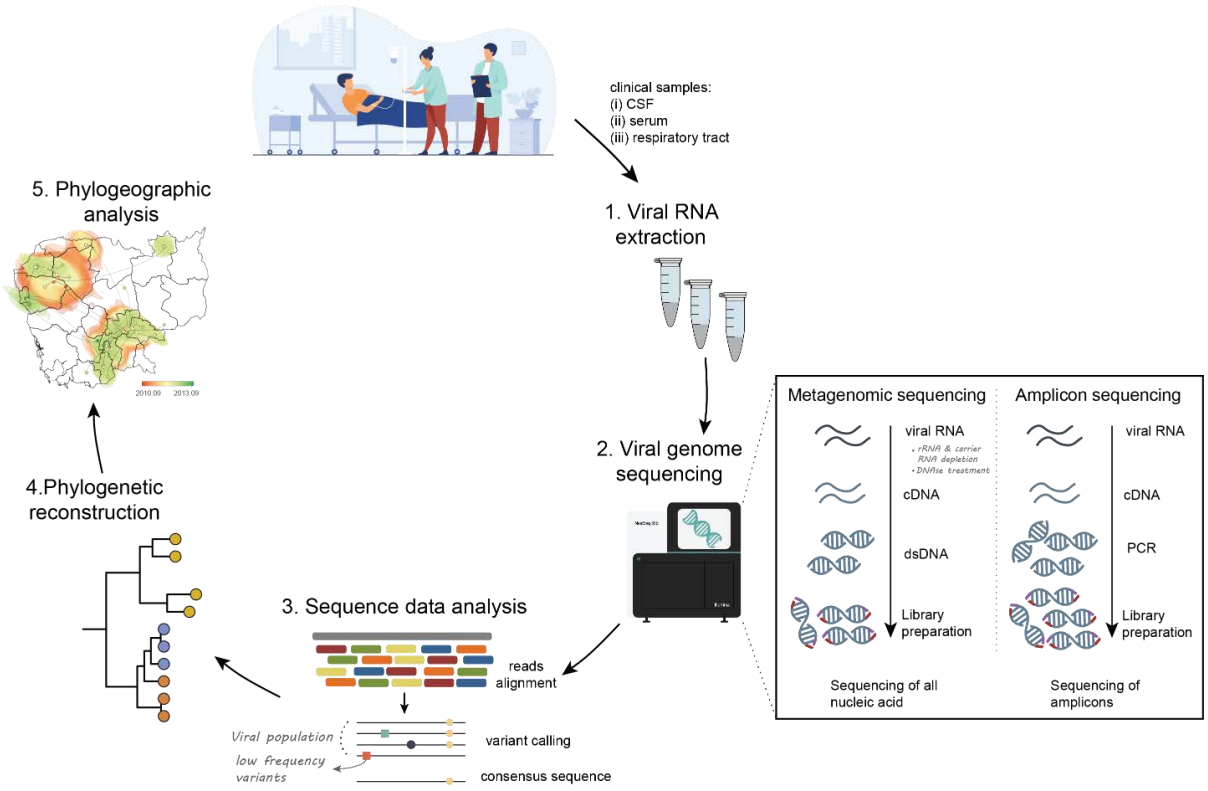


***Figure 4-1**: Schematic representation of the methods implemented in the thesis, which are one of the common threads that link the three chapters of this thesis. Together these methods aim to identify the virus responsible for the disease, generate complete viral genomes from clinical samples, assess viral genetic diversity, including low frequencies variants (when applicable), and when the data made it possible, to implement phylodynamic and phylogeographic approaches.*

We sequenced viral genomes from clinical samples (CSF, sera, or respiratory samples) using two approaches. Initially, we pursued an untargeted metagenomic sequencing, following the outlines previously described by Matranga et al. (171). This method uses a selective depletion step to remove unwanted carrier RNA (generally added during the RNA extraction to increase the yield) and host ribosomal RNA, enriching in proportion the sample with viral material. Despite the major improvements of this protocol to generate virus genomes from clinical and biological samples, it might fail to generate complete viral genomes from low-concentrated or low-quality samples. We designed an amplicon-based sequencing approach to overcome this challenge, following the protocol described by Quick et al. (172).

Next, using in-house bioinformatics pipelines, we proceed to generate the consensus sequence of the virus and, when applicable, to analyze the viral population, including low frequency variants.

Armed with our viral genomes, we implemented different genomic tools to shed light on the epidemiology and evolution of these RNA viruses. In particular, we inferred phylogenetic trees, a natural means to illustrate and study the evolutionary relationship among our viral genomes and closely related sequences found in public databases, such as Virus Pathogen Resource (ViPR) (173). When available, we integrated genomic with temporal and geographic data to reconstruct time-calibrated trees and perform phylogeographic analysis.

5 CHAPTER 1: Using mNGS to identify and characterize potential RNA virus causing meningitis

The following pages present our work concerning the study of known and unknown meningitis cases collected from Southern Spain between 2015 and 2019. The chapter starts by providing a short introduction to the use of mNGS to diagnose neurological infections such as meningitis. It continues outlining the details of our work, and finally, the chapter finishes with a general discussion and conclusion about the study.

5.1 Metagenomic sequencing for the diagnosis of neurological infections

As already discussed in the main introduction of the thesis, one of the most significant use cases of NGS technology is metagenomic sequencing in clinical samples, through which total DNA or RNA material in the sample can be sequenced in an untargeted manner (111, 112). Given the continuous enrichment of microbial genomic databases through pathogen discovery studies, mNGS allows for identification of potentially any pathogen (excluding prions) present in the sample, including novel recombinant forms or emerging drug-resistant genetic variants (112). Being such a powerful, mNGS has already helped diagnose several infectious diseases with severe symptoms such as meningitis and encephalitis. In these cases, samples from the CSF or brain tissue obtained by biopsy have been sequenced allowing the identification of viruses (e.g., WNV, CHIKV), bacteria (e.g., *Salmonella enterica*) or fungi (e.g., *Candida dubliniensis*) as potential etiologic agent (174-177).

Neurological infections are a significant cause of morbidity and mortality worldwide, with 10.6 million cases of viral or bacterial meningitis alone in 2017 (178). Despite this high prevalence, the etiologic agent responsible for meningitis often remains unknown (179, 180), reaching 81% in one study (181). Indeed, several difficulties are associated with diagnosing patients with such neurological infections. First, meningitis and encephalitis are caused by various pathogens, including viruses, bacteria, fungi, and parasites (182). Studies have estimated that over one hundred viral and bacterial species can lead to meningitis, most of which are viruses (180, 183). For instance, among the viruses causing neurological infections, EVs are recognized as the most common cause of meningitis. Several members of *Enterovirus* genus have been implicated, such as EV-A71, EV-D68, and coxsackievirus B (182, 184). Numerous viruses belonging to the *Herpesviridae* family are commonly implicated in the development of meningitis or encephalitis: namely, Herpes Simplex virus type 1 and type 2, Varicella zoster virus, Epstein-Barr virus, and Cytomegalovirus (184). Within the *Paramyxoviridae* family, infections with members such as Mumps virus (MuV) have been reported to cause meningitis (185). In addition, the infection with arboviruses can lead to meningitis and other neurological disorders. These arboviruses include members of several families: *Flaviviridae* (e.g., WNV, Japanese encephalitis virus, and DENV), *Togaviridae* (e.g., Eastern equine encephalitis virus and CHIKV), and *Phenuiviridae* (e.g., Toscana virus). Second, specific-pathogen diagnostic assays such as PCR may fail to detect the pathogen in case of genetic divergence. Lastly, several emerging and re-emerging pathogens might present new neurological manifestations, for instance, ZIKV. During the

outbreak in Brazil in 2015, ZIKV was, for the first time, associated with several cases of microcephaly and Guillain–Barré syndrome (59).

However, some of the challenges in diagnosing neurological infections can be overcome using mNGS. For example, mNGS allows the identification of several different pathogens in a single assay, thereby saving the need for having a broad panel of pathogen-specific tests to diagnose neurological infections. Furthermore, its unbiased nature can identify novel pathogens to a specific region or population, highly divergent from previously known pathogens or presenting atypical neurological manifestations (186).

5.1.1 Viral meningitis in Spain

There are three major groups of viruses responsible for most meningitis cases in Spain. EV occupies the first place, given that they are the primary cause of meningitis and other neurological manifestations, with most studies reporting on EV-A71 (187, 188), EV-D68 (189) and EV-B (e.g., Echovirus 9 and 30) (190, 191). These are followed by arboviruses such as WNV or Toscana virus (TOSV) (192). In particular, TOSV has gained much attention given that since its identification, it has been associated with sporadic but increasing numbers of meningitis and encephalomeningitis cases in southern Spain (193). Lastly, lymphocytic choriomeningitis virus (LCMV) is a rodent-borne virus, and it has been associated with several sporadic cases of meningitis (192).

In the following two subsections, a brief introduction will be provided about the biology and epidemiology of EV and TOSV, the two viruses on which our work, described in the next section of this chapter, has focused.

5.1.2 Enteroviruses

EVs are non-enveloped viruses with a positive-sense single-stranded RNA genome of approximately 7,500 nucleotides long (194). Their genome contains a single open reading frame that encodes a polyprotein that is subsequently subdivided into three other segments: P1, P2, and P3. The P1 region encodes the structural proteins (VP1, VP2, VP3 and VP4), while the P2 and P3 regions encode the non-structural proteins associated with replication (194).

EVs belong to the Enterovirus genus within the *Picornaviridae* family, and those infecting humans have been assigned to four species: A (EV-A) to D (EV-D) (195). Numerous EVs types have been characterized by phylogenetic clustering; EV-A: 25 types, EV-B: 63 types, EV-C: 23 types, and EV-D: 5 types (196).

Beyond the name EV, old naming conventions are still used, which terms EVs as coxsackieviruses A or B (CAV or CBV) or echoviruses (E) according to their biological properties rather than their genetic relationship; therefore, these viruses are found distributed across the four EV species. In addition, numerous serotypes have been assigned within each group, which is named by consecutive numbering (e.g., EV-A71, EV-D68, CAV-16, E11, E30) (195).

While millions of people get infected with EV every year, most cases remain subclinical. Nevertheless, symptomatic infections show a broad spectrum of clinical manifestations, from fever to severe respiratory or/and neurological diseases, including meningitis, encephalitis, paralysis, and myocarditis (197, 198).

EVs show seasonal incidence patterns in temperate and tropical regions with a higher number of cases during summer and early autumn (198). In particular, in Europe, EVs are more commonly detected in late summer and autumn (199). Several different EV serotypes may circulate at different frequencies during a season, and their prevalence may change from season to season, with different EV serotypes replacing each other. For instance, before 2008 in Spain, the most prevalent EV was E30, replaced the following year by E6 (200). Between 2016 and 2019, E30 and E6 were the most prevalent EV types in Spain, together with E7, E9, and EV-A71, to name a few (201). Furthermore, certain EVs have been associated with recent outbreaks, resulting in significant morbidity and mortality, and for this considered emerging pathogens (197). For example, EV-71 has been responsible for a significant outbreak of hand, foot and mouth disease with neurological affliction in Asia. The EV-D68 caused a large outbreak associated with respiratory and neurological disease in children in North America in 2014 (202), and the wave of cases continued in 2016 and 2018 (203). More recently, an increasing number of EV-D68 infections has been reported in Europe, peaking in September 2021 (204).

As noted in the main introduction, recombination is an important mechanism of evolution for EV (93, 94). During epidemics, recombinant EVs forms are frequently observed, especially those forms arising from the genetic exchange between viral strains within the same serotype (96, 205, 206). The high frequency of recombination among EVs might be due to two main reasons: (i) the extensive co-circulation of different genotypes and serotypes of EV in a specific geographic region, and (ii) the possibility of human co-infections. Regarding the first, in Europe, several studies have reported the co-circulation of multiple EV-D68(207), EV-A71 strains (206), or serotypes A, B and C (208). Regarding the latter, human co-infections with different EV strains have been detected (209).

Studies analyzing the genetic diversity of circulating EVs have revealed junctions between the 5'UTR and the structural region and junctions between the structural and non-structural regions as potential recombination hotspots (94). While EVs with chimeric protein capsids have been reported (e.g., as a result of the recombination between CV-B3 and CV-B4 (210)), intertypic recombination occurs mostly outside the structural region (the P1 region), particularly throughout the P2 region (211). The low frequency of recombination within the capsid region can be primarily explained by structural constraints between the capsid proteins from different EV types during virus assembly or receptor binding (94).

These observations were later on confirmed by experimental recombination studies, which identified six recombination hotspots within the EV genome. The first three are located in the 5'UTR and the other three in the non-structural region (P2 and P3 regions), at the junction between two viral genes: VP1-2A, 2A-2B, and 2C-3A (Figure 5-1) (94). In this way, while structural regions (P1) appear to be cold-spots, P2 and P3 regions are hotspots for recombination. This observation has led to the concept of separate and modular evolution of structural and non-structural regions of the EV genome, where in contrast to the capsid-encoding regions, non-structural regions are subject to frequent recombination (93, 94, 212).

These recombinant events have the potential to generate recombinant variants which might have different properties (i.e., pathogenicity) than those of the parental strains. This, together with the increasing incidence of non-polio EVs and their ability to cause severe respiratory and neurological disease calls for strong genomic surveillance programs to monitor and control these viruses.

Despite this, at the moment, non-polio EVs are not included in the global virus surveillance guidelines by WHO (213, 214). Before 2020, EV surveillance relied on voluntary and laboratory-based systems, where laboratories would report EV detection to reference centers. Although, in theory, these programs would provide relevant information about the prevalence and the disease burden of circulating EVs, this type of surveillance is biased (214). The reason is that the most reported cases are severe, requiring hospitalization; therefore, only a small proportion of all cases are diagnosed (198, 214). Recently, a surveillance network was established to improve EV surveillance in Europe: the European Non-Polio Enterovirus Network (215). Through the collaboration of public health institutions, national reference laboratories and research centers, this network aims to establish standardized surveillance for respiratory and neurological infections caused by non-polio EV. More efforts like this one would be needed to monitor the co-circulation of EVs as well as the emergence of novel recombinant forms.

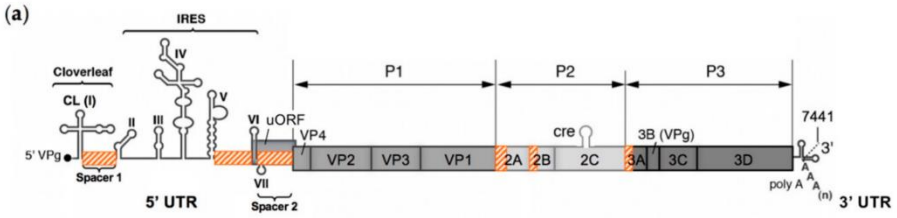
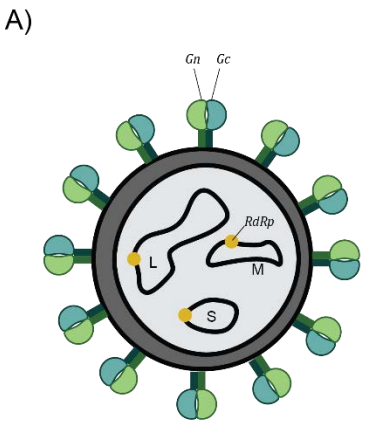


Figure 5-1: *Schematic representation of the EV genome and the recombination hotspot which are indicated with the hatched orange rectangles. Image extracted from Muslin et al. 2019.*

5.1.3 Toscana virus

TOSV is a negative-sense and segmented RNA virus which belongs to the *Phlebovirus* genus within the *Bunyaviridae* family. TOSV is an arbovirus transmitted to humans mainly through the bites of an infected sandfly of the genus *Phlebotomus,* which is widely distributed in the Mediterranean Sea (216). Like other bunyaviruses, the TOSV genome consists of three segments termed L, M, and S. The S segment, about 1900 nucleotides long, uses an ambisense strategy to encode non-structural proteins (N and NS). The M segment has approximately 4200 nucleotides in length and encodes a polyprotein processed by the host protease to generate glycoproteins Gn and Gc and non-structural proteins. The longest segment, the L segment, is about 6400 nucleotides long, and it encodes the viral polymerase (Figure 5-2).



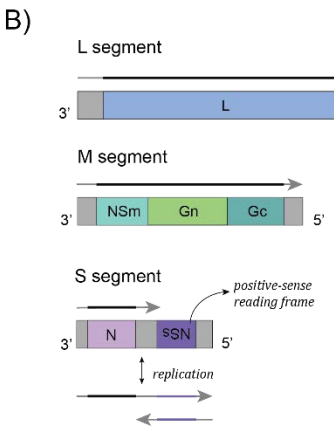


Figure 5-2: *TOSV structure and genome organization. A) TOSV particles contain three negative-sense single stranded RNAs, each one associated with the RdRp (yellow). The envelope of the virion contains heterodimers of the glycoproteins Gn abd Gc. B).The segments L and M are purely negative-sense while the segment S ambisense. Image adapted from ViralZone, Phlebovirus.*

TOSV was initially identified in Italy in 1971 from the sandfly *Phlebotomus perniciosus* in central Italy. In the past few years, the geographic distribution of TOSV has been expanding across the Mediterranean basin, with cases reported in several countries, including Algeria, Croatia, Cyprus, France, Italy, Malta, Morocco, Portugal, Spain, Tunisia, and Turkey (216, 217). The expansion of TOSV is supported by an increasing number of human cases and the detection of the virus in sandflies trapped in the wild (218, 219) (Figure 5-3).

Genetic studies have initially identified two main lineages: lineage A and B. While lineage A circulates in Algeria and Tunisia, lineage B circulates in Portugal, Spain, France, Morocco, and Croatia (216). More recently, a third lineage has been proposed, lineage C, which appears to circulate in Croatia (220) and Greece (221, 222). However, at the moment, there are only a few partial sequences belonging to the lineage C (two partial sequences of the L segment and one partial sequence of the M segment of approximately 200 and 500 nucleotides long, respectively), and the virus has not yet been isolated nor characterized (216). The co-circulation of lineages has been detected in different countries: lineages A and B in Spain and Turkey, and lineages B and C in Croatia (Figure 5-3).

While TOSV causes asymptomatic infection in most cases, symptomatic infections impose a significant burden on individuals. Symptoms generally are fever, intense headache, vomiting, and more severe clinical presentation, including acute aseptic meningitis, encephalitis, and meningoencephalitis (216). Although rare, life-threatening and fatal diseases might occur (223, 224).

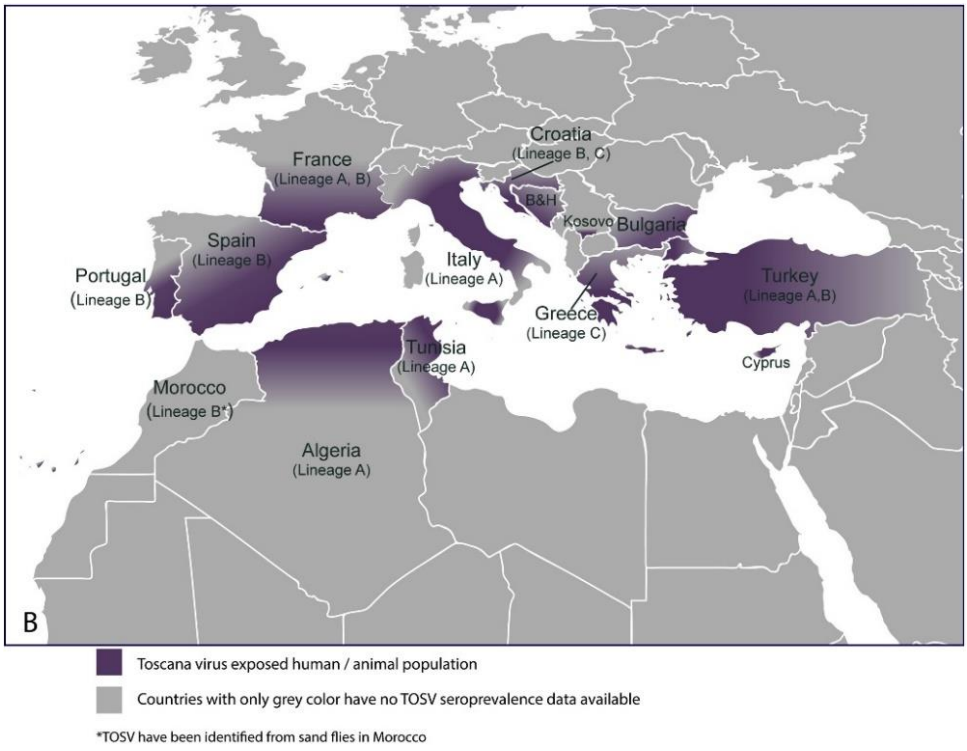


Figure 5-3: *Region of the Mediterranean basin in which TOSV has been detected in either human or animal population. Image extracted from Ayhan and Charrel 2020.*

5.2 Metagenomic sequencing of known and unknown meningitis cases in Southern Spain, 2015–2018

In the present study, we collaborated with the Enterovirus and Viral Gastroenteritis Unit from the "Instituto San Carlos III" and the "Instituto Maimónides de Investigación Biomédica de Córdoba" in Spain. As a part of a meningitis surveillance program, our colleagues received CSF samples from meningitis cases collected from the Southern region of Spain, the region of Andalucía, between 2015 and 2018. All these samples were tested for enteroviruses, herpesvirus types 1 and 2, varicella zoster by RT-qPCR, and the presence of bacterial pathogens (e.g., *Neisseria meningitidis*, *Streptococcus pneumoniae)* by culture. Additionally, two samples were tested by RT-qPCR for MuV and TOSV. While most samples were negative, several EVs and one MuV were detected.

In this context, the goal of the present work was (i) to detect in the CSF samples from patients with idiopathic meningitis the presence of RNA viruses known to cause meningitis, providing a possible explanation for the disease outcome observed, and (ii) to generate full-length genomes of EV and MuV in the CSF samples in which the RT-qPCR test previously detected them. To achieve this, we planned to do metagenomic sequencing on the following selected sample cohort: positive controls – samples from patients with meningitis with the previously identified pathogen by RT-qPCR test (enteroviral meningitis n = 12 and mumps meningitis n=1); negative controls (n=8) – samples from patients with no meningitis but in whom another diagnosis was made; and idiopathic meningitis samples (n=23) – samples from patients with aseptic meningitis with an unknown etiologic agent.

The following study is being been prepared as:

Fabiana Gámbaro, Ana Belén Pérez, Matthieu Prot, Eduardo Agüera, Luis MartínezMartínez, Artem Baidliuk1, Maria Paz Sanchez-Seco, Ana Vazquez, Maria Dolores Fernandez-Garcia, Etienne Simon-Loriere**. Untargeted metagenomic sequencing identifies Toscana virus in aseptic meningitis cases from Southern Spain between 2015 and 2019.**

The genomic characterization and phylogenetic analysis of the EVs generated in this study has been published as follows:

Fabiana Gámbaro, Ana Belén Pérez., Eduardo Agüera, Matthieu Prot., Luis MartínezMartínez, Maria Cabrerizo, Etienne Simon-Loriere, Maria Dolores Fernandez-Garcia (2021). **Genomic surveillance of enterovirus associated with aseptic meningitis cases in Southern Spain, 2015–2018**. Scientific reports

Supplementary information associated to this chapter can be found in the following link

Untargeted metagenomic sequencing identifies Toscana virus in patients with idiopathic meningitis, Southern Spain, 2015 - 2019

Fabiana Gámbaro[1,2], Ana Belén Pérez[3,4], Matthieu Prot[1], Eduardo Agüera[3,4], Luis Martínez-Martínez[3,4], Artem Baidliuk[1], Ana Vazquez(?), Maria Dolores Fernandez-Garcia[5]*, Etienne Simon-Loriere[1]*

[1] Institut Pasteur, Université Paris Cité, G5 Evolutionary Genomics of RNA Viruses, 75015 Paris, France

[2] Université Paris Cité, Paris, France

[3] Hospital Universitario Reina Sofía, Córdoba, Spain

[4] Instituto Maimónides de Investigación Biomédica de Córdoba (IMIBIC), Córdoba, Spain

[5] National Centre for Microbiology, Instituto de Salud Carlos III, Madrid, Spain

* These authors jointly supervised this work

ABSRACT

Viruses are the leading cause of meningitis; however, the diverse landscape of possible causes of meningitis poses a significant challenge for pathogen-specific diagnostic assays to identify the etiologic agents, which remain unknown in most cases. Metagenomic next-generation sequencing (mNGS) of cerebrospinal fluid (CSF) has the potential to detect for nearly all possible neurological infections and can identify novel or unexpected pathogens. Here we applied RNA mNGS on CSF samples from patients with idiopathic aseptic meningitis (n = 23) collected in Southern Spain between 2015 and 2019, and used identified neurologic infections (n = 13) and non-infectious (n = 8) cases as controls. We identified the Toscana virus (TOSV) in 8 idiopathic cases, and we developed an amplicon-based sequencing approach to help sequence low concentrated or partially degraded samples corresponding to TOSV genotype B, improving the detection and generation of genomic data. This study also highlights that patients with aseptic meningitis from Southern Spain or with travel history to areas where TOSV is known to circulate should be tested for the virus regardless of the history to insect bites referred by the patient.

INTRODUCTION

Neurological infections are a significant cause of morbidity and mortality worldwide, with 10.6 million viral or bacterial meningitis cases alone in 2017 (1). Despite this high prevalence, the etiologic agent responsible for meningitis remains unknown in a large proportion of cases (2, 3), reaching up to 81% in one study (4). Multiple factors likely contribute to explain this relatively high proportion.

First, various pathogens, including viruses, bacteria, fungi, and parasites, can cause meningitis and encephalitis (5). It has been estimated that over one hundred viral and bacterial species can lead to meningitis, and most of them are viruses (3, 6). For instance, among the viruses causing neurological infections, enteroviruses (EVs) are recognized as the most common cause of meningitis (5, 7). Numerous viruses belonging to the *Herpesviridae* family may cause infection of the CNS, leading to diverse neurological diseases, including meningitis (7). Within the *Paramyxoviridae* family, Mumps virus (MuV) is an old and well-known cause of meningitis cases (8-10). In addition, several arboviruses of different families cause meningitis and other neurological disorders. These viruses include members of several families: *Flaviviridae* (e.g., WNV, Japanese encephalitis virus, and DENV), *Togaviridae* (e.g., Eastern equine encephalitis virus and CHIKV), and *Phenuiviridae* (e.g., Toscana virus). Second, pathogen-specific diagnostic assays (molecular or antigenic) may fail to detect a given pathogen due to genetic divergence. Third, the poor degree of suspicion for some neurological infections with a viral etiology and lack of testing. Fourth, the possible presence of a novel pathogen. Lastly, several emerging and re-emerging pathogens could have new neurological manifestations. Such was the case of ZIKV during the outbreak in Brazil in 2015, in which ZIKV was associated with several cases of microcephaly and Guillain–Barré syndrome (11).

In particular in Spain, there are three major groups of viruses in Spain responsible for most meningitis cases. EVs are the major cause of meningitis, with most studies reporting on EV-A71 (12, 13), EV-D68 (14), and EV-B (e.g., Echovirus 9 and 30) (15, 16). These are followed by arboviruses such as WNV or Toscana virus (TOSV) (17). TOSV has gained much attention lately, given that, since its identification, it has been associated with sporadic but increasing numbers of meningitis and encephalomeningitis cases in Southern Spain (18). Lastly, lymphocytic choriomeningitis virus is a rodent-borne virus for which sporadic cases of meningitis have also been reported (17).

Metagenomic next-generation sequencing (mNGS) is a promising approach for diagnosing infectious diseases since it can overcome some of the challenges faced by conventional diagnostic techniques such as PCR. For instance, mNGS allows the identification of several different pathogens in a single assay, thereby saving the need for having a broad panel of pathogen-specific tests to diagnose neurological infections. Furthermore, due to its unbiased nature, mNGS can identify novel pathogens to a specific region or population, highly divergent from previously known pathogens, including new recombinant forms or presenting atypical neurological manifestations (19). In 2012, a study aimed to determine viruses causing CNS infections in Spain by conventional testing, concluded that a significant number of cases (43% meningitis, 60% meningoencephalitis and 72% encephalitis) remained with no etiological diagnosis (20).

This study applies mNGS to explore the presence of RNA viruses in CSF samples from patients with aseptic meningitis with unknown etiology in Southern Spain for which no cause has been found after routine clinical testing. We use RNA mNGS and amplicon-based sequencing to reconstruct multiple complete genomes of TOSV to determine phylogenetic relationships with already known TOSV genomes.

RESULTS

Patient characteristics

A total of 44 previously tested CSF clinical samples were collected for viral metagenomics analysis. The sample cohort included CSF samples from patients with: (i) known neurologic infections (n=13), (ii) idiopathic aseptic meningitis where no etiology was identified after conventional routine laboratory testing in the source hospital (n=23), and (iii) with no infection (n=8). All idiopathic samples were tested by clinicians for herpesvirus types 1, 2, 3, 4, 5 and 6 and enterovirus, 21 (91%) for bacterial pathogens (*Neisseria meningitidis*, *Streptococcus pneumoniae, Hemophilia influenza type b* and Listeria), 10 (43%) for Coxiella, Borrelia, Rickettsia and Brucella, 9 (39%) for HIV and 8 (35%) for Treponema pallidum (Table S1). Additionally, idiopathic sample LCR_1152 was tested by qPCR for TOSV. All patients were from Cordoba province, 59% male (mean age 41 years [range 14 - 95 years]). Patients from urban Córdoba city were slightly predominant (59%) compared to remaining cases living in villages (populations between 900-9000 inhabitants). Of the idiopathic patients, 11 (48%) referred information on insect bites in the medical record. Of these 11 patients, 2 referred explicitly insect bites in the last days. With regard to laboratory results, noninfectious-negative controls presented lower CSF WBC compared to the remaining patients (Figure 1). CSF from infectious-positive controls and idiopathic patients showed a high WBC count (>100mm^3) in 28 (78%) of these patients. In all of them, differential CSF count revealed lymphocyte predominance (mean 83%). Epidemiological, clinical details and routine laboratory testing performed in all selected samples are provided in Table S1. Details about the laboratory test used can be found in Table S2.

Metagenomic sequencing of cerebrospinal fluid

We constructed cDNA libraries from 44 CSF samples for metagenomic next-generation sequencing (mNGS) on the Illumina platform, resulting in an average depth of 13.9 million reads/sample (interquartile range: 9.9 to 16.5 million). There was no significant difference in the total number of reads obtained between the three types of samples (Figure 1).

In this work, our analysis aimed to investigate in idiopathic meningitis cases the presence of RNA viruses, particularly those known to cause this type of clinical manifestation, such as EV, MuV, and arboviruses.

Figure 1 summarizes the result of such an analysis. Following the PCR test, using mNGS, we correctly identified EV (n=12) and MuV (n=1) in the positive control samples. In the negative control samples, we did not detect the presence of any of the target viruses. We identified TOSV in 8 out of 23 idiopathic aseptic meningitis samples. In 7 of these samples, we managed to reconstruct from partial to almost complete TOSV genomes (between 40 – 97% of the genome covered). However, for sample LCR_1152, from which we obtained a very partial fragment of the S and L segments (Table 1 and Figure S1).
Interestingly, sample LCR_1152 was the only one tested by RT-qPCR for TOSV, and the result was negative. Therefore, to rule out possible cross-contamination with the TOSV-positive samples, we re-extracted RNA from sample LCR_1152 alone and repeated the experiment. The re-sequencing of this sample (LCR_1152_v2) verified our previous results.

The patients in whom TOSV was detected were admitted to the hospital with intense headaches and fever, and their CSF contained high levels of WBC (in average ~ 416 cells/µl) (Figure 1). Some also presented vomiting, sensitivity to light, and neck stiffness (Table S1).

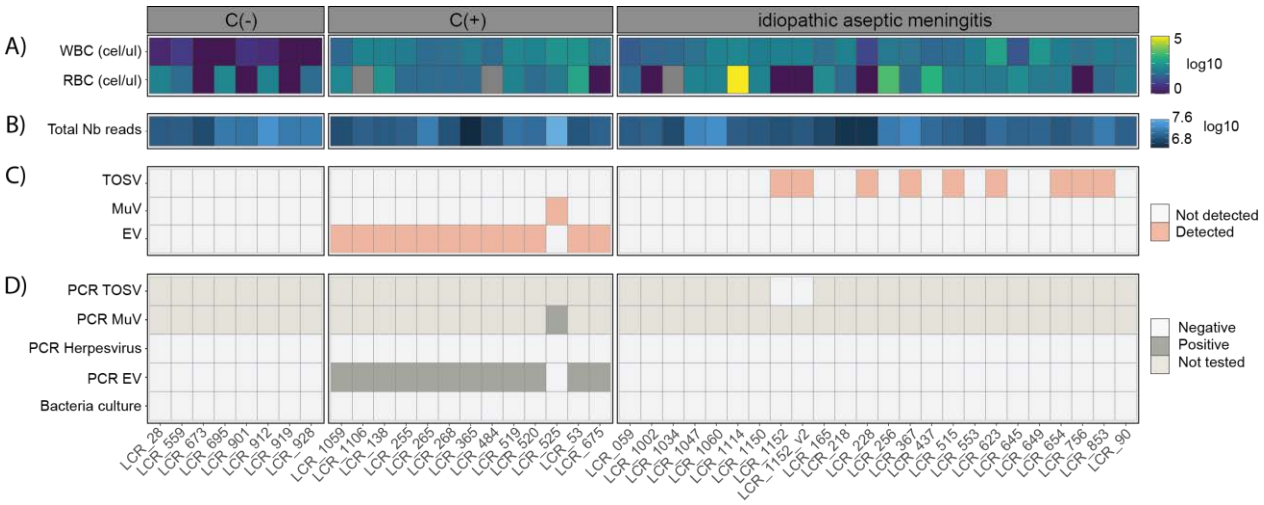


Figure 1: Summary of the diagnostic test and mNGS shown for all the samples. From the top, panel A represents white blood cells (WBC) and red blood cells (RBC) counts measured at the hospital were plotted as a heatmap (log10). Panel B represents the total number of raw reads (as log10) for each sample with a blue gradient. Panel C – results of the mNGS for the target viruses: gray squares indicate no detection, and pink squares indicate detection of the pathogen. Panel D provides a summary of the qPCR results for TOSV, MuV, Herpesvirus, EVs, and culture of bacterial pathogens (*Salmonella enterica, Mycobacterium tuberculosis).* Gray squares indicate that the qPCR test was negative, green positive, and light beige that no test was performed.

Table 1: TOSV mNGS results from CSF samples

SAMPLE	SEGMENT	TOTAL NB OF READS	AVERAGE COVERAGE	MAPPED READS (%)	GENOME COVERED (%)
LCR_228	Segment L	5270466	34.21	0.0569	98.3
	Segment M	5270466	36.84	0.0404	93.3
	Segment S	5270466	21.83	0.0107	81.1
LCR_367	Segment L	21978596	19.92	0.0079	64.7
	Segment M	21978596	5.89	0.0015	29.4
	Segment S	21978596	19.58	0.0023	71.9
LCR_515	Segment L	11158116	23.46	0.0185	96.5
	Segment M	11158116	20.9	0.0109	90.6
	Segment S	11158116	10.24	0.0024	66.5
LCR_623	Segment L	13147500	62.58	0.0419	99.0
	Segment M	13147500	42.64	0.0189	94.4
	Segment S	13147500	61.72	0.0121	91.9
LCR_654	Segment L	9122218	4.84	0.0047	38.4

	Segment M	9122218	NA	0.0030	39.6
	Segment S	9122218	9.86	0.0028	65.5
LCR_756	Segment L	11329328	5.05	0.0039	38.8
	Segment M	11329328	1.62	0.0008	9.6
	Segment S	11329328	9.87	0.0022	63.3
LCR_853	Segment L	17472260	201.38	0.1014	98.1
	Segment M	17472260	118.39	0.0393	96.8
	Segment S	17472260	274.18	0.0405	96.7
LCR_1152	Segment L	8170920	0.44	0.0005	2.9
	Segment M	8170920	0.28	0.0002	1.0
	Segment S	8170920	2	0.0006	13.0
LCR_1152_V2	Segment L	10369600	3.51	0.0030	18.3
	Segment M	10369600	0.05	0.0000	0.0
	Segment S	10369600	5.31	0.0013	37.9

TOSV amplicon-based sequencing

To capture sufficient TOSV content for complete genome reconstruction, we turned to set up an amplicon-based sequencing approach as the one described by Quick, J. et al. (1) while validating the results of the metagenomic sequencing. We designed the TOSV PCR primer scheme using the web-based multiplex primer design tool for amplicon-based sequencing named Primal Scheme (https://primalscheme.com/). To test and optimize the generated primer scheme named “pool_v1”, we used a TOSV genomic standard belonging to the genotype B. We optimized this initial primer scheme by (i) optimizing primer concentration, (ii) testing different primer pairs for specific genomic regions, and (iii) identifying the optimal amount of starting cDNA material. In this way, we created two alternative primer schemes that varied in primer’s composition and concentration (see Table S3 and methods section for description) named "pool_v2" and "pool_v3". We tested the three versions of the primer pool scheme using different input concentrations of the TOSV RNA standard for the cDNA synthesis. By doing so, we performed the PCR using 200, 100, and 10 copies of cDNA. We generated the tiled virus amplicons for each condition and sequenced each sample using the Illumina NextSeq500 platform. Primer pool "pool_v2" yielded better genome coverage for the different amounts of cDNA input (Figure S2).

Subsequently, we re-extracted RNA from those CSF samples in which we had detected TOSV by mNGS, and we used the pool_v2 primer scheme to generate tiled TOSV amplicons. We successfully generated nearly complete TOSV genomes in 7 out of the eight samples, largely improving mNGS results. However, despite these additional efforts, we did not succeed in sufficiently amplifying TOSV and preparing libraries for sample LCR_1152. Completeness and coverage for these genomes are described in Table 2 and shown in Figure S3.

Table 2: TOSV amplicon-based sequencing results from CSF samples

SAMPLE	SEGMENT	TOTAL NB OF READS	AVERAGE COVERAGE	MAPPED READS (%)	GENOME COVERED (%)
LCR_228	L	26,036,074.00	104,289	59.6	92.7
	M	26,036,074.00	82,122	30.9	93.4
	S	26,036,074.00	49,627	8.3	93.1
LCR_367	L	14,332,700.00	39,732	42.2	78.7
	M	14,332,700.00	57,434	40.2	79.1
	S	14,332,700.00	40,549	12.6	93.1
LCR_515	L	13,504,654.00	51,314	57.7	96.4
	M	13,504,654.00	43,734	32.4	97.0
	S	13,504,654.00	21,050	6.9	93.1
LCR_623	L	10,416,918.00	36,578	53.4	96.4
	M	10,416,918.00	33,657	32.4	97.0
	S	10,416,918.00	23,004	9.8	93.1
LCR_654	L	15,046,180.00	60,901	61.6	90.4
	M	15,046,180.00	34,548	23	92.1
	S	15,046,180.00	34,514	10.2	93.1
LCR_756	L	5,841,460.00	21,491	56	94.8
	M	5,841,460.00	16,182	27.8	97.0
	S	5,841,460.00	16,692	12.7	93.1
LCR_853	L	5,234,954.00	18,406	53.4	96.4
	M	5,234,954.00	18,570	35.5	97.0
	S	5,234,954.00	9,257	7.9	93.1

Characteristics of TOSV-positive patients

All TOSV-infected patients were male. Ages of case-patients ranged from 15 to 78 years (median 39 years). Length of hospital stay ranged from 2 to 16 days (median 7 days). Toscana cases were detected in the period comprised from July to November. These patients in whom TOSV was detected were admitted to the hospital with strong headaches and fever, and their CSF contained high levels of WBC (in average ~ 416 cells/µl) (Figure 1). Some of them also presented vomiting, sensitivity to light and neck stiffness (Table S1).

Phylogenetic analysis of TOSV

In combination with a set of publicly available partial and complete TOSV sequences with representatives of all the three different lineages of TOSV (A, B and C), we generated a maximum likelihood (ML) phylogeny for each segment separately. The estimated phylogenies placed the novel sequences within the genotype B with high support for the three segments (bootstrap node support of 99%, 100% and 98% for the segment S, M and L respectively) (Figure 2). Within the genotype B we found sequences collected from Spain (1998 – 2004), Portugal (1983), Switzerland (2018) and France (2004 – 2015).

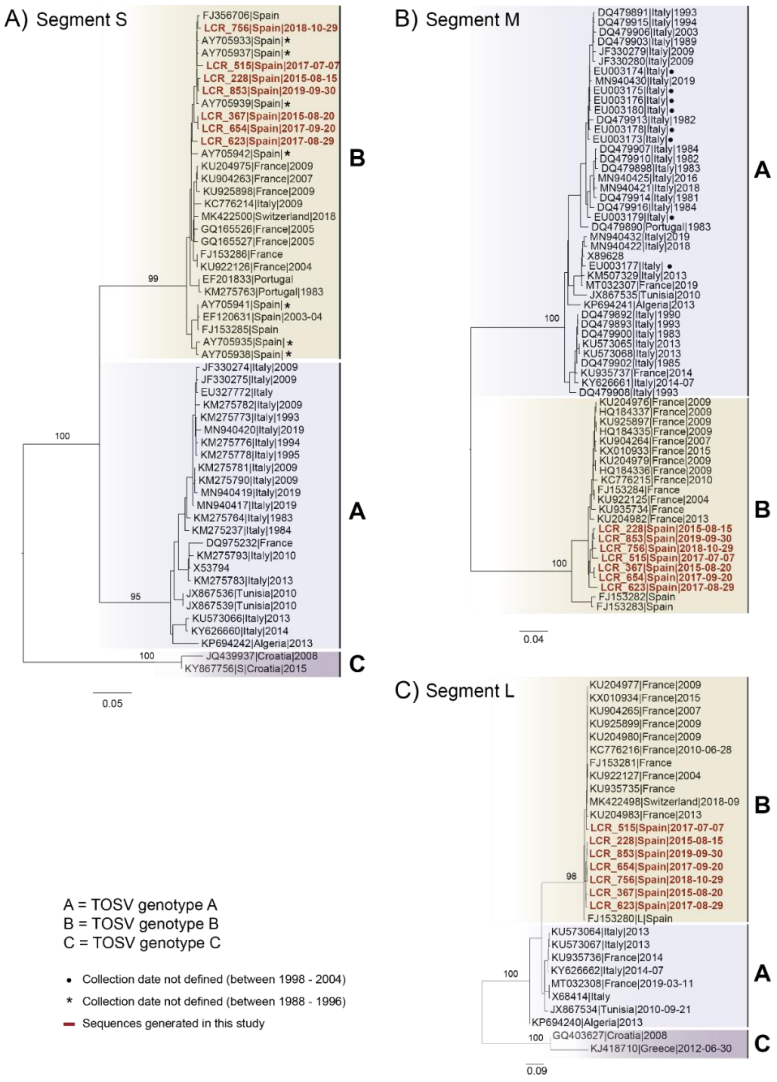


Figure 2: Phylogenetic trees inferred by IQ-TREE v.2.0.6 for the three segments of TOSV: segment S (A), M (B), and L (C). We indicated the different TOSV genotypes with different colors: genotype A in blue, B in beige, and C in purple. We highlighted in red sequences reported in our study. Node support values for the main lineages are ultrafast bootstrap percentages, and the scale bars represent the number of nucleotide substitutions per site.

Characteristics of the MuV-positive sample and phylogenetic analysis

The CSF sample LCR_525 belonged to a woman admitted to the hospital with parotitis, strong headaches, and sensitivity to light. The CSF specimen tested in this study contained 747 WBC/µl and tested positive for MuV by qPCR. The mNGS validated this result and successfully generated a nearly complete MuV genome. To characterize the resulting MuV genome, we constructed a phylogeny using full-length MuV genomes, including the WHO reference genomes for the different genotypes (A – N), totaling 211 sequences. The resulting phylogeny placed our MuV genome within the genotype G, together with

sequences collected from the USA during 2015 – 2017 and Canada in 2017. The two immediately sequences basal to LCR_525 are genomes collected in the USA in 2011 (accession number KY969482) and the Netherlands in 2010 (accession number MW261742) (Figure 3 and Figure S4).

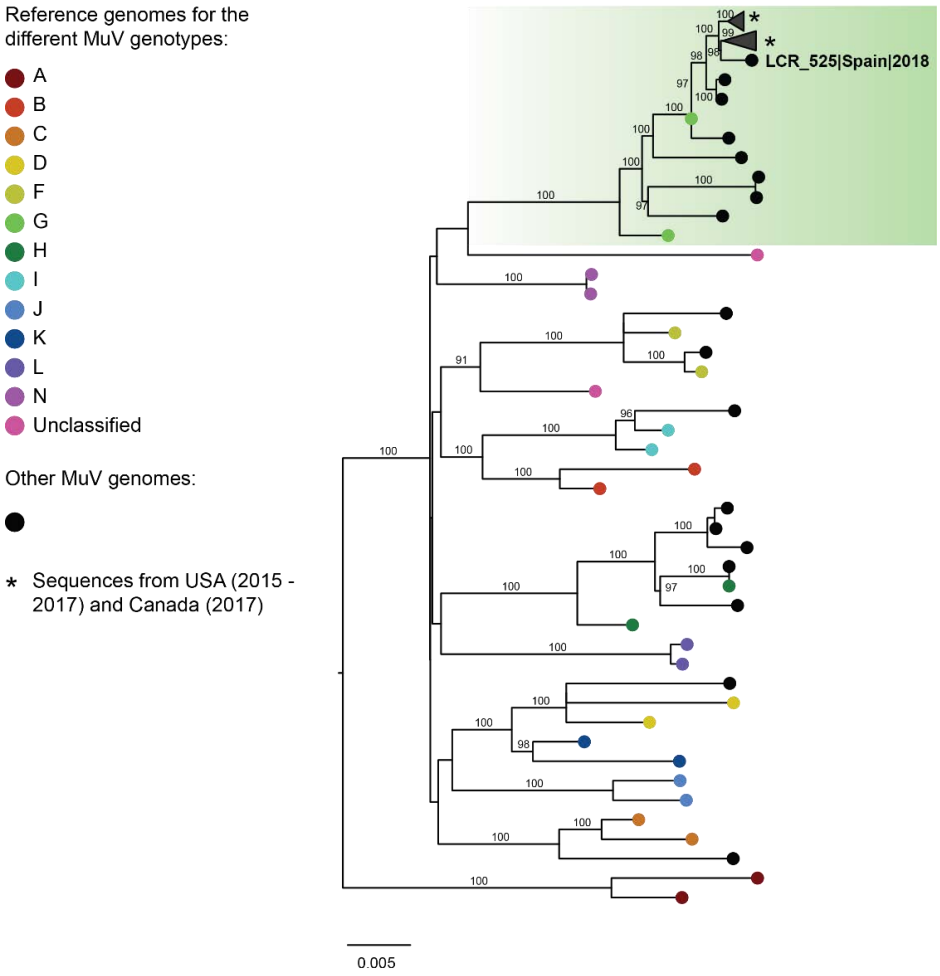


Figure 3: ML tree of MuV full-length genomes and the WHO reference strains for the different genotypes. The tip nodes for the reference strains are colored according to the MuV genotype they represent. Black nodes represent non-reference MuV sequences. The light green box highlights the MuV genotype G, the genotype to which the genome generated in this study (LCR_525) belongs. The expanded tree can be found in Figure S4. Ultrafast bootstrap values above 80% are indicated above the branches leading to corresponding nodes. Scale bar indicates nucleotide substitutions per site.

DISCUSSION

Metagenomic sequencing has become a powerful tool as it allows the identification of potentially any pathogen (excluding prions) present in the sample, including new recombinant forms or new drug-resistant genetic variants (21, 22). Indeed, mNGS has already helped diagnose several infectious diseases with severe symptoms such as meningitis and encephalitis, identifying viruses (e.g., WNV, CHIKV), bacteria (e.g., *Salmonella enterica*) or fungi (e.g., *Candida dubliniensis*) as potential etiologic agent (23-26). Here we used mNGS to study the presence of RNA virus known to cause neurological infection in a series of idiopathic meningitis from Southern Spain. To validate our findings, we also include positive and negative control samples.

In our study, mNGS correctly identified the viral pathogen in all CSF samples with known etiology previously determined by clinical qPCR (positive controls), and succeeded in generating full-length EV and MuV genomes; two RNA viruses well-known to cause meningitis. The 100% concordance between mNGS and routine laboratory testing, the recovery of the viral full genomes and their utility to identify for instance a new recombinant form of the echovirus 13 included in this study (published as part of a separate study (27)), highlights the utility of mNGS to identify viral pathogens in CSF, for virus characterization studies and genomic surveillance.

Among the samples with unknown aseptic meningitis etiology, we detected TOSV in 8 out of 23 samples by mNGS. We used two sequencing methods to generate complete TOSV genomes from these samples: an untargeted metagenomic and an amplicon-based sequencing approach. The first method allowed us to detect TOSV but failed to obtain full-length genomes. Nevertheless, mNGS identified TOSV with enough read coverage to determine that it belonged to the genotype B. This allowed us to design primers to implement a highly multiplexed PCR amplicon approach to obtain nearly complete viral genomes from these challenging samples (28). This amplicon-based approach was also used as a confirmatory test to validate the results of the metagenomic sequencing. Combining these two approaches, we generated nearly full-genome sequences of the TOSV-positive samples except for sample 1152. This sample was the only one tested for TOSV by qPCR, resulting in a negative.

For this reason, the mNGS of this sample was repeated by a different experimenter, starting from the RNA extraction and in the absence of any other TOSV-positive CSF sample. By repeating the sequencing of this sample, we identified TOSV, verifying our previous results. By analyzing the target region of the primers and probes of the TOSV qPCR, we found that the mNGS did not recover such a region, providing a possible explanation for the discrepancy between the two methods. Therefore, possibly the sample had degraded viral RNA, which hampered its detection by RT-qPCR and amplification by the multiplexed PCR. Further experiments and/or analyses would be necessary to estimate the total level of RNA degradation in comparison to other samples. Nevertheless, the fact that information obtained from the mNGS (positive for TOSV) exceeded that of conventional RT-qPCR (negative for TOSV), proves again mNGS utility to detect pathogenic viruses in CSF clinical specimens.

TOSV is an arbovirus transmitted to humans mainly through the bites of an infected sandfly (*Phlebotomus* and *Sergentomyia* genera). Interestingly, only one TOSV case had insect bites reported. In particular, this patient (corresponding to sample LCR_654) reported several insect bites during a trip to Portugal shortly before the development of symptoms. On the ViPR database, we only found three sequences sampled in Portugal: a partial sequence of the M segment collected in 1983 (accession number DQ479890), a partial sequence of the S segment collected in 1983 (accession number KM275763), and a complete sequence of

the S segment with no collection date available (accession number EF201833). Because of this, our analysis could neither show nor rule out the phylogenetic proximity between the TOSV detected in this particular patient with travel history to Portugal and TOSV circulating there. More recent genomic data from the travel destination could potentially elucidate whether the patient got infected in Portugal. Our phylogenetic analysis showed that for the three different segments, the TOSV collected from cases in Southern Spain between 2015 and 2018 fall within the genotype B and close to sequences collected from Spain (1998 – 2004) and France (2004 – 2015). Nevertheless, we could make no further phylogenetic inferences due to the fractionated TOSV genomic data available.

In the last few years, TOSV has been responsible for increasing infections in countries of the Mediterranean basin, where its vector, the sandfly, is widely distributed (29). Indeed, cases have been reported in several countries across this region, including Algeria, Croatia, France, Italy, Portugal, Spain, and Turkey, to name a few (26, 27). Although seroprevalence studies suggest that TOSV causes an asymptomatic infection in most cases, some people might experience severe disease with fever, intense headache, vomiting, or neurological diseases such as meningitis or encephalitis (29). Indeed, a recent study showed that TOSV is the most common cause of summer viral meningitis in central Italy, outnumbering EV cases (28). A study recently reported a young man returning from Elba Island, Italy, with TOSV meningitis and viral RNA persistence in semen up to 59 days after symptom onset (30). In Spain in particular, the virus circulates for decades, and it has been mostly associated with sporadic neurological cases in the Andalusia region in Southern Spain. A recent work revised all documented cases of TOSV neurological infection detected in this region from 1988 to 2020, totaling 107 cases diagnosed by viral culture or RT-qPCR (31).

Despite the increasing number of TOSV cases and its increasing association with neurological disorders such as meningitis, this study highlights the lack of clinical suspicion and therefore lack of laboratory testing in patients with idiopathic meningitis. In this study only one patient was tested by clinicians for TOSV who surprisingly referred no insect bites in the last days but referred being a cattle farm worker, an occupation which has been associated with an increased risk for TOSV infection (32). The lack of suspicion could be due to the fact that in all TOSV-positive patients referred explicitly no insect bites to the doctor except the patient who travelled to Portugal. This highlights the need to include TOSV-testing in the routine diagnosis at the hospitals in southern Spain for cases of aseptic meningitis, regardless the history of insect bites referred by the patient. This could result in shortened length of hospital stay (median 7 days in this study for TOSV-positive patients), reduced associated costs and improved patient care.

In addition, there is little information available about the genomic diversity of TOSV. For instance, as of March 2022, the most recent TOSV sequences in the ViPR database collected from Spain are from 2005, and most of them are partial sequences from the S and L segments. Similarly, the global diversity of TOSV circulating in the Mediterranean Sea is understudied, with less than 20 complete sequences from this region in the ViPR database with the complete metadata. Our mNGS findings led us to design a specific TOSV genotype B amplicon-based sequencing approach to help sequence partially degraded or low concentrated. In this way, our work provides an update on TOSV circulating in Spain and a methodological improvement to facilitate further detection and genome sequencing of TOSV. We hope this will help enlarge the current TOSV genomic database necessary for high-resolution phylogenetic analysis to ultimately contribute to understanding the epidemiology and evolution of this emerging virus.
Nonetheless, our work has several limitations. The storage conditions of the samples used in this study were not optimal: the samples were collected from 2015–2018, stored at -45°C until processed in 2019, and subjected to more than one freeze-thaw cycle. This fact could have affected the quality of the RNA in the samples and thus limited our findings. Another limitation is that our study, starting from the processing of the samples to the analysis of the mNGS, only focused on RNA viruses known to cause neurological

infections. For this reason, we performed RNA extraction from the CSF samples followed by treatment with Turbo DNase to digest contaminating DNA, enriching the samples with viral RNA content as suggested by Matranga et al. (33). We did not explore viral contigs beyond the known human-infecting viruses that cause meningitis; however, our data might be useful in future virus discovery efforts, such as the recent one by Edgar et al. (34).

Lastly, although mNGS for applications in clinical microbiology laboratories remains very challenging (19, 22, 35), we provide further evidence that it could be a significant addition to a surveillance program to monitor emerging viruses and update diagnostic tools.

Our work suggests that TOSV should be considered when diagnosing patients with meningitis of unknown etiology from Southern Spain or with a travel history to locations where the virus is known to circulate (i.e., the Mediterranean basin). Additionally, in this work, we set up an amplicon-based sequencing approach to help sequence TOSV genotype B from samples with varying RNA quality and concentration, providing a methodological improvement to facilitate complete genome sequencing.

MATERIALS AND METHODS

Ethics statement

Ethical approval for the experimental protocol was given by the Ethical Review Board of Cordoba (Comité de Ética de Investigación de Córdoba) under protocol 201999903552445. Written informed consent was obtained from all patients/legal guardians. The procedures were carried out in accordance with approved guidelines, regulations and the principles of the Declaration of Helsinki.

Sample selection and laboratory methods

Clinical CSF specimens used in this study were collected for routine clinical care from patients with suspected viral CNS infection in the Department of Neurology in the University Hospital Reina Sofia (Cordoba, Spain) between 2015 and 2019. The Department of Neurology is responsible for the treatment of patients from 14 years old and is the reference unit for patients with suspected CNS infections in the province. CSF samples were obtained at patient admission according to standard procedures. These CSF samples were analyzed in the Department of Microbiology for a range of pathogenic agents according to the clinical suspicion of the attending physician. Included tested pathogens were enteroviruses, herpesvirus types 1, 2, 3, 4, 5 and 6, and bacterial pathogens (*Neisseria meningitidis*, *Streptococcus pneumoniae, Hemophilia influenza type b*, Listeria, Mycobacterium tuberculosis). For some patients, serum samples were also tested for Treponema pallidum, Coxiella, Borrelia, Ricketsia, Brucella, HIV, Mycoplasma pneumoniae, Cryptococcus, hepatitis viruses, parvovirus, herpesvirus such as VZV, EBV or CMV. For one patient LCR was also tested for Toscana and West Nile Viruses in the Virus Reference Laboratory of Andalucia in the University Hospital Virgen de las Nieves from Granada. Routine laboratory testing performed in all selected patients (in CSF and serum samples) are provided in table S1. The laboratory methods used are summarized in table S2. Residual CSF were stored at -45ºC. Those with sufficient volume (1 mL) were included in the study. Samples included in this study were: A) positive controls which were randomly chosen CSF samples from patients with meningitis with previously identified pathogen by qPCR test; B) negative controls which were randomly chosen CSF samples from patients with no meningitis but in whom another diagnosis was made (epilepsy or cognitive impairment); and C)

idiopathic samples which were randomly chosen CSF samples from patients diagnosed with aseptic meningitis with unknown etiologic agent after conventional routine laboratory testing in the hospital. Meningitis-suspected cases were identified as described previously (27).

RNA extraction

RNA extraction was performed at the Institut Pasteur Paris. In short, RNA was extracted from 140 ul of CSF using the QIAamp Viral RNA Mini Kit (Qiagen) following the manufacturer's recommendations. This was followed by Turbo DNAse treatment (Ambion) and purification with Agencourt RNAClean XP beads (Beckman Coulter).

Metagenomic next-generation sequencing

We used a generic protocol for untargeted metagenomic sequencing of clinical samples previously described (33). Briefly, prior to library construction, poly-A carrier RNA and host rRNA was depleted using oligo (dT) and custom probes, respectively, to form RNAse H target DNA-RNA hybrids. The RNA resulting from selective depletion was used for random-primed cDNA synthesis using the SuperScript IV (Invitrogen). Second-strand cDNA was generated using a cocktail of enzymes, including *Escherichia coli* DNA ligase, RNAse H, and DNA polymerase (New England Biolabs), and then it was purified using Agencourt AMPure XP beads (Beckman Coulter). We prepared libraries from the dsDNA using the Nextera XT kit (Illumina).

Metagenomic next-generation sequencing data analysis

Raw reads were trimmed using Trimmomatic v0.36 (36) to remove low-quality reads. Reads were *de novo* assembled using the metaspades option from SPAdes v3.12 (37), and the contigs obtained were used as queries for blastx using DIAMOND v2.0.7 against version 18.0 of the RVDB protein database (38). Taxonomy was assigned to each conting with DIAMOND and in-house R (39) scripts were used to analyze the resulting output.

For the pathogens identified by DIAMOND blastx, chimeric contigs were constructed using the contig assembled with metaspades and a reference genome. For TOSV, this was done for the three segments separately, and the reference genome used was a consensus sequence result of an alignment of TOSV genotype B. For MuV as a reference genome, we used the Mumps reference genome NC_002200. In all cases, nucleotide divergence between the contig and the reference genome was less than 2%. These chimeric contigs were used as scaffolds to map the reads, using clc-assembly-cell v5.1.0. The virus consensus sequence generation was performed with ivar v1.0 using a minimum of 5X read depth coverage. In case of lower read coverage, we added an N. Samtools v1.10 (40) was used to sort the aligned BAM files and generate alignment statistics. We manually inspected all alignments and consensus sequences using Geneious Prime 2020.2 (https://www.geneious.com/).

We observed in several samples the presence of relevant human pathogens. In particular, nine samples were contaminated with samples from the same run containing high pathogen loads (e.g., viral stocks, samples with amplified viral content). This can occur due to "index hopping," a well-known problem with Illumina sequencers (41). In addition, in sample LCR_1150, we reconstructed a contig of 256 nucleotides, which mapped against the Yellow Fever virus (YFV), a virus we frequently sequence in our laboratory (Table 3). We discarded these viruses from the analysis, but we mention this here to avoid misinterpretation from future researchers who might use the raw fastq file generated in this study for analysis.

Table 3: Relevant pathogens detected as contaminants in several of our samples.

Number	Sample	Contaminant
1	LCR_28	CHIKV
2	LCR_695	CHIKV
3	LCR_912	CHIKV
4	LCR_928	CHIKV + DENV-4 + DENV-3
5	LCR_520	CHIKV
6	LCR_1047	CHIKV
7	LCR_1114	CHIKV
8	LCR_1152_s5 (but no LCR_1152_s20)	CHIKV
9	LCR_525	ZIKV
10	LCR_1150	YFV

Amplicon-based sequencing

To obtain complete TOSV genomes, we implemented a highly multiplexed short PCR amplicon approach (42). The primer scheme was designed using PrimalScheme: primer panels for multiplex PCR to generate ~ 400 nucleotide long overlapping amplicons to cover the entire length of the three TOSV segments. The primers are divided into two separate primer pools (pool1 and pool2), generating non-overlapping amplicons pooled in the following protocol step to cover the entire genome. We followed the protocol generated by Quick, J. et al. (28) to generate the tiled virus amplicons. Briefly, two microliters of viral cDNA were used in the two multiplexed PCR reactions using Q5 DNA High-fidelity Polymerase (New England Biolabs) to obtain ~ 400 nucleotide long amplicons in 35 cycles. Amplicons were purified using Agencourt AMPure XP beads (Beckman Coulter) and combined to 50ng. Libraries were constructed using the NexteraXT kit.

Regardless of the sequencing protocol used, library quality and quantification was assessed using Qubit 4 (Thermo Fisher), Bioanalyzer (Agilent), and qPCR (NebNext Library Quant Kit, Illumina) and sequenced using the paired-end strategy on an Illumina NextSeq500 platform (2x75 cycles)

Primer pool scheme optimization

We constructed three primer pool schemes that tested and validated using a TOSV genomic standard for the genotype B (Toscana Standard#1, strain MRS2010 4319501) obtained from the European Viral Archive.

The first primer scheme named “pool_v1” contained the 48 pairs of primers covering the three segments of TOSV designed directly by the Primal Scheme. Due to divergence observed between the TOSV genomes retrieved from GenBank, some primers were modified with degenerate nucleotides. We verify that this does not cause a change in the annealing temperature beyond that recommended one (28). Regardless of the different template concentrations, when sequencing the TOSV standard using the “pool_v1”, we observed an uneven coverage of reads in some areas of the L segment and a drop of coverage in a specific region of the S segment containing a high GC content. We observed no coverage of reads in three specific regions for the M segment. Looking into the primers covering such specific regions, we noticed that they

corresponded to the primers #1, #3, and #5 within the pool1. Although not tested, we hypothesized that primer interference could have played a role. We designed the second primer scheme, “pool_v2”, containing a different primer set for these genomic regions within the M segment to overcome this situation. Also, in this primer scheme, we increase the concentration of those primers covering regions within the L and S segments that we observed to be less efficient. Following the same reasoning, we designed “pool_v3", which contains the same set of primers as “pool_v2” but with an even higher concentration of primers for the M and S segments.

Phylogenetic analysis

To build the TOSV dataset for phylogenetic reconstruction, sequences generated during this study were combined with partial and full-length sequences, including representatives of the different TOSV genotype, available in ViPR (43) and Genbank (44) as of October 2021. Partial sequences less than 300 bp in length were excluded. We put together a total of 53, 44, and 29 sequences for the S, M, and L segments, respectively. The resulting dataset was aligned using MAFFT v7.467 (45) for each segment separately and alignments were visually inspected in Geneious Prime 2020.2. We constructed maximum-likelihood phylogenies (ML) using IQ-TREE v2.0.6 (46). Tree reconstruction was performed using the default settings and the best-fitted model provided by ModelFinder (47), followed by 1000 ultrafast bootstrap (48) implemented in IQ-TREE software.

We proceeded similarly to build the MuV dataset to perform the phylogenetic analysis. We retrieved all available full-length genomes of MuV from 2008 to date from ViPR (43) as of October 2021, and combined with the complete genome generated in this study and the WHO reference strains for the different MuV genotypes. This dataset contained 211 sequences and was used to construct the phylogeny with ML.

SUPPLEMENTARY FIGURE LEGENDS

Figure S1: TOSV genome coverage for the three segments (L, M, and S) obtained using an untargeted metagenomic sequencing. The genome coverage is represented as the logarithm function of the total number of reads + 1 (to avoid conflicts with regions with cero coverage) for the eight different samples in which TOSV was detected. Sample LCR_1152 has been sequenced twice, and here we show the result of both runs: LCR_1152_S5 and LCR_1152_S20.

Figure S2: Optimization of TOSV amplicon-based approach. Three different schemes of primers (pool_v1, pool_v2, and pool_v3), which generate short tiled amplicons for sequencing, were tested on the TOSV B reference standard. Three different numbers of cDNA copies were used as input for the multiplex PCR reaction: 10, 100, and 200. The sequencing depth obtained for the different conditions and for the three segments (L, M, and S) is represented as the logarithm function of the total number of reads + 1.

Figure S3: TOSV genome coverage for the three segments (L, M, and S) obtained using the amplicon-based sequencing approach.

Figure S4: Phylogenetic divergence tree of all labeled MuV sequences used. The maximum-likelihood tree includes MuV full-length genomes retrieved from ViPR as of October 2021, the WHO reference strains for the different genotypes, and the MuV generated in this study (LCR_525).

REFERENCES

1. James SL, Abate D, Abate KH, Abay SM, Abbafati C, Abbasi N, et al. Global, regional, and national incidence, prevalence, and years lived with disability for 354 diseases and injuries for 195 countries and territories, 1990–2017: a systematic analysis for the Global Burden of Disease Study 2017. The Lancet. 2018;392(10159):1789-858.
2. McGill F, Griffiths MJ, Bonnett LJ, Geretti AM, Michael BD, Beeching NJ, et al. Incidence, aetiology, and sequelae of viral meningitis in UK adults: a multicentre prospective observational cohort study. The Lancet Infectious Diseases. 2018;18(9):992-1003.
3. Hasbun R, Rosenthal N, Balada-Llasat JM, Chung J, Duff S, Bozzette S, et al. Epidemiology of Meningitis and Encephalitis in the United States, 2011–2014. Clinical Infectious Diseases. 2017;65(3):359-63.
4. Shukla B, Aguilera EA, Salazar L, Wootton SH, Kaewpoowat Q, Hasbun R. Aseptic meningitis in adults and children: Diagnostic and management challenges. Journal of Clinical Virology. 2017;94:110-4.
5. John CC, Carabin H, Montano SM, Bangirana P, Zunt JR, Peterson PK. Global research priorities for infections that affect the nervous system. Nature. 2015;527(7578):S178-S86.
6. McGill F, Griffiths MJ, Solomon T. Viral meningitis: current issues in diagnosis and treatment. Current opinion in infectious diseases. 2017;30(2):248-56.
7. Kohil A, Jemmieh S, Smatti MK, Yassine HM. Viral meningitis: an overview. Archives of Virology. 2021;166(2):335-45.
8. Gailson T, Vohra V, Saini AG, Bhatia V. Mumps infection with meningoencephalitis and cerebellitis. BMJ case reports. 2021;14(11).
9. Johnstone J, Ross CA, Dunn M. Meningitis and encephalitis associated with mumps infection: a 10-year survey. Archives of Disease in Childhood. 1972;47(254):647-51.
10. Bockelman C, Frawley TC, Long B, Koyfman A. Mumps: an emergency medicine-focused update. The Journal of emergency medicine. 2018;54(2):207-14.
11. Faria NR, Azevedo RDSDS, Kraemer MUG, Souza R, Cunha MS, Hill SC, et al. Zika virus in the Americas: Early epidemiological and genetic findings. Science. 2016;352(6283):345-9.
12. Casas-Alba D, De Sevilla MF, Valero-Rello A, Fortuny C, García-García JJ, Ortez C, et al. Outbreak of brainstem encephalitis associated with enterovirus-A71 in Catalonia, Spain (2016): a clinical observational study in a children's reference centre in Catalonia. Clinical Microbiology and Infection. 2017;23(11):874-81.
13. Leal Barceló AM, Carrascosa García P, Rincón López EM, Herrero M, Navarro ML, editors. Brote de infección por enterovirus causantes de afectación neurológica grave en un hospital terciario. Anales de Pediatría; 2018.
14. Cabrerizo M, García-Iñiguez JP, Munell F, Amado A, Madurga-Revilla P, Rodrigo C, et al. First cases of severe flaccid paralysis associated with enterovirus D68 infection in Spain, 2015–2016. The Pediatric infectious disease journal. 2017;36(12):1214-6.
15. III IdSC. Vigilancia de la Parálisis Flácida Aguda y Vigilancia de Enterovirus, Informe año 2018 2018 [Available from: https://www.isciii.es/QueHacemos/Servicios/VigilanciaSaludPublicaRENAVE/EnfermedadesTransmisibles/Documents/archivos%20A-Z/POLIO/Resultados_Vigilancia_Polio/Informes_Anuales_Polio/Informe_PFA_y_Enterovirus_2018.pdf.
16. III IdSC. Vigilancia de la Parálisis Flácida Aguda y Vigilancia de Enterovirus, Informe año 2020 2020 [Available from: https://www.isciii.es/QueHacemos/Servicios/VigilanciaSaludPublicaRENAVE/EnfermedadesTransmisibles

/Documents/archivos%20A-Z/POLIO/Resultados_Vigilancia_Polio/Informes_Anuales_Polio/Informe_PFA_EV_2020_web.pdf.
17. Antón AIN, de Ory Manchón F, Fariñas MPS-S, Narváez LF, Cámara MIG, Mari JMN, et al. Microbiological diagnosis of emerging arboviral and rodent borne diseases. Enfermedades infecciosas y microbiologia clinica. 2015;33(3):197-205.
18. Centro de Coordinación de Alertas y Emergencias sanitarias MdS, Consumo y Bienestar Social. Informe de situación y evaluación del riesgo de enfermedad por flebovirus trasmitidos por flebotomos en España 2019 [Available from: https://www.sanidad.gob.es/ca/profesionales/saludPublica/ccayes/analisisituacion/doc/ER_Flebovirus.pdf.
19. Ramachandran PS, Wilson MR. Metagenomics for neurological infections—expanding our imagination. Nature Reviews Neurology. 2020;16(10):547-56.
20. De Ory F, Avellón A, Echevarría J, Sanchez-Seco M, Trallero G, Cabrerizo M, et al. Viral infections of the central nervous system in Spain: a prospective study. Journal of medical virology. 2013;85(3):554-62.
21. Houldcroft CJ, Beale MA, Breuer J. Clinical and biological insights from viral genome sequencing. Nature Reviews Microbiology. 2017;15(3):183-92.
22. Chiu CY, Miller SA. Clinical metagenomics. Nature Reviews Genetics. 2019;20(6):341-55.
23. Wilson MR, Sample HA, Zorn KC, Arevalo S, Yu G, Neuhaus J, et al. Clinical metagenomic sequencing for diagnosis of meningitis and encephalitis. New England Journal of Medicine. 2019;380(24):2327-40.
24. Saha S, Ramesh A, Kalantar K, Malaker R, Hasanuzzaman M, Khan LM, et al. Unbiased metagenomic sequencing for pediatric meningitis in Bangladesh reveals neuroinvasive chikungunya virus outbreak and other unrealized pathogens. MBio. 2019;10(6):e02877-19.
25. Wilson MR, Zimmermann LL, Crawford ED, Sample HA, Soni PR, Baker AN, et al. Acute West Nile Virus Meningoencephalitis Diagnosed Via Metagenomic Deep Sequencing of Cerebrospinal Fluid in a Renal Transplant Patient. American Journal of Transplantation. 2017;17(3):803-8.
26. Wilson MR, O'Donovan BD, Gelfand JM, Sample HA, Chow FC, Betjemann JP, et al. Chronic Meningitis Investigated via Metagenomic Next-Generation Sequencing. JAMA Neurology. 2018;75(8):947.
27. Gámbaro F, Pérez AB, Agüera E, Prot M, Martínez-Martínez L, Cabrerizo M, et al. Genomic surveillance of enterovirus associated with aseptic meningitis cases in southern Spain, 2015–2018. Scientific Reports. 2021;11(1).
28. Quick J, Grubaugh ND, Pullan ST, Claro IM, Smith AD, Gangavarapu K, et al. Multiplex PCR method for MinION and Illumina sequencing of Zika and other virus genomes directly from clinical samples. Nature Protocols. 2017;12(6):1261-76.
29. Ayhan N, Charrel RN. An update on Toscana virus distribution, genetics, medical and diagnostic aspects. Clinical Microbiology and Infection. 2020;26(8):1017-23.
30. Matusali G, D'Abramo A, Terrosi C, Carletti F, Colavita F, Vairo F, et al. Infectious Toscana Virus in Seminal Fluid of Young Man Returning from Elba Island, Italy. Emerging Infectious Diseases. 2022;28(4):865-9.
31. Sanbonmatsu-Gámez S, Pedrosa-Corral I, Navarro-Marí JM, Pérez-Ruiz M. Update in Diagnostics of Toscana Virus Infection in a Hyperendemic Region (Southern Spain). Viruses. 2021;13(8):1438.
32. Cusi MG, Savellini GG, Zanelli G. Toscana virus epidemiology: from Italy to beyond. The open virology journal. 2010;4:22.
33. Matranga CB, Andersen KG, Winnicki S, Busby M, Gladden AD, Tewhey R, et al. Enhanced methods for unbiased deep sequencing of Lassa and Ebola RNA viruses from clinical and biological samples. Genome biology. 2014;15(11):1-12.
34. Edgar RC, Taylor J, Lin V, Altman T, Barbera P, Meleshko D, et al. Petabase-scale sequence alignment catalyses viral discovery. Nature. 2022:1-6.

35. Goldberg B, Sichtig H, Geyer C, Ledeboer N, Weinstock GM. Making the leap from research laboratory to clinic: challenges and opportunities for next-generation sequencing in infectious disease diagnostics. MBio. 2015;6(6):e01888-15.
36. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics. 2014;30(15):2114-20.
37. Nurk S, Meleshko D, Korobeynikov A, Pevzner PA. metaSPAdes: a new versatile metagenomic assembler. Genome research. 2017;27(5):824-34.
38. Bigot T, Temmam S, Pérot P, Eloit M. RVDB-prot, a reference viral protein database and its HMM profiles. F1000Research. 2020;8:530.
39. Ihaka R, Gentleman R. R: a language for data analysis and graphics. Journal of computational and graphical statistics. 1996;5(3):299-314.
40. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The sequence alignment/map format and SAMtools. Bioinformatics. 2009;25(16):2078-9.
41. Gu W, Miller S, Chiu CY. Clinical metagenomic next-generation sequencing for pathogen detection. Annual Review of Pathology: Mechanisms of Disease. 2019;14:319-38.
42. Grubaugh ND, Gangavarapu K, Quick J, Matteson NL, De Jesus JG, Main BJ, et al. An amplicon-based sequencing framework for accurately measuring intrahost virus diversity using PrimalSeq and iVar. Genome Biology. 2019;20(1).
43. Pickett BE, Greer DS, Zhang Y, Stewart L, Zhou L, Sun G, et al. Virus pathogen database and analysis resource (ViPR): a comprehensive bioinformatics database and analysis resource for the coronavirus research community. Viruses. 2012;4(11):3209-26.
44. Benson DA, Cavanaugh M, Clark K, Karsch-Mizrachi I, Lipman DJ, Ostell J, et al. GenBank. Nucleic Acids Research. 2012;41(D1):D36-D42.
45. Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Molecular biology and evolution. 2013;30(4):772-80.
46. Nguyen L-T, Schmidt HA, Von Haeseler A, Minh BQ. IQ-TREE: A Fast and Effective Stochastic Algorithm for Estimating Maximum-Likelihood Phylogenies. Molecular Biology and Evolution. 2015;32(1):268-74.
47. Kalyaanamoorthy S, Minh BQ, Wong TKF, Von Haeseler A, Jermiin LS. ModelFinder: fast model selection for accurate phylogenetic estimates. Nature Methods. 2017;14(6):587-9.
48. Hoang DT, Chernomor O, Von Haeseler A, Minh BQ, Vinh LS. UFBoot2: Improving the Ultrafast Bootstrap Approximation. Molecular Biology and Evolution. 2018;35(2):518-22.

scientific reports

Check for updates

OPEN

Genomic surveillance of enterovirus associated with aseptic meningitis cases in southern Spain, 2015–2018

Fabiana Gámbaro[1,9], Ana Belén Pérez[2,3,9], Eduardo Agüera[2,3], Matthieu Prot[1], Luis Martínez-Martínez[2,3,4], María Cabrerizo[5,6,7], Etienne Simon-Loriere[1,10]✉ & Maria Dolores Fernandez-Garcia[3,8,10]✉

New circulating Enterovirus (EV) strains often emerge through recombination. Upsurges of recombinant non-polio enteroviruses (NPEVs) associated with neurologic manifestations such as EVA71 or Echovirus 30 (E30) are a growing public health concern in Europe. Only a few complete genomes of EVs circulating in Spain are available in public databases, making it difficult to address the emergence of recombinant EVs, understand their evolutionary relatedness and the possible implication in human disease. We have used metagenomic (untargeted) NGS to generate full-length EV genomes from CSF samples of EV-positive aseptic meningitis cases in Southern Spain between 2015 and 2018. Our analyses reveal the co-circulation of multiple Enterovirus B (EV-B) types (E6, E11, E13 and E30), including a novel E13 recombinant form. We observed a genetic turnover where emergent lineages (C1 for E6 and I [tentatively proposed in this study] for E30) replaced previous lineages circulating in Spain, some concomitant with outbreaks in other parts of Europe. Metagenomic sequencing provides an effective approach for the analysis of EV genomes directly from PCR-positive CSF samples. The detection of a novel, disease-associated, recombinant form emphasizes the importance of genomic surveillance to monitor spread and evolution of EVs.

Enteroviruses (EVs) belong to the family *Picornaviridae*, genus *Enterovirus*. Their positive single-strand RNA genome is about 7,500 nucleotides (nt), and is composed of a large open reading frame (ORF) flanked by 5' and 3' untranslated regions (UTRs)[1]. The 5' part of the ORF encodes the structural proteins that form the capsid (among them the VP1 which is the most external), while the 3' part of the ORF encodes the non-structural proteins. There are more than 100 types of EVs infecting humans which are classified into four species A to D[2]. EVs are associated with a wide spectrum of clinical symptoms ranging from nonspecific febrile illness or upper respiratory illness to severe neurological conditions, including meningitis, encephalitis or acute flaccid paralysis[1].

Genetic recombination is a major process in EVs evolution[3]. Since new recombinant EVs may present different properties (variations in virulence or transmissibility), it is essential to capture complete genomes to detect the occurrence of these recombination events for surveillance and public health purposes. Application of metagenomic next-generation sequencing (mNGS) approaches to characterize the whole-genome sequence of EVs has become a valuable tool for detecting multiple EVs that may be present during co-infection, especially caused by two EV types of same species. Currently, Sanger sequencing of the VP1 capsid protein gene is the gold standard for EV genotyping. But as mNGS becomes increasingly affordable, accessible and cost-effective, sequencing the whole genome of EVs will likely allow improved typing in the near future[4]. Therefore, it is crucial that microbiology reference laboratories progress towards implementation of NGS methods to fully characterize and respond to EVs and publicly share these sequence data[4]. In most countries in Europe, non-polio EV (NPEV) infections are not notifiable and surveillance is mainly passive[5]. Because there is no specific EV surveillance system, surveillance captures primarily EVs detected from hospitalized patients with neurological infections[5]. Identification and full

[1]Institut Pasteur, Paris, France. [2]Hospital Universitario Reina Sofía, Córdoba, Spain. [3]Instituto Maimónides de Investigación Biomédica de Córdoba (IMIBIC), Córdoba, Spain. [4]Universidad de Córdoba, Córdoba, Spain. [5]National Centre for Microbiology, Instituto de Salud Carlos III, Madrid, Spain. [6]CIBER de epidemiología y Salud Pública (CIBERESP), Madrid, Spain. [7]Red de Investigación Translacional en Infectología Pediátrica (RITIP), IdiPaz, Madrid, Spain. [8]Present address: National Centre for Microbiology, Instituto de Salud Carlos III, Madrid, Spain. [9]These authors contributed equally: Fabiana Gámbaro and Ana Belén Pérez. [10]These authors jointly supervised this work: Etienne Simon-Loriere and Maria Dolores Fernandez-Garcia. ✉email: etienne.simon-loriere@pasteur.fr; mdfernandez@isciii.es

EV type	Sample N°	Age (y)	Sex	Clinical features	Date sample collection	Days of hospitalization	CSF analysis			
							WBC mm^3	Lymph %	Gluco (mg/dl)	Prot (mg/dl)
E6	138	19	F	Fever, headache, vomiting, neck stiffness, GI	May-2015	2	300	48	94	41
	255	31	M	Fever, headache, nausea, GI	Apr-2015	4	233	42	59	69
	268	27	F	Fever, headache, nausea, vomiting	Apr-2015	6	118	63	55	46
	365	23	F	Fever, headache	May-2015	2	203	61	63	41
E11	484	29	F	Fever, headache, nausea	Nov-2015	2	86	90	53	47
	1059	29	M	Fever, headache, photophobia, GI	Dic-2018	3	383	72	NA	68
	1106	36	F	Fever, headache	Dic-2018	2	312	90	55	63
E13	53	35	F	Fever, headache, nausea, GI	May-2016	10	812	72	50	113
E30	265	30	F	Fever, headache, neck stiffness	Feb-2016	1	101	90	53	45
	519	31	M	Fever, headache, vomiting	Mar-2018	2	422	91	NA	63
	520	25	M	Headache, vomiting, photophobia, GI	May-2018	4	310	82	NA	41
	675	17	M	Fever, headache, vomiting, photophobia	Mar-2017	3	156	63	60	58

Table 1. Clinical features and laboratory findings for the cases of enteroviral meningitis, Córdoba, Spain, 2015–2018. *EV* enterovirus, *CSF* cerebrospinal fluid, *GI* gastrointestinal symptoms, *Gluco* glucose, *NA* not available, *Prot* proteins, *WBC* White Blood Cells.

characterization of the EV types involved in these patients are essential to (1) identify EV types associated with neurological infections and estimate better their associated disease burden, (2) monitor the emergence of new EV strains or new recombinant forms, and (3) gain a better understanding of circulating NPEVs[4].

Despite the common occurrence of EV-associated meningitis cases in Spain[6–8], there are limited studies defining their molecular epidemiology and addressing the emergence of recombinant EVs. These studies have sequenced either multiple parts of the genome (the complete VP1 and 3Dpol proteins)[9–15] or the whole genome[16,17]. However, very few full genomes of EVs circulating in Spain apart from EVA71 or EVD68 are available in public databases (Echovirus 30 [n = 5], Coxsackie B2 [n = 1] and Coxsackie A6 [n = 3]) making it difficult to understand the evolutionary relatedness of the different EV types and the possible implication in human disease.

Here, we aimed to characterize the genetic diversity of EVs circulating in Southern Spain during 2015–2018. We performed mNGS on 12 CSF samples of laboratory-diagnosed enteroviral meningitis. Our viral genomic analysis revealed the co-circulation of E6, E11, E13 and E30 types including a novel E13 recombinant form.

Results

Clinical, epidemiological and laboratory features of patients with enteroviral meningitis. From 2015 to 2018, twelve adult patients were laboratory-diagnosed with enteroviral meningitis at the University Hospital Reina Sofia (Córdoba, Spain). All samples were identified as EVs positive but no other typing data was available for these strains. In this work, we investigated the twelve CSF samples collected from these patients by using ribosomal RNA depletion before mNGS on total extracted RNAs. Epidemiological and clinical information about each patient is listed in Table 1. The median age of the patients was 27.6 years (range 17–36 years) with a female-to-male ratio of 1:0.7. All patients were living in the Cordoba Province, located in southern Spain. Cases showed a seasonal pattern, with E6, E13 and E30 infections occurring during spring (March–May) months, while E11 cases occurred during winter (November and December). The most common symptoms in patients were headache (100%) and fever (90.9%). Seven patients (58.3%) reported contact with potential EV-infected children at home that presented fever or gastrointestinal symptoms. All patients had pleocytosis, defined as CSF WBC counts of > 5/mm^3, with a median WBC count of 286/mm^3 and predominance of lymphocytes (from 42 to 91% of WBC) (Table 1).

Complete genome analysis. After quality trimming, we assembled the reads using MetaSPAdes v3.12, and used Diamond to query the contigs against the non-redundant protein database (NCBI). For each sample, we obtained a large contig corresponding to a near-complete enterovirus genome, or fragments that could be assembled into a genomic scaffold. We did not detect co-infection with other pathogens, including those that have been found in the CSF of patients with meningoencephalitis such as Dengue virus, Chikungunya virus or West Nile virus[18]. Results of the metagenomic analysis can be found in Table 2. De novo assembly resulted in the reconstruction of complete or near-complete EVs genome for all 12 samples. All viruses presented high sequence coverage throughout the genome as shown in Supplementary Fig. S1. Overall, 4 different EV types belonging to species-B were identified. Four samples were found to contain E6, three samples E11, four samples

	Sample N°	Number of mapped viral reads	Mapped viral reads (%)	Genome covered	Average coverage depth
E6	138	127,810	1.2	99.30%	1273
	255	70,030	0.6	99.88%	697
	268	56,357	0.59	99.86%	560
	365	6524	0.14	99.31%	64
E11	484	1046	0.01	95,76%	10
	1059	240,901	2.76	99.40%	2383
	1106	366,776	3.13	99.70%	3624
E13	53	159,567	1.61	100%	1593
E30	265	68,949	0.35	98.91%	684
	519	54,571	0.33	99.30%	537
	520	628,623	4.39	99.60%	6214
	675	91,887	0.76	99.50%	909

Table 2. RNA-sequencing results using mNGS from CSF samples.

E30 and one sample E13. Genomes ranged from 7213 to 7451 nt in length, encoding polyproteins of 2188, 2191, 2194 and 2195 aa for E13, E6, E30 and E11, respectively.

Phylogenetic analysis. In order to characterize the study strains and evaluate their evolutionary relationship to previously characterized homotypic strains detected in Spain and globally in the last decades, phylogenetic analyses were performed for the four encountered genotypes first using the VP1 region (Fig. 1).

E6 lineages were assigned according to Smura et al. and Cabrerizo et al. into lineages A-C[10,19]. Phylogenetic analysis showed that all E6 strains detected in 2015 clustered within sublineage C1 (bootstrap value of 97%). Previously reported E6 strains from Spain (detected between 2004 and 2010) fell into the sublineage C9, clustering into three distinct groups (C9a, C9f. and C9h) (Fig. 1A). All four E6 study strains showed together high nucleotide similarity in the complete VP1 region (mean 97.8% (96.2–99.2%) nt and 99.2% (98.6–100%) aa identity) and branched together within the diversity of strains detected in France and Poland in 2011.

E30 lineages were defined according to Bailly et al. into lineages A-H[20]. The E30 study strains clustered in two different lineages. Sequences LCR265 (2016) and LCR675 (2017) clustered within lineage F, a group that contained most of E30 strains previously identified in Spain (from 1996 to 2018). Analysis with complete VP1 sequences (Supplementary Fig. S2) showed that while LCR675 strain forms a subcluster (bootstrap value of 100%) with other E30 strains from Spain isolated in 2016 in Catalonia (displaying identities of at least 96.4% and 91.1% at nucleotide and peptide levels, respectively), LCR265 cluster with viruses detected in France, Luxembourg and the Netherlands in 2016–2017 suggesting the circulation of two different sublineages. In contrast, strains LCR519 and LCR520 from 2018 grouped into a distinguishable lineage, tentatively called here "lineage I" (corresponding to clade G6 in the classification of Benschop et al.[21] which contains sequences from Netherlands, France, Ireland and Spain, sampled in the same time period (2017–2018) (Fig. 1B). Similar pattern of clustering was obtained when the complete VP1 region was used for phylogenetic inference (Supplementary Fig. S2). A notable observation is that "lineage I" is more closely related to lineage B, which was not detected beyond the 1970s[20] (mean *p*-distance of 0.17), than to more recent lineages like E and F which contain worldwide E30 strains including all E30 strains previously reported in Spain (mean *p*-distance between lineage I and lineages E and F of 0.25 and 0.26, respectively) (Supplementary Table S1).

All E11 strains obtained in this study belong to subgenotype D5 according to the classification introduced by Li et al.[22] which included stains detected in Europe and Asia that have circulated for over 20 years (Fig. 1C). E11 study strains were all highly similar (nt identity ≥ 95.3%; aa identity ≥ 97.6%) and were closely related to strains obtained in France in 2014 from meningitis cases in children (mean 96.5% (95.7–97.8%) nt and 98.6% (98.2–99.3%) aa similarities in the complete VP1 region).

Finally, Fig. 1D shows that E13 study strain LCR53 from 2016 cluster closely (bootstrap value of 100%) with two strains collected from children with meningitis in France in 2015 (nt identity ≥ 98.8%; aa identity ≥ 99.6%).

Recombination analysis. Considering the prevalence of recombination in the evolutionary history of enteroviruses, we next looked for evidence of recombination in the study strains. Recombination analyses were carried out using a combination of six methods implemented in RDP5.5[23]. We detected multiple events of recombination in the evolutionary history of the EV types studied. This was also evidenced when comparing the topologies of phylogenies build using the 3 coding regions (P1, P2 and P3) of the sequences from this study and representative EV-B reference genomes available in GenBank (Fig. 2). The phylogenetic analysis using the P1 capsid coding region showed that the E6, E11, E13 and E30 sequences obtained from this study cluster together with their respective reference genome which coincides with what we observed using the partial VP1 region (Fig. 2A). However, this was no longer observed when constructing the phylogeny based on the P2 and P3 non-capsid genomic regions (Fig. 2B,C). These incongruent tree topologies between the capsid and non-capsid regions indicate that genetic exchanges through intertypic recombination with other EV-B types might have occurred. We also used similarity plot and bootscanning analyses comparing the study strains and closely related types to visualize the mosaic genomes reported here. The sequences of closely related types included in

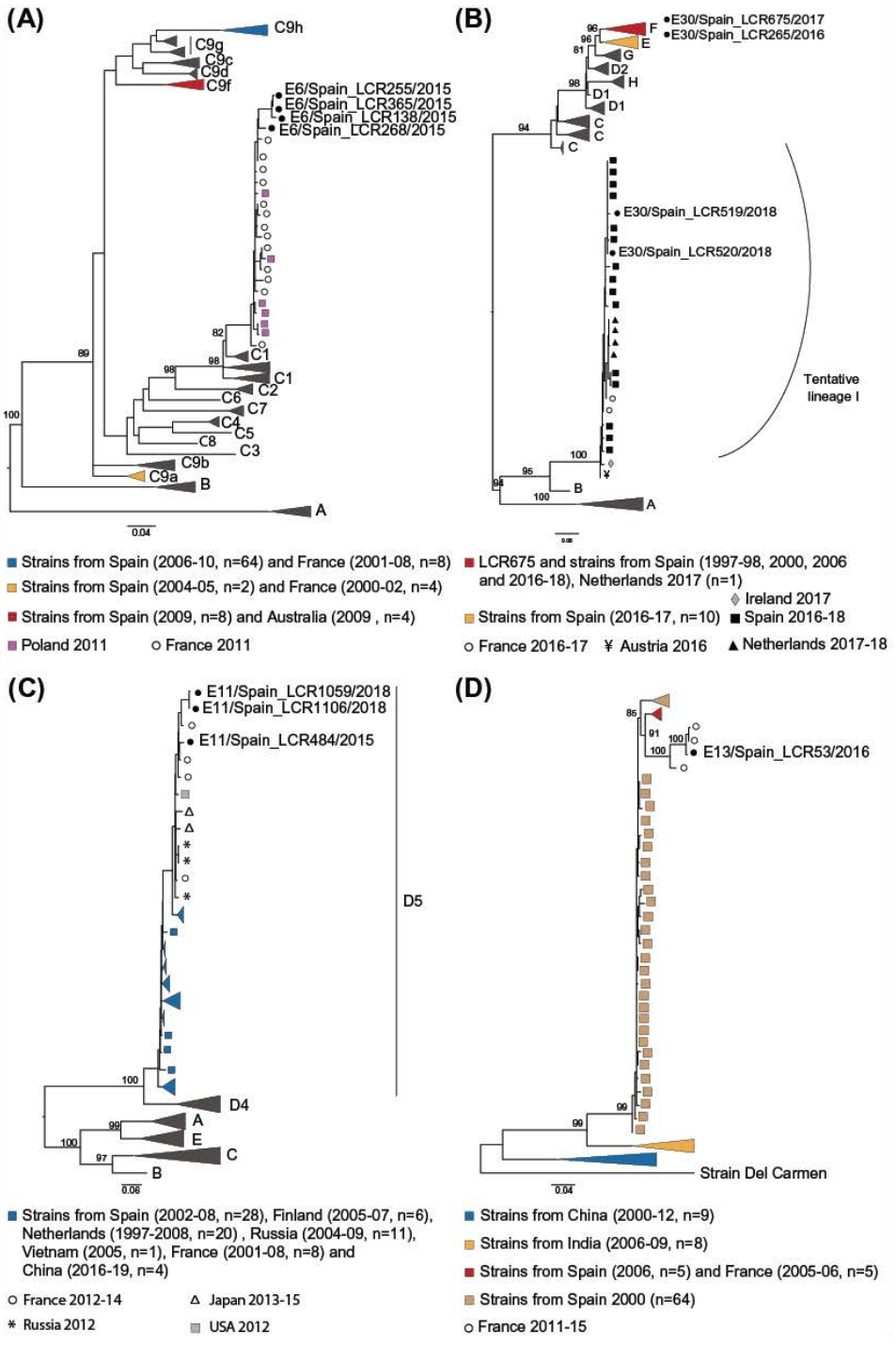


Figure 1. Maximum-likelihood trees based on a partial VP1 inferred by IQ-TREE v.2.0.6 for the four genotypes study here: E6 (**A**), E30 (**B**), E11 (**C**) and E13 (**D**). Sequences reported in our study are indicated by black circles. Numbers on nodes indicate the bootstrap support of the node (> 80) and the scale bars represents the expected number of nucleotide substitutions per site. Non collapsed trees can be found in Supplementary Fig. 4A–D. Trees are annotated with the classification proposed by Smura et al. and Cabrerizo et al.[10,19] for E6, the classification proposed by Bailly et al.[20] for E30, and the classification proposed by Li et al.[22] for E11. Only bootstrap values above 80% are indicated in branch nodes. Scale bars indicate nucleotide substitutions per site.

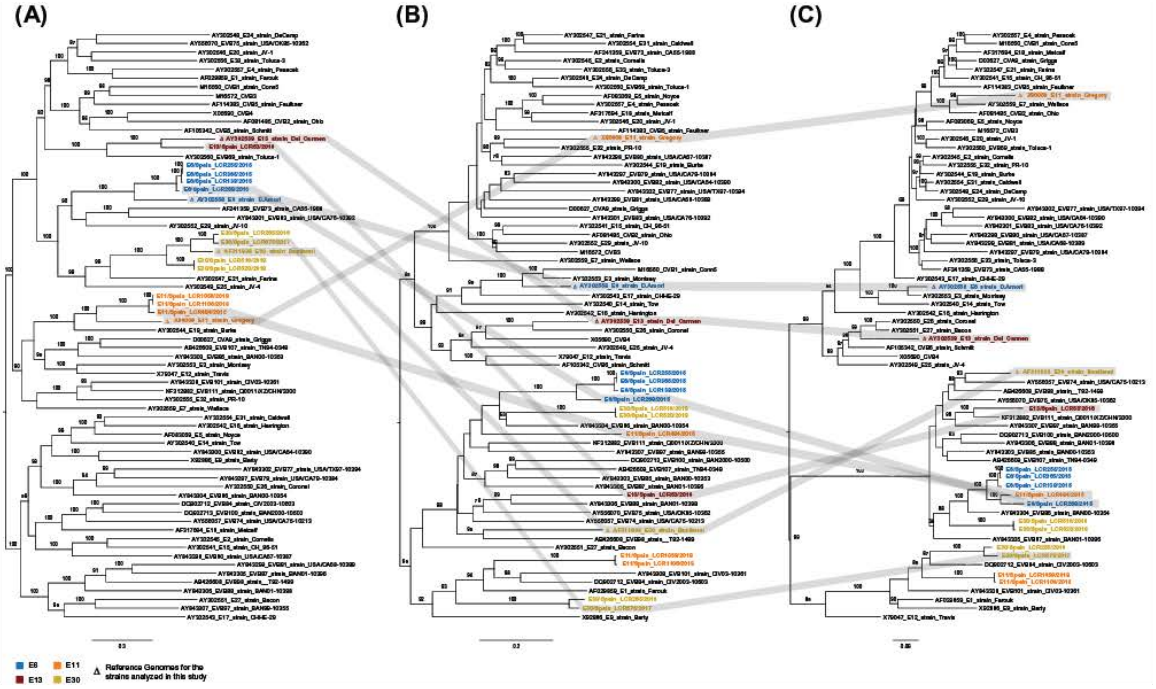


Figure 2. Phylogenetic analysis based on the P1, P2 and P3 coding sequences of the 12 study strains and other fully sequenced EV-B global strains. Maximum-likelihood phylogenetic trees were inferred using IQ-TREE2 on an alignment of ~ 2565, 1734 and 2268 nt sequences corresponding to the P1 (**A**), P2 (**B**), and P3 (**C**) coding regions respectively. The numbers at the nodes indicate bootstrap support values > 80 for that node. Scale bar represents nucleotide substitutions per site. GenBank accession numbers for published sequences are shown in the tree.

the analysis were those with highest score when using the genomic regions identified in the RDP5.5 analysis as of recombinant origin as queries for BLASTn (http://www.ncbi.nlm.nih.gov/) (Supplementary Table S2).

Genome sequences of E11 study strains were compared with all E6 study strains, the E11 prototype strain and sequences closely related to different parts of LCR484 which was identified as mosaic by all 6 methods within RDP5.5. While most of LCR484 genome presented high identity with the other E11 strains reported here (LCR1106, LCR1059), the 3' end of the genome (from the 3C region) presented high identity with non-E11 sequences. Genomes with similar recombinant genomic organization have been detected previously (Supplementary Fig. S3).

The LCR53 (E13) genome also presented strong signal of recombination, both intra and intertypic. In particular, we noted signal of recombination with other E13 strains in the 5' (P1 and P2 regions) of the genome. Most importantly, we detected intertypic recombination signal in the 3'end of the genome (from the 3C region) (Fig. 3). No similar recombinant genomes have been reported.

Among the E30 strains reported here, those from lineage I (LCR 519 and LCR520) and those from lineage F (LCR675 and LCR265) were detected as of recombinant origin, with signal in the 3'end of the genome (Supplementary Fig. S3). However, like for E11, genomes with similar mosaicism have been reported in the Netherland during the same year (Supplementary Fig. S3).

Discussion

Application of mNGS methods to characterize the whole-genome sequence of viruses has become a valuable tool for tracking viral evolution, to elucidate the mechanisms shaping their diversity and to better understand the biology of viruses associated with human disease[24]. In the present study, we used a previously described[25] protocol for mNGS of clinical samples on twelve CSF samples confirmed as EV-positive and collected from patients with aseptic meningitis during 2015–2018 in Southern Spain. It is important to note that the clinical samples were previously frozen and thawed at least once and stored at − 45 °C. Our approach allowed to obtain complete or near-genomes despite the length of storage (samples were collected from 2015–2018 and processed in 2019), freeze–thaw cycles and the storage temperature above − 480 °C. This suggests that this method can be effectively implemented in public health reference laboratories to detect and reconstruct complete viral genomes from CSF samples even if there has been a progressive degradation of RNAs. Moreover, unlike PCR amplicon-based or capture-based enrichment methods, this metagenomic approach does not require previous knowledge of the pathogen sequence, making this approach suitable to characterize novel viruses or recombinant forms. This is important as upsurges of emerging recombinant NPEVs associated with neurologic manifestations such

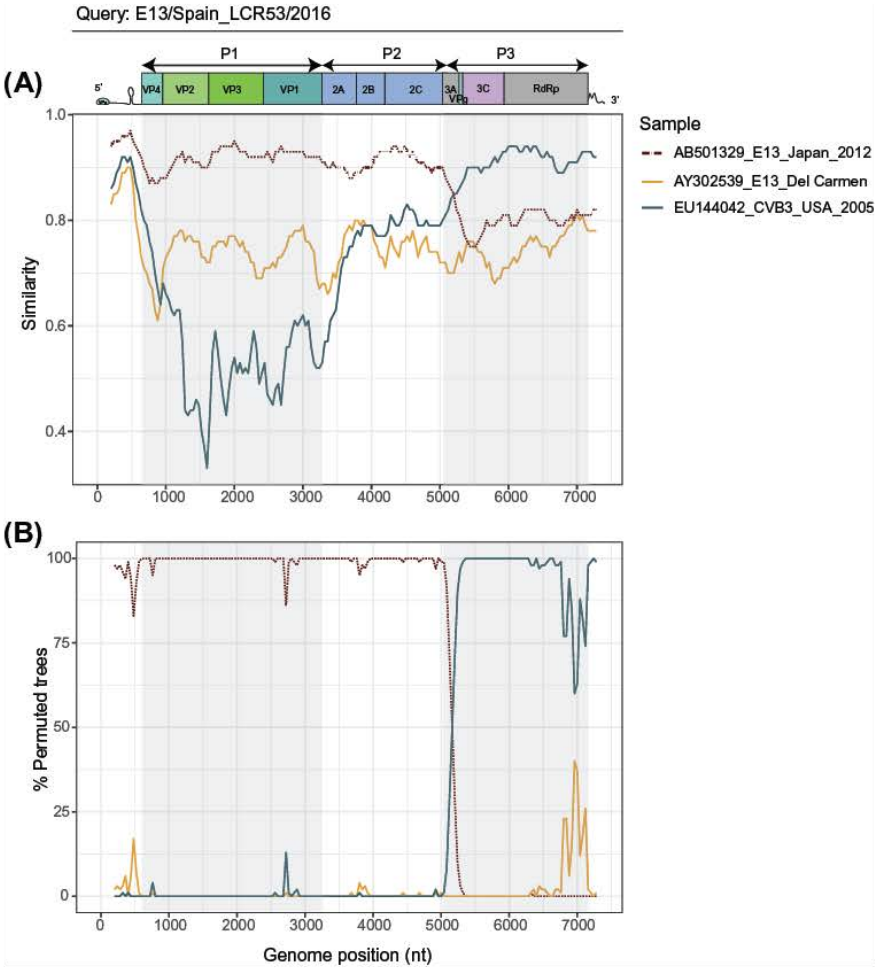


Figure 3. Plot of similarity (**A**) and bootscanning analysis (**B**) of E13/Spain_LCR53/2016 study strain with prototype and closely related strains. The enterovirus genetic organization is shown in the top panel. Analyses were conducted by using SimPlot 3.5.1 (Kimura distance model, window size 400 bp moving in 40 nucleotides steps).

as EVA71 or E30 are a growing public health concern in Europe[26,27]. A real-time detection and rapid sharing of whole-genome sequence information on these emerging NPEVs from public health reference laboratories may help detect outbreaks of severe neurologic disease and implement early intervention strategies ensuring rapid control of disease spread.

The types identified in this study were E6, E11, E13 and E30. Echoviruses belong to the EV-B species and constitute the largest group of the EV genus, with 28 types. A meta-analysis of the worldwide distribution of EVs showed that E30, E6 and E13 were the most commonly detected human EV types in meningitis cases[28]. In Europe, between 2015 and 2017, E30, E6 and E11 were among the ten most frequently reported types representing 12%, 12% and 4%, respectively, of all typed EV-positive samples[5]. In 2018, an E30 upsurge was observed in Denmark, Germany, The Netherlands, Norway and Sweden associated with meningitis or meningoencephalitis[26]. A recent study identified that this upsurge was caused by co-circulation of E30 viruses from two different lineages: E and I (corresponding to clades G1 and G6, respectively) that replaced E30 viruses from lineage F (corresponding to clade G2) which predominated in 2016 and 2017[21]. We similarly note that the E30 strains detected in Spain in 2017 are from lineage F, and that the subsequent strains (from 2018) correspond to the emerging lineage I. The close genetic relatedness of the E30 sequences reported in lineage I during 2017–2018 in the Netherlands, France, Ireland and Spain is further indicative of rapid widespread transmission of this lineage[29].

In Spain, from 1988 to 2008, E30 has been the predominant EV type (33.7% of the total typed EVs)[8,30]. However, in 2009, an upsurge of E6 (60%) replaced E30 as the predominant type. In more recent years (2016–2019), E30 and E6 were among the ten most prevalent types of all typed EVs in Spain[6,31]. All E6 study strains analyzed in this study were from 2015 and segregated into the same sublineage C1, together with previously described

strains from France and Poland isolated in 2011, indicating that related C1 strains were circulating during those years over a broad geographic area in Europe. Interestingly, E6 strains circulating in Spain (n = 86) in the previous decade (2000–2010), all clustered into sublineage C9 suggesting again, as was observed for E30, a genetic turnover where C9 viruses might have subsequently be replaced by C1 viruses.

In this study, phylogenetic analysis showed that E11 study strains from 2015 and from 2018 belong to the same genomic cluster (subgenotype D5) which is distributed worldwide. Study strains have close relationship with each other (> 95% identity) suggesting sustained local circulation in the region during the study period. The fact that strains detected in Spain from 2002 to 2008 also cluster in D5 supports the idea that this subgenotype has persisted for almost two decades in the country and is in agreement with previous studies describing relative genetic stability of E11 strains, associated with long-term endemic patterns[13,32]. Of note, although E11 study strains grouped within the previously described D5 cluster, they formed a separate subcluster suggesting genetic evolution of this endemic E11 viral population from a common ancestor.

Consistently with Europe where E13 is among the less reported types (< 1%), E13 is also rarely reported in Spain[5,6,33]. It appears to circulate in a cyclical fashion with epidemic years followed by years with low detection rates[33]. In Spain, E13 reporting increased in 2000 and 2019 (45% and 8% of the total typed EVs, respectively) with almost no detection in between[6,33]. In this study, we detected a sporadic case of aseptic meningitis associated with an E13 strain from 2016 closely related to E13 strains from France also collected in 2016 from children with meningitis. This close genetic relatedness between E13 sequences reported in 2016 in both France and Spain suggests rapid widespread transmission.

Recombination is known to constitute a widespread mechanism of EV evolution, through the continual exchange of fragments between co-circulating viruses which may impact viral replication and pathogenicity. Full genome sequencing is crucial to detect these recombination events. Our analysis revealed that the E6, E11, E13 and E30 circulating types present complex mosaic genomes constituted of fragments of different origins, implicating intertypic recombination events.

Our analysis highlights the continued circulation of recombinant E11 viruses carrying non-structural sequences similar to other (non-E11) types. Evidence of evolution of E6 and E11 through intertypic recombination has been suggested before[34], and the detection of strains with similar mosaic organization in multiple parts of the world (USA, Japan, Italy) indicates that this lineage has been circulating worldwide in recent years, in parallel to other E11 lineages.

Within E30, the lineage F strains reported here (LCR265, 2016 and LCR675, 2017) present a mosaic pattern previously detected in genomes from Spain (Catalonia) in 2016, as well as genomes noted in the Netherlands in 2017. In a similar way, the lineage I study strains (LCR519 and LCR520, 2018) also present mosaic patterns previously detected in genomes from the Netherlands and Ireland in 2017 and from Austria in 2016. These results suggest that these recombinant lineages have been circulating for some time and started to diversify. Multiple reports highlight the involvement of genetic recombination in the evolutionary process of E30 with other EV-B strains in the non-structural region[13,29,35].

The E13 study strain (LCR53) presents a mosaic organization with no reported homologs, suggesting that the recombination event leading to the circulation of this virus might be recent, or that it has been circulating at low noise and was not captured up to now. The detected recombination breakpoint lies at the beginning of the P3 region, providing further support to the idea that the structural and non-structural genome regions of EV-B types evolve independantly[36]. To better monitor recombination events and more accurately identify potential recombination partners or location of emergence (including potential geographical hotspots), stronger genomic surveillance programs for EVs are needed.

In Europe, EVs are more commonly detected in late summer and autumn[5]. Our study revealed that most types (E6, E13 and E30) were detected during the spring season (March–May) which is consistent with previous reports describing the incidence of EV-B in Spain peaking during spring[8,30,37]. Rather unexpectedly, all three cases associated to E11 occurred in winter. Different seasonality in circulating EV types and the fact that seasonality in temperate zones such as Spain is less marked than in tropical climates could explain these differences.

One limitation of this study is the limited number of samples here analyzed within a 3-year period (only severe cases requiring hospitalization). For this reason, we are likely underestimating the genetic diversity of EVs in Southern Spain. This emphasizes the need to improve the genomic surveillance of EVs in Spain by screening larger cohorts of patients and obtaining detailed epidemiological and clinical data. This could be complemented with increased nationwide surveillance and further exploration of environmental samples using mNGS to monitor emergent and/or novel recombinant types and to better understand their circulation patterns and potential recombination events among co-circulating EVs.

In conclusion, this study documents for the first time the whole-genome sequences of E6, E11 and E13, as well as E30 circulating in Spain. The detection of a novel, disease-associated, and recombinant form highlights the importance of genomic surveillance characterizing full-length EV genomes and public sharing of whole-genome sequence data.

Materials and methods

Ethics statement. Ethical approval for the experimental protocol was given by the Research Ethics Committee from Cordoba (Comité de Ética de Investigación de Córdoba, reference 201999903552445). The procedures were carried out in accordance with approved guidelines, regulations and the principles of the Declaration of Helsinki. Written informed consent was obtained from all patients or from a parent and/or legal guardian.

Study background. From January 2015 to December 2018, eleven adult patients were diagnosed in the Department of Neurology in the University Hospital Reina Sofia (Córdoba, Spain) with a laboratory-confirmed

enteroviral meningitis. The Department of Neurology is responsible for the treatment of patients from 14 years old and is the reference unit for patients with suspected CNS infections in the province. The University Hospital Reina Sofia is a 1280-bed tertiary referral hospital for the southwestern Spanish province of Córdoba (population 461.078). A case was defined as meningitis-suspected based upon the presence of these symptoms: fever, headache, vomiting, neck stiffness, nausea and sometimes accompanied by photophobia, abdominal pain or diarrhea. CSF samples were obtained at patient admission according to standard procedures and analyzed at the Department of Microbiology. Samples were tested for the presence of EVs using a qualitative multiplexed PCR (FilmArray® Meningitis/Encephalitis Panel, BioFire Diagnostics) or RT-PCR (Xpert EV®, Cepheid). In all CSF samples, bacterial, herpes simplex virus types 1 and 2, and varicella-zoster virus infections of the CNS were excluded by culture or PCR. Bacterial screening was performed by culture on chocolate agar, blood agar and thioglycollate broth, while herpes simplex virus and varicella zoster virus screening was performed by a qualitative multiplexed PCR (FilmArray® Meningitis/Encephalitis Panel, BioFire Diagnostics) or RT-PCR (RealCycler® herpesvirus type 1 + herpesvirus type 2 + varicella-zoster virus, Progenie Molecular). Sample remnants were stored at −45 °C. Clinical, epidemiological and laboratory data were collected retrospectively from medical records.

Library preparation and metagenomic sequencing. RNA was extracted from 140 µl of CSF using the QIAamp Viral RNA Mini Kit (Qiagen) followed by Turbo DNAse treatment (Ambion) and purification with Agencourt RNAClean XP beads. We next depleted host rRNA using custom probes and RNAse H treatment as previously described[25]. Depleted samples were purified using AMPure RNA clean beads (Beckman Coulter Genomics) and eluted in 10 ul of RNase-free water for cDNA synthesis. RNA was converted to double-stranded cDNA in two steps. First, RNA was retro transcribed using random primers and SuperScript IV (Invitrogen). Second-strand cDNA was generated using E. coli DNA ligase, RNAse H and DNA polymerase (New England Biolabs) and purified using Agencourt AMPure XP beads (Beckman Coulter). Libraries were then prepared using the Nextera XT DNA Library Prep Kit (Illumina) and sequenced on an Illumina NextSeq500 (2 × 75 cycles).

Genome assembly and analysis of sequence data. Raw reads were trimmed using Trimmomatic v0.36 to remove low-quality reads and Illumina adaptors. Reads were de novo assembled using the metaspades option from SPAdes v3.12 and the contigs obtained were used as blast queries on Virus Pathogen Resource (ViPR)[38]. We used clc-assembly-cell v5.1.0 to perform mapping and extract consensus. A minimum of 3X read depth coverage was used, and N were added in case of lower coverage. Samtools v1.3 was used to sort the aligned BAM files and to generate alignment statistics. All alignments and consensus sequences were manually inspected using Geneious Prime 2020.2 (https://www.geneious.com/).

Recombination analysis. Recombination analyses were carried out on the study strain using a background of randomly sampled EV genomes retrieved from GenBank on May 2021, using a combination of six methods implemented in RDP5.5[23] (RDP, GENECONV, MaxChi, Bootscan, SisScan and 3SEQ) and we considered recombination signals detected by more than three methods for breakpoint identification. Except for specifying that sequences are linear, all settings were kept to their defaults. In addition, similarity plot and bootscanning analysis were performed by using the SimPlot program, version 3.5.1, with a 400-nt window moving in 40-nt steps and using a Kimura 2-parameter method with a transitions-transversions ratio of 2 with 1000 resampling.

Phylogenetic analysis. We performed phylogenetic analysis of the 12 genomes from our dataset, together with published genomes available on NCBI GenBank retrieved on May 2021. Because the majority of the reported sequences from strains circulating in Spain consisted of partial VP1 sequences, we first performed phylogenetic analysis based on the alignment of partial 3'-half VP1 region (E6, E30, and E13) or in the 5'-half VP1 region (E11). Nt sequence alignment was performed by using ClustalW multiple alignment program[39] within the BioEdit sequence Alignment Editor package, version 7.0.9.0. Maximum-likelihood trees were constructed using IQ-TREE2 (v.2.0.6)[40] with 1000 bootstrap replicates[41]. We constructed the trees using the best-fit model as determined by ModelFinder[42] implemented in IQ-TREE2. Genetic distances and sequence divergence were calculated using the Molecular Evolutionary Genetics Analysis (MEGA-X) software package (http://megasoftware.net/).

Data availability

Raw sequencing reads were deposited on the European Nucleotide Archive (https://www.ebi.ac.uk/ena/browser/home) with study number PRJEB45068. The GenBank (https://www.ncbi.nlm.nih.gov/genbank/) accession numbers of the assembled virus genomes are MZ389224—MZ389234 and MZ436966. Both ENA and GenBank are part of the International Nucleotide Sequence Database Collaboration (INSDC https://www.insdc.org/).

Received: 16 July 2021; Accepted: 20 October 2021
Published online: 02 November 2021

References

1. Palacios, G. & Oberste, M. S. Enteroviruses as agents of emerging infectious diseases. *J. Neurovirol.* **11**, 424–433 (2005).
2. The Online (10th) Report of the International Commitee on Taxonomy of Viruses (2017). Available at: https://talk.ictvonline.org/ictv-reports/ictv_online_report/positive-sense-rna-viruses/picornavirales/w/picornaviridae/681/genus-enterovirus.
3. Oberste, M. S., Maher, K. & Pallansch, M. A. Evidence for frequent recombination within species human enterovirus B based on complete genomic sequences of all thirty-seven serotypes. *J. Virol.* **78**, 855–867 (2004).

4. Harvala, H. *et al.* Recommendations for enterovirus diagnostics and characterisation within and beyond Europe. *J. Clin. Virol.* https://doi.org/10.1016/j.jcv.2018.01.008 (2018).
5. Bubba, L. *et al.* Circulation of non-polio enteroviruses in 24 EU and EEA countries between 2015 and 2017: A retrospective surveillance study. *Lancet. Infect. Dis* https://doi.org/10.1016/S1473-3099(19)30566-3 (2020).
6. Annual Epidemiological reports of Acute Flaccid Paralysis Surveillance and Enterovirus Surveillance, Spain, 2016–2019. Available at: https://www.isciii.es/QueHacemos/Servicios/VigilanciaSaludPublicaRENAVE/EnfermedadesTransmisibles/Paginas/Resultados_Vigilancia_Polio.aspx.
7. Cabrerizo, M., Echevarria, J. E., González, I., De Miguel, T. & Trallero, G. Molecular epidemiological study of HEV-B enteroviruses involved in the increase in meningitis cases occurred in Spain during 2006. *J. Med. Virol.* https://doi.org/10.1002/jmv.21197 (2008).
8. Trallero, G. *et al.* Enteroviruses in Spain over the decade 1998–2007: Virological and epidemiological studies. *J. Clin. Virol.* https://doi.org/10.1016/j.jcv.2009.11.013 (2010).
9. Andrés, C. *et al.* Recombinant CV-A6 strains related to hand-foot-mouth disease and herpangina at primary care centers (Barcelona, Spain). *Fut. Microbiol.* https://doi.org/10.2217/fmb-2018-0336 (2019).
10. Cabrerizo, M., Trallero, G. & Simmonds, P. Recombination and evolutionary dynamics of human echovirus 6. *J. Med. Virol.* https://doi.org/10.1002/jmv.23741 (2014).
11. Calderón, K. I. *et al.* Molecular epidemiology of coxsackievirus B3 infection in Spain, 2004–2014. *Adv. Virol.* https://doi.org/10.1007/s00705-016-2783-1 (2016).
12. Leitch, E. C. M. *et al.* Transmission networks and population turnover of echovirus 30. *J. Virol.* https://doi.org/10.1128/jvi.02109-08 (2009).
13. McWilliam Leitch, E. C. *et al.* Evolutionary dynamics and temporal/geographical correlates of recombination in the human enterovirus echovirus types 9, 11, and 30. *J. Virol.* https://doi.org/10.1128/jvi.00783-10 (2010).
14. McWilliam Leitch, E. C. *et al.* The association of recombination events in the founding and emergence of subgenogroup evolutionary lineages of human enterovirus 71. *J. Virol.* **86**, 2676–2685 (2012).
15. González-Sanz, R. *et al.* Molecular epidemiology of an enterovirus A71 outbreak associated with severe neurological disease, Spain, 2016. *Eurosurveillance* https://doi.org/10.2807/1560-7917.ES.2019.24.7.1800089 (2019).
16. Leon, K. E. *et al.* Genomic and serologic characterization of enterovirus A71 brainstem encephalitis. *Neurol. Neuroimmunol. Neuroinflam.* https://doi.org/10.1212/NXI.0000000000000703 (2020).
17. Puenpa, J. *et al.* Molecular epidemiology and the evolution of human coxsackievirus A6. *J. Gen. Virol.* https://doi.org/10.1099/jgv.0.000619 (2016).
18. de Lima, S. T. S. *et al.* Fatal outcome of chikungunya virus infection in Brazil. *Clin. Infect. Dis.* https://doi.org/10.1093/cid/ciaa1038 (2020).
19. Smura, T. *et al.* Molecular evolution and epidemiology of echovirus 6 in Finland. *Infect. Genet. Evol.* **16**, 234–247 (2013).
20. Bailly, J. L. *et al.* Phylogeography of circulating populations of human echovirus 30 over 50 years: Nucleotide polymorphism and signature of purifying selection in the VP1 capsid protein gene. *Infect. Genet. Evol.* https://doi.org/10.1016/j.meegid.2008.04.009 (2009).
21. Benschop, K. *et al.* Molecular epidemiology and evolutionary trajectory of emerging echovirus 30. *Eur. Emerg. Infect. Dis.* **27**(6), 1616–1626. https://doi.org/10.3201/eid2706.203096 (2021).
22. Li, J. *et al.* Multiple genotypes of Echovirus 11 circulated in mainland China between 1994 and 2017. *Sci. Rep.* https://doi.org/10.1038/s41598-019-46870-w (2019).
23. Martin, D. P. *et al.* RDP5: A computer program for analyzing recombination in, and removing signals of recombination from, nucleotide sequence datasets. *Virus Evol.* https://doi.org/10.1093/ve/veaa087 (2021).
24. Houldcroft, C. J., Beale, M. A. & Breuer, J. Clinical and biological insights from viral genome sequencing. *Nat. Rev. Microbiol.* https://doi.org/10.1038/nrmicro.2016.182 (2017).
25. Matranga, C. B. *et al.* Enhanced methods for unbiased deep sequencing of Lassa and Ebola RNA viruses from clinical and biological samples. *Genome Biol.* **15**, 519 (2014).
26. Broberg, E. K. *et al.* Upsurge in echovirus 30 detections in five EU/EEA countries, April to September, 2018. *Eurosurveillance* https://doi.org/10.2807/1560-7917.ES.2018.23.44.1800537 (2018).
27. Ngangas, S. T. *et al.* Multirecombinant enterovirus A71 subgenogroup C1 isolates associated with neurologic disease, France, 2016–2017. *Emerg. Infect. Dis.* https://doi.org/10.3201/eid2506.181460 (2019).
28. Suresh, S., Rawlinson, W. D., Andrews, P. I. & Stelzer-Braid, S. Global epidemiology of nonpolio enteroviruses causing severe neurological complications: A systematic review and meta-analysis. *Rev. Med. Virol.* https://doi.org/10.1002/rmv.2082 (2020).
29. Mirand, A., Henquell, C., Archimbaud, C., Piegue-Lafeuille, H. & Bailly, J. L. Emergence of recent echovirus 30 lineages is marked by serial genetic recombination events. *J. Gen. Virol.* https://doi.org/10.1099/vir.0.82146-0 (2007).
30. Cabrerizo, M., Echevarria, J. E., Gonzalez, I., de Miguel, T. & Trallero, G. Molecular epidemiological study of HEV-B enteroviruses involved in the increase in meningitis cases occurred in Spain during 2006. *J. Med. Virol.* **80**, 1018–1024 (2008).
31. Cabrerizo, M. *et al.* Molecular epidemiology of enterovirus and parechovirus infections according to patient age over a 4-year period in Spain. *J. Med. Virol.* https://doi.org/10.1002/jmv.24658 (2017).
32. Oberste, M. S., Nix, W. A., Kilpatrick, D. R., Flemister, M. R. & Pallansch, M. A. Molecular epidemiology and type-specific detection of echovirus 11 isolates from the Americas, Europe, Africa, Australia, southern Asia and the Middle East. *Virus Res.* https://doi.org/10.1016/S0168-1702(02)00291-5 (2003).
33. Trallero, G. *et al.* First epidemic of aseptic meningitis due to echovirus type 13 among Spanish children. *Epidemiol. Infect.* https://doi.org/10.1017/S0950268802008191 (2003).
34. Su, T. *et al.* Molecular characterization of a new human echovirus 11 isolate associated with severe hand, foot and mouth disease in Yunnan, China, in 2010. *Adv. Virol.* https://doi.org/10.1007/s00705-015-2496-x (2015).
35. Lukashev, A. N., Ivanova, O. E., Eremeeva, T. P. & Gmyl, L. V. Analysis of echovirus 30 isolates from Russia and new independent states revealing frequent recombination and reemergence of ancient lineages. *J. Clin. Microbiol.* https://doi.org/10.1128/JCM.02386-06 (2008).
36. Lukashev, A. N. *et al.* Recombination in circulating Human enterovirus B: Independent evolution of structural and non-structural genome regions. *J. Gen. Virol.* https://doi.org/10.1099/vir.0.81264-0 (2005).
37. Rodà, D. *et al.* Clinical characteristics and molecular epidemiology of Enterovirus infection in infants <3 months in a referral paediatric hospital of Barcelona. *Eur. J. Pediatr.* https://doi.org/10.1007/s00431-015-2571-z (2015).
38. Pickett, B. E. *et al.* ViPR: An open bioinformatics database and analysis resource for virology research. *Nucleic Acids Res.* **40**, 2 (2012).
39. Thompson, J. D., Higgins, D. G. & Gibson, T. J. CLUSTAL W: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.* https://doi.org/10.1093/nar/22.22.4673 (1994).
40. Minh, B. Q. *et al.* IQ-TREE 2: New models and efficient methods for phylogenetic inference in the genomic era. *Mol. Biol. Evol.* https://doi.org/10.1093/molbev/msaa015 (2020).
41. Hoang, D. T., Chernomor, O., Von Haeseler, A., Minh, B. Q. & Vinh, L. S. UFBoot2: Improving the ultrafast bootstrap approximation. *Mol. Biol. Evol.* https://doi.org/10.1093/molbev/msx281 (2018).

42. Kalyaanamoorthy, S., Minh, B. Q., Wong, T. K. F., Von Haeseler, A. & Jermiin, L. S. ModelFinder: Fast model selection for accurate phylogenetic estimates. *Nat. Methods* https://doi.org/10.1038/nmeth.4285 (2017).

Acknowledgements
This work was funded by the Junta de Andalucía, Consejería de Salud y Familias (Project number PI-0216-2019) and by the Acción Estratégica en Salud Intramural (Project number PI20CIII/00005). MD Fernandez-Garcia received a Miguel Servet Research Contract (CP18/00067) from the Strategic Action in Health 2018 and funded by National Institute of Health Carlos III (ISCIII). ESL acknowledges funding from the INCEPTION programme (Investissements d'Avenir grant ANR-16-CONV-0005).

Author contributions
Conceptualization, M.D.F.G.; Coordinated and supervised the study, E.S.L., M.D.F.G.; Funding acquisition, E.S.L., M.D.F.G.; Collection of clinical samples and clinical information, E.A.; Performed experiments, F.G., A.B.P., M.P., M.D.F.G.; Analyzed and interpreted the data, E.S.L., F.G., M.D.F.G.; Writing—Original draft preparation, E.S.L., M.D.F.G., F.G.; Writing—Review and editing, L.M.M., M.C., A.B.P., E.A.; All authors have read and agreed to the published version of the manuscript.

Competing interests
The authors declare no competing interests.

Additional information
Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-021-01053-4.

Correspondence and requests for materials should be addressed to E.S.-L. or M.D.F.-G.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

5.3 Discussion and conclusion about the studies

mNGS has become increasingly important for basic research and clinical medicine (111). That is because we can use mNGS for a wide range of applications: from the identification of pathogens that are novel or highly divergent from previously catalogued pathogens, including new recombinant forms of the virus; to the detection of different known drug-resistant variants in a single test which might be present as majority variants or at low levels before the resistance becomes clinically apparent.

In particular, in our study, we attempt to identify the potential etiologic agent for a series of aseptic meningitis cases in Southern Spain and generate, in an unbiased manner, full-length genomes from EV-positive aseptic meningitis cases. Because the vast majority of viruses known to be associated with meningitis have RNA genomes (e.g., EVs) (182), we decided to implement a mNGS protocol targeting in particular RNA. Such a sequencing protocol is also known as RNA-sequencing (RNA-seq). Although this method could also detect viral mRNAs, DNA viruses that might experience low-level transcription might be poorly detected. For this reason, our analysis focused on RNA viruses known to cause meningitis, and we recognize it as one of the limitations of our study.

Regarding our first goal, we identified TOSV in several samples. Although we could not generate complete genomes from all of them, the mNGS findings allowed us to establish an amplicon-based sequencing approach to obtain complete viral genomes from samples of variable viral load and quality. Such sequencing protocol will hopefully be helpful to the scientific community as it allows to detect and generate complete TOSV genomes from clinical samples of suboptimal quality for total RNA-seq. Such addition might be of great importance as it would allow the TOSV genomic database to enlarge at a lower cost. Currently, in the ViPR database, there are less than 20 full-length genomes with complete associated metadata. Having more genomic data available would enable us to perform further analysis (e.g., phylodynamic studies) to better address the epidemiology and evolution of this emerging virus.

Among the enteroviral meningitis cases, we succeeded in reconstructing full-length genomes in 100% of the cases, documenting the first complete genomes of E6, E11, E13, and E30 co-circulating in Spain. This result aligns with a previous study that reported that the most common human EV types detected in meningitis cases are E6, E11, E13, and E30 (225). In addition, generating full-length genomes allowed us to look for evidence of recombination: an important mechanism of evolution for EVs. Interestingly, this analysis showed strong support for intertypic recombination in one of the genomes, leading to the description of a novel E13 recombinant form. The detection of this novel and disease-associated recombinant form was an important finding of this study as it illustrates how the co-circulation of several EV types in a region during a single year, as also demonstrated here, can provide a suitable environment for the appearance of novel recombinant variants. New emerging EVs have shown the potential to cause epidemics with devastating consequences, for instance, the EV-A71 epidemics in the Asia-Pacific Region (226). Altogether, this highlights the need to detect and monitor these viruses closely.

Although mNGS offers numerous advantages, like any other test, it also has drawbacks. Most human clinical samples have high levels of host genetic material. Therefore a key disadvantage of using mNGS on such samples is that, given its unbiased nature, the vast majority of the reads will be derived from the human host, limiting pathogen detection (227). However, this disadvantage can be mitigated, for example, by using host depletion methods. In particular, we followed a protocol described by Matranga et al. (171). Briefly, after RNA extraction, samples were treated to remove double-stranded DNA. Next, host ribosomal

RNA present in human clinical samples and poly (A) carrier RNA (usually added to enhance RNA extraction) are depleted using custom probes and oligo (dT), respectively, to form RNase H-cleavable DNA-RNA hybrids. These two steps helped eliminate the human host background, enriching the samples with RNA viral content.

As already mentioned, mNGS is a powerful and highly sensitive technique to detect RNA or DNA. However, this high sensitivity is a double-edged sword as the smallest amount of exogenous DNA or RNA introduced from the environment, reagents, handlers, or equipment; will also be detected. Therefore, another potential drawback of mNGS is the detection of microbial contaminants in the samples, which can obscure the analysis and interpretation of the results (171). Sample contamination is not a new issue; several articles have comprehensively described it (112, 186, 228). Nevertheless, is important to highlight two points. First, the contaminants can arise at different stages of the mNGS protocol: during the sample extraction and aliquoting, nucleic acid extraction, library preparation, or the sequencing run. Second, in low-biomass samples, such as CSF samples, contaminants can be detected at the same or even higher level than bona fide pathogens, making it difficult to interpret the data (229).

Despite all the precautions taken, we observed contamination among our samples. The most common source of contamination observed was the sequencing run. Several samples (9/44) were contaminated with samples from the same run containing high pathogen loads (e.g., viral stocks, samples with amplified viral content). This well-known problem might occur when using Illumina sequencers, named “index-hopping” (227). In short, index-hopping can occur due to two phenomena: (i) failure to remove free adapters of prepared libraries before proceeding to the pooling and sequencing; (ii) high-frequency indices might be incorrectly assigned during scanning of the flow cell (227, 230). Consequently, after the demultiplexing, reads belonging to a library could be incorrectly assigned to another, leading to inaccurate sequencing results. Although we followed most of the recommended practices to reduce index hoping (e.g., removing adapters after library preparation, using dual indexed libraries), we sequenced together libraries from very different samples: CSF and viral stocks. We could plan the pooling of the samples differently to improve future sequencing.

Nevertheless, contaminations occurring during the sequencing run are easy to deal with as one can quickly identify them. More difficult is to deal with contaminations that might come from the skin flora or the hospital facility, especially when working with CFS samples collected in different establishments and through various years, as is our case. As previously discussed, pathogen loads are typically low in CSF and, therefore, there are usually very few reads aligning to the potential etiologic agent. Under these circumstances, even minimal quantities of environmental contaminants can be amplified (e.g., during library preparation), increasing the proportional representation of ‘pollutants’ in the final dataset. For this reason we included non-meningitis CSF samples as negative controls to differentiate potential infectious pathogens from contaminants. In this way, we could construct a background dataset to determine if the pathogen detected in the idiopathic samples could be the potential etiologic pathogen or not. This approach worked well when investigating the presence of RNA viruses of interest, such EVs, MuV, or TOSV, among our samples.

We detected RNA of TOSV in 8/23 idiopathic samples, which raises the following question: what is the potential etiologic agent for the 15 other cases?

First of all, like every other method, mNGS has a detection limit[1]. Therefore, the potential etiologic agent for these other cases might be any pathogen present in the sample below this detection limit. In this regard, the quality of the samples might be an important factor influencing the detection of pathogens. We recognized that this is one of the limitations of our study as the samples used in this study were not stored in optimal conditions (stored at -45°C until processed in 2019 and subjected to more than one freeze-thaw cycle), probably affecting the quality of the RNA.

Nonetheless, to answer this question, the next was to make a more comprehensive analysis of all the pathogens identified by the amino acid sequence similarity search. However, the number of taxonomic units was quite large (n=1525). The first approach focused on the taxonomic identifiers present in the idiopathic and positive control samples, which reduced the number to 839. Taking a closer look at these identifiers, we observed that in some cases, they belonged to non-human viruses (e.g., avian leucosis virus). Using a Virus-Host database could help address this issue quickly by removing those viruses whose hosts are non-human. Such a database exists, called "Virion" (232). However, we realized that some NCBI taxonomic identifiers are not present in Virion —for example, TOSV, a virus essential for our study, rendering the approach relying on purely taxonomic identifier matches challenging. More in-depth, comprehensive, and exhaustive (with many synonymous species identifiers) taxonomic annotation of contigs would be necessary to address this issue. In addition, in other cases, despite using a viral database to perform the similarity search of the contigs, there were non-viral taxons detected (e.g., Acinetobacter). Specifically, the viral database used was RVDB (Reference Viral Database) (233). It consists of a collection of all currently known viral genomes, virus-related and virus-like nucleic sequences retrieved from NCBI using keywords (e.g., viral). In addition, negative keywords are used to exclude non-viral sequences (233). The RVDB is a curated database subjected to manual and computational reviewing processes to eliminate non-viral sequences such as cloning vectors and wrongly annotated sequences. Despite this reviewing process, non-viral sequences may remain, explaining why we detected non-viral taxon, highlighting the difficulties associated with dealing with databases.

Two other strategies could have been explored to facilitate the discrimination between bona fide pathogens versus contaminants.

The first one is adding an internal spike-in control. Spike-in control may consist of whole organisms or synthetic DNA or RNA sequences added to the original sample or at different workflow stages (254). One example of RNA spike-in control is the use of External RNA Controls Consortium spike-ins, which are commercially available. As suggested by Zinter et al. (228), including a series of spike-in controls in the samples would allow the calculation of total sample input mass. Once this quantity is known, the correlation between total sample input mass and the number of reads targeting a given pathogen can be studied. If the result is an inverse correlation, the given pathogen is suspicious for contamination. Additionally, this method allows for detecting outliers; this means identifying samples in which the pathogen is a bona fide pathogen (228).

Second, use machine learning methods to distinguish potential etiologic pathogens from ubiquitous environmental contaminants and commensal flora. This method has already been published (177) and implemented in other mNGS studies (234). In short, this method generates a "pathogen database" using

[1] Limit of detection is defined as the lowest concentration of the analyte that can be reliably detected. 231. Lindon JC, Tranter GE, Koppenaal D. Encyclopedia of spectroscopy and spectrometry: Academic Press; 2016.

no-template and healthy patients control samples. According to this database, each pathogen gets a score. Those with a score below a certain threshold are removed from the analysis.

While a comprehensive and exhaustive analysis of all the taxonomic identifiers detected in the samples discriminating between bona fide pathogens versus contaminants is an interesting continuation of the project, our study focused on studying RNA viruses known to cause meningitis.

Nevertheless, we hope that our analysis will be a good contribution in the future to enlarge databases of patient mNGS, thereby enhancing the ability to discriminate between spurious sequences and potential pathogens. Ultimately, this will broaden the list of pathogens known to be detectable by the mNGS, providing a more solid ground for pathogen identification and reporting.

Key points of Chapter 1

- We implemented RNA mNGS on CSF samples with known neurological infections (n = 13), with idiopathic meningitis (n = 23), and with no infection (n=8). We generated full-length genomes from EV-positive and MuV-positive meningitis cases and detected TOSV in 8 out of 23 idiopathic samples.
- Our study documents the circulation of several EV-B types (E6, E11, E13, and E30), including a novel E13 recombinant form associated with meningitis in the Spanish population.
- We developed an amplicon-based sequencing approach to help sequence low concentrated or partially degraded samples corresponding to TOSV genotype B, improving the detection and generation of genomic data.

6 CHAPTER 2: Understanding CHIKV re-emergence and spread in Cambodia during the 2011-2013 and 2020 outbreaks

The following chapter reports on our work on CHIKV in Cambodia. The chapter begins with a brief introduction to CHIKV biology, epidemiology, and evolution. The subsequent two sections outline our work concerning the emergence and diffusion of CHIKV in Cambodia during the 2011-2013 three-year epidemic and the 2020 outbreak. Lastly, the chapter is wrapped up with a general discussion and conclusions about the two studies.

6.1 Background: Chikungunya virus biology

6.1.1 Genome organization and life cycle

CHIKV is a single-stranded, positive sense and non-segmented RNA virus which belongs to the alphavirus genus, within the *Togaviridae* family, together with other known viruses such as O'nyong-nyong virus or Sindbis virus. CHIKV can be transmitted to humans mainly by *Aedes aegypti* and *Aedes albopictus* mosquitoes (235).

CHIKV genome is approximately 12Kb long and it encodes four non-structural proteins (nsP1–4) that constitute the replication complex, and five structural proteins (C, E3, E2, 6K, and E1) that give rise to the mature virion (236, 237) (Figure 6-1). The genome is encapsulated in an enveloped icosahedral particle of approximately 70 nm in diameter formed by host-cell derived lipid bilayer in which heterodimers of the E1 and E2 glycoproteins are assembled, forming trimmers. They constitute the spikes on the virus surface (Figures 6-1 B and C) and thus mediate contact between the virus and the host cell. In particular, the E1 protein contains a hydrophobic fusion peptide necessary for cellular and viral membrane fusion, whereas the E2 protein is thought to be responsible for receptor binding as it is the main target of neutralizing antibodies (237, 238).

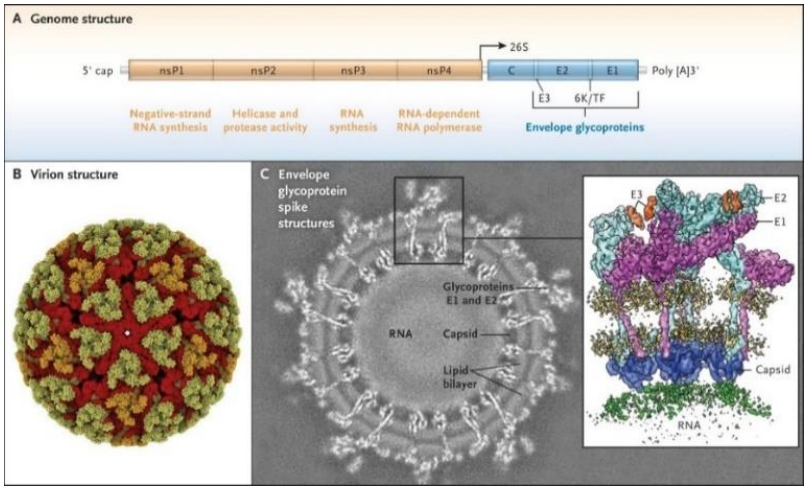


Figure 6-1: *CHIKV genome and virion structure. A) Schematic organization of CHIKV genome, including its non-structural proteins and five structural proteins. B) Structure of the virion. C) Structure of envelope glycoprotein spike proteins predicted by atomic resolution and cryoelectron microscopic reconstructions. Image extracted from Weaver and Lecuit, 2015.*

After recognition of host receptors by E2, CHIKV enters the host cells through clathrin-mediated endocytosis (236, 237, 239) (Figure 6-2). The endosome containing the viral particle subsequently matures, triggering a drop in pH which leads to a conformational change of E1-E2 heterodimers exposing the fusion loop. This fusion loop interacts with the endosomal membrane, provoking membrane fusion and the release of the genome into the host-cell cytoplasm (236, 237, 239). As the genome is capped and positive sense, it is directly translated into the cytoplasm by the cellular host machinery. A conserved opal stop codon (UGA) is located between the nsP3 and nsp4 gene in several alphaviruses including CHIKV. The non-structural polyprotein P123 is produced when translation is stopped at the opal termination codon, whereas the polyprotein P1234, which contains the RNA-dependent RNA polymerase (nsP4), is generated when read-through of the stop codon occurs. *In vitro* experiments have determined that this read-through occurs in 5 to 20% of the cases. As a consequence, stoichiometric concentrations of nsP4 are much lower than those of the other non-structural proteins (240, 241). The cleavage of the polyprotein to generate individual proteins is subsequently performed by nsP2. After production of the proteins of the replication complex (nsP1-4), viral RNA is replicated in negative-strand RNA from which several positive genomic and subgenomic RNAs are generated (242). Translation of the subgenomic RNA produces the C–pE2–6K–E1 polyprotein precursor and subsequently the structural proteins, which are essential for virion formation and genome encapsidation. The pE2 and E1 proteins are subjected to further processing in the Golgi complex and exported to the plasma membrane. Finally, the budding of chikungunya virions occurs at the plasma membrane (236, 239).

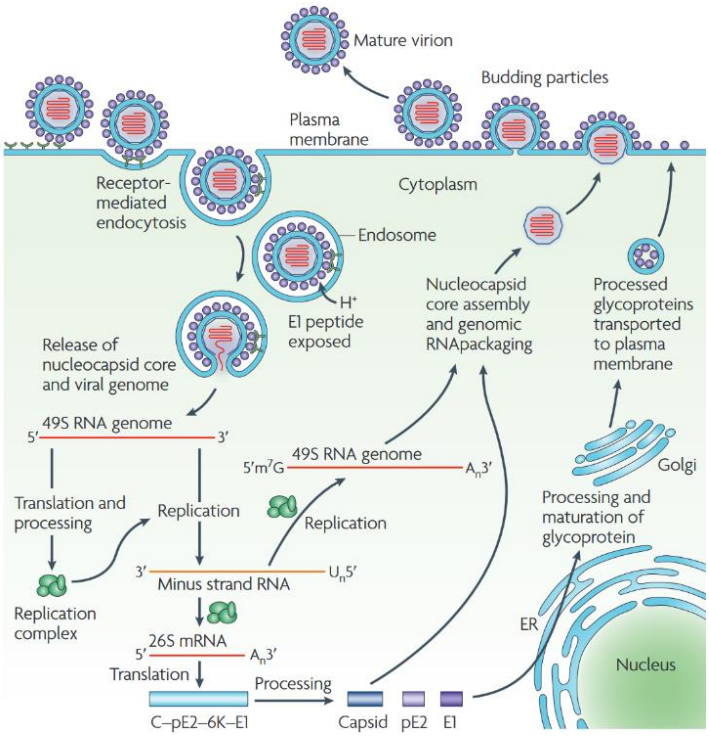


Figure 6-2: *Alphavirus life cycle. After entry by endocytosis, the viral genome is released in the cytoplasm of the cell. Following the production of the proteins of the replication complex (nsP1–4), viral RNA is replicated in minus strand RNA and a subgenomic RNA, coding the structural proteins. Viral RNA is subsequently enwrapped by nucleocapsid proteins and released from the cell through budding. The figure was extracted from Schwartz and Albert, 2010.*

6.1.2 Epidemiology and evolution

CHIKV was first identified somewhere between 1952 and 1953 during an outbreak in Tanzania (243). Since its identification, CHIKV has been classified into different genotypes according to its genomic diversity: West African, East-Central-South-African (ECSA), Asian and, more recently, Indian Ocean lineage (IOL), which emerged as an independent clade from the ECSA genotype (244).

The virus is believed to have originated in Africa (245), where it is mainly maintained in a sylvatic cycle between non-human primates (NHP) and arboreal *Aedes* mosquitoes such as *Aedes furcifer* or *Aedes africanus.* Indeed, CHIKV has been detected in multiple NHPs' species across different countries all over the continent, including Senegal, South Africa, Zimbabwe, Uganda, Gabon and Kenya (246). In Africa, the ECSA and West African genotypes are endemic and cause epidemics through spillover from the sylvatic cycles into human populations and by urban transmission cycles between *humans and* urban mosquitoes, such as *Aedes aegypti* (247) (Figure 6-3). The virus quickly spread to Asia, probably through shipping, with

the first CHIKV case reported in the Philippines in 1954. This introduction gave rise to the Asian genotype, and it is believed that the virus introduced belonged to the ECSA genotype (247). The Asian genotype is maintained in Asia through *Aedes aegypti*-human urban transmission cycles (247). From the 1950s to 2005, sporadic CHIKV emergences in Asia were attributed to the Asian genotype. However, at the end of 2005, a strain of the ECSA lineage spread to the Indian Ocean islands and the Indian continent, causing a massive outbreak, and giving rise to the Indian Ocean Lineage (IOL) (248). This outbreak was of an unprecedented magnitude for CHIKV, with approximately 6.5 million documented cases (249). The epidemic also spread to the United States (250) and Europe (251) through infected travelers returning from these affected areas. In Europe, this introduction led to the first autochthonous CHIKV outbreak in Italy in early 2007, with more than 200 documented cases (153) and, on a smaller scale, to local transmission of CHIKV in Southern France, in 2010 (252).

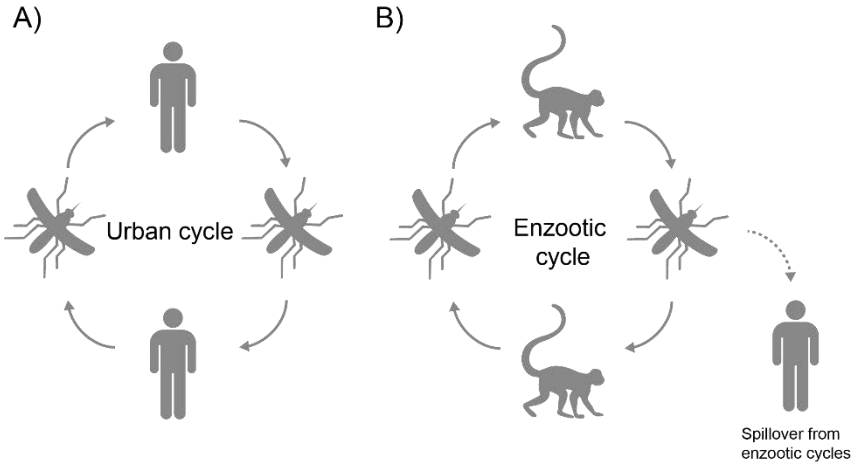


Figure 6-3: *Urban and enzootic CHIKV transmission cycles. A) CHIKV is transmitted by urban transmission cycles between urban mosquitoes such as Aedes aegypti or Aedes albopictus and humans B) The enzootic transmission cycle is maintained among NHP as reservoir and mosquitoes Aedes furcifer or Aedes africanus as primary vectors. Human infection can arise from direct spillover of these enzootic cycles. This figure has been adapted from Weaver and Barret.*

This unprecedented outbreak caught the attention of many researchers, who tried to find out why a CHIKV outbreak of such magnitude took place. The first thing they observed was a potential change of vector. Indeed, although *Aedes aegypti* is considered the classical CHIKV vector, it is believed that the IOL strains might have been primarily transmitted by *Aedes albopictus,* a more abundant mosquito species at that time. Extensive genomic analysis of the CHIKV IOL strains identified previously undescribed mutations (253). In particular, it was observed that in later stages of the outbreak, the IOL strains acquired an AA change in the E1 glycoprotein (E1-A226V). *In vivo* studies later confirmed that this mutation increased the infectivity, dissemination, and transmission for *Aedes albopictus*, allowing CHIKV to be transmitted by a new vector (80, 254). Interestingly, this mutation had no effect with *Aedes aegypti* (80). Several other mutations were later observed to have a positive epistatic effect on the E1-A226V substitution. For instance, in the background of E1-A226V, the substitutions E2-L210Q or E2-K252Q provided an additional fitness increase (255, 256). Interestingly, epidemiological data showed that these second-step *Aedes albopictus* adaptive mutations were simultaneously detected in viruses circulating in India. Laboratory

investigations confirmed that in the background of E1-A226V, the combination of both substitutions (E2-L210Q and E2-K252Q) had an additive effect, further increasing the fitness and dissemination of the virus in *Aedes albopictus* in comparison to the E2-L210Q and E2-K252Q single-mutant competitors (255). All this suggested that the evolution of the IOL conferred CHIKV a selective advantage to infect, replicate, and disseminate in *Aedes albopictus.*

This gave rise to the following question: why were mutations allowing CHIKV to be transmitted through *Aedes albopictus* selected? A possible answer is that several factors make *Aedes albopictus* a suitable vector for viruses such as CHIKV. First, as mentioned in the main introduction, its distribution has been expanding over the years, conquering tropical and temperate regions including Europe and the United States (257, 258). Second, *Aedes albopictus* can tolerate lower temperatures than *Aedes aegypti* (66, 67), and its eggs can remain viable throughout the dry season (239). Third, it is smaller than domestic mosquitoes, which makes it more furtive, and, unlike many other mosquito species, it is diurnal (239). In sum, the mutations present in the IOL allowed the transmission of CHIKV by two different mosquito vectors, *Aedes aegypti* and *Aedes albopictus,* resulting in an efficient human–mosquito transmission cycle. Furthermore, the adaptation to a new vector resulted in a geographic expansion of CHIKV, leading the virus to infect immunologically naïve human populations, ultimately resulting in an outbreak of chikungunya disease of unseen dimensions.

The outbreak in India continued into 2010, with new cases appearing in areas that were not previously affected and since then, the virus has been re-emerging causing several outbreaks almost continuously in Southern and Southeast Asia countries like Bangladesh, Bhutan, Cambodia, China, India, Sri Lanka, Myanmar and Thailand (259-261).

Very interestingly, since 2006 in Southeast Asia, no virus belonging to the Asian genotype has been associated with epidemic activity, suggesting that the newly introduced IOL genotype outcompeted the Asian genotype. Furthermore, although *Aedes albopictus* is very abundant in Asia, there is no evidence that it has played a significant role in transmitting CHIKV strains of the Asian genotype. These two facts raised another interesting question: why mutations that would confer adaptation of the Asian genotype to *Aedes albopictus* were not selected? Also, very interestingly, despite the adaptive advantage of the E1-A226V substitution, this mutation was not detected in Asian lineages. Based on this information, Weaver et al. addressed this question and showed that the single residue E1-98T, present in all Asian CHIKV strains characterized so far, prevents the acquisition of the E1-A226V substitution, and ultimately the adaptation of Asian CHIKV to *Aedes albopictus* (262). This study brought new information regarding (i) the non-implication of *Aedes albopictus,* despite its abundance, as a vector of CHIKV before 2007, and (ii) the shift in CHIKV genotypes in Southeast Asia from the Asian to the IOL genotype, as the latter can exploits the human-*Aedes albopictus* transmission cycle.

In the Americas, the virus was first detected on Saint Martin Island at the beginning of October 2013, and then quickly spread to several countries (263). One year later, in October 2014, autochthonous cases were confirmed in 50 territories across the continent (264). Genetic characterization showed that the circulating CHIKV strains belonged to the Asian genotype and were closely related to the viruses detected in China in 2012 and the Philippines in 2013 (263). Such evidence was not expected as, at least in Southeast Asia, the IOL genotype seemed to have been replacing the Asian genotype. In Brazil, the first local cases of CHIKV were detected by September 2014, and phylogenetic analyses later revealed the co-circulation of both

Asian and ECSA genotypes introduced almost simultaneously in the country (265). The latter was linked to a traveler who had recently returned from Angola (265).

As of December 2021, CHIKV is still circulating in many parts of the world. In the Americas, most cases were reported in Brazil where, according to the PAHO (Pan American Health Organization), there was an increase of 32% of the number of cases between 2020 and 2021 (266). In Southeast Asia, in late 2018, a large CHIKV outbreak started in Thailand (267), and by the end of 2020, 27,000 confirmed cases were documented (268). Almost simultaneously, after nearly ten years of no reported cases, outbreaks of CHIKV infection were identified by the end of 2019 in Myanmar (269) and in 2020 in Cambodia (270). Phylogenetic analyses revealed that the 2018–2020 strains were similar to those that caused the recent outbreak in Bangladesh and Pakistan between 2016 and 2017. They were all characterized by two novel mutations, E1-K211E and E2-V264A, and the lack of the E1-A226V substitution. (268).

6.1.3 CHIKV pathogenesis

In contrast to other arboviruses such as DENV, CHIKV causes symptomatic disease in the majority of infected individuals (237). Most cases are characterized by fever, rash, and intense joint and muscle pain. The joint pain is usually symmetric and localized in the large joints of the body (i.e., arms and legs). The word “Chikungunya” comes from the African Makonde language and means “bent over in pain”, one of the hallmarks of CHIKV disease (237). While this acute stage generally only lasts for 1 to 2 weeks, severe joint pain may be prolonged for months or even years in some cases, leading to a chronic stage (271). Furthermore, severe chikungunya fever, manifested mostly as encephalitis, has been reported during large outbreaks such as the one on La Reunion Island in 2006 (272). More recently, de Souza et al. reported the detection of CHIKV by RT-qPCR in the CSF of several patients with fatal outcomes in Brazil (273). Nonetheless, these severe forms of the disease are rare. They can be seen in aged or immunosuppressed patients or people with coexisting disorders such as diabetes or cardiovascular, neurological, or respiratory conditions (274). Newborns are another group at risk for severe infection associated with neurological signs (237). While fetal infection seems rare, mother-to-neonate transmission from viremic mothers can lead to severe disease and neurological sequelae in newborn babies (275, 276). These sequelae could have long-term effects, including microcephaly, cerebral palsy (277), and cardiovascular problems (278).

6.2 Neurological Chikungunya and molecular epidemiology of the 2011-2013 outbreak in Cambodia

CHIKV re-emerged in Cambodia in 2011 (259). This outbreak was followed by two subsequent waves of cases in 2012 and 2013. Multiple cases – including rare neurological presentations - were captured by syndromic surveillance of Institut Pasteur du Cambodge; however, no information was gathered about the genetic diversity of CHIKV circulating in Cambodia. Furthermore, despite the increasing reports of neurological disease associated with CHIKV, no study has analyzed the virus present in the CSF and sera within the same patient.

In this work, we study the emergence and spread of CHIKV in Cambodia between 2011 and 2013. Additionally, to gain insights into the difference in disease outcomes upon CHIKV infection, we sequenced in parallel serum and CSF samples from infected patients with classic chikungunya symptoms or with neurological affliction (encephalitis).

6.2.1 Results

Two approaches to sequence low amount and/or degraded clinical samples

Since its reemergence in Cambodia in early 2011 (259), CHIKV continued to circulate and spread over the next two years, reaching 14 different provinces. We sought to characterize the virus underpinning the outbreak by sequencing the virus directly from clinical samples. Interestingly, this cohort of 39 patients showed different clinical presentations, with 5 of them suffering from encephalitis. An exciting feature of this cohort is that we also received samples from the CSF from these 5 patients.

Initially, using untargeted metagenomic sequencing, we generated complete genomes for samples with Ct values lower than 30 (Figure 6-4A). However, this technique failed to obtain full-length genomes for samples where the amount of viral RNA was lower, particularly for the samples derived from the CSF. We implemented a highly multiplexed PCR amplicon approach (279) to circumvent this situation, which allowed us to obtain complete viral genomes from these challenging samples. The primers scheme was designed using Primal Scheme (280) to generate ~ 400 nucleotide long overlapping amplicons to cover the entire length of the CHIKV genome. This method allowed us to recover almost full-length genomes from samples with very high Ct values, such as those derived from the CSF from encephalitic patient numbers 1 and 3 corresponding to E1_CSF and E3_CSF in Figure 6-4A, respectively. Using these two approaches, we generated full-genome sequences from 39 serum samples and 5 CSF samples from the encephalitic patients totaling 44 new CHIKV genomes. Completeness and coverage for these genomes are shown in Figure 6-4A and B, respectively.

No genomic differences in CHIKV populations sampled from matched sera and CSF

We compared the consensus sequence of the virus detected in the sera and the CSF within each encephalitic patient and found no differences (Figure 6-4C and D). Next, we used a non-encephalitic

sequence to reveal encephalitic-specific mutations shared by all five encephalitic patients. Using one of the genetically closest non-encephalitic sequences as a reference allowed us to find only one genetic change shared across all encephalitic patients. However, we also observed this mutation in sequences from other non-encephalitic patients, which indicated that viruses collected in the sera from patients with different clinical presentations did not present any unique genomic mutation associated with encephalitis (Figure 6-4D). Additionally, we identified only one mutation between the virus collected from patient E1 and our reference, further highlighting the genetic proximity between viruses associated with different disease outcomes (Figure 6-4D).

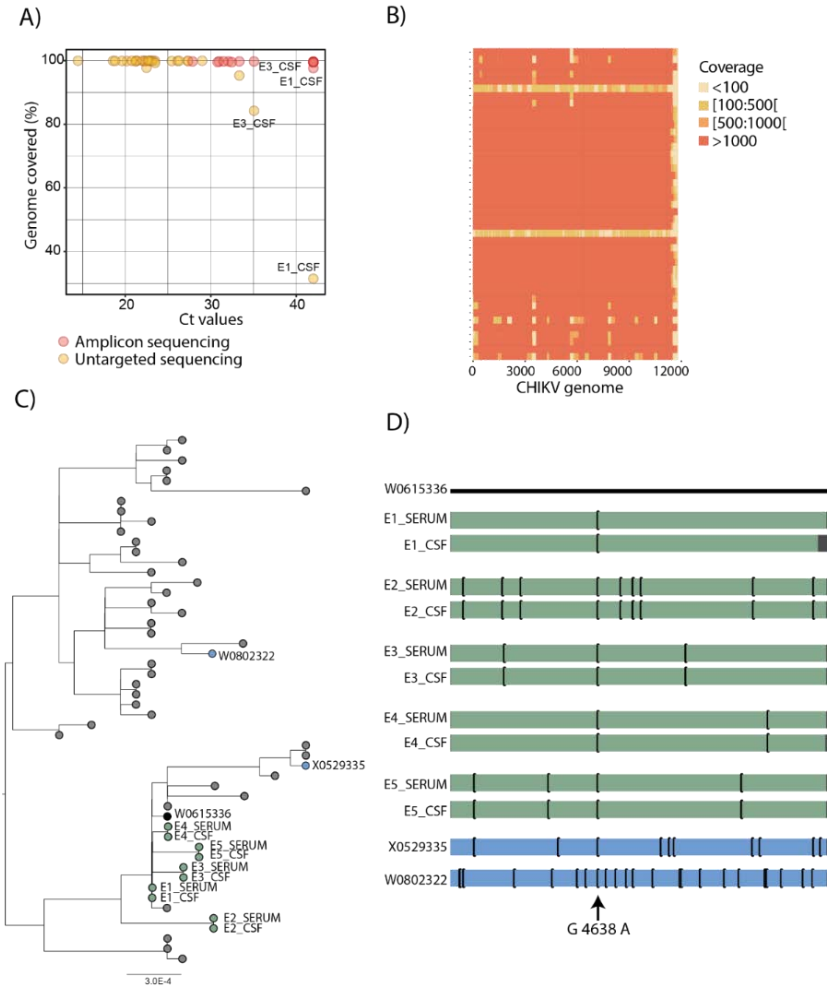


Figure 6-4: *Genomic data from clinical samples. No mutation is associated with different clinical presentations in CHIKV infected patients. A) Percentage of CHIKV genome according to the Ct values obtained from RT-qPCR on the 44 samples reported here. We specify the two different sequencing techniques implemented with the different colors: amplicon-based sequencing (red) and untargeted metagenomics (yellow). B) Completeness and coverage of the CHIKV genome sequences. C) ML tree of sequences from Cambodia highlighting sequences obtained from encephalitic (green nodes) and non-encephalitic patients (black and blue nodes). D) Schematic representation of CHIKV genomes sampled from matched serum and CSF within each encephalitic patient (green) and serum of non-encephalitic patient (blue). According to the reference (W0615336, in black), single nucleotide variants are shown as black vertical bars. As a reference, we used a genetically close sequence to the sequences from the encephalitic cases to avoid displaying irrelevant mutations. Sequences from non-encephalitic patients were chosen randomly from different parts of the tree (W0802322 and X0529335). With the arrow, we highlight the only nucleotide change between the virus sampled from matched serum and CSF within encephalitic patient 1 (E1), which is also present in the virus from the three non-encephalitic patients displayed in this panel*.

Phylogenetic and phylogeographic analysis of CHIKV circulation in Cambodia during the three-year epidemic

Combined with a set of publicly available partial and complete CHIKV genome sequences (n = 798) collected until 2015, we generated a maximum likelihood phylogeny (ML). Such phylogenetic analysis placed the novel sequences with sequences from Cambodia reported in May and August 2011 (n = 8) (259) in a single clade within the IOL lineage (bootstrap node support = 100%) (Figure 6-5).

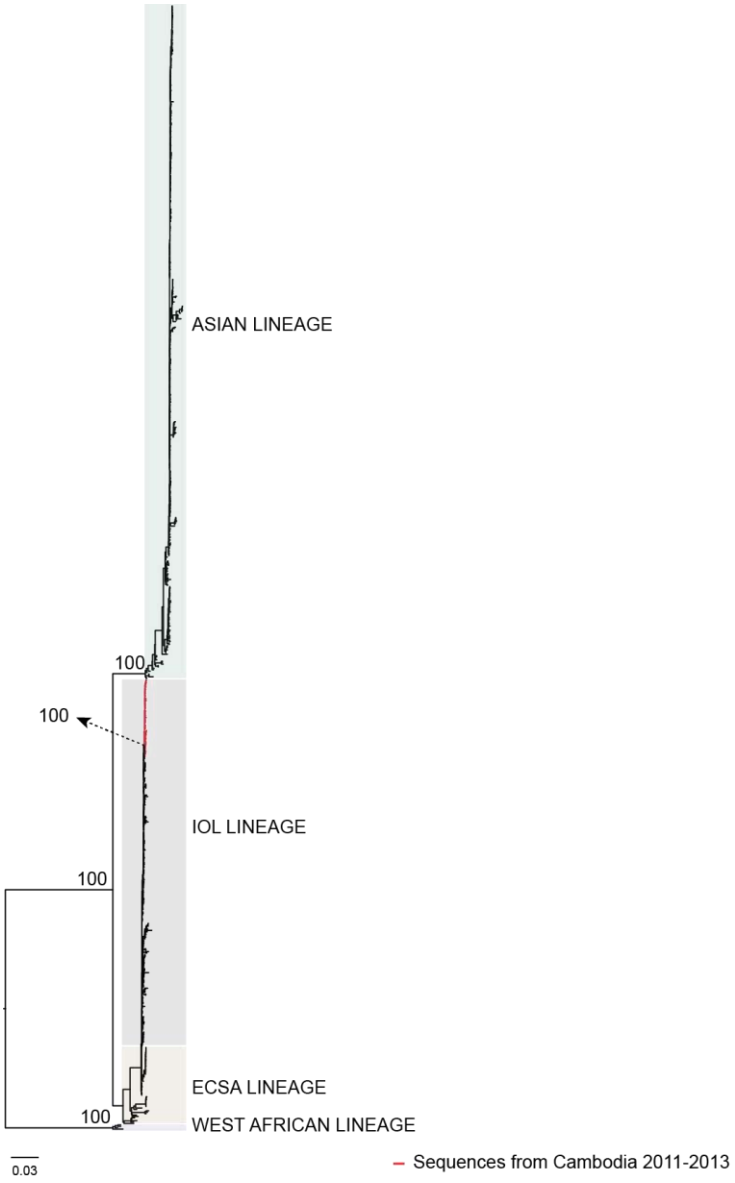


***Figure 6-5**: Global phylogenetic tree of complete and partial CHIKV genome with representatives of the four lineages (n=789). We highlighted in red sequences from Cambodia collected between 2011 and 2013. Node support values for the main lineages are ultrafast bootstrap percentages, and the scale bars represent the number of nucleotide substitutions per site. The tree with the accession number as tip label can be found in Figure S1.*

To reconstruct the dynamics of CHIKV circulation in Cambodia from 2011 to 2013, we performed a phylogeographic analysis of a subsampled dataset consisting of 121 partial and complete CHIKV genome sequences. We performed the subsampling on the initial tree according to the phylogenetic proximity to the sequences collected in Cambodia from 2011-2013 (see Methods section). We implemented a Bayesian discrete phylogeographic approach (150), testing different model combinations of tree priors and molecular clock using BEAST v10.4. According to the Bayes factor obtained, the best model that fitted our data was the combination of Skyride as the tree prior and relaxed as the molecular clock (Table 2). We used this model for inferring the evolutionary parameters (i.e., substitution rate and time to the most recent ancestor).

***Table 2**: Model selection. Marginal likelihoods were calculated with path-sampling (PS) and stepping-stone sampling (SS) for a total of six model combinations using three coalescent tree priors (Bayesian Skyride, exponential growth, and constant size) and two clock models (uncorrelated relaxed clock with log-normal distribution, UCLN, and strict clock and). The Bayes factor is calculated against the baseline model, a constant size tree prior, and a strict clock.*

	log (marginal likelihood)		**log (Bayes Factor)**	
MODEL	PS	SS	PS	SS
Skyride, Relaxed	-22346.0972	-22347.7107	13	12
Skyride, Strict	-22350.8134	-22351.6453	8	8
Constant, Relaxed	-22354.935	-22355.7196	4	4
Exponential, Relaxed	-22355.3797	-22356.4513	4	3
Constant, Strict	-22359.0396	-22359.8633	0	0
Exponential, Strict	-22366.7885	-22368.1268	-8	-8

We estimated the substitution rate to be 5.47×10^{-4} substitutions per site per year (95% HPD [4.5×10^{-4}, 6.7×10^{-4}]) (Table 3 and 4), which resulted similarly (i.e., with overlapping HPD intervals) to previous estimates for the IOL lineage (247, 281, 282) and consistent among the different coalescent models and molecular clock tested.

The time to the most recent ancestor (tMRCA) for CHIKV in Cambodia was estimated to be during May 2010 (95% HPD November 2009 – October 2010). The estimated tMRCA for the sequence outside the outbreak clade is October 2008. Thus, CHIKV was likely introduced into Cambodia between October 2008 and October 2010, clearly before the first documented cases in May 2011.

***Table 3**: Evolutionary parameters retrieved from each model*

Summary Statistic	Skyride, Strict	Exponential, Strict	Constant, Strict	Statistic
median	$4.56x10^{-4}$	$4.74 x10^{-4}$	$4.73 x10^{-4}$	Clock rate
95% HPD interval	[$3.90 x10^{-4}$, $5.23 x10^{-4}$]	[$4.06 x10^{-4}$, $5.36 x10^{-4}$]	[$4.10 x10^{-4}$, $5.39 x10^{-4}$]	Clock rate
ESS	2443.9	3824.4	3932.5	Clock rate
median	2004.9676	2004.6328	2004.6106	tMRCA Root
95% HPD interval	[2004.4236, 2005.4933]	[2003.8507, 2005.2351]	[2003.8841, 2005.2363]	tMRCA Root
ESS	3091.6	5104	4975.8	tMRCA Root
median	2010.1978	2010.1263	2010.1323	tMRCA Cambodia
95% HPD interval	[2009.8188, 2010.6247]	[2009.7063, 2010.5833]	[2009.6939, 2010.5848]	tMRCA Cambodia
ESS	4512.2	5605.8	5971.8	tMRCA Cambodia
median	2012.3597	2012.3355	2012.3373	tMRCA Laos1
95% HPD interval	[2012.0559, 2012.5777]	[2012.0453, 2012.5474]	[2012.0559, 2012.5405]	tMRCA Laos1
ESS	4799.4	6689.4	6289	tMRCA Laos1
median	2012.445	2012.4379	2012.4387	tMRCA Laos2
95% HPD interval	[2012.1878, 2012.5908]	[2012.2069, 2012.5868]	[2012.211, 2012.5891]	tMRCA Laos2
ESS	7739	8707.1	7740.7	tMRCA Laos2

***Table 4**: Evolutionary parameters retrieved from each model*

Summary, Statistic	Skyride, Relaxed	Exponential, Relaxed	Constant, Relaxed	Statistic
median	$5.47 x10^{-4}$	$5.66 x10^{-4}$	$5.66 x10^{-4}$	Clock rate
95% HPD interval	[$4.47 x10^{-4}$, $6.66 x10^{-4}$]	[$4.61 x10^{-4}$, $6.94 x10^{-4}$]	[$4.59 x10^{-4}$, $6.90 x10^{-4}$]	Clock rate
ESS	1793.2	1931.5	2413.3	Clock rate
median	2005.2411	2004.6104	2004.569	tMRCA Root
95% HPD interval	[2004.6604, 2005.7273]	[2003.4767, 2005.4942]	[2003.4221, 2005.4287]	tMRCA Root
ESS	2267.2	3811.7	3354.9	tMRCA Root
median	2010.3073	2010.2033	2010.2285	tMRCA Cambodia
95% HPD interval	[2009.8893, 2010.7886]	[2009.6984, 2010.7267]	[2009.7054, 2010.7244]	tMRCA Cambodia
ESS	2803.5	4046.9	3783.4	tMRCA Cambodia
median	2012.4223	2012.3941	2012.4072	tMRCA Laos1
95% HPD interval	[2012.124, 2012.6191]	[2012.1099, 2012.6032]	[2012.1368, 2012.5979]	tMRCA Laos1
ESS	3995.5	5646.2	5149.7	tMRCA Laos1
median	2012.4612	2012.4524	2012.4572	tMRCA Laos2
95% HPD interval	[2012.1958, 2012.5944]	[2012.211, 2012.5963]	[2012.2281, 2012.5941]	tMRCA Laos2
ESS	7358	7808.1	7287.1	tMRCA Laos2

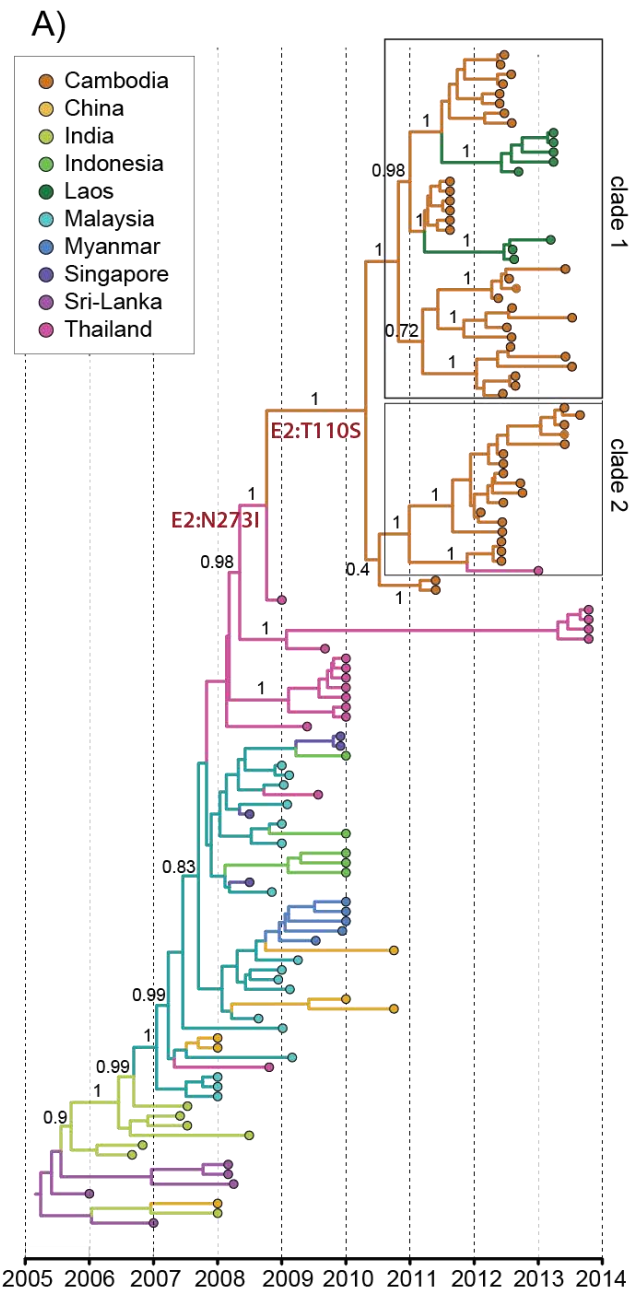
The estimated MCC tree (Figure 6-6A) shows that the Cambodian clade shares a common ancestor with viruses circulating in Thailand in 2009 (accession number KU561434 (283) and GU301779 (284) corresponding to partial and complete genomes, respectively). We used the Bayesian stochastic search variable selection (BSSVS) analysis (150) followed by the Bayes factor (BF) test using SpreaD3 (285) to quantify and identify well-supported transitions – defined as transitions with BF > 3 (150) – between the sampled countries. In line with the topology of the MCC tree, such analysis identified Thailand to Cambodia as a well-supported transition with a BF of 148 and posterior probability for the transition (PP) of 0.95, supporting here again Thailand as the source of the introduction (Figure 6-6A and B).

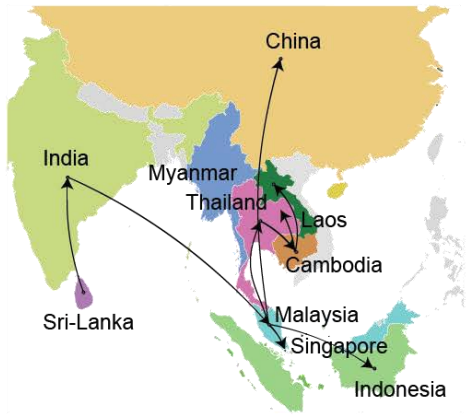
The phylogeny also shows that CHIKV genomes detected in Cambodia fell into one single clade (posterior probability = 1), segregating into two well-supported clades, clade 1 and clade 2, with high support (posterior probability = 1). However, the node basal to clade 2 is not well resolved (posterior probability = 0.4). This suggests that the smallest branch from this node which contains the earliest sequences from our dataset, sampled in the province of Battambang in early May 2011, had too few mutations that did not allow to distinguish if they fell in clade 1, 2, or outside of these clades. This poor resolution suggests that we probably did not capture a representative snapshot of the CHIKV genetic diversity circulating at the beginning of the outbreak. ML tree obtained using IQ-TREE (Figure S2) also placed these two sequences outside clades 1 and 2. From these phylogenies, we can hypothesize at least three introductions in Cambodia. A first introduction led to the sequences captured in May 2011 in Battambang. As shown by the MCC tree, these viruses did not spread further in the population, or if so, we did not capture them. A second introduction was responsible for the clade 1 sequences, and a final third introduction that led to the sequences of clade 2. However, the absence of sequences sampled closer to the root of the Cambodian clade hampers a confident inference of the location state at the root. Hence, two scenarios could explain the genetic diversity of CHIKV observed in Cambodia: either a single introduction in late 2008 followed by silent or cryptic circulation and viral divergence until the wave of documented cases started in May 2011 or multiple independent introductions.

The BSSVS analysis followed by BF test also identified ‘Cambodia to Laos’ and ‘Cambodia to Thailand’ as well-supported location transitions with BF of 3381 and 19 respectively, with posterior probabilities higher than 0.7 (Figure 6-6C).

CHIKV genomes sampled in Laos during 2012-2013 (286) cluster within the Cambodian clade, forming two other well-supported clades (posterior probability = 1). The estimated date for tMRCA of the first Laos clade is June 2012 (95% HPD March 2012–September 2012; Table 3 and 4). The estimated date of divergence of this clade from the sequences sampled outside Laos is July 2011. Similarly, for the second Laos clade, the estimated date for tMRCA is June 2012 (95% HPD March 2012–September 2012; Table2). The common ancestor between the second Laos’ clade and sequences sampled outside Laos is inferred to be March 2011. All this indicates that CHIKV in Laos was most likely introduced through two different but contemporary introductions, probably from Cambodia, the first between July 2011 and September 2012 and the second between March 2011 and September 2012.

The tMRCA between the sequence sampled in Thailand 2013 falling within the Cambodian clade 2 and viruses captured in Cambodia is estimated to be the end of November 2011 (95% HPD May 2011–May 2012). Although there is only one sequence from Thailand branching from this clade, this suggests that virus exchange from Thailand to Cambodia could have occurred in both directions during the outbreak.





C)

FROM	*TO*	*BF*	*PP*
Cambodia	Laos	3381	1.0
Malaysia	Thailand	1057	0.99
Malaysia	China	695	0.99
Malaysia	Singapore	444	0.98
India	Malaysia	309	0.97
Sri-Lanka	India	158	0.95
Thailand	Cambodia	148	0.95
Malaysia	Indonesia	37	0.82
Cambodia	Thailand	19	0.70

Figure 6-6: *Discrete phylogeography showing the spread of CHIKV in Southeast Asia. A) Time-scaled maximum clade credibility (MCC) tree of CHIKV circulating Southeast Asia from 2006 to 2015 (n=121) obtained by discrete phylogeographic inference. Node labels are posterior probabilities indicating support for the main nodes. Branches and tips nodes are colored according to the sampling location. We highlighted in red two amino acid substitutions in the E2 envelope protein (E2: T110S and E2: N273I) carried by all the viruses collected in Cambodia and Laos. The substitution E2: N273I was detected in one virus sequence sampled in Thailand in 2009. B) The Southeast Asia map showing only transitions among locations yielding posterior probability > 0.7. C) Table showing the Bayes Factor (BF) and posterior probability (PP) associated with each transition. Only transitions yielding posterior probability > 0.7 are displayed.*

To quantify and map the spatial spread of CHIKV within Cambodia, we extracted the spatiotemporal information embedded in 1000 subsampled posterior trees obtained from a continuous phylogeographic inference using only the sequences collected from Cambodia. We then used Seraphim to estimate key dispersal parameters (162). We estimated a mean lineage dispersal velocity of ~147.3925 km/year, which remained relatively constant over time (Figure 6-7B).

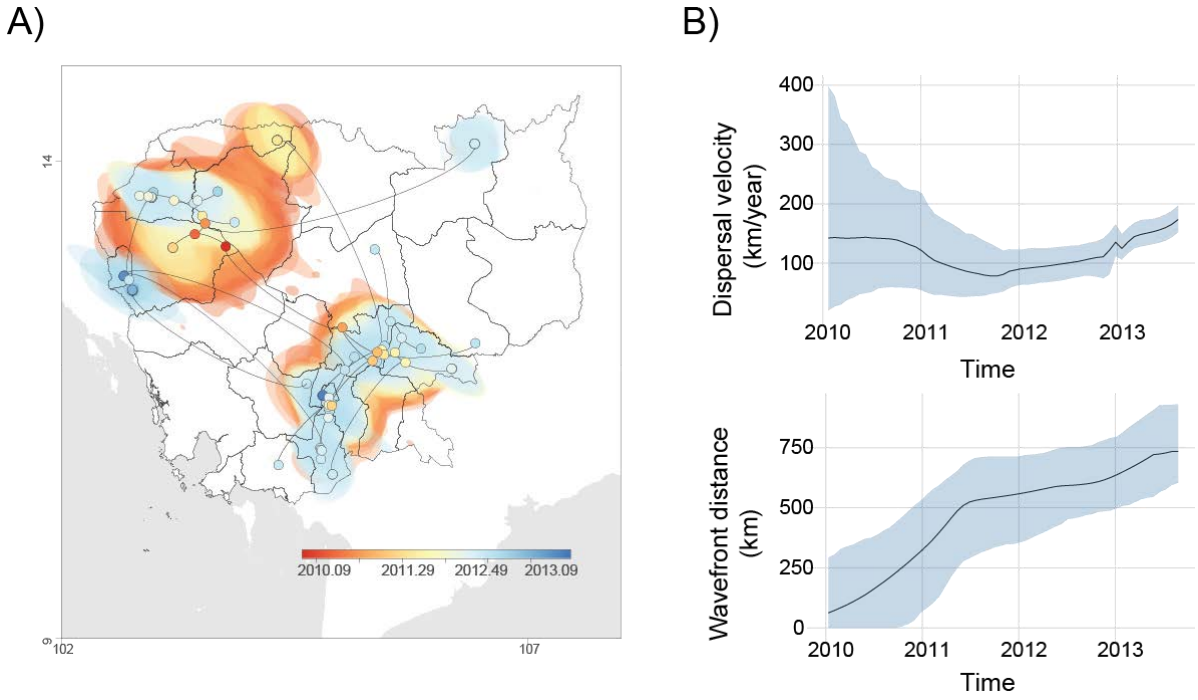


***Figure 6-7**: Reconstructed spatiotemporal diffusion of CHIKV in Cambodia. A) Continuous phylogeography showing the local spread of CHIKV in Cambodia. Shaded areas are colored according to time, and they show the 80% highest posterior density (HPD) of the possible locations of origin of viral ancestors. B) Weighted lineage dispersal velocity through time (top) and spatial wavefront distance from the epidemic origin over time (bottom).*

Genomic characterization of CHIKV circulating in Cambodia

We noted that all the genomes of our dataset carried the mutation E1: A266V associated with increasing CHIKV dissemination in *Aedes Albopictus,* and two other mutations leading to non-synonymous substitutions in the E2 envelope protein, E2: T110S and E2: N273I. Both substitutions were present in the viruses collected in Laos between 2012 and 2013 (accession number MF076568 to MF076576). However, the sequence from the virus collected in Thailand in 2009, which is basal to the sequences from Cambodia, only harbors the E2: N273I substitution.

We observed that the E2: T110S and E2: N273I substitutions were located in the domain A and C of the E2 protein (Figure S3). Interestingly, these two domains strongly interact with E1, particularly domain A, where most of the reported mutations conferring escape to neutralizing antibodies or affecting cell attachment are located (238). Nevertheless, more studies will be needed to address this.

Additional changes differentiate the viruses detected within Cambodia. Clade 1 sequences harbored two common mutations (C301T and T3988G), leading to synonymous substitutions in the nsp1 and nsp2, respectively, and a third in the non-coding region (A11366G). Clade 2 sequences shared three mutations (C4894T, C6358T, and A10383G), resulting in synonymous substitutions in the nsp3, nsp4, and E1 proteins, respectively. Furthermore, Battambang sequences also carried three changes (C478T, C3082T, and T11215G) that led to synonymous substitution in the nsp1, nsp2, and the E1 protein, respectively.

6.2.2 Discussion

As a result of the syndromic surveillance system of CHIKV in Cambodia, we obtained samples from CHIKV infected patients from several provinces from 2012 until 2013.

In recent years, there have been an increasing number of studies reporting neurological complications associated with CHIKV infection (234, 287). In combination with other reports showing the capacity of CHIKV to reach and disseminate in the CNS in animal models (288), these studies provided strong evidence of the neurotropic role of CHIKV. However, this is the first report generating the complete CHIKV genome from the CSF and sera from the same encephalitic patient. This allowed us to compare the virus sequence found in the two compartments and look for any genomic mutation that could explain the observed phenotypes. We found that the consensus sequence of the viruses within each patient was identical between the two compartments. As such, our work suggests that no additional mutations are required for CHIKV to reach the CNS and cause neurological disease.

Nevertheless, we recognize two major limitations. First, the low number of samples obtained from patients suffering encephalitis upon CHIKV infection (n=5) prevented us from having statistical support in our findings. Second, the lack of medical information from these patients prevented us from studying whether other factors could have been involved in determining different chikungunya disease outcomes. Indeed, CHIKV infection leading to neurological disease is often seen in aged, immunosuppressed, or people with coexisting disorders such as diabetes, cardiovascular or respiratory disease (237, 274). Additionally, a higher viral load has been associated with increased disease severity for Influenza virus (289) and SARS-CoV-2 (290, 291). Although we performed RT-qPCR from all these samples (Figure 6-4A), due to the several rounds of thawing and freezing that these samples suffered from the initial extraction at the IPC and the subsequent manipulations at the Institut Pasteur Paris, the Ct values do not reflect the patients' viral load at the moment of the sampling.

Therefore, while our observations suggest that no virus genetic mutation was associated with different clinical presentations, we cannot conclude whether other factors may have played a role in the development of different disease outcomes.

As previously suggested, this current dataset suggests that CHIKV in Cambodia could have been introduced from Thailand (259). Nevertheless, we acknowledge that one limitation of our phylogeographic analysis is the little to no genomic or epidemiological information on CHIKV circulating in neighboring countries in those years. In addition, while genetic proximity might not give enough evidence to claim a direct source, two facts might support Thailand as the source of introduction. First, it has been reported that in early 2008 initial cases of CHIKV infection were detected in southern Thailand (292). By December 2009, a large

outbreak had already been spreading throughout Thailand, reaching a total of 46,000 documented cases, the highest number of infections in any country in Southeast Asia at that moment (293). Secondly, there is an intense migration flux between the two countries due to the cultural similarities and the easy flow through the Thailand-Cambodian border. Indeed, many Cambodians travel to Thailand to work temporarily or permanently (294).

Interestingly, sequences from Laos sampled during 2012 -2013 also fall into the Cambodian clade within clade 2. MCC tree topology and BSSVS analysis with BF test suggested Cambodia as the source of the outbreak in Laos in 2012. This was already suggested by Somlor et al. (286), who also emphasized that the reported cases were detected next to the Laos-Cambodian border in the province of Champasak, located in the south part of the country, neighboring the Cambodian province of Preah Vihear. Further supporting this claim, we found that the Laos sequences branch from sequences collected by August 2011 in Preah Vihear, forming two different clades with high support (posterior probability = 1), indicating at least two introductions. Our data suggest that these two independent introductions occurred almost simultaneously, the first between July 2011 and September 2012 and the second between March 2011 and September 2012.

Our analyses also suggest that the genetic diversity of CHIKV in Cambodia captured during the three-year epidemic was not the result of successive reintroductions from neighboring countries but rather the result of one or multiple initial introductions before the first wave of documented cases, followed by cryptic and local transmission and viral persistence during the dry seasons. The inter-seasonal maintenance of CHIKV was possibly achieved through mosquito vertical transmission cycles (295), or low transmission rate leading to a small and undetected number of cases. However, more data is required to further investigate and explore these two scenarios.

6.2.3 Conclusion

To the best of our knowledge, this work provides the first analysis of complete CHIKV genomes detected in both, the serum and CSF of encephalitic patients. Our analyses suggest that no additional mutations are required for CHIKV to reach the CNS. Phylogenetic and phylogeographic analysis showed that the CHIKV circulating in Cambodia from 2011 to 2013 belonged to the IOL lineage introduced in the country, probably from Thailand, somewhere between 2009 and 2011. Overall, this work contributes to the understanding of CHIKV pathogenesis and provides an update on CHIKV evolution in Southeast Asia.

6.2.4 Methods

Patients and sampling collection

We report the findings of samples obtained from 47 patients with CHIKV infection confirmed by the reference laboratory of the Institut Pasteur of Cambodia. The samples correspond to 8 cases reported in May and August 2011 (259) and 39 cases collected from 14 provinces across Cambodia from May 2012 to

August 2013 (Table S1). This cohort of patients showed different clinical presentations; in particular, 5 of these patients suffered encephalitis from whom we received samples from the serum and the CSF.

Viral RNA extraction and real-time PCR

The samples selected for sequencing consisted of clinical specimens (n=33) or serum samples that had been subjected to one passage on C6/36 cell line (n=16) (Table S1). To obtain a quantitative measurement of the viral RNA, we performed RT-qPCR using custom-designed primers. Briefly, RNA was extracted using the QIAamp Viral RNA Mini Kit (Qiagen) according to the manufacturer's recommendations, followed by Turbo DNAse treatment (Ambion) and purification with Agencourt RNAClean XP beads. The extracted RNA was subjected to retro transcription using the SuperScript IV (Invitrogen). The cDNA was used as a template for the PCR, which was performed using customs primers and Taqman probes targeting the E1 gene of CHIKV. Ten-fold dilutions of a plasmid containing the E1 gene of CHIKV were used as a standard curve.

Viral genome sequencing

a. Untargeted sequencing protocol

We used a general protocol for untargeted sequencing of clinical samples previously described (171). In short, prior to library construction, poly-A carrier RNA and host rRNA were depleted using oligo (dT) and custom probes, respectively. The RNA resulting from selective depletion was used for random-primed cDNA synthesis using the SuperScript IV (Invitrogen). Second-strand cDNA was generated using a cocktail of enzymes, including *Escherichia coli* DNA ligase, RNAse H, and DNA polymerase (New England Biolabs), and purified using Agencourt AMPure XP beads (Beckman Coulter). The libraries were prepared using the Nextera XT kit and sequenced using a paired-end strategy on an Illumina NextSeq500 platform from the dsDNA.

b. Amplicon-based sequencing protocol

We developed a highly multiplexed short PCR amplicon panel following the approach described in (296) for those samples in which the untargeted sequencing failed. We used the Primal scheme web-server (https://primalscheme.com) (280) to design a set of 43 primer pairs that generate overlapping products along the CHIKV genome. To design the primers, we used as a scaffold a CHIKV sequence that we generated using the untargeted sequencing approach. The primer pool scheme and protocol can be found here.

Two microliters of viral cDNA were used in the two multiplexed PCR reactions using Q5 DNA High-fidelity Polymerase (New England Biolabs) to obtain ~ 400 nucleotides long amplicons in 35 cycles. Amplicons were purified using Agencourt AMPure XP beads (Beckman Coulter) and combined to 50ng. Libraries were constructed using the NEBNext Ultra II DNA Library Prep Kit for Illumina (New England Biolabs) and sequenced on an Illumina MiSeq at the Biomics platform (Institut Pasteur Paris).

Regardless of the sequencing protocol used, library quality and quantification were assessed using Qubit 4 (Thermo Fisher), Bioanalyzer (Agilent), and qPCR (NebNext Library Quant Kit, Illumina).

Sequence data analysis

After sample demultiplexing and adapter trimming, the quality of the fastq files was assessed with FastQC v.0.11.9 (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/). For the untargeted sequencing, reads were filtered by quality using Trimmomatic v.0.39 and *de novo* assembled using metaspades option from SPAdes v3.1.0. The contigs obtained were used as blast queries on Virus Pathogen Resource (ViPR) (173) nucleotide database. A CHIKV isolate from Cambodia 2011 (GenBank: JQ861260) was the closest to our samples and, thus, used to align reads from the untargeted sequencing approach. For the amplicon-based sequencing approach, we used Trimmomatic v.0.39 to filter out low-quality reads and remove primer sequences (first 27 nucleotides from 5' end of reads, which is the maximum length of primers used for multiplexed PCRs). CHIVK sequence used to design the set of primers was used as a reference for the mapping. Then, for either untargeted or amplicon-based sequencing methods, the alignment and the consensus sequences were called using CLC genomics suite v5.1.0 (QIAGEN). We used a minimum of 5X read depth coverage to generate the consensus sequence. In case of lower coverage, Ns were added in each position. All alignments and consensus sequences were manually inspected using Geneious Prime 2020.2 (https://www.geneious.com/). Then, SAMtools v1.3 was used to sort the aligned BAM files and to generate alignment statistics. Variants were called using iVar v1.0 (296).

Viral sequences selection and alignment

To build our initial dataset, sequences generated during this study were combined with partial and complete CHIKV genomes dated until 2015 available on ViPR (173) as of August 2021 (n= 798) (Figure S1). Genomes without collection dates and location information were excluded. The resulting dataset was aligned using MAFFT v7.467 (297) and inspected manually. We constructed preliminary maximum-likelihood phylogenies (ML) using IQ-TREE v2.0.6 (298). Tree reconstruction was performed using the default settings and the best-fitted model provided by ModelFinder (299) followed by 1000 ultrafast bootstrap (300) implemented in IQ-TREE software. A smaller dataset was obtained based on this dataset by subsampling according to the phylogenetic proximity to the viruses detected in Cambodia from 2011 to 2013, and subsampling overrepresented clades. Then, we used TempEst v1.4 to inspect these genomes and identify major molecular clock outliers that we removed from downstream analyses. This yielded a smaller and good quality dataset (n=121) with which we continue the following analysis (Figure S2)

Discrete phylogeographic analysis

To investigate the origin and the spread of CHIKV in Cambodia, time-structured phylogenies were inferred using a discrete asymmetric diffusion model (150) available in BEAST v1.10.4. The analysis was run for 150 million Markov chain Monte Carlo steps using the BEAGLE library v3.1.0 to accelerate the computation process. The parameters and trees were sampled every 15,000 generations. We used the general time-reversible (GTR) nucleotide substitution model and gamma-distributed sites. We tested three trees prior: Bayesian Skyride, Exponential growth and Constant population size, and two clock models: an uncorrelated relaxed clock with log-normal distribution (UCLN) and a strict clock. For the six resulting combinations of models, we performed a model selection test using path sampling (PS) and stepping stone (SS) to estimate marginal likelihoods (301, 302). We used the default parameters, sampling for 100 path steps with a chain of 1 million steps. We calculated the BF against the baseline model: constant population

size and strict clock. The best supported model was the combination of Bayesian Skyride and UCLN (Table 1). The discrete phylogeographic analysis was also performed using BSSVS, followed by a BF test using SpreaD3 (285) to identify and quantify the support for the transition between the sampled countries. The maximum clade credibility (MCC) tree was built under this model using TreeAnnotator v.10.4 and visualized in FigTree v1.4. Evolutionary rates and tMRCA were extracted using Tracer v1.4 for tested model combinations (Table2).

Continuous phylogeographic analysis

To infer the dispersal history of CHIKV in Cambodia, we implemented a continuous phylogeographic model focusing only on the CHIKV genomes detected in Cambodia from 2011 to 2013. We used the Cauchy relaxed random walk diffusion model (151) available in BEAST v1.10.4 to infer ancestral locations. The MCMC chain was run for 150 million generations and sampled every 15,000 generations. As before, we performed a model selection test using path sampling (PS) and stepping stone (SS). According to the BF obtained, the best model was the combination exponential as tree prior and strict as a molecular clock. Convergence was inspected with Tracer v1.4. After discarding 10% of the sampled trees as burn-in, the MCC was built using TreeAnnotator v.10.4. We extracted and plotted the spatio-temporal information embedded in 1000 subsampled trees from the posterior using Seraphim R package (162).

Protein structure modeling

The substitutions T110S and N273I were mapped onto the CHIKV E1-E2-E3 heterodimer (Protein Data Bank ID: 3N42) using UCSF Chimera.

Supplementary information associated with this work can be found in the following link.

6.3 Chikungunya virus outbreak in Cambodia in 2020: re-emergence after a decade of absence

After the 2011 - 2013 epidemic, no CHIKV cases were reported in Cambodia until 2020. Following the outbreaks reported in Thailand (267) and Myanmar (269), the first cases were reported in Cambodia by the end of June 2020. From there, the virus quickly spread throughout the country, reaching 23 out of 25 provinces by October 2020, with 1,258 CHIKV positive cases detected by our colleagues from Institut Pasteur du Cambodge in the context of the National Dengue Control Program, Ministry of Health Cambodia (Figure 6-8). In this work, we aimed at characterizing the genetic diversity, emergence, and spread dynamics of CHIKV during the 2020 outbreak.

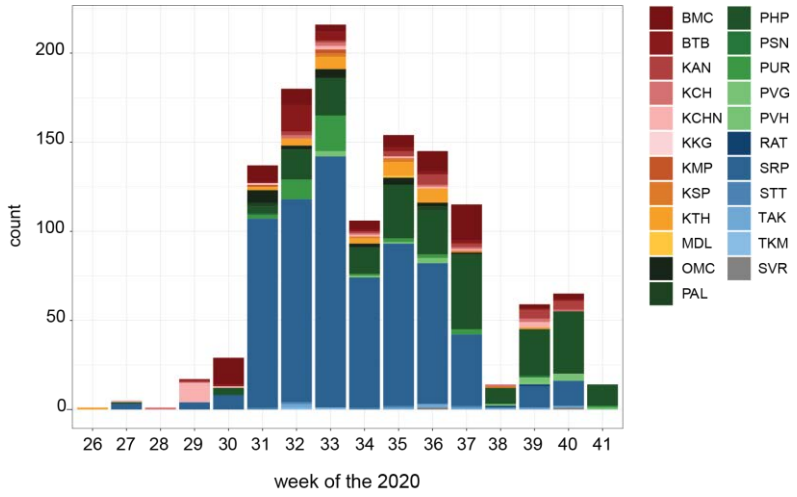


Figure 6-8: *Number of CHIKV cases documented by our colleagues from the Institut Pasteur du Cambodge.*

Phylogenetic studies

Among the samples found positive for CHIKV by real-time PCR, 73 samples were selected for sequencing according to their sampling location, date and Ct value.

We combined our CHIKV genomes sequences with complete CHIKV genomes published and available as of November 2021, totaling 775 CHIKV full-length genomes encompassing the four main CHIKV lineages: WA, Asian, ECSA and IOL. This dataset was used to perform phylogenetic analysis using the Nextstrain pipeline, and the resulting phylogeny can be visualized at https://nextstrain.org/community/Simon-LoriereLab/ChikungunyaCambodia2020@main. From this global analysis, we found that CHIKV captured from cases in Cambodia during the 2020 outbreak belonged to the IOL lineage but fell distant from those captured in the 2011 outbreak. Indeed, in contrast to the CHIKV captured in 2011, viruses reported from Cambodia in 2020 lack the E1:A226V substitution associated with an increase of CHIKV dissemination by *Aedes Albopictus* mosquitoes. Instead, this lineage harbors two other substitutions in the surface proteins (E1:K211E and E2:V264A). Phylogenetic analysis shows that CHIKV from this outbreak falls into five groups

(clades A to E), being closely related to viruses circulating in Thailand between 2018 and 2019 and in China and Myanmar in 2019, suggesting multiple introductions in the country.

To investigate the different possible introductions of CHIKV in Cambodia, we performed a phylodynamic and phylogeographic analysis of a dataset consisting of 155 partial and complete CHIKV genome sequences. This dataset was obtained by subsampling the initial global tree (n=775, only complete genomes) according to the phylogenetic proximity, resulting in 127 complete CHIKV genomes. To this dataset, we also added partial CHIKV genomes (larger than 9000 bp) collected from 2016 to 2021 as we thought that they could add a phylogenetic signal. We implemented a Bayesian discrete phylogeographic approach, testing different tree priors and molecular clock combinations using BEAST v10.4. According to the Bayes factor obtained, the best model that fitted our data was the combination of Skyride as tree prior and relaxed as the molecular clock (Table 5).

Table 5: *Model Selection. Marginal likelihoods were calculated with path-sampling (PS) and stepping-stone sampling (SS) for a total of six combinations using three coalescent tree priors (Bayesian skyride, exponential growth and constant size) and two clock models (uncorrelated relaxed clock with log-normal distribution [UCLN] and strict clock). The Bayes factor is calculated against the baseline model, a constant size tree prior and strict clock.*

Model Combination	**log (marginal likelihood)**		**log (Bayes Factor)**	
	PS	SS	PS	SS
Constant.Relaxed	-25730	25732	52	52
Constant.Strict	-25782	25784	0	0
Exponential.Relaxed	-25718	25720	64	64
Exponential.Strict	-25773	25775	9	9
Skyride.Relaxed	-25695	25697	87	87
Skyride.Strict	-25734	25736	48	48

With this dataset, we estimated the substitution rate to be 9.76×10^{-4} substitutions per site per year (95% HPD [7.96×10^{-4}, 1.18×10^{-3}]), which resulted in having overlapping HPD intervals to previous estimates for the IOL lineage (23-25), including the CHIKV epidemic in Cambodia during the 2011-2013 period, and it was consistent among the different coalescent models and molecular clock tested.

Using time-structured phylogenies, we estimated that at least five introductions were responsible for the local cases detected in Cambodia (Figure 6-9). While clades A, C and D had well-supported nodes (posterior probabilities higher than 0.9), the node basal to the clade E (which consists of a single CHIKV sequence) had a posterior probability of 0.3. However, the internal node from which branch the CHIKV sequence from Cambodia, together with sequences from China, Thailand, and Myanmar, is well-supported (posterior probability 0.94). This phylogenetic placement of the clade E suggests an additional introduction independent from those of the other clades. In addition, while the support for the clade B is high (posterior probability 1), the node connecting clade B sequences with sequences outside this clade is not well-supported (posterior probability 0.1). This suggests that although these sequences form a monophyletic clade, their position along the tree is unclear.

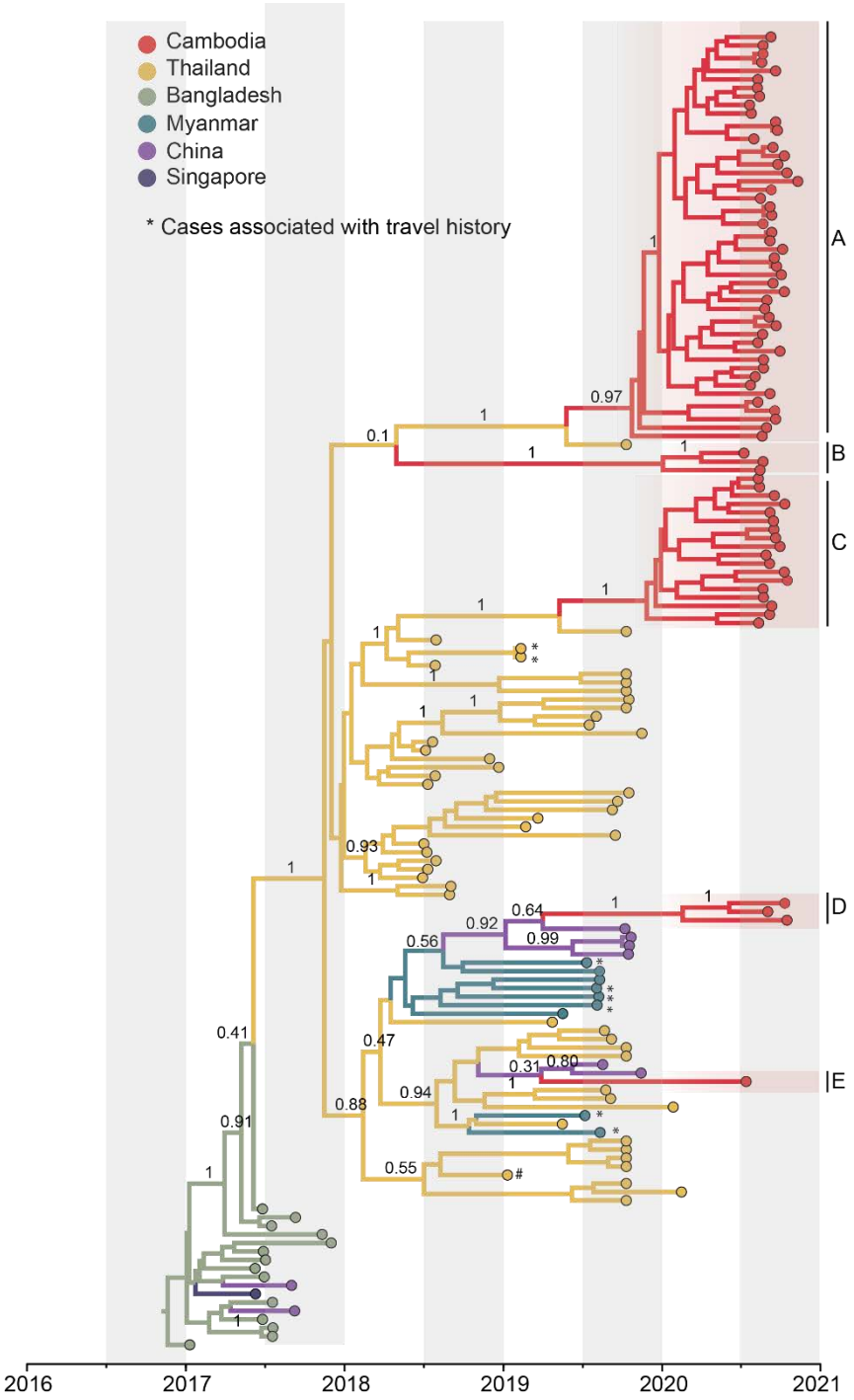


Figure 6-9: *Time-scaled maximum clade credibility (MCC) tree of CHIKV circulating Southeast Asia from 2017 to 2021 (n=151) obtained by discrete phylogeographic inference. Node labels are posterior probabilities indicating support for the main nodes. Branches and tips nodes are colored according to the sampling location. Genomes associated with cases with travel history have been colored according to the travel location and highlighted with an asterisk. The five possible independent introductions were labeled from A to E.*

We implemented the discrete diffusion model extended with the BSSVS and followed by a BF test to identify relevant transition rates between the sampled countries. This analysis identified 'Thailand to Cambodia' and 'China to Cambodia' as well-supported transitions with BF results of 291 and 9, respectively (Table 6). Additionally, we extracted from the MCC tree the estimated dates (mean and 95% HPD) for each clade, and the node connecting the sequences from each clade with sequences outside the clade. This allowed us to estimate a date range for each introduction (Table 6). Based on the genetic diversity sampled here, we observed three introductions from Thailand with overlapping date ranges: the first between November 2018 and December 2019; the second between November 2017 and April 2020; and the third between October 2018 and January 2020. Of note, for the third introduction observed (clade C), the support of the node is weak (posterior probability 0.1). However, the probability for the inferred location (Thailand) is close to one. Additionally, we observed two other introductions from China happening almost simultaneously: one introduction between October 2018 and July 2020 and a second between September 2018 and May 2020.

Table 6: *Inferred introductions to Cambodia. The BF and posterior probability (PP) associated with each transition were obtained under the BSSVS analysis. The estimated date for each transition was extracted from the MCC tree and calculated considering the higher bound value of the 95% HPD for the node associated with the Cambodian clade (node y) and the lower bound value of the 95% HPD for the node outside that clade (node x). The location probability and posterior probability for the internal node x are also shown.*

	FROM (node x)	TO (node y)	BF	PP	Estimated date	loc.prob.x	posterior.x
clade A	Thailand	Cambodia	291.3	0.99	2018 Nov - 2019 Dec	1.00	1.00
clade B					2017 Nov - 2020 Apr	0.99	0.10
clade C					2018 Oct - 2020 Jan	0.98	1.00
clade D	China	Cambodia	9.6	0.69	2018 Oct - 2020 Jul	0.91	0.64
clade E					2018 Sep - 2020 May	0.59	0.31

Ongoing analyses

Currently, we are trying to implement a continuous phylogeographic analysis to reconstruct the spatiotemporal dispersion of the virus within Cambodia. One of the reasons for this is because we noted that while the virus in 2011 spread to 19 provinces in 73 weeks, this time, it took only 14 weeks to reach 23 out of 25 provinces. Implementing a continuous phylogeographic analysis as the one we set up for the 2011 outbreak will allow us to calculate summary statistics of spatial spread such as dispersal velocity or evolution of the maximal wavefront distance. These statistics will help to further describe the dynamics of the outbreak and to compare the mode and rate of spatial spread among the two different outbreaks.

6.4 Discussion and conclusions of the studies

This chapter reported on CHIKV emergence and spread in Cambodia during the 2011-2013 and the 2020 outbreaks, and several conclusions can be drawn.

First of all, we successfully implemented two different sequencing techniques (metagenomics and PCR amplicons) that allowed us to obtain the complete sequence of the virus from samples of varying viral load and quality (degraded or partially degraded). By combining both techniques, we obtained complete CHIKV genomes from the 2011-2013 and 2020 outbreaks. The protocol and PCR primers panel optimized to obtain the complete CHIKV genomes were quickly shared with the community (https://github.com/Simon-LoriereLab/ChikungunyaCambodia2020) and, very importantly, with our collaborators from Institut Pasteur du Cambodge. This allowed our colleagues to implement their own sequencing while avoiding the delay of sending new samples to our institute for sequencing, allowing them to act quickly on the ongoing CHIKV outbreak. This highlights the importance of creating scientific networks and research collaborations between different academic institutions that allow (i) the generation and analysis of genomic data and thereby, the possibility of informing about the outbreak in real-time, and (ii) to build research capacity enabling countries to implement their own solutions to their specific problems.

Second, our work provided an update on CHIKV genetic diversity and evolution in Cambodia, contributing to the general understanding of CHIKV circulation in Southeast Asia. Overall, our phylogenetic analysis showed that the CHIKV strains circulating in Cambodia from 2011 to 2013 belonged to the IOL lineage, most likely introduced in the country from Thailand somewhere between 2009 and 2011. The CHIKV from this outbreak harbored the E1:A226V substitution that is associated with an increase of CHIKV dissemination in *Aedes albopictus* mosquitoes, along with two other mutations in the E2 glycoprotein that seemed to be characteristic of the sequences falling in this clade. In contrast, viruses reported from Cambodia in 2020 lacked the E1:A226V substitution but harbored a double substitution in the surface proteins, which were recently associated with increased infectivity and transmission by *Aedes aegypti* (E1:K211E and E2:V264A) (303). Our analysis also showed that the CHIKV strains from the 2020 outbreak were phylogenetically closer to CHIKV circulating in Southeast Asia. This suggests that the recent outbreak was not seeded from previously CHIKV circulating in Cambodia, but instead from the introduction of the virus from neighboring countries. Indeed, our phylogenetic analysis inferred at least five different introductions, possibly from Thailand and China. Nonetheless, we should interpret these results carefully, in light of the genetic diversity sampled.

As noted in the main introduction of this thesis, a common problem in phylogenetic and phylogeographic inference is sampling bias (122, 137, 158). Having an unbiased sampling can be very hard to achieve as it requires (i) knowing the geographic extension of the outbreak, (ii) having access to the specific regions for sampling, and (iii) extensive sequencing efforts. Therefore, we might have an uneven sampling in most cases, leading to the absence or overrepresentation of samples collected from specific regions. In turn, this can lead to inadequate phylogenetic reconstruction, resulting in misleading conclusions about the geographical source of the outbreak. This might be the case for the 2011-2013 outbreak, from which we acknowledge that there is little to no genomic or epidemiological information on CHIKV circulating in neighboring countries in those years. For example, there is no data on CHIKV circulating in Vietnam or Laos between 2009 and 2011; therefore, we could completely exclude the possible "true" origin of the outbreak. Consequently, we acknowledge that our results rely on the available data giving us a picture

that might not have captured the complete information about the outbreak and might change if more information becomes available.

Still, we believe that this update on the CHIKV diversity is important because although CHIKV infection is not life-threatening, with no vaccine or treatment available, CHIKV imposes a significant burden on the health of the infected individuals. Currently, CHIKV circulates in several parts of the world, particularly in Africa and Asia. However, with increasing globalization, climate change, expansion of *Aedes* mosquitoes and possibly new adaptive viral mutations, CHIKV will probably continue to circulate and expand its global distribution, becoming an increasing threat to public health. For this reason, continuous surveillance and genomic investigations to identify genetic variations are crucial for developing effective diagnostics, treatments and future vaccines.

Finally, while we hope both studies bring new information to understanding CHIKV spread and evolution in Cambodia, our work leaves the door open to other interesting questions.

The first one concerns the inter-seasonal quiescence of CHIKV during the 2011-2013 three-year outbreak. Arbovirus dynamics are very seasonal and temperature dependent as they rely on mosquitoes to be transmitted, which are themselves highly influenced by temperature (304). In Cambodia, there are two main seasons: a dry season from October to April and a rainy season from May to September, when mosquitoes, and hence mosquito-borne viruses, are more likely to circulate. The typical seasonality of arboviruses in Cambodia is described with an increasing number of cases during the rainy season, peaking during July/August (305). Indeed, no inter-seasonal cases were documented during the 2011-2013 outbreak. However, our phylogenetic analysis suggests that the three-year epidemic was not the result of successive reintroductions but rather the result of one (or multiple) initial introductions in 2011 followed by local transmission. Therefore, how was the virus maintained throughout the dry season? Was it through mosquito vertical transmission cycles? Or maybe it was maintained thanks to a low transmission rate leading to a small and undetected number of cases? In this regard, there is evidence supporting both hypotheses. First, vertical transmission is considered a possible mechanism for the persistence of arboviruses during unfavorable periods (295). In these vertical transmission cycles, arboviruses such as CHIKV would be transmitted to the eggs from an infected female mosquito. Very importantly, this mechanism is possible as mosquitoes such as Aedes *Aegypti* exploit different strategies to resist hostile conditions. One of them is laying eggs that can withstand the desiccation caused by low humidity and high temperatures for long periods of time. Indeed, it has been estimated that Aedes *Aegypti* eggs can survive under unfavorable conditions for up to one year (306). Therefore, it could be hypothesized that during periods of drought, CHIKV persisted in the deposited eggs, which remained viable until they hatched at the beginning of the rainy season in May, leading to a subsequent period of CHIKV circulation. Second, while it is true that the intensity of transmission would probably wane as the dry season begins, mosquitoes are getting more and more adapted to urban settings, very often breeding in water-holding containers that people keep around their homes (307). Therefore, despite drier conditions, CHIKV transmission might not have stopped entirely, leading to a low and undetected but sufficient number of transmissions cycles that contributed to the persistence of CHIKV during the unfavorable periods.

The second question is about the inter-epidemic evolution and spread of CHIKV. Indeed, after nearly ten years of absence, we observed the re-emergence of CHIKV in Cambodia and neighboring countries (Thailand, Myanmar). In light of the genetic diversity of CHIKV sampled, our phylogenetic analyses clearly showed that the viruses captured in 2020 belonged to a separate cluster from those captured in 2011-

2013 while being genetically closer to the viruses circulating in Thailand, Bangladesh, China, Myanmar, and India between 2016 and 2019, all characterized for harboring the two AA substitutions E1-K211E and E2-V264A, in the background of E1-226A. Notably, this was also the case for the 2019 outbreak in Thailand and Myanmar, with recent and previously circulating viruses (in 2009 - 2013) falling distant in the global CHIKV phylogeny (Figure 6-10). Furthermore, in contrast to other arboviruses circulating in these countries, such as DENV, CHIKV does not circulate endemically but instead causes episodic outbreaks. This brings the following question: where does this CHIKV variant come from? Very interestingly, CHIKV strains harboring either the substitution E1-K211E and E2-V264A in the background of E1-226A (as we observed for the 2020 outbreak in Cambodia), or the E1-A226V substitution (as in the 2011-2013 outbreak) were detected co-circulating in India between 2010 and 2012 (56, 57). In the following years, cases have been reported annually in India (58), with viruses collected from 2012 until 2016 harboring E1-K211E and E2-V264A in the background of E1-226A. This fact suggests that CHIKV responsible for the recent epidemics in Cambodia, Thailand, and Myanmar probably evolved from viruses circulating in India in 2010. Therefore, based on the current genomic and epidemiological data available for the Southeast Asia region, a parsimonious explanation for the inter-epidemic evolution of CHIKV could be that after the epidemics were brought under control in 2013, the herd immunity in these populations started to wane slowly. Given that CHIKV was still circulating in countries like India, all this, in combination with other possible ecological and social factors taking place (e.g., travel, urbanization, climate change), led to the upsurge of CHIKV cases observed in 2018 – 2020, first in Thailand and then in Myanmar and Cambodia.

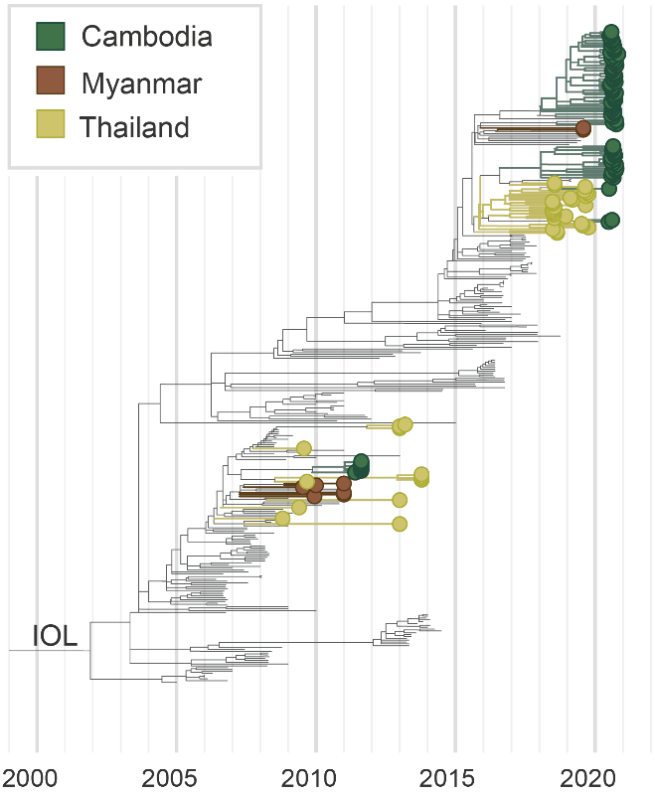


Figure 6-10: *Time calibrated phylogeny focusing on CHIKV belonging to the IOL. Genomes were collected from the two epidemics taking place during 2009-2013 and 2018-2020 in Cambodia, Myanmar, and Thailand. With different colors we highlight the different countries. Image taken from the Nextstrain build which is fully available at: https://nextstrain.org/community/Simon-LoriereLab/ChikungunyaCambodia2020@main*

The third open question is whether the different mutational profiles in the CHIKV strains collected in the 2011-2013 and 2020 outbreaks can be associated with a shift in the vector preferentially used for transmission. In this regard, a study suggested that the combination of both substitutions E1-K211E and E2-V264A in the background of E1-226A increased virus infectivity and dissemination in *Aedes aegypti* while having no significant effect on the fitness in *Aedes albopictus* (303). However, there is little information about the geographic distribution and relative abundance of *Aedes aegypti* and *Aedes albopictus*. For instance, only one recent study addressed the relative abundance of these mosquito vectors. Although the authors found that *Aedes aegypti* was slightly more prevalent than *Aedes albopictus*, the study limited the sampling and screening of mosquitoes to 24 schools located in two out of twenty-five Cambodian provinces (308). Our colleagues from the Institut Pasteur du Cambodge have suggested that both mosquito species circulate in the country almost homogenously.

Another exciting question related to the above is whether there is a link between the geographic distribution of *Aedes* mosquitoes and CHIKV strains with different genomic signatures: E1-226A (+E1-K211E and E2-V264A) versus E1-A226V. This question arises from previous publications where it was reported that CHIKV strains containing the *Aedes albopictus* adaptive mutation E1-A226V were detected in Southern India (309), where the relative abundance of *Aedes albopictus* is higher (310). In contrast, CHIKV harboring the substitutions E1-K211E and E2-V264A in the background of E1-226A were detected in Northern India (311), where *Aedes aegypti* is reported to be highly abundant. While this correlation is very intriguing, I believe that for addressing this question, we need (i) more *in vivo* evidence about the role of these mutations in the modulation of infectivity, dissemination, and transmission by the two different *Aedes* species, and (ii) more studies addressing the geographic distribution and relative abundance of these two mosquito species.

The last open question is related to the differences in chikungunya disease outcomes observed during the 2011-2013 outbreak in Cambodia. Although our work suggests that no additional mutations are required for CHIKV to reach the CNS, we are aware that our results do not show significant statistical support, mostly due to the low number of samples obtained from patients suffering encephalitis upon CHIKV infection (n=5). Consequently, the question remains, and our work only scraped the surface of this intriguing question. More work investigating larger cohorts, associating clinical and genomic data of patients with different clinical presentations, should shed light on the factors involved in the disease outcomes.

Key points of chapter 2

- We successfully set up an amplicon-based sequencing approach that allowed us to obtain the complete viral genomes from samples of varying viral load and quality.
- We shared the PCR amplicon scheme with the community and, very importantly, with our colleagues from Cambodia, providing building capacity.
- Both the 2011 and 2020 outbreaks in Cambodia appear to have been seeded from introductions from neighboring countries, particularly Thailand. We found that China was also a possible origin for the 2020 epidemic.
- We observed a different mutational profile in the CHIKV strains collected in the 2011-2013 and 2020 outbreaks, which can be associated with a shift in vector usage for transmission.

7 CHAPTER 3: Studying the epidemiology and intra-host evolution of SARS-CoV-2

7.1 Biology, epidemiology, and evolution of SARS-CoV-2

Coronaviruses (CoVs) are a large group of enveloped positive sense and single-stranded RNA viruses. Their name "corona" is due to the crown-like appearance of the spike protein on the virion surface when viewed under an electron microscope (Figure 7-1) (312).

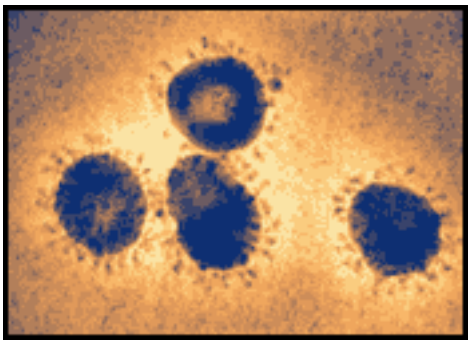


Figure 7-1: *Electron microscopy of Human 229E coronavirus. Image taken from CDC.*

CoVs cause mild to severe respiratory disease in humans and animals, including livestock, making them a challenge for public and animal health and an economic concern (313).

Based on genetic and serological characterization, CoVs fall into four different genera: Alphacoronavirus, Betacoronavirus, Gammacoronavirus, and Deltacoronavirus, with the first two mainly infecting mammals and the other two primarily birds (314)

Before the ongoing COVID-19 pandemic caused by SARS-CoV-2, we have witnessed in the last decades at least five significant outbreaks of coronaviruses which all had important consequences for our society:

1. Infectious bronchitis virus: it is a gammacoronavirus that causes infectious bronchitis in poultry, and although it was first identified in 1931, it still is one of the most important causes of economic loss within the poultry industry (315).
2. Transmissible gastroenteritis virus: it is a porcine enteropathogenic coronavirus belonging to the alphacoronavirus genera, which causes enteritis in pigs and can be lethal in piglets. It was first identified in the United States in 1946 and subsequently identified in Europe, Africa, South America, and China (316).
3. Porcine epidemic diarrhea virus: porcine alphacoronavirus emerged in 1971 in Belgium, and from there, it spread to Europe, Asia, and North America, where it has circulated ever since. Like TGEV,

it causes acute diarrhea and high mortality in piglets. However, while TGEV infections are currently controlled, the recent emergence of new PEDV strains resulted in a significant burden (317).

4. Severe acute respiratory syndrome coronavirus (SARS-CoV): is a zoonotic virus that emerged in humans, causing fatal respiratory illness. It was first detected in 2002 in China, and from there, it spread to North America, South America, Europe, and Asia before it was contained.
5. Middle East respiratory syndrome coronavirus (MERS-CoV): like SARS-CoV, MERS-CoV is a betacoronavirus belonging to the subgroup known as Sarbecovirus. It causes severe pneumonia with a higher fatality rate than SARS-CoV. Humans get infected through direct or indirect contact with infected dromedary camels or human-to-human transmission. MERS-CoV was first reported in Saudi Arabia in 2012, and since then, cases have been registered in 27 countries. (318)

In contrast to SARS-CoV, MERS-CoV, and SARS-CoV-2, there are four other strains of coronavirus causing seasonal and mild symptoms in humans, namely, HCoV-229E and HCoV-NL63 belonging to Alphacoronavirus genera and HCoV-OC43 and HCoV-HKU1, which belong to the Betacoronavirus genera but a different subgroup of Betacoronavirus named Embecovirus (319).

7.1.1 Emergence and spread of SARS-CoV-2

In late December 2019, a cluster of patients with pneumonia of unknown cause was detected in Wuhan City, Hubei Province, China (320). Very similar to the SARS-CoV and MERS-CoV outbreaks, these patients had symptoms of viral pneumonia such as fever, difficulty in breathing, cough, or chest pain (320). As was the case of the SARS-CoV outbreak in China in 2002, the first 27 cases of SARS-CoV-2 infection were epidemiologically linked to a traditional market, this time the Huanan Seafood Wholesale Market (321). These markets, sometimes also called "wet markets[2]," sell fresh meat, fish, agriculture products, and live animals, including poultry and wildlife (321). Interestingly, according to a summary report from the Chinese Center for Disease Control and Prevention, the symptoms of the first known case, identified retrospectively in Wuhan City, started on the 8th of December 2019 (322). Although this information does not coincide with the patient's interview in which he states that the symptoms began on the 16th of December (323), it means that the virus was circulating even before the first documented cases. By the end of December, the number of cases of pneumonia associated with the market increased, prompting the Wuhan Health authorities to communicate to the general public and the WHO on the 31st of December, about a pneumonia outbreak of unknown etiology. Consequently, on the 1st of January 2020, the market in Wuhan was closed because of sanitary reasons (324).

Subsequently, bronchoalveolar samples from these infected patients were sequenced using an untargeted metagenomic approach, and it was identified that the etiologic agent was an unknown betacoronavirus (320, 325, 326). On the 10th of January 2020, the first complete genome of the novel coronavirus was made available on the Virological website (327), and later on, more complete genome sequences from different research centers were released on the GISAID database (328). Later, patients with no relation to the market in Wuhan and several familial and healthcare facilities clusters were identified, which proved clear evidence of human-to-human transmission (329, 330). The virus spread quickly throughout China,

[2] The “wet markets” are also called public, informal and traditional markets and it refers to a marketplace selling perishable goods as distinguished from “dry markets” that sell durable goods such as fabric and electronics.

and by the end of January 2020, cases were reported in 34 Chinese provinces. Therefore, on the 30th of January, WHO declared the novel coronavirus outbreak an international health emergency (331). Very importantly, on the 11th of February, the previously known "2019 novel coronavirus" was given the official name of severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) by the International Committee on Taxonomy of Viruses, and the WHO named the disease COVID-19 (332). In order to control the spread of the virus, the Chinese government established strict public health measures: cities were closed, and outdoor activities and social gatherings were restricted. In parallel, several countries implemented specific syndromic surveillance to identify SARS-CoV-2 infections. Through such national surveillance efforts, the first COVID-19 cases in Europe were detected in France and Germany on the 24 and 28 January 2020, respectively (333). Despite many efforts, the virus spread quickly, leading the WHO to declare a pandemic on the 11th of March 2020. Since then, SARS-CoV-2 has been circulating, wreaking havoc worldwide.

7.1.2 Origins of SARS-CoV-2 (as of December 2021)

Since the identification of the first COVID-19 cases, one of the questions that have generated considerable discussion among researchers and the non-scientific community has been the origin of SARS-CoV-2. While all the possible theories about the origin of SARS-Cov-2 have already been reviewed (334, 335), here I will only describe the most outstanding genomic features of SARS-CoV-2 and how these features support the animal origin as the most plausible and parsimonious scenario.

Similar to other coronaviruses, SARS-CoV-2 has an RNA genome of approximately 30 kb long, encoding four structural proteins, including the spike protein (S), an envelope protein (E), membrane protein (M), and nucleocapsid protein (N). The four structural genes of SARS-CoV-2 share more than 90% AA identity with those of SARS-CoV except for the S gene, which is more divergent. Nevertheless, similarly to SARS-CoV, the S protein of SARS-CoV-2 plays a key role in receptor recognition and cell membrane fusion. It is functionally divided into two subunits, with the S1 subunit containing the receptor-binding domain (RBD), thus responsible for the receptor binding, whereas the S2 domain mediates the cell membrane fusion process (336). In addition, several open reading frames lead to non-structural and accessory proteins. Within the non-structural proteins, the nsp14 stands out. It is part of the replication-transcription complex, and very interestingly, it has 3'-to-5' exonuclease activity, assisting RNA synthesis with RNA proofreading activity (337). Such a proofreading complex might explain why coronaviruses generally have a low mutation rate compared to other RNA viruses.

Initial phylogenetic analysis of the full-length genome of betacoronaviruses shows that while SARS-CoV-2 clusters together with SARS-CoV and SARS-related coronaviruses (SARSr-CoVs) within the subgenus Sarbecovirus, it forms a distinct cluster with coronaviruses detected in horseshoe bat (RaTG13, RmYN02, ZC45 and ZXC21) and pangolins (Figure 7-2) (313). Indeed, RaTG13, a bat coronavirus identified in *Rhinolophus affinis* from Yunnan province in China, was initially the closest relative to SARS-CoV-2 sharing 96.2% of nucleotide identity (326). More recently, three coronaviruses were identified in three different *Rhinolophus* bat species collected from northern Laos in the Indochinese peninsula (338). These bat viruses exhibit high nucleotide identity, particularly the virus named BANAL-52, which presents 96.8% of nucleotide identity (338). Notably, while the RBD of RaTG13 differs in five out of the six key contact residues for binding the human receptor ACE2, BANAL-52 differs in only one. This high genetic similarity

makes BANAL-52 the closest virus to SARS-CoV-2, as of December 2022, while supporting a bat origin of SARS-CoV-2 (326).

Despite the relevance of bats in coronavirus outbreaks, two facts might suggest that another animal may have acted as an intermediate host between bats and humans. First, the genetic distance with the Wuhan-Hu-1 reference sequence of RaTG13 and BANAL-52 is 3.8% and 3.2%, respectively. This means a difference in 1150 and 967 sites across the genome, respectively. Considering an evolutionary rate of 0.80×10^{-3} substitutions/site/year (339), this reflects decades (~ 40 years) of evolutionary divergence, as also suggested by Boni et al. (340). This evolutionary gap demonstrates that neither RaTG13 nor BANAL-52 is the direct progenitor of SARS-CoV-2 (334). Second, the SARSr-CoVs found in horseshoe bats were sampled very distant from the first COVID-19 detected cases.

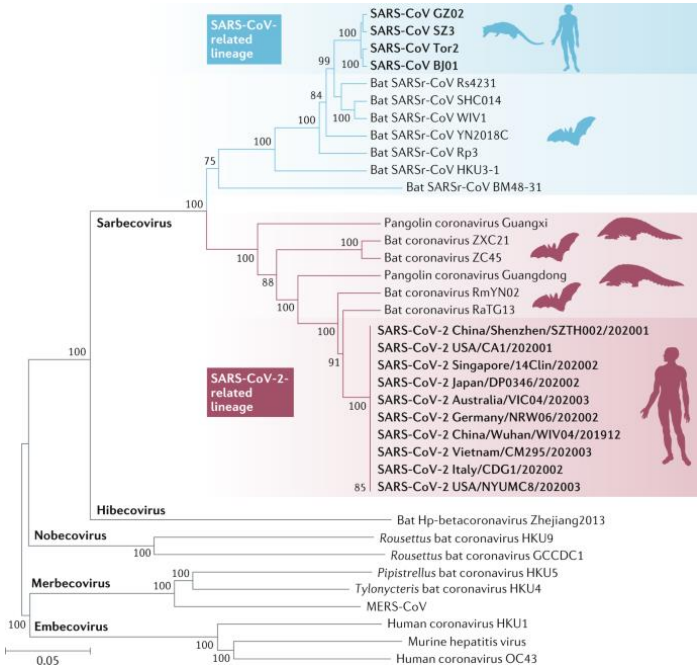


Figure 7-2: *Phylogenetic tree of the full-length genome sequences of SARS-CoV-2, SARSr-CoVs and r representative members of the different subgenera of betacoronaviruses. Figure extracted from Ben et al. (2021).*

Beyond bats, SARS-CoV-2 related viruses have been identified in pangolins (Pangolin-CoV) (341). Remarkably, while RaTG13 has the highest average genetic similarity to SARS-CoV-2, the S protein's receptor-binding domain (RBD) of SARS-CoV-2 showed higher AA similarity with the one of Pangolin-CoV (341, 342). More specifically, within the RBD, Pangolin-CoV has only one AA variation from SARS-Cov-2, and this difference is not of the six key residues constituting the receptor-binding motif involved in the interaction with the human receptor ACE2 (341). Nevertheless, pangolins infected with coronaviruses do not stay healthy. Generally, they exhibit signs of respiratory disease, including shortness of breath, lack of appetite, and weight loss (342). This fact suggests that the pangolins are not the natural reservoir of these coronaviruses and likely acquired the viruses from other hosts.

Furthermore, Pangolin-CoVs identified to date have an overall genome identity of 92% with SARS-CoV-2 (341), indicating that pangolins would not be the intermediate host of SARS-CoV-2. Additionally, because SARS-CoV-2 is closer to the RaTG13 bat virus in all genomic regions but in the RBD, which is genetically closer to a pangolin virus, it has been hypothesized that SARS-CoV-2 could be a recombinant of an ancestor of Pangolin-CoV and RaTG13 (343, 344). However, Boni et al. presented evidence against that hypothesis and proposed that RaTG13 is the product of recombination from an unsampled bat coronavirus (340). Therefore, the authors suggested that the most parsimonious explication is that the shared RBD residues binding to ACE2 were present in the common ancestor leading to RaTG13, Pangolin-CoV and SARS-CoV-2 but were lost in RaTG13 through recombination (340).

So, how did the virus get into humans? There is not yet a clear answer for that. While more animals are being found to be susceptible to SARS-CoV-2 infection, such as raccoon dogs, civet cats, and minks, all for sale in the Hunan market (26), there is no clear evidence of which could have acted as an intermediate host. Nevertheless, this highlights the importance of continuing sampling animals and suggests a central role of SARSr-CoV–susceptible animals that might have been a direct SARS-CoV-2 progenitor and, therefore, the primary source of human infection.

The direct progenitor of SARS-CoV-2 has not been found yet, leading to the rise of alternative hypotheses such as the “laboratory escape” scenario (334, 335). Supporters of such scenarios also rely on the fact that notable genomic features distinguish SARS-CoV-2 from SARS-CoV and other betacoronaviruses. Nevertheless, as I will explain more in detail in the following lines, all notable SARS-CoV-2 features have been already observed in nature and very importantly, in related coronaviruses (334, 335).

Two major genomic features distinguish the SARS-CoV-2 genome from SARS-CoV and other betacoronaviruses. First, despite binding to the same cellular receptor, ACE2 (345), the SARS-CoV-2 RBD sequence is different from that of SARS-CoV. Specifically, they differ in the six key residues on the RBD, critical for human ACE2 binding. Remarkably, biochemical studies have shown that these changes have stabilized the interaction between the two RBD binding hotspots on the surface of the ACE2 receptor (346) and have strengthened the binding affinity to ACE2 compared to that of SARS-CoV (346, 347). Nevertheless, as already mentioned, the RBD region of SARS-CoV-2 is almost identical at the AA level, including key residues, to the one in pangolin coronaviruses (Figure 7-3) or BANAL-52 (338). Therefore, the presence in pangolins and bats of an RBD very similar to that of SARS-CoV-2 proves that the “alternative solution” of the SARS-CoV-2 spike protein for binding the human ACE2 already existed in nature and probably is the result of natural selection (335). In addition, despite SARS-CoV-2 RBD’s high binding affinity for human ACE2, ongoing SARS-CoV-2 evolution has led to the emergence of several mutations in the RBD region (e.g., D614G, N501Y), enhancing its binding affinity to the human ACE2 receptor (88, 348), suggesting that there was and may still be room for further SARS-CoV-2 evolution and human adaptation. In this way, the hypothesis of the “creation” or “laboratory escape” of a humanized virus, highly adapted to infect humans, is weakened (334, 335).

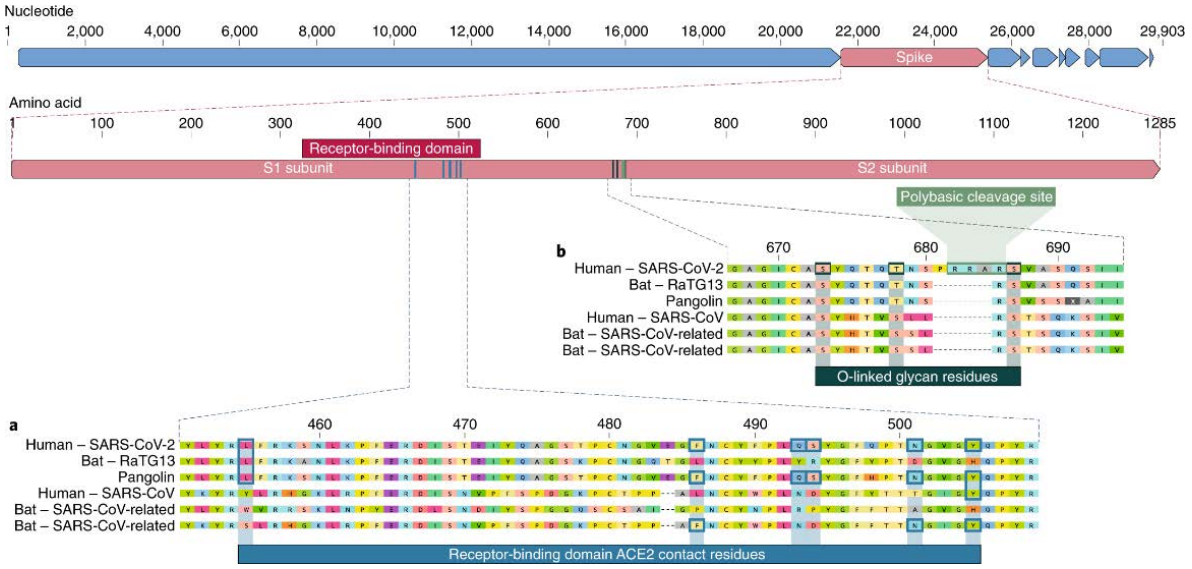


Figure 7-4: *Features of the spike protein in human SARS-CoV-2 and related coronaviruses; a) RBD sequence of SARS-CoV-2 and the most closely related SARS-CoV-like coronaviruses and SARS-CoV is shown. Key residues in the spike protein in contact with the ACE2 receptor are marked with blue boxes in both SARS-CoV-2 and related viruses; b) S1/S2 polybasic furin cleavage site and the three adjacent predicted O-linked glycans are present only in SARS-CoV-2 and were not previously observed in lineage B betacoronaviruses. Figure extracted from Andersen et al. (2020).*

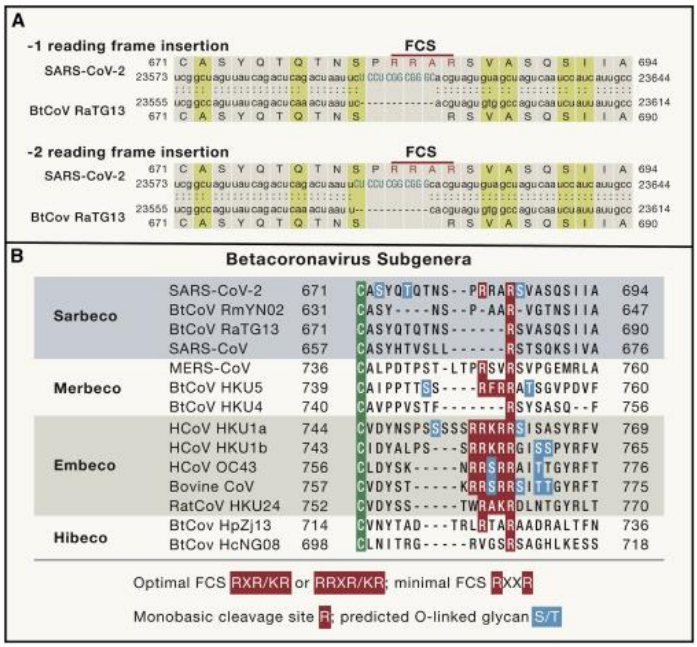


Figure 7-3: *Evolution of the furin cleavage site in the spike protein of betacoronaviruses. A) Sequence alignment of the region around the furin cleavage site (FCS) in SARS-CoV-2 and RaTG13. B) AA sequence alignment of the FCS region in representative members of the different subgenera of betacoronaviruses. The less functional AA motif (RRAR) in SARS-CoV-2 is highlighted in opposition to the more functional motif RXR/KR or RRXR/KR present in other coronaviruses. Extracted from Holmes et al., (2021).*

The second specific genomic feature of SARS-CoV-2 is a four AA insertion (PRRA) between the subunits S1 and S2 of the S protein, which generates a polybasic cleavage site (RRAR). This site enables an efficient cleavage by furin and proteases, essential for spike-driven viral entry into cells (349). Such a cleavage site is not present in any member of the subgenus Sarbecovirus identified, including BANAL-52. However, although it does not constitute a functional polybasic cleavage site, RmYN02 contains three AA insertions (PAA) (350), and a canonical furin cleavage site (RRKR) has been described in betacoronaviruses belonging to the subgenus Embecovirus, namely HKU-1 (335) (Figure 7-4). This clearly shows that insertions and deletions near the S1/S2 junction are recurrent among coronaviruses and not entirely exclusive to SARS-CoV-2 as first thought. Additionally, as the furin cleavage site is present across the coronavirus family tree, the site is probably the result of convergent evolution. Therefore, the polybasic furin cleavage site in SARS-CoV-2 could have arisen naturally (334, 335).

Overall, while the ‘‘laboratory escape’’ scenario cannot be completely ruled out, the epidemiological history and genomic features of SARS-CoV-2 strongly support that SARS-CoV-2 is a zoonotic virus. Furthermore, the similarity between SARS-CoV and SARS-CoV-2 emergence suggests they might share similar origins: a virus from bats that spread to humans through an intermediate host (civets for SARS-CoV). While it is true that the specific zoonotic origin of SARS-CoV-2 has not been yet determined, this is often the case. For example, it took several years to determine the origin of SARS-CoV, and for many other well-known viruses such as EBOV, the zoonotic origin is still unclear. The main reason might be that the right animals might not have been sampled yet. Therefore, more investigations, including sequencing SARS-CoV-2 from very early human cases and different animals, could shed light on the origin of the virus, providing a better understanding of the emergence of the pandemic.

7.1.3 SARS-CoV-2 evolution – (as of January 2022)

Understanding the ongoing evolution of SARS-CoV-2 is crucial for establishing accurate public health responses (e.g., vaccination), ultimately controlling the pandemic. So far, we have witnessed more than two years of SARS-CoV-2 evolution, leading us to complete half the Greek alphabet.

While this pandemic has provided us with an extraordinary example of viral evolution in real-time, significant efforts from national authorities, institutions, expert networks, and researchers were needed to monitor SARS-CoV-2 evolution. This extensive genomic surveillance has popularized the use of genomic tracking tools to analyze genomes in real-time such as Nextstrain (351), the creation of new platforms for efficient sharing of the information such as CoVariants or CoV-Lineages, and a formal system to classify and identify SARS-CoV-2 variants. Initially, three nomenclature systems were developed for epidemiological surveillance. First, the GISAID nomenclature assigns SARS-CoV-2 genomes within 11 major clades (352). Currently, the GISAID clades include the S and L clades, the two earliest clades to be identified, with the SARS-CoV-2 reference strain belonging to clade L; the V and the G clades, which evolved from the L clade. It also includes the GH, GR, and GV clades, all three descendants from the G clade; the GRY clade, which split from the GR clade around September 2020 and is today best known as the Alpha variant; the GK clade, also known as the Delta variant, evolving from the G clade; and lastly, the GRA clade which

emerged from the G clade with no clear progenitor and today is best known as the Omicron variant (352) (Figure 7-5).

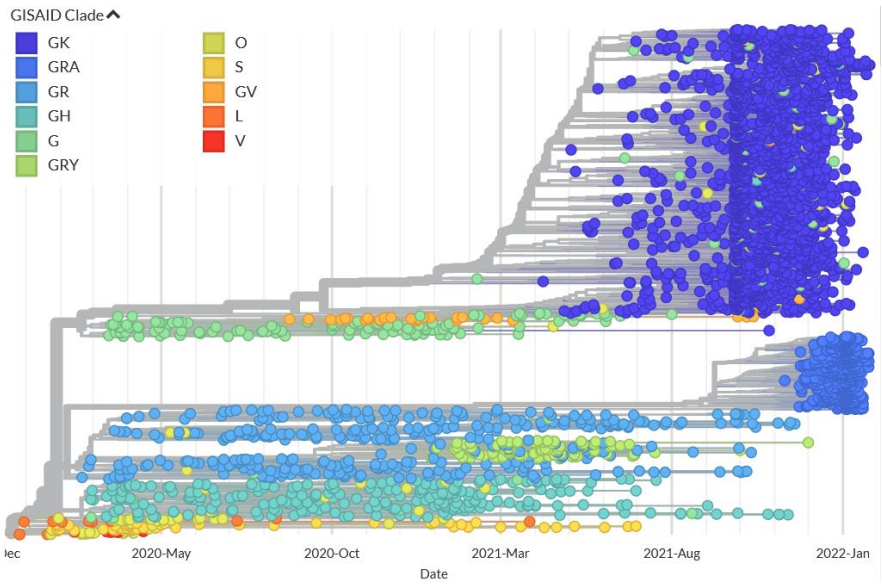


Figure 7-5: *Global phylogeny of 3223 SARS-CoV-2 genome from the origin of the pandemic to January 2022. The different GISAID clades in which SARS-CoV-2 genomes fall into are shown with the different colors. Screenshot was taken from Nextstrain website as January 2022.*

Second, the PANGO nomenclature consists of a dynamic and tree-based nomenclature implemented through the Phylogenetic Assignment of Named Outbreak LINeages (PANGOLIN) software developed by Rambaut et al. (353). The PANGOLIN tool assigns to SARS-CoV-2 query sequences the most likely Pango lineage, which was first described in Rambaut et al. (353) and now is constantly updated by the Pango Network Team (https://www.pango.network). Further details on the PANGOLIN tool are described in the recent publication by O'Toole et al. (354). The PANGO nomenclature offers a more detailed lineage classification aiming to track and understand the patterns of the global spread of the different variants driving the pandemic. Each label consists of an alphabetical prefix and a suffix containing up to three numbers separated by periods indicating the descendent sub-lineages, such as B.1.1.7. A complete list of the current PANGO lineages can be found at (355). Third, there is the Nextstrain nomenclature in which a new clade is defined when it reaches more than 20% of frequency in a representative global sample and differs by at least two mutations from its parent clade (356). Each clade's label consists of two numbers representing the year of emergence followed by a capital letter.

Furthermore, given the continuing emergence of variants and the continuing change in our understanding of their impact, variants have been further classified as variants of concern (VOC), variants of interest (VOI), and variants under monitoring (VUM) (357). The WHO has defined these variants as follows:

1. A VOC is a variant with evidence of increased transmissibility, more severe disease (e.g., increased hospitalizations), reduction in neutralization by antibodies generated either by natural infection or vaccination, decreased effectiveness of public health measures, or diagnostic detection failure.

2. A VOI is a variant responsible for significant community transmission, suggesting an emerging global risk, and its mutations might affect the transmission, diagnostics, and therapeutics. A VOI would therefore require enhanced surveillance and further laboratory characterization.

3. VUM is a variant with mutations that are suspected to affect the properties of the virus, imposing future global risk; however, evidence of phenotypic or epidemiological impact is unknown. Similar to VOIs, VUM requires enhanced monitoring and laboratory analysis.

Furthermore, to achieve a more straightforward and practical discussion with the general public, the WHO recommended using Greek Alphabet letters to label key VOCs and VOIs (358).

The first evidence of SARS-CoV-2 mutations having a substantial effect on the virus was the emergence of the substitution D614G in the S protein. It was first detected early, in March 2020, and it was later observed to emerge multiple times independently and simultaneously in the global SARS-CoV-2 population, suggesting convergent evolution and an adaptive benefit of this mutation (82). Supporting the latter, by June 2020, the prevalence of the D614G substitution rose to nearly 100%, and now several lines of evidence suggest that SARS-CoV-2 variants carrying this substitution have increased transmissibility (84-86). Shortly after, several SARS-CoV-2 variants emerged associated with increased transmissibility, decreased sensibility for antibodies generated either from previous natural infection or vaccination, and increased risk of reinfection; and therefore, they have been identified as VOCs. As of the 30th of January 2022, the Alpha, Beta, Gamma, Delta, and Omicron variants have been designated VOCs. According to the PANGO nomenclature, the first of these variants to emerge was the Alpha variant or B.1.1.7. The earliest known cases of the Alpha variant were detected in England in late September 2020, and by January 2021, the variant was being detected in 45 countries across the globe (87). The variant carried an unusual number of genetic mutations, including 14 non-synonymous substitutions and three deletions (359). Among these mutations, several were located in the S protein, and in particular, three of them showed to have critical biological effects: the N501Y substitution, which increased the viral binding affinity to the human ACE2 receptor (88); the 69/70 deletion, which affected PCR assays targeting the S gene and was potentially associated with immune evasion (360); and the P681H mutation, located in the S1/S2 furin cleavage site and it has been suggested to facilitate viral fusion and entry (361). After the Alpha variant emergence, two other novel and rapidly growing lineages with many genetic changes were identified in South Africa (362) and Brazil (363). These lineages were given the name of Beta and Gamma variants or B.1.351 and P1, according to the PANGO nomenclature, respectively. The Beta variant was initially reported in South Africa in October 2020, where it spread rapidly, accounting for 87% of the sequenced COVID-19 cases at the beginning of December 2020, and by February 2021, the variant had been reported in 45 countries (364). Early estimations indicated that the variant was 50% more transmissible than preexisting variants in South Africa, causing a rapid saturation of public health services (365). The Beta variant showed decreased neutralization activity from previous infection and vaccination with Pfizer, Moderna, and AstraZeneca vaccines (364). The characteristics of the Beta variant can be attributed to the set of 23 mutations present throughout the genome, including 17 AA changes. Among these substitutions, three are located in the RBD of the spike protein and have been associated with (i) higher binding affinity to the human ACE2 receptor (N501Y) (88), (ii) reduced antibody neutralization (E484K) (366, 367), (iii) and reduced sensibility to antiviral treatments (K417N) (368).

Interestingly, the Gamma variant was reported almost simultaneously in a different continent and contained almost the same RBD mutations: the N501Y, the E484K, and the K417T substitutions. For the latter, in the Beta variant, an N was observed instead of a T. Again, convergent evolution highlights the adaptive benefit of these substitutions. The Gamma variant was first detected in Manaus, Brazil, with a surge of new cases at the beginning of November 2020 (363), and like the other VOCs, it quickly spread throughout the country and internationally (369). Furthermore, it was estimated that this variant was between 1.7 to 2.4-fold more transmissible than preexisting circulating variants and associated with higher mortality risk (363). As observed during the emergence of the other variants, the increased transmissibility of the Gamma variant led to an overload of the health systems. Therefore, it is difficult to determine whether a higher mortality risk is due directly to the infection with this new variant or to the consequent saturation of health systems.

Although these variants disseminated rapidly across the globe, they were quickly displaced in many parts of the world by the next emerged VOC: the Delta variant (370). It was first detected in India by the end of September 2020, and by April 2021, it had already replaced previously circulating variants, the Alpha and the Kappa variants, causing a major surge of infections, which peaked at 400,000 new cases per day (370, 371). A similar increase was later observed in the UK and the USA (372, 373), and by July 2021, the Delta variant was present in 130 countries across the globe (374). Studies have shown that the Delta variant has increased transmissibility and hospitalization rates (375, 376) and reduced vaccine effectiveness (377-379). These characteristics were attributed to the set of 30 novel mutations present in the Delta variant compared to the SARS-CoV-2 reference strain. Among those mutations, several are located in the S protein, including two substitutions in the RBD, L452R and T478K, thought to reduce antibody neutralization (380) and the substitution P681R in the S1/S2 furin cleavage site, thought to facilitate the cleavage of the full-length spike to S1 and S2, enhancing viral fusogenicity (381-383) During its spread, the Delta variant has been evolving and diverging into multiple sublineages or clades (384). According to the Nextstrain nomenclature, the Delta variant has been classified into three major clades, 21A, 21I, and 21J. In the PANGO nomenclature, the Delta variant includes the Pango lineage B.1.617.2 and 180 descendent sublineages, all named AY, as aliases of B.1.617.2, followed by a number (or several) for distinction (355). These features of the Delta variant, together with its overwhelming circulation, prompted the idea that the next VOC would evolve from this lineage.

Nevertheless, the next VOC that emerged, Omicron, is phylogenetically distinct from any VOC or VOI or any other SARS-CoV-2 variant previously circulating in Southern Africa, where it was first detected (385). This genetic distance is reflected by the length of the branch, which is rooted in the B.1.1 lineage (Nextstrain clade 20B) and gives rise to the Omicron clade (Figure 7-6). As expected from such a long branch, the Omicron variant carries a high number of mutations, with 47 lineage-defining mutations (386). Notably, it carries 30 mutations in the spike protein compared to the ancestral virus, 15 of which are located in the RBD (385). Among which, the substitutions G339D, N440K, S477N, T478K and N501Y have been associated with an increase of the binding affinity to the ACE2 receptor (88). In addition, three other substitutions, H655Y, N679K, and P681H, are located next to the S1/S2 furin cleavage site. While P681H has already been detected in the Alpha variant and is predicted to enhance the speed and efficiency with which viral and membrane fusion occurs (361), there is no data available for the other two. However, the H655Y has been reported in the Gamma variant, suggesting that it might provide an adaptive advantage.

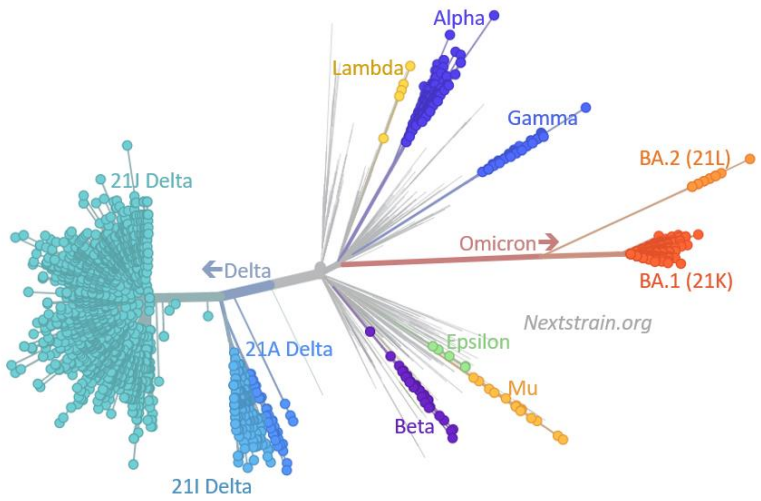


Figure 7-6: *SARS-CoV-2 family tree, generated from the data available on GISAID and diagrams created by Nextstrain and modified by Emma Hodcroft to visualize how distant the Omicron family is from the rest of SARS-CoV-2 lineages.*

Furthermore, the remarkable genetic distance, especially in the RBD of the spike protein, between the Omicron variant and the previously circulating variants, including the SARS-CoV-2 reference strain used in the current vaccines, correlates with current evidence regarding the new variant: decreased susceptibility to monoclonal antibodies (387); strong immune evasion from neutralizing antibodies conferred by prior infection or vaccination (387-389) and a higher frequency of reinfections (390). Therefore, this increase in infections rates and vaccines breakthroughs, and probably, an inherent increase in transmissibility, might explain why Omicron has been spreading very rapidly worldwide despite the high levels of circulation of the Delta variant. Early evidence also suggests that the Omicron variant's infection causes less severe disease than infection with Delta, although some people may still develop severe symptoms. For this reason, despite the strong immune evasion from vaccine-induced protection, vaccines are still expected to protect against severe illness, hospitalizations, and deaths and are thus highly recommended.

Based on the diversity of the first sampled genomes, it has been estimated that the Omicron variant has been circulating since around mid-October 2020 (391). This observation created much discussion among the scientific and non-scientific communities about the possible origin of the Omicron variant. Three hypotheses have been postulated: first, one year of undocumented circulation; second, continual evolution in a chronically infected individual (e.g., an immunocompromised patient) and subsequent spill back to the human population; and third, reverse zoonosis followed by new zoonosis (391, 392). There is evidence to support all three hypotheses. First, although undocumented circulation in the general population for almost one year seems unlikely given the great sequencing efforts worldwide, there are significant disparities in genomic surveillance across countries. In particular, Omicron was first detected in South Africa and Botswana, countries in which only a small fraction of the documented cases are sequenced – a ratio that is even lower for South Africa neighboring countries such as Tanzania or Mozambique. Hence, the accumulation of the high number of mutations giving rise to the Omicron variant could have occurred undetected (392). In support of the second hypothesis, several studies have reported a substantial accumulation of mutations in the SARS-Co-V-2 genome during long-term infections (393-

395). As such, the heavily mutated RBD of the Omicron variant has led to the belief that it could have emerged as a result of the antibody-mediated selective pressure during long-term infection. Nonetheless, previous reports of chronically SARS-CoV-2 infected individuals actually found a much lower number of mutations than observed in the Omicron variant. For instance, while the virus described by Choi et al., (394) after 146 days of evolution in an immunocompromised patient harbored 7 non-synonymous mutations and two deletions in the spike protein, the Omicron variant presented 15 mutations in the RBD only. Lastly, the zoonotic origin of the Omicron variant is supported given (i) the prior knowledge that SARS-CoV-2 is capable of infecting several animal species, including cats and dogs, (ii) the observation of human-animal-human transmission (396), and (iii) the detection of several Omicron variants in SARS-CoV-2 sampled in animals including rodents (397) or mouse-adapted strains (398).

The Omicron variant was initially designated as belonging to a new PANGO lineage, the B.1.1.529; more recently, it has been classified into three sublineages: BA.1 (the initial clade), BA.2, and BA.3, which corresponds to the 21K, 21L and 21M according to the Nextstrain nomenclature, respectively. Phylogenetic analyses have shown that these three sublineages have emerged from the B.1.1 node and subsequently evolved independently from one another (385). This divergence is such that BA.1 and BA.2 differ in at least 40 AA sites, being as divergent as the Alpha, Beta, and Gamma are between each other. Given the large number of genetic changes differentiating BA.1, BA.2, and BA.3 between each other and from other SARS-CoV-2 lineages, it was considered possible that: (i) the three lineages descended from a recombinant ancestor, (ii) one or more viruses from the BA lineage emerged from the recombination with BA and non-BA lineage viruses, and (iii) one of the viruses from the BA lineage is the result from the recombination of the other two BA lineage viruses. Despite one initial study, there is still no clear answer (385).

Interestingly, it has been pointed out that most of the numerous mutations present in the RBD of the spike protein which distinguishes the Omicron variant from the Delta variant are shared by both the BA.1 and BA.2. Indeed, it is the N-terminal domain of the spike protein and other genomic regions such as the ORF1a, that differ the most (399). While BA.1 was the primary and first sublineage of Omicron in circulation, the BA.2 currently appears to be the major Omicron lineage in India and is growing in frequency compared to the BA.1 in the UK, Denmark, USA, Germany, and Sweden (399).

The emergence of the Omicron variant generated much debate about whether we will reach the end of the pandemic. However, no one anticipated the Delta nor the Omicron variant, so probably Omicron will not be the last variant we will hear about. The big question now is whether our vaccines and countermeasures will be less effective against future variants.

7.1.4 Clinical presentation

Patients infected with SARS-CoV-2 can experience various clinical manifestations, from no symptoms to severe illness depending on: (i) the infecting variant (e.g., Alpha, Delta, and Omicron), (ii) the presence of natural-induced immunity, (iii) the degree of vaccination (e.g., non-vaccinated people, fully vaccinated people), and (iv) the health status of the infected person.

Patients with comorbidities are more likely to develop severe respiratory diseases (400). Such comorbidities include being older than 65, having cardiovascular disease, diabetes, cancer, or obesity, being a smoker, and receiving immunosuppressive therapy (401).

Upon infection, the most common symptoms are fever, fatigue, cough, loss of taste or smell, and dyspnea in severe cases (400). In general, symptoms appear after an incubation period of 1-14 days with an average of 5 days. The severe disease usually develops on day eight after symptoms onset and critical disease and death on day 16 (313). Asymptomatic SARS-CoV-2 infection can occur, in particular in children and young adults. Two meta-analysis studies indicate that asymptomatic infections among COVID-19 patients are around 40% (402, 403). However, it is unclear what percentage of individuals who present an asymptomatic infection stay asymptomatic throughout the infection or progress towards clinical disease (400). There are conflicting opinions regarding the impact of asymptomatic infections on the progression and dynamics of the pandemic. Some consider that asymptomatic infections are a significant public health risk as asymptomatic individuals would more likely be out in the community than confined at home, hence a significant source of viral transmission. In contrast, other researchers believe that asymptomatic infections are not the primary drivers of the pandemic as asymptomatic patients would not be sneezing or coughing as much as symptomatic people do (404). In general, acute SARS-CoV-2 infections are resolved in 1-2 weeks; however, multiple reports have shown that people with compromised immune systems can remain infected for a longer period. In addition, there are increasing reports of patients who experience persistent and prolonged symptoms after acute COVID-19. This condition is usually referred to as “long COVID” (400).

Reinfections with SARS-CoV-2 after recovery from prior infection have been widely reported (400). Such reinfections may occur because of the waning of the immunity against the infection over time or the emergence of new variants that evade preexisting immunity. Although the prevalence of reinfection is unknown, it might increase with the circulation of the new variants (405). Alarmingly although expected, reinfections have been seen in fully vaccinated people. An infection in a completely vaccinated person is a “vaccine breakthrough infection”(406). Vaccine breakthrough infections are expected because while the vaccines had never been 100% effective against the original strain they were designed, vaccines’ efficacy is significantly lower against the new variants (407). This is because new variants, such as the Delta or Omicron variant, blunt the potency of antibodies developed against older versions of SARS-CoV-2. Furthermore, two studies have shown that after the two doses of Pfizer–BioNTech vaccine, one of the most widely used COVID-19 vaccines, the humoral response significantly decreases within months (408, 409). Such waning of the vaccine-induced protection would help to explain why vaccine breakthrough infections occur. Despite the reduction of circulating antibodies, vaccines are still effective in preventing people from developing severe symptoms and hospitalization. Indeed, fully vaccinated people are less likely to get infected than those who are unvaccinated, and if they do, they tend to develop less severe diseases than unvaccinated people (406). For this reason, the WHO and CDC still recommend a booster dose of Pfizer-BioNTech, Moderna, and the second shot of Johnson & Johnson’s Janssen. As of January 2022, there are ten vaccines against COVID-19 approved by the WHO, and several companies are already working on producing variant-specific vaccines to protect against the different SARS-CoV-2 variants (170).

7.2 Studies carried out in the context of the "Corona Task Force" at the Institut Pasteur

Following the emergence of SARS-CoV-2 in China, a syndromic surveillance was implemented in France on the 10th of January 2020 to identify cases and prevent onward transmission in the community. Shortly after, to respond to the global sanitary crises, the Institut Pasteur set up a "Corona Task Force" with the National Reference Center for Respiratory Viruses (NRC) at its heart.

At the end of February 2020, I volunteered to join the "Corona Task Force" to be part of the many efforts of the Institut Pasteur's mobilization against COVID-19. Specifically, in collaboration with the NRC, I have contributed to the implementation of an amplicon-based high throughput sequencing technique which has been the technique used by the NRC to ensure genomic surveillance during the surge of SARS-CoV-2 infections in France. I have also actively contributed to the genomic analysis of the first introductions and spread of SARS-CoV-2 in France. Still, in support of the NRC, I have contributed to other studies such as the characterization of SARS-CoV-2 detected in a mink farm in France; the genomic epidemiology of SARS-CoV-2 in Guadeloupe, Saint Barthélemy, and Saint-Martin from February to April 2020 (a work in collaboration with Institut Pasteur of Guadeloupe); and a longitudinal follow-up of an immunocompromised patient infected with SARS-CoV-2.

This chapter reports on two of the works mentioned above. The first work involves the study of the initial introductions and early spread of SARS-CoV-2 in the Northern regions of France. The samples were obtained within the syndromic surveillance framework implemented in France early after the disease reports in China. From such surveillance, we obtained 100 complete genome sequences of SARS-CoV-2 from the earliest cases detected in France and reconstructed their phylogenetic relationships in the context of sequences drawn globally at the time. Such analysis allowed us to get insights into the initial introductions in France and the later spread of the virus at the local level, and the impact of the containment measures imposed on early detected (imported) cases.

The present work has been published as:

Fabiana Gámbaro , Sylvie Behillil , Artem Baidaliuk, Flora Donati , Mélanie Albert , Andreea Alexandru , Maud Vanpeene , Méline Bizard, Angela Brisebarre , Marion Barbet , Fawzi Derrar , Sylvie van der Werf, Vincent Enouf , Etienne Simon-Loriere . **Introductions and early spread of SARS-CoV-2 in France, 24 January to 23 March 2020**. Euro Surveill. 2020;25(26): pii=2001200. https://doi.org/10.2807/1560-7917.ES.2020.25.26.2001200

RAPID COMMUNICATION

Introductions and early spread of SARS-CoV-2 in France, 24 January to 23 March 2020

Fabiana Gámbaro[1,2,3], Sylvie Behillil[3,4,5], Artem Baidaliuk[1,3], Flora Donati[4,5], Mélanie Albert[4,5], Andreea Alexandru[6], Maud Vanpeene[6], Méline Bizard[6], Angela Brisebarre[4,5], Marion Barbet[4,5], Fawzi Derrar[7], Sylvie van der Werf[4,5,8], Vincent Enouf[4,5,6,8], Etienne Simon-Loriere[1,8]

1. Evolutionary genomics of RNA viruses, Institut Pasteur, Paris, France
2. Université de Paris, Paris, France
3. These authors contributed equally
4. National Reference Center for Respiratory Viruses, Institut Pasteur, Paris, France
5. Molecular Genetics of RNA Viruses, CNRS - UMR 3569, University of Paris, Institut Pasteur, Paris, France
6. Mutualized Platform of Microbiology, Pasteur International Bioresources Network, Institut Pasteur, Paris, France
7. National Influenza Centre, Viral Respiratory Laboratory, Algiers, Algeria
8. These authors co-supervised this work

Correspondence: Etienne Simon-Loriere (etienne.simon-loriere@pasteur.fr) and Sylvie van der Werf (sylvie.van-der-werf@pasteur.fr)

Citation style for this article:
Gâmbaro Fabiana , Behillil Sylvie , Baidaliuk Artem , Donati Flora , Albert Mélanie , Alexandru Andreea , Vanpeene Maud , Bizard Méline , Brisebarre Angela , Barbet Marion , Derrar Fawzi , van der Werf Sylvie , Enouf Vincent , Simon-Loriere Etienne . Introductions and early spread of SARS-CoV-2 in France, 24 January to 23 March 2020. Euro Surveill. 2020;25(26):pii=2001200. https://doi.org/10.2807/1560-7917.ES.2020.25.26.2001200

Article submitted on 15 Jun 2020 / accepted on 02 Jul 2020 / published on 02 July 2020

Following SARS-CoV-2 emergence in China, a specific surveillance was implemented in France. Phylogenetic analysis of sequences retrieved through this surveillance suggests that detected initial introductions, involving non-clade G viruses, did not seed local transmission. Nevertheless, identification of clade G variants subsequently circulating in the country, with the earliest from a patient who neither travelled to risk areas nor had contact with travellers, suggests that SARS-CoV-2 might have been present before the first recorded local cases.

Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) was identified as the cause of an outbreak of severe respiratory infections in Wuhan, China in December 2019 [1]. Despite strict quarantine measures in Wuhan and surrounding areas, the virus, responsible for coronavirus disease (COVID-19), rapidly spread across the globe, leading the World Health Organization (WHO) to declare a pandemic on 11 March 2020. Soon after the emergence of the virus, a specific syndromic surveillance for COVID-19 was implemented in France. Because viral genomics, coupled with modern surveillance systems can help to understand outbreak dynamics [2], we sequenced SARS-CoV-2 genomes from clinical cases sampled through the surveillance.

Surveillance of COVID-19 in northern France

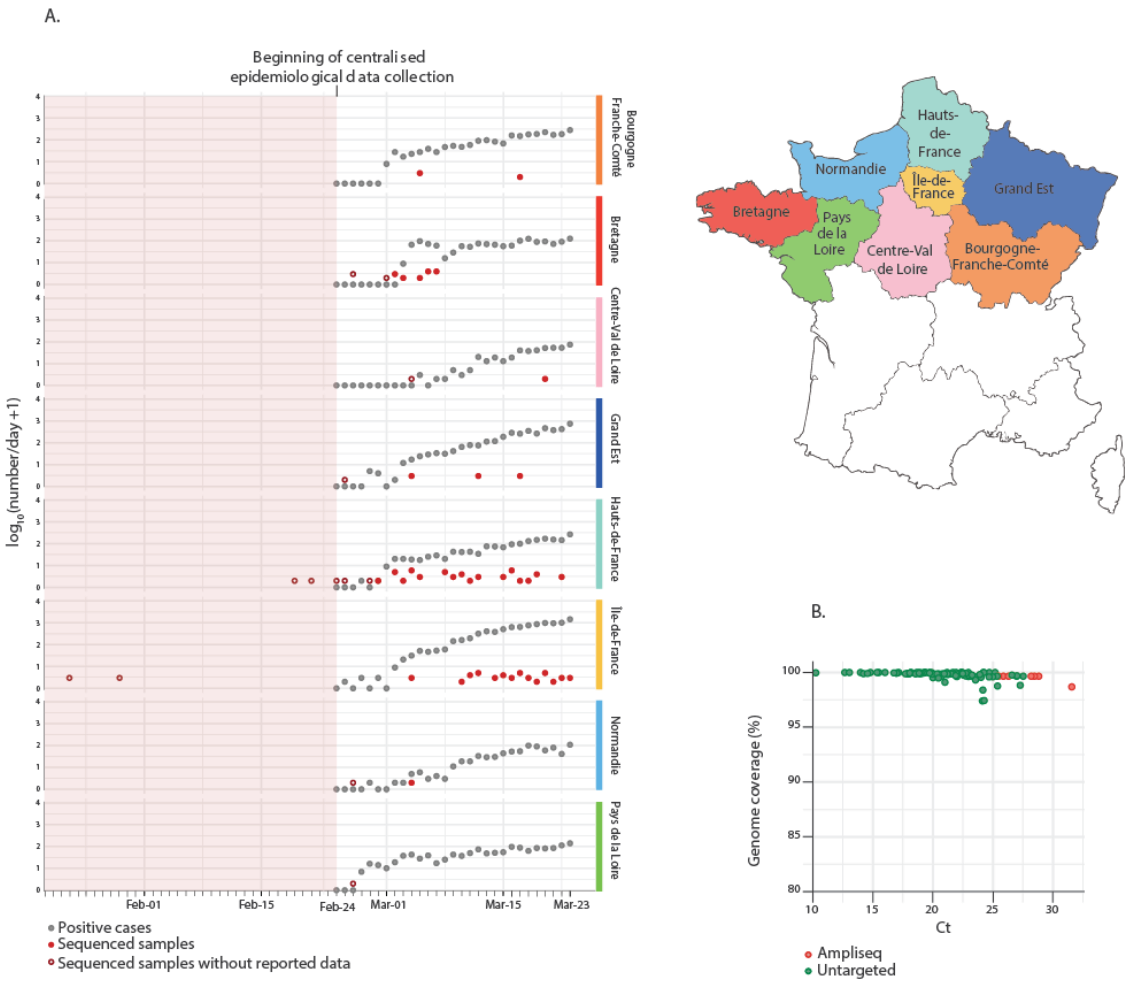
Strengthened surveillance of COVID-19 cases was implemented in France on 10 January 2020, with the objective of identifying imported cases early to prevent secondary transmission in the community. In this context, the first cases detected by the National Reference Center for Respiratory Viruses (NRC) hosted at Institut Pasteur, Paris, happened to be the first identified in Europe. As the COVID-19 epidemic progressed in the country, the task of identifying SARS-CoV-2 infections was shared with the NRC-associated laboratory in Lyon and then extended to first line hospital laboratories in the whole country, with the NRC at Institut Pasteur focusing on the northern part of France, including the densely populated capital. Screening and sampling for SARS-CoV-2 was targeted towards individuals who had symptoms (fever and/or respiratory problems) or a travel history to risk areas for infection [3]. As the virus continued to spread, it became clear that COVID-19 patients could exhibit greatly variable clinical characteristics [4], including a proportion presenting with asymptomatic infection or mild disease [5].

Patient sampling and analysis of retrieved SARS-CoV-2 genomes

We generated complete SARS-CoV-2 genome sequences from nasopharyngeal or sputum samples sent to the NRC at the Institut Pasteur as part of the ongoing surveillance (Figure 1A). We combined the SARS-CoV-2 genome sequences generated here, including 97 from northern France and three from Algeria with recent history of travel to France, with 338 sequences published and freely available from the Global Initiative on Sharing All Influenza Data (GISAID) EpiCoV database and/or GenBank. This dataset enabled to perform a phylogenetic analysis to gain more insight into the initial introductions and spread of the virus in France. More details on the methods used can be found in the Supplementary Material.

FIGURE 1

SARS-CoV-2 genome sequencing effort in northern French regions, 24 January–23 March 2020



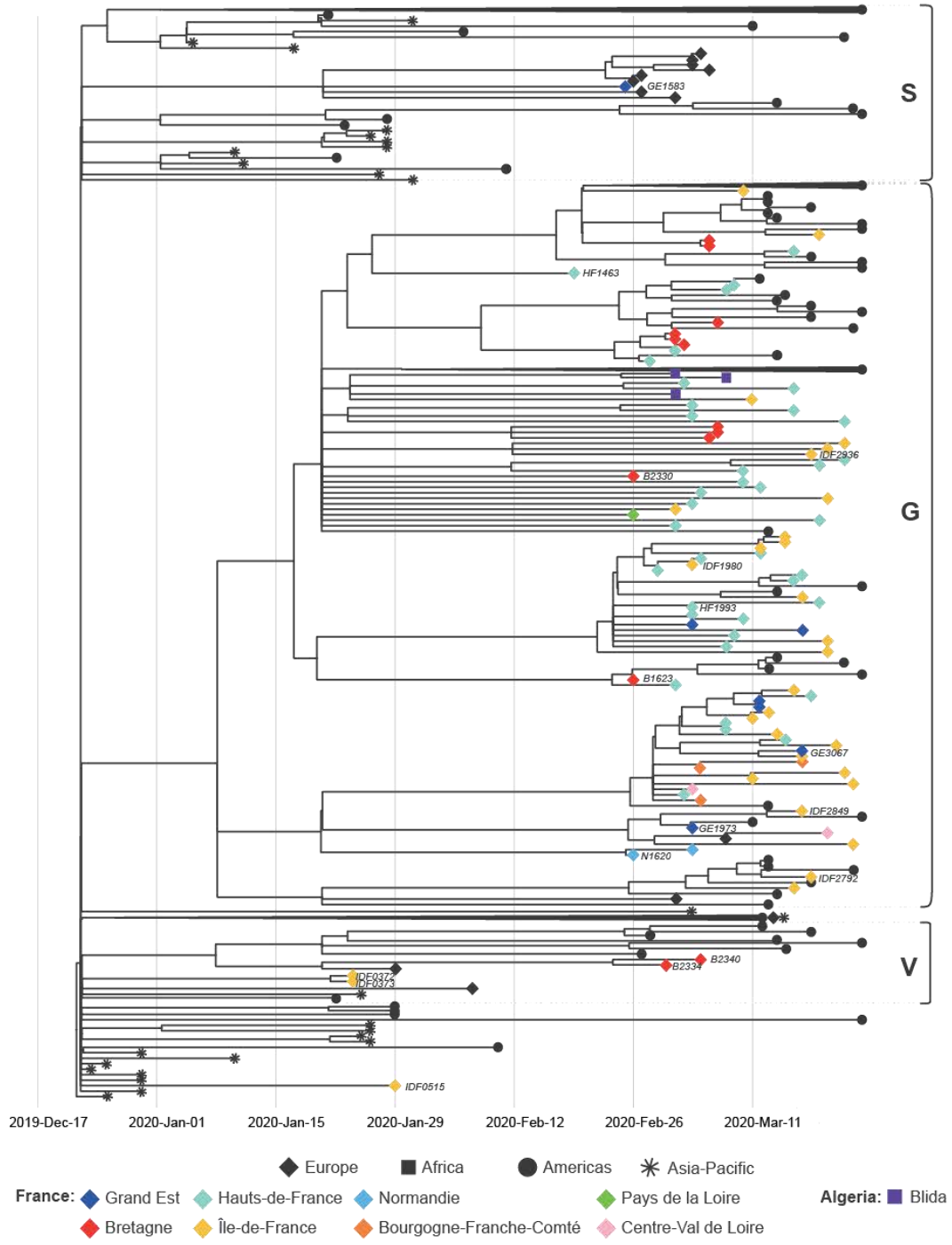
Ampliseq: amplicon-based sequencing; Ct: cycle threshold; number/day: number of laboratory-confirmed cases per day; SARS-CoV-2: severe acute respiratory syndrome coronavirus 2.

A. The plot represents the numbers of daily sequenced genomes in this study (red filled or hollow circles) overlaid with the number of reported positive cases (grey circles) obtained from Santé Publique France (www.santepubliquefrance.fr). Hollow circles indicate samples obtained on dates with zero reported SARS-CoV-2 positive cases. The data are shown separately for each region of northern France as indicated on the map on the right.

B. Percentage of SARS-CoV-2 genome coverage in relation to the Ct values obtained from the SARS-CoV-2 real-time reverse transcription PCR on the original samples, for the 97 genomes reported here. For reliability, amplicon-based sequencing was implemented for samples with Ct values higher than 25. Colours indicate sequencing approach: untargeted metagenomics (green) or amplicon-based sequencing (red).

FIGURE 2

Phylogenetic analysis of sequences of early introductions and subsequently circulating SARS-CoV-2 in northern France, 24 January–23 March 2020

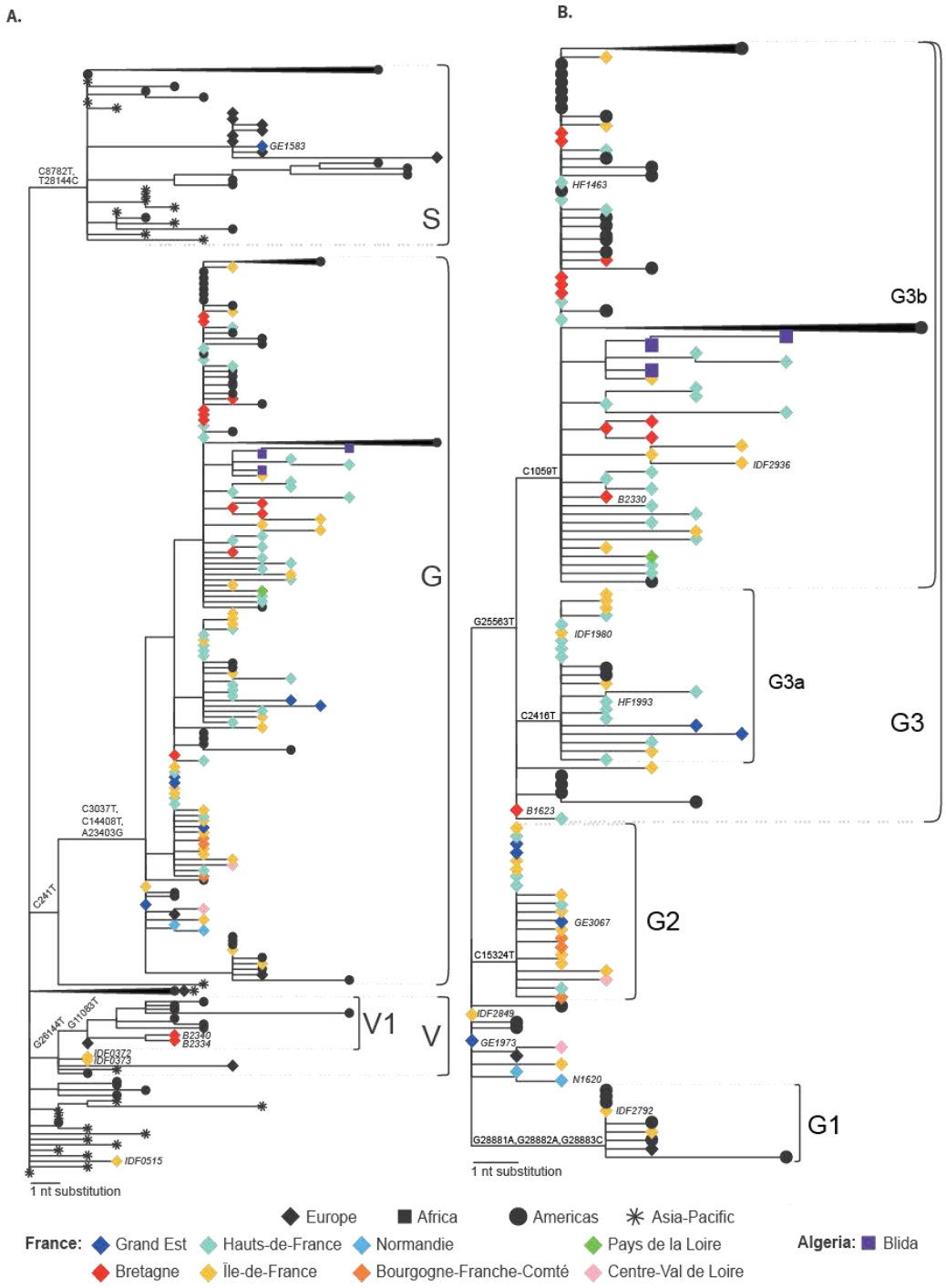


GISAID: Global Initiative on Sharing All Influenza Data; SARS-CoV-2: severe acute respiratory syndrome coronavirus 2.

Time calibrated tree of 438 SARS-CoV-2 sequences including northern France, Algeria and publicly available global sequences. The tree is rooted using the reference strain Wuhan/Hu-1/2019n (GenBank accession number: MN908947). The tips of the tree are shaped and coloured according to sampling location. Branch lengths are proportional to the time span from the sampling date to the inferred date of the most recent common ancestor. The three major clades according to GISAID nomenclature are indicated. Strain names of the sequences discussed in this study are indicated next to the corresponding tips.

FIGURE 3

Phylogenetic trees with SARS-CoV-2 sequences showing (A) clades S,G,V and (B) clade G, with details on corresponding lineages, northern France, 24 January–23 March 2020



GISAID: Global Initiative on Sharing All Influenza Data; SARS-CoV-2: severe acute respiratory syndrome coronavirus 2.

The tips of the trees are shaped and coloured according to sampling location. Branch lengths are proportional to the number of nucleotide substitutions from the reference and tree root Wuhan/Hu-1/2019 (GenBank accession number: MN908947). GISAID clades and putative lineages are indicated on the right of each panel. Strain names of the sequences discussed in this study are indicated next to the corresponding tips in italic. Nucleotide substitutions shared among all the sequences of each clade or lineage are indicated next to the corresponding nodes. Some monophyletic lineages are collapsed for ease of representation. A complete tree is shown in Fig. S1.

Ethical statement

Samples used in this study were collected as part of approved ongoing surveillance conducted by the NRC at Institut Pasteur (WHO reference laboratory providing confirmatory testing for COVID-19). The investigations were carried out in accordance with the General Data Protection Regulation (Regulation (EU) 2016/679 and Directive 95/46/EC) and the French data protection law (Law 78–17 on 06/01/1978 and Décret 2019–536 on 29/05/2019).

Detected early viral introductions appear not to have seeded local transmission

Our analysis indicates that the quarantine imposed on the initial imported COVID-19 cases, who were captured by the syndromic surveillance in France, appears to have prevented local transmission. The first European cases, who were originally in Île-de-France (IDF) and who were previously described elsewhere [6], were direct imports from Hubei, China. They were sampled on 24 January 2020 and the two derived respective viral genomes, IDF0372 and IDF0373, fall accordingly near the base of the tree, within clade V, according to GISAID nomenclature (Figure 2, Figure 3A). Clade V is characterised by sequences with a T nucleotide at position 26144, instead of a G, corresponding to a V amino acid, rather than a G, at position 251 of non-structural protein 3a. The IDF0372 and IDF0373 genomes were identical and both harboured a G22661T non-synonymous mutation (V367F) in the receptor-binding domain of the spike protein, not observed in other genomes. Similarly, IDF0515, obtained from a 29 January sample, corresponds to a traveller from Hubei, China. This basal genome falls outside of the three major GISAID proposed clades V, G, and S (Figure 2), but carries the G11083T mutation associated with putative lineage V1 (Fig. S2), suggesting convergent evolution or a reversion of the V-clade defining G26144T change. Subsequent early cases detected in February in the West (Bretagne; B) or East (Grand Est; GE) of France (B2334/B2340, clade V and GE1583, clade S), all with recent history of travel to Italy, add to the genomic diversity of viruses from northern Italy, but also do not appear to have seeded local transmission within the current sampled sequence set (Figure 2).

Clades and lineages of SARS-CoV-2 further circulating in northern France

All other sequences from northern France fall in clade G, defined by two synonymous mutations (C241T, C3037T) and one non-synonymous substitution (A23403G) corresponding to a D614G mutation in the spike protein (Figure 3), and this includes sequences captured during the steep increase of reported cases in many strongly affected regions (Figure 1). While a more thorough sampling will be needed to confirm this hypothesis, the phylogenetic analysis of sequences recovered in the current study suggests that the French outbreak was mainly seeded by one or several variants of this clade, unlike what is observed for many other European countries (https://nextstrain.org/ncov/europe?f_region=Europe) [7,8]. This clade can be further classified into lineages (putatively named G1, G2, G3, G3a, G3b), albeit supported again by only one to three substitutions. The lineages are for the most part respectively represented by sequences from several regions. Several genomes correspond to patients in GE, Normandie (N), IDF, Hauts-de-France (HF) and B with recent history of travel in Europe (GE3067, N1620, IDF2792), United Arab Emirates (IDF2936), Madagascar (HF1993), Egypt (B1623, B2330) or linked to Paris airports (IDF1980). These genomes might represent additional introductions of the same clade, since the respective cases tested positive for the virus when other local G-clade-virus infections had already been detected in the north of France. On the other hand, in lineage G3b, three sequences sampled in Algeria are closely related to sequences from northern France and likely represent exported cases in light of recent history of travel to France.

The syndromic surveillance allowed to capture one of the earliest representatives of clade G (HF1463, sampled on 19 February) (Figure 2). Importantly, this sequence carries two additional mutations compared with the reconstructed ancestral sequence of this clade (Figure 3B). Other sequences sampled weeks later (IDF2849, GE1973) are more basal to the clade, highlighting the complexity and risk of inferences based on 1 or 2 nucleotide substitutions. Because of this, and the scarcity of early sequences in many countries in Europe, country and within-country level phylogeographic estimations are unreliable with the current dataset. It is thus impossible to infer with confidence how the virus was introduced to France from the epicentre of the outbreak, and multiple routes are possible.

Discussion

The generated genomes in this study provide more insight into the SARS-CoV-2 clades and variants circulating in northern France at the beginning of the outbreak and later during the pandemic. Results of the analyses seem to indicate that, at least for the first imported cases who could be captured by the surveillance, these introductions did not lead to further transmission of the virus in the community. Indeed, sequences from imported cases detected early in the outbreak did not belong to clade G, a clade identifying all the genomes retrieved later in the epidemic. Within clade G, a number of variants could be observed. Moreover, the earliest patient infected with a representative of clade G (HF1463) had no history of travel or contact with returning travellers, suggesting that the virus was silently circulating in France in February, a scenario compatible with the large proportion of persons with mild disease or asymptomatic infections [5], and observations in other European countries [9,10]. While this is also compatible with the time to the most recent common ancestor estimate for clade G (Figure 2), the current sampling clearly prevents reliable inference for the timing of introduction in France. Moreover

while the current data may lead to hypothesize that the French outbreak could have been mainly seeded by one or several variants of the G clade, more data will be needed to confirm this. Another explanation would be that while the outbreak began with viruses belonging to various clades, the clade G might have become dominant in the north of France as the epidemic progressed.

Crucially, while all early symptomatic suspected COVID-19 cases samples were sent to the NRC for testing, this was no longer the case as the epidemic developed (Figure 1A). In addition, pauci- or asymptomatic cases are scarcely represented in our dataset. This study also reveals areas for potential improvement of SARS-CoV-2 genomic surveillance in France as several regions are poorly represented (Figure 1A). This is likely due to the heavy burden on hospitals, which while being able to perform local testing owing to the rapid sharing of molecular detection tools by the NRC, might have had to reduce the number of positive samples sent for confirmation and sequencing to the NRC. Because of this, and of the syndromic-only based surveillance, we likely underestimate the genetic diversity of SARS-CoV-2 circulating in France.

In conclusion, our study sheds light on the origin and diversity of the COVID-19 outbreak in France with insights for Europe, and highlights the challenges of containment measures when a significant proportion of cases are asymptomatic.

Data and materials availability

The assembled SARS-CoV-2 genomes generated in this study were deposited on the GISAID database (https://www.gisaid.org/) as soon as they were generated, accession numbers can be found in Data S1 (Table S2).

Acknowledgements

We would like to thank all of the healthcare workers, public health employees, and scientists involved in the COVID-19 response.

We acknowledge the hospital laboratories from the RENAL network in the north of France (list of names in Data S1, Table S4).

We acknowledge the authors, originating and submitting laboratories of the sequences from GISAID and GenBank (Data S1, Table S2). We avoided any direct analysis of genomic data not submitted as part of this paper and used this genomic data only as background.

We thank Laurence Ma (Biomics Platform, C2RT, Institut Pasteur, Paris, France) for the MiSeq sequencing.

This work used the computational and storage services (TARS cluster) provided by the IT department at Institut Pasteur, Paris.

FG is part of the Pasteur-Paris University (PPU) International PhD programme, BioSPC doctoral school.

Funding statement

This study has received funding from Institut Pasteur, CNRS, Université de Paris, Santé publique France, the French Government's Investissement d'Avenir programme, Laboratoire d'Excellence "Integrative Biology of Emerging Infectious Diseases" (grant n°ANR-10-LABX-62-IBEID), REACTing (Research & Action Emerging Infectious Diseases), France Génomique (ANR-10-INBS-09-09), IBISA, and the RECOVER project funded by the European Union's Horizon 2020 research and innovation programme under grant agreement No. 101003589. ESL acknowledges funding from the INCEPTION programme (Investissements d'Avenir grant ANR-16-CONV-0005).

Conflict of interest

None declared.

Authors' contributions

SB, FDo, MA, AA, MV, MBi, ABr, MBa, FG – investigation; FDe – resources; FG, ABa, ESL – data curation and analysis, visualization, writing original draft; SB, VE, ESL, SvdW – writing, review and editing; SvdW, VE, ESL – study conceptualization, resources, supervision; SvdW, ESL –funding acquisition.

References

1. Zhu N, Zhang D, Wang W, Li X, Yang B, Song J, et al. , China Novel Coronavirus Investigating and Research Team. A novel coronavirus from patients with pneumonia in China, 2019. N Engl J Med. 2020;382(8):727-33. https://doi.org/10.1056/NEJMoa2001017 PMID: 31978945
2. Grubaugh ND, Ladner JT, Lemey P, Pybus OG, Rambaut A, Holmes EC, et al. Tracking virus outbreaks in the twenty-first century. Nat Microbiol. 2019;4(1):10-9. https://doi.org/10.1038/s41564-018-0296-2 PMID: 30546099
3. Bernard Stoecklin S, Rolland P, Silue Y, Mailles A, Campese C, Simondon A, et al. , Investigation Team. First cases of coronavirus disease 2019 (COVID-19) in France: surveillance, investigations and control measures, January 2020. Euro Surveill. 2020;25(6):2000094. https://doi.org/10.2807/1560-7917.ES.2020.25.6.2000094 PMID: 32070465
4. Guan WJ, Ni ZY, Hu Y, Liang WH, Ou CQ, He JX, et al. , China Medical Treatment Expert Group for Covid-19. Clinical characteristics of coronavirus disease 2019 in China. N Engl J Med. 2020;382(18):1708-20. https://doi.org/10.1056/NEJMoa2002032 PMID: 32109013
5. Li R, Pei S, Chen B, Song Y, Zhang T, Yang W, et al. Substantial undocumented infection facilitates the rapid dissemination of novel coronavirus (SARS-CoV-2). Science. 2020;368(6490):489-93. https://doi.org/10.1126/science.abb3221 PMID: 32179701
6. Lescure F-X, Bouadma L, Nguyen D, Parisey M, Wicky P-H, Behillil S, et al. Clinical and virological data of the first cases of COVID-19 in Europe: a case series. Lancet Infect Dis. 2020;20(6):697-706. https://doi.org/10.1016/S1473-3099(20)30200-0 PMID: 32224310
7. Gudbjartsson DF, Helgason A, Jonsson H, Magnusson OT, Melsted P, Norddahl GL, et al. Spread of SARS-CoV-2 in the Icelandic Population. N Engl J Med. 2020;382(24):2302-15. "https://doi.org/10.1056/NEJMoa2006100" https://doi.org/10.1056/NEJMoa2006100 PMID: 32289214
8. Zehender G, Lai A, Bergna A, Meroni L, Riva A, Balotta C, et al. Genomic characterization and phylogenetic analysis of SARS-COV-2 in Italy. J Med Virol. 2020. https://doi.org/10.1002/jmv.25794 PMID: 32222993
9. Gudbjartsson DF, Helgason A, Jonsson H, Magnusson OT, Melsted P, Norddahl GL, et al. Spread of SARS-CoV-2 in the Icelandic Population. N Engl J Med. 2020;382(24):2302-15. https://doi.org/10.1056/NEJMoa2006100 PMID: 32289214

10. Onder G, Rezza G, Brusaferro S. Case-Fatality Rate and Characteristics of Patients Dying in Relation to COVID-19 in Italy. JAMA. 2020. https://doi.org/10.1001/jama.2020.4683 PMID: 32203977

License, supplementary material and copyright

In the second study, we monitored the evolution of SARS-CoV-2 during a long-term infection in an immunocompromised patient. As mentioned in the introduction of this chapter, it has been proposed that some of the SARS-CoV-2 variants, in particular the Alpha and the Omicron variant – which are characterized by a high number of mutations in comparison to their close relatives – may have emerged during long-term infections of immunocompromised individuals. Indeed, while most SARS-CoV-2 infections are generally resolved within 1 to 2 weeks, multiple reports have shown that immunocompromised individuals can remain infected for more extended periods, sometimes noting rapid accumulation of changes in the viral genome.

In an attempt to eliminate the infection, immunocompromised patients are often treated with monoclonal antibodies or convalescent plasma, which may sometimes shape the evolutionary dynamics of the virus. In this work, we describe the case of an immunocompromised patient in remission of non-Hodgkin lymphoma who suffered from severe COVID-19 upon infection by SARS-CoV-2. Despite being treated with convalescent plasma, the patient remained positive for the virus for over 131 days before recuperating. Using an unbiased NGS approach, we were able to capture snapshots of the genetic diversity of the virus at regular intervals all along the course of the infection. We tracked the dynamics of the viral population over time, including low frequency variants. To ensure accurate variant calling, we sequenced samples in duplicate. In addition, in the analysis, we only included variants with a minimum frequency threshold of 1% and minimum coverage of 1000X. We noted a significant shift in the viral population after the patient was treated with convalescent plasma but no virus clearance. We found that the long-term evolution of SARS-CoV-2 in an immunocompromised individual is not always associated with a substantial accumulation of changes, particularly in the spike protein, and highlights the challenges of managing persistently infected immunocompromised individuals.

We are preparing this work as:

Fabiana Gámbaro, Stéphanie Marque-Juillet, Melanie Albert , Flora Donati , Artem Baidaliuk, Sylvie Behillil, Vincent Enouf, Sylvie Van der Werf, Fabrice Bruneel, Etienne Simon-Loriere **. Snapshots of SARS-COV-2 population diversity during a long-term infection in an immunocompromised host.**

Supplementary information associated to this chapter can be found in the following link

Snapshots of SARS-COV-2 population diversity during long-term infection of an immunocompromised host.

Fabiana Gámbaro[1,2], Stéphanie Marque-Juillet[3], Flora Donati[4,5], P Bargain[3], S Rigaudeau[3], A Henry[3], AC Gauchie[3], Sylvie Behillil[4,5], Vincent Enouf[4,5], Sylvie Van der Werf[4,5], Fabrice Bruneel[3], Etienne Simon-Loriere[1].

1 G5 Evolutionary Genomics of RNA viruses, Institut Pasteur, Paris, France
2 University of Paris, France
3 Centre Hospitalier de Versailles, Versailles, France.
4 Molecular Genetics of RNA Viruses, CNRS - UMR 3569, University of Paris, Institut Pasteur, Paris, France.
5 National Reference Center for Respiratory Viruses, Institut Pasteur, Paris, France

ABSTRACT

The long-term infections have been proposed as plausible contributors to the emergence of new SARS-CoV-2 variants. In this work, we sequenced duplicates respiratory samples using an unbiased metagenomic deep sequencing technique to capture snapshots of the genetic diversity of SARS-CoV-2 and track the dynamics of the viral population during a long-term infection in an immunocompromised patient in the presence of convalescent plasma therapy. We found that long-term infections are not always associated with a high accumulation of changes in the viral genome and that the convalescent plasma therapy did not result in viral clearance but a replacement of the viral population within the respiratory tract.

INTRODUCTION

In late 2019 we witnessed the emergence of severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), leading to one of the largest and fast-spreading pandemics in modern history (1, 2). Two years into the pandemic and SARS-CoV-2 infections are still raging in many parts of the world, fueled by more transmissible SARS-CoV-2 variants of concern. With five different variants of concern around the corner, there is an urge to understand the threat these variants might impose on our society. In this context, much discussion has been generated about the possible origin of these variants. Determining under which circumstances these more transmissible variants emerged could help understand the risk of new variants emerging, suggesting ways to prevent it from happening again.

It has been proposed that some of these variants, in particular the Alpha (B.1.1.7) and the heavily mutated Omicron variant (B.1.1.529), are characterized by a high number of mutations in comparison to their close relatives, notably in the spike protein, could have emerged during long-term infections in chronically infected individuals (3). Indeed, while most SARS-CoV-2 infections are generally resolved within 1-2

weeks, multiple reports have shown that people with compromised immune systems can remain infected for a longer period, sometimes noting rapid accumulation of changes in the viral genome. In an attempt to eliminate the infection, immunocompromised patients are often treated with monoclonal antibodies or convalescent plasma (CP). Although the use of CP showed initial promising results in severe COVID-19 patients (4, 5), evidence supporting its clinical efficacy is still ambiguous (6, 7). While it is true that immunocompromised patients would benefit from CP therapy, given the underlying deficits in B or T cell immunity, recent studies have shown that CP treatment might not be successful in clearing the infection resulting, on certain occasions, in the emergence of escape mutations in SARS-CoV-2 (8, 9).

In this work, we describe a case of an immunocompromised patient facing a non-Hodgkin lymphoma in remission who suffered from severe COVID-19 upon infection by SARS-CoV-2. We recovered infectious virus at least until day 86 since the onset of the symptoms. We successfully sequence samples in duplicates at different time points using an unbiased next-generation sequencing approach. In this way, we were able to capture snapshots of the genetic diversity of the virus, including low frequency variants, at regular intervals all along the course of the infection. We observed that the persistent infection led to the within-patient evolution of SARS-CoV-2 and a significant shift in the viral population when the patient was treated with convalescent plasma but with no virus clearance. Our data provide insights into the intrahost viral dynamics while showing that the long-term evolution of SARS-CoV-2 in an immunocompromised individual is not always associated with a substantial accumulation of changes in the spike protein. Our work also highlights the challenges of managing persistently infected immunocompromised individuals.

RESULTS

Clinical presentation of the immunocompromised patient infected with SARS-CoV-2

A 70-year old-year-old immunosuppressed patient facing a lymphoma who received an autologous transplant in late December 2019 started with fever and cough on the 23rd of April 2020 (Day 0). On the 29th of April, tested positive for SARS-CoV-2 by reverse-transcription polymerase chain reaction (RT-qPCR) from a nasopharyngeal swab specimen. On the 7th of May (day 14 of illness), he was admitted to the hospital with fever, cough, and pneumonia. The patient tested positive for SARS-CoV-2 until day 138 (except for day 55), and he was finally discharged on the 6th of November (Figure 1)

In an attempt to treat the persistent SARS-CoV-2 infection, the patient was subjected to CP therapy for four days, starting on day 90. He was transfused with 200 mL/day of SARS-CoV-2 convalescent plasma. Pre- and post-treated serum samples were tested by neutralization assay to detect SARS-CoV-2 specific antibodies. While no antibodies were detected prior to the transfusion (days 35 and 85), four days after the treatment (day 97), antibodies against SARS-CoV-2 were detected (Table S1). The individual remained positive for SARS-CoV-2 until day 131 before recuperating.

Virus isolation was attempted on RT-qPCR-positive samples. Individual's respiratory sample collected on day 89 was successfully cultured on Vero E6 cells supporting persistent SARS-CoV-2 infection with the shedding of infectious virus at least until day 89 since the onset of symptoms (Figure 1).

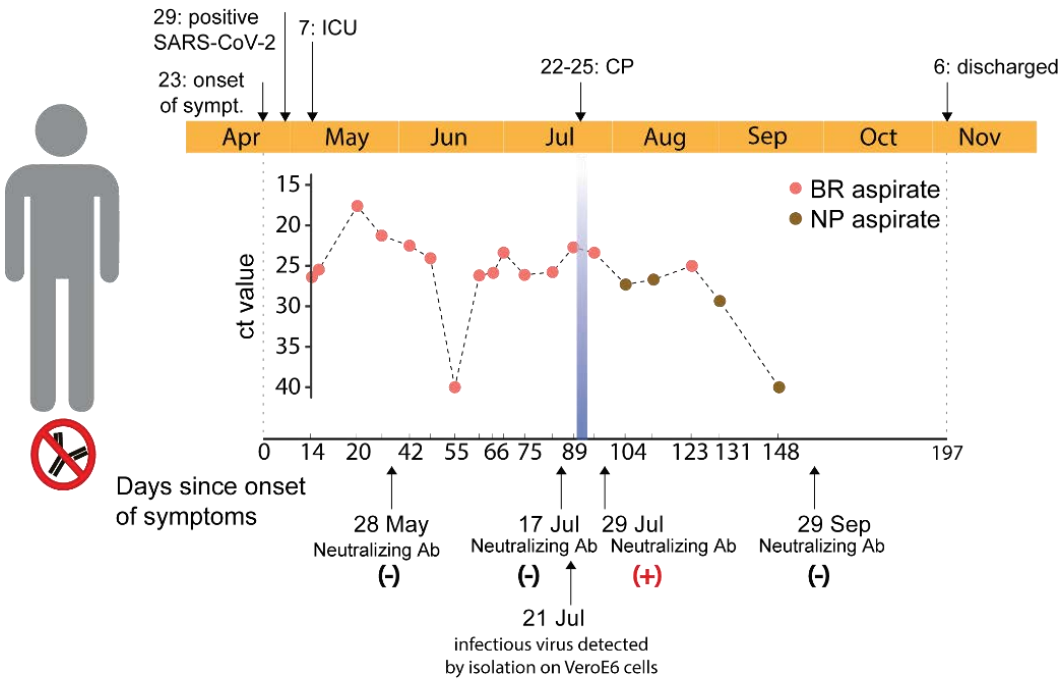


Figure 1: Timeline of clinical presentation and diagnostic tests of the immunocompromised individual. Treatments received by the individual such as intravenous immunoglobulin (IGIV), corticoids, and convalescent plasma (CP) therapy are indicated. RT-qPCR was done on bronchoalveolar (BR) or nasopharyngeal (NP) aspirates at different time points and the corresponding Ct values are expressed over time (days since onset of the symptoms). Neutralizing antibodies detection performed on serum samples pre, and post-CP transfusion is indicated.

Deep sequencing and intra-host low frequency variants calling.

To ensure accurate variant calling for the intra-host variant analysis, samples included in this analysis were sequenced using an untargeted metagenomics sequencing approach that allowed us to assess the viral population unbiased (no PCR-based sequencing). In addition, each sample was prepared in duplicates starting from the extracted RNA and sequenced in different batches. The resulting genomes had an average coverage of over 90% and except for one sample, all had an average coverage of at least 1000X in one of the duplicates (Table S2). We considered for the analysis only variants found across replicates. We found that only near 60% of the variants (interquartile range: 55.7 to 74.5%) were conserved in both replicates for each sample (Figure S1-A). For most variants, we observed a difference of less than five percent in the frequency detected in both replicates, showing a high concordance between replicates (Figure S1-B). Lastly, we selected low frequency variants with a minimum frequency threshold of 1% and a minimum read depth of 1000X at each site to minimize false positives.

Genomic investigation

We generated a maximum-likelihood tree using subsampled SARS-CoV-2 sequences around the globe from the GISAID database (https://www.gisaid.org/) and sequences derived from this individual and the other three patients treated in the same hospital ward. The sequences from the other three patients fell within the 20B Nextstrain clade. According to the PANGO lineage, the sequence from patients 1 and 2 (same familial cluster) belonged to the B.1.1.317 while patient 3 to B.1.1.209 lineage. The sequences from the immunocompromised patient fell within the 20C Nextstrain clade, or the B.1 PANGO lineage, forming a monophyletic group compatible with infection and subsequently virus persistence (Figure 2).

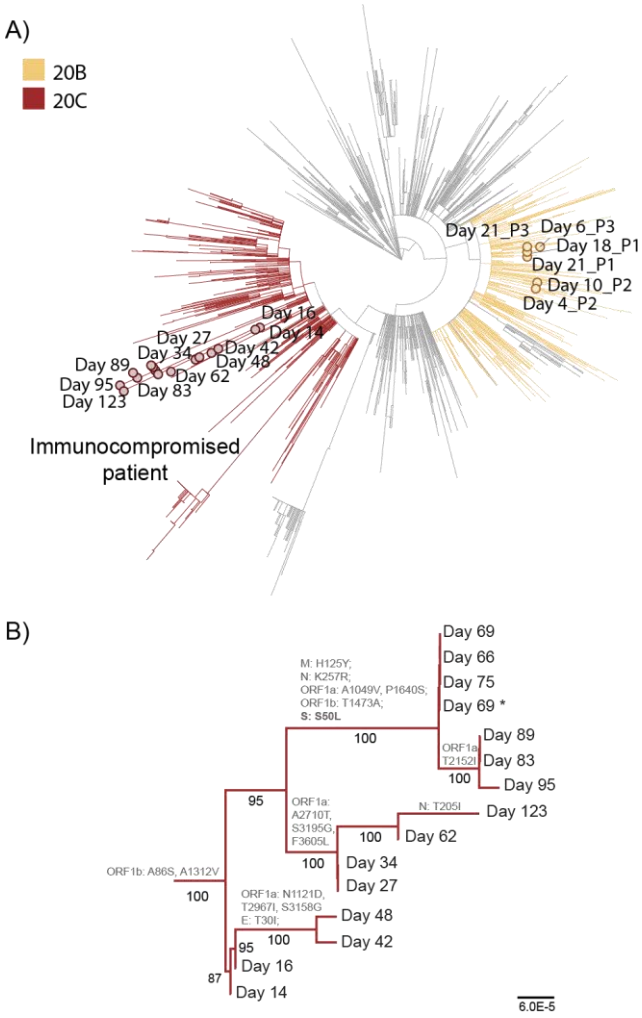


Figure 2: A) Global SARS-CoV-2 phylogeny. Major clades are colored according to Nextstrain classification. The patient SARS-CoV-2 sequences are shown with red dots, and sequences from three other patients (P1, P2, and P3) treated in the same hospital ward are shown with yellow dots. B) Zoom in into the patient SARS-CoV-2 sequences. In the branches, we show AA changes and the support values, which are ultrafast bootstrap percentages. The scale bars represent the number of nucleotide substitutions per site.

In addition to capturing the consensus sequences, we monitored the viral population on the lower respiratory tract, including low frequency variants. Samples from the lower respiratory tract were collected by bronchoalveolar aspirates collected between days 14 and 123 since the onset of symptoms, totaling 14 time points (Figure 3). Over the course of the infection, we observed the within-patient evolution of the virus without substantial accumulation of amino acid changes, particularly in the spike protein. In contrast to the early period of infection, this within-patient evolution led to the emergence of a viral population bearing the A1049V, P1640S, and T2152I substitutions in the ORF1a polyprotein; the T1473A substitution in the ORF1b polyprotein; the H125Y substitution in the M protein; the K257R substitution in the N protein; and the S50L in the S1 N-terminal domain of spike protein, becoming the dominant population at day 69. After the CP treatment administration, on day 97, we observed on day 123 a major shift in the viral population in which the substitutions mentioned above were reverted, including the S50L substitution in the spike protein.

Additionally, variants previously detected on viruses sampled on day 62, namely the V86F substitution in the ORF1a polyprotein, reappeared. These changes led to a viral population carrying a variant signature similar to the dominant population in the early stages of the infection, particularly the one captured on day 62 (Figure 3). Indeed, except for three synonymous changes located in the 5'UTR, in the spike, and the N gene, corresponding to the nucleotide positions 208, 22345, and 28891, respectively, the variants detected on day 123 were the dominant variants on day 62. Furthermore, the nucleotide change leading to the non-synonymous change in the N protein (N: T205I), which further characterizes the genome captured at day 123, was observed as a low frequency variant (around 1%) on viruses collected on day 14. This dynamic structure of the viral population is also observed in the phylogenetic tree, with the consensus sequence of the genome captured on day 62 being basal to y the sequence of the genome collected on day 123, while sequences of the genomes from days 69 to 95 branching differently within the monophyletic group (bootstrap value = 100%) (Figure 2B).

We found it very interesting that the amino acid sequence of the replaced and novel dominant population only differs in one position (S: S50L), given that the majority of antibodies generally target the spike protein. Additional information from other samples from the upper respiratory tract (nasopharyngeal aspirates) collected after the transfusion with the CP confirmed this observation. While on day 104, the spike variant S50L was still dominant in the upper respiratory tract, on the other two samples collected on days 112 and 131, the viral population was replaced by a population bearing a serine residue at position 50 in the spike protein. Indeed, consensus sequences of genomes captured on days 123 and 131, the latest samples collected from the lower and the upper respiratory tract, are identical (Figure S2). In addition, the spike protein of the viruses captured on the upper respiratory tract on days 104 and 112 are further characterized by additional non-synonymous substitutions, specifically the Q321R substitution and the double mutation A942S and G1124V, respectively (Figure S2). Nevertheless, having no samples collected from the upper respiratory tract prior to the transfusion with the CP prevents us from drawing further conclusions about these variants.

Interestingly, viruses collected on days 42 and 48 were characterized by three non-synonymous changes on the ORF1a polyprotein (N1121D, T2967I, and S3158G) and one non-synonymous change on the E protein (T30I). Although these variants were dominant, the alternative variants were found to have close to 20 -30% frequency. The divergence of these samples from the rest of the samples collected from the lower respiratory tract is also shown by the phylogenetic tree (bootstrap value = 100%) (Figure3B).

This divergence might suggest that these viruses represent a compartmentalized subpopulation within the low respiratory tract.

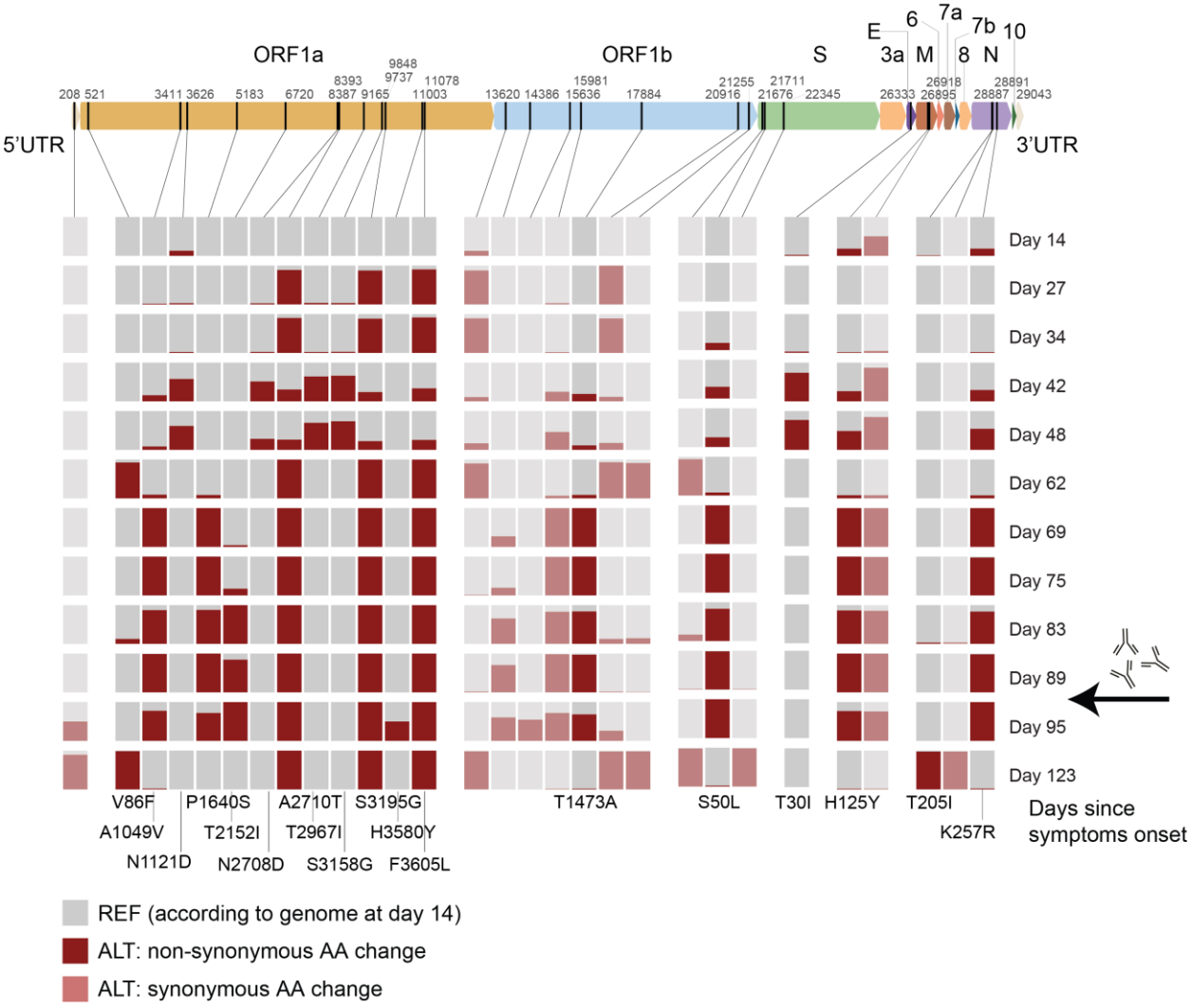


Figure 3: Schematic representation of the variables site across the SARS-CoV-2 genome collected from the lower respiratory tract at the different time points compared to the reference genome collected on day 14. Each bar plot represents a specific position in which the nucleotide frequency. With grey, we colored the nucleotide matching the one present in the consensus sequence of the reference genome ("REF") and with red when the nucleotide is different from the reference ("ALT"). We verified that for every position among the different genomes, there was only one alternative nucleotide in comparison to the reference genome. Additional information from samples from the upper respiratory tract can be found in Figure S2.

DISCUSSION

Here we report the long-term evolution of SARS-CoV-2 in an immunocompromised patient with non-Hodgkin lymphoma in the presence of CP therapy.

O(10, 11) (12)ne of the strengths of our study is that each sample was sequenced in duplicates using an untargeted metagenomics sequencing which allowed us to follow the dynamics of virus populations, including low frequency variants.

Similar to previous studies, the CP therapy provided the patient with neutralizing antibodies, but it did not result in viral clearance (8, 9, 11, 12). Throughout the infection, we did not observe a strong accumulation of amino acid changes, particularly in the spike protein, which aligns with previous studies (19, 20) but contrasts with other similar studies (21). For instance, other studies have reported substantial selection upon CP therapy, leading to the emergence of mutations in the spike believed to affect SARS-CoV-2 affinity to ACE-2 receptors such as Y453F (13) and N501Y (8, 12) or believed to be involved in immune evasion (ΔH69-V70 and E484K) (8, 13). We observed a significant shift in the viral population that restored a genotype similar to the dominant one in the early stage of the infection characterized by the spike protein variant S50. Such differences in the evolution of SARS-CoV-2 during long-term infections may reflect the different processes affecting each individual's evolution, for example, the degree of immunodeficiency or the different treatments schemes.

Indeed, while most antibodies generally target the spike protein, the amino acid sequence of both the replaced and novel dominant population only differs at position 50 of the spike protein. This observation might have two explanations. First, the population emerged under positive selection: the antibodies in the CP cleared viruses harboring the variant S:L50, leading to replacing a viral population with an escape genotype (S:S50). Second, the antibodies managed to eliminate the dominant population in the lower respiratory tract, regardless of the genotype of the spike protein (i.e., S:S50 or S:L50), which was subsequently replenished by a viral population less accessible to the administered antibodies, possibly located in another compartment.

Although two studies have reported the S50L substitution in long-term infections (10, 13), and *in silico* analysis suggests that the substitution S50L may stabilize SARS-CoV-2 spike protein (14), there is no evidence that this substitution affects immune evasion. In addition, both the phylogenetic tree and the analysis of low frequency variants show that the replaced population is genetically closer to the one detected in the early stage of the infection, with most of the variants previously observed in the viral population collected at day 62, including the spike protein variant S50. Remarkably, spike protein variants bearing the serine or leucine residue at position 50 were detected coexisting in the viral population from samples collected on days 34, 42, 48, 62, and 83, suggesting the presence of distinct viral populations. In this context, the second option seems like the most parsimonious scenario. Nevertheless, more data will be needed to address this.

Indeed, we observed signs of possible compartmentalization within the lower respiratory tract, with samples collected from days 42 and 48 probably representing a distinct subpopulation. Although these samples are genetically distant from the rest, they still fall within the monophyletic group suggesting they all arise from the same viral population. Kemp et al. (8) has already addressed (8) the same observation in a similar study of long-term SARS-CoV-2 infection.

MATERIAL AND METHODS

EXPERIMENTAL PROCEDURES

Ethical statement

Informed consent was obtained from this patient to participate in the present study conducted by the Institut Pasteur and the Hospital in Versailles, France.

Sample collection

Sera samples (n= 6) and respiratory samples from the upper respiratory tract (nasopharyngeal aspirates, n=3) and lower respiratory tract (bronchoalveolar aspirates n= 15) were collected at the Versailles Hospital and shipped to the National Reference Center for Respiratory Viruses (NRC) hosted at the Institut Pasteur.

RNA Extraction and RT-qPCR

RNA extraction was performed with the Extraction NucleoSpin Dx Virus kit (Macherey Nagel). Briefly, RNA was extracted from 100µl of the specimen, eluted in 100 µl of water, and used as a template for RT-qPCR. Samples were tested with a one-step RT-qPCR using three sets of primers as described on the WHO website (15).

Next-generation sequencing of patient clinical samples

We followed a protocol for untargeted metagenomic sequencing of clinical samples previously described by Matranga et al. (16). Extracted RNAs were first treated with Turbo DNase (Ambion), followed by purification using SPRI beads (Agencourt RNA clean XP, Beckman Coulter). Host ribosomal RNA was depleted using custom probes to form RNAse H target DNA-RNA hybrids. The RNA from the selective depletion was used for cDNA synthesis using random primers and SuperScript IV (Invitrogen). Second-strand cDNA was generated using a cocktail of enzymes, including *Escherichia coli* DNA ligase, RNAse H, and DNA polymerase (New England Biolabs), and subsequently purified using Agencourt AMPure XP beads (Beckman Coulter). From the dsDNA, libraries were prepared using the Nextera XT kit and sequenced using a paired-end strategy on an Illumina NextSeq500 platform (2X75 cycles).

Genome assembly

Raw reads were trimmed using Trimmomatic v0.36 (17) to remove Illumina adaptors and low-quality reads. We assembled using the metaspades option from SPAdes v3.12 (18), and the contigs obtained were used as queries for blastx using DIAMOND v2.0.7 against version 18.0 of the RVDB protein database (19). Direct mapping was also performed against reference genome Wuhan/Hu-60 1/2019 (NCBI Nucleotide – NC_045512, GenBank – MN908947) using the CLC Genomics Suite v5.1.0 (QIAGEN). The virus consensus was generated with iVar v.1.0 using a minimum of 5X read depth coverage. We added an N for such a position in case of lower read coverage. Samtools v1.10 (20) was used to sort the aligned BAM files and generate alignment statistics. We manually inspected all alignments and consensus sequences using Geneious Prime 2020.2 (https://www.geneious.com/).

Phylogenetic analysis

All SARS-CoV-2 sequences available on the GISAID EpiCov database (21) as of October 2020 were retrieved (until the most recent sample of our dataset was collected). The global SARS-CoV-2 phylogeny was reconstructed using the Nextstrain pipeline, version from February 2022 (22). Within the Nextstrain pipeline, high-coverage sequences were randomly subsampled to contain up to (i) ten sequences per month collected in France, (ii) five sequences per month per region (Europe), and (iii) two sequences per month collected in countries outside Europe (to avoid resampling). This dataset was combined with the sequences generated in this study. The resulting dataset (n=2206) was analyzed using augur and visualized with auspice as implemented in the Nextstrain pipeline. Acknowledgment of the contributing and originating laboratories for all sequences used in the analysis is provided in Supplementary Table S3.

Low frequency variants detection and filtering

Samples were sequenced in duplicates using an untargeted metagenomics sequencing approach. We did not manage to sequence one sample in duplicates (Table S2); hence, that sample was not included in subsequent analyses. Next, we used iVar v.1.0 (github.com/andersen-lab/ivar) (23) to call variants and to filter out variants that were not found across replicates. On average, nearly 60% of the variants were found in both replicates. We observed that there was a difference of less than five percent in the frequency detected in both replicates for most variants. However, in some cases, that difference reached 10 or 15% (Figure S1). For this reason, a mean of frequency and coverage depth was calculated for each variant in each sample. Lastly, we picked a variant frequency threshold of 1% and a minimum read depth of 1000X.

SUPPLEMENTARY FIGURE LEGENDS

Figure S1: Low frequency variants across replicates. A) Graphic showing the percentage of variant conserved across replicates for the 15 samples collected from the upper respiratory tract (nasopharyngeal aspirates, n=3) and lower respiratory tract (bronchoalveolar aspirates n= 12). B) Standard deviation of the frequency of each specific variant found in both replicates per sample.

Figure S2: Genomic analysis of viruses collected from the upper (nasopharyngeal aspirates, NP) and lower (bronchoalveolar aspirates). A) Schematic representation of the variables site across the SARS-CoV-2 genome collected from the lower and upper respiratory tract at the different time points compared to the reference genome collected on day 14. Each bar plot represents a specific position in which the nucleotide frequency. With grey, we colored the nucleotide matching the one present in the consensus sequence of the reference genome ("REF") and with red when the nucleotide is different from the reference ("ALT"). B) Zoom in into the phylogeny where the SARS-CoV-2 sequences of the immunocompromised patient fall. In the branches, we show AA changes and the support values, which are ultrafast bootstrap percentages. The scale bars represent the number of nucleotide substitutions per site.

REFERENCES

1. Wu F, Zhao S, Yu B, Chen Y-M, Wang W, Song Z-G, et al. A new coronavirus associated with human respiratory disease in China. Nature. 2020;579(7798):265-9.
2. Zhu N, Zhang D, Wang W, Li X, Yang B, Song J, et al. A Novel Coronavirus from Patients with Pneumonia in China, 2019. New England Journal of Medicine. 2020;382(8):727-33.
3. Andrew Rambaut NL, Oliver Pybus, Wendy Barclay, Jeff Barrett, Alesandro Carabelli, Tom Connor, Tom Peacock, David L Robertson, Erik Volz. Preliminary genomic characterisation of an emergent SARS-CoV-2 lineage in the UK defined by a novel set of spike mutations 2020 [Available from: https://virological.org/t/preliminary-genomic-characterisation-of-an-emergent-sars-cov-2-lineage-in-the-uk-defined-by-a-novel-set-of-spike-mutations/563.
4. Shen C, Wang Z, Zhao F, Yang Y, Li J, Yuan J, et al. Treatment of 5 Critically Ill Patients With COVID-19 With Convalescent Plasma. JAMA. 2020;323(16):1582.
5. Duan K, Liu B, Li C, Zhang H, Yu T, Qu J, et al. Effectiveness of convalescent plasma therapy in severe COVID-19 patients. Proceedings of the National Academy of Sciences. 2020;117(17):9490-6.
6. Li L, Zhang W, Hu Y, Tong X, Zheng S, Yang J, et al. Effect of Convalescent Plasma Therapy on Time to Clinical Improvement in Patients With Severe and Life-threatening COVID-19. JAMA. 2020;324(5):460.
7. Simonovich VA, Burgos Pratx LD, Scibona P, Beruto MV, Vallone MG, Vázquez C, et al. A Randomized Trial of Convalescent Plasma in Covid-19 Severe Pneumonia. New England Journal of Medicine. 2021;384(7):619-29.
8. Kemp SA, Collier DA, Datir RP, Ferreira IATM, Gayed S, Jahun A, et al. SARS-CoV-2 evolution during treatment of chronic infection. Nature. 2021.
9. Avanzato VA, Matson MJ, Seifert SN, Pryce R, Williamson BN, Anzick SL, et al. Case Study: Prolonged Infectious SARS-CoV-2 Shedding from an Asymptomatic Immunocompromised Individual with Cancer. Cell. 2020;183(7):1901-12.e9.
10. Vítor Borges JPG. Long-term evolution of SARS-CoV-2 in an immunocompromised patient with non-Hodgkin lymphoma virological.org2021 [Available from: https://virological.org/t/long-term-evolution-of-sars-cov-2-in-an-immunocompromised-patient-with-non-hodgkin-lymphoma/621.
11. Baang JH, Smith C, Mirabelli C, Valesano AL, Manthei DM, Bachman MA, et al. Prolonged Severe Acute Respiratory Syndrome Coronavirus 2 Replication in an Immunocompromised Patient. The Journal of Infectious Diseases. 2021;223(1):23-7.
12. Choi B, Choudhary MC, Regan J, Sparks JA, Padera RF, Qiu X, et al. Persistence and Evolution of SARS-CoV-2 in an Immunocompromised Host. New England Journal of Medicine. 2020;383(23):2291-3.
13. Bazykin GA SO, Danilenko D, Fadeev A, Komissarova K, Ivanova A, Sergeeva M, Safina K, Nabieva E, Klink G, Garushyants S, Zabutova J, Kholodnaia A, Skorokhod I, Ryabchikova VV, Komissarov A, Lioznov D. Emergence of Y453F and Δ69-70HV mutations in a lymphoma patient with long-term COVID-19 2021 [Available from: https://virological.org/t/emergence-of-y453f-and-69-70hv-mutations-in-a-lymphoma-patient-with-long-term-covid-19/580.
14. Teng S, Sobitan A, Rhoades R, Liu D, Tang Q. Systemic effects of missense mutations on SARS-CoV-2 spike glycoprotein stability and receptor-binding affinity. Briefings in Bioinformatics. 2021;22(2):1239-53.
15. Pasteur I. Protocol: Real-time RT-PCR assays for the detection of SARS-CoV-2, Institut Pasteur, Paris 2020 [Available from: https://www.who.int/docs/default-source/coronaviruse/real-time-rt-pcr-assays-for-the-detection-of-sars-cov-2-institut-pasteur-paris.pdf?sfvrsn=3662fcb6_2.
16. Matranga CB, Andersen KG, Winnicki S, Busby M, Gladden AD, Tewhey R, et al. Enhanced methods for unbiased deep sequencing of Lassa and Ebola RNA viruses from clinical and biological samples. Genome biology. 2014;15(11):1-12.

17. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics. 2014;30(15):2114-20.
18. Nurk S, Meleshko D, Korobeynikov A, Pevzner PA. metaSPAdes: a new versatile metagenomic assembler. Genome research. 2017;27(5):824-34.
19. Bigot T, Temmam S, Pérot P, Eloit M. RVDB-prot, a reference viral protein database and its HMM profiles. F1000Research. 2020;8:530.
20. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The sequence alignment/map format and SAMtools. Bioinformatics. 2009;25(16):2078-9.
21. GISAID. GISAID database [Available from: https://www.gisaid.org/.
22. Hadfield J, Megill C, Bell SM, Huddleston J, Potter B, Callender C, et al. Nextstrain: real-time tracking of pathogen evolution. Bioinformatics. 2018;34(23):4121-3.
23. Grubaugh ND, Gangavarapu K, Quick J, Matteson NL, De Jesus JG, Main BJ, et al. An amplicon-based sequencing framework for accurately measuring intrahost virus diversity using PrimalSeq and iVar. Genome Biology. 2019;20(1).

7.3 Discussion and conclusions of the studies

From the report on the early introductions and spread of SARS-CoV-2 in the northern regions of France there are key messages I would like to further address.

As noted in the main introduction of this thesis, once an outbreak has been detected, several questions regarding the nature and the origins of the disease arise. These questions can be addressed, at least to some extent, by genomic epidemiology if applied comprehensively. Critically, answering some questions (e.g., when or where did the outbreak begin) using genomic data requires knowing the evolutionary rate of the pathogen or of a related pathogen to be use instead. As explained in the main introduction of the thesis, to estimate the evolutionary rate, we make use of molecular clocks methods. Briefly, this process requires to calibrate the molecular clock, which can be done using the sampling times and to choose the molecular-clock model (e.g., strict, relaxed). By implementing these methods, we incorporate time information into the phylogenies, rescaling them into units of time. The result is a time-calibrated tree with branch lengths representing time units. In this way, we can also estimate the time of divergence between two sequences and the overall timescale of the outbreak.

Importantly, molecular clocks methods have two main assumptions: (i) evolution occurs at a predictable rate over time, and (ii) sampled genomes capture a measurable amount of evolutionary change (136)However, this might not always be the case. We can face two main problems when estimating the evolutionary rate early during an emerging outbreak. First, sampled genomes collected over a short period could lead to inflated evolutionary rates. This situation could occur if sampling was done early in an epidemic over a limited timescale. In that case, the genomes could harbor an excess of mildly deleterious mutations, which would have eventually been removed over time from the population by selection. For this reason, evolutionary rates are generally time-dependent (410). For instance, this situation has been reported in the context of the 2013–2016 EBOV outbreak in West Africa. Early estimations led to an evolutionary rate twice as high compared to what was estimated for previous EBOV outbreaks, leading to speculations about how this would impact the transmissibility and virulence of the virus (411). Second, genomes captured early on an outbreak might contain low genetic variation leading to poor temporal signal in the data. This might occur because the sampling window might not be yet large enough to capture sufficient evolutionary change in the viral genomes. If this is the case, the data is too little informative about the evolutionary process shaping the outbreak, and hence, inferences from the molecular clock, inadequate. In this case, evolutionary rates should be based on previous outbreaks or closely related pathogens. For this reason, it is very important to determine if an outbreak has reached the “phylodynamic threshold,” that is, if, within the available genomes, there are enough molecular changes to calculate robust phylodynamic estimates (136). During the early phase of the COVID-19 pandemic, the pace of SARS-CoV-2 sequencing was swift, with genomic data being publicly available as soon as the virus started spreading, but with geographic inequalities.

Consequently, during the first months of SARS-CoV-2 circulation, a plethora of phylogenetic analyses were carried out to investigate SARS-CoV-2 origins, genetic diversity, and spatiotemporal spread. However, at the same time, there were a lot of concerns raised about the interpretation of real-time molecular epidemiology as inferred rates could significantly change over time before converging into stable and long-term estimates. And indeed, it was observed that the substitution rate for SARS-CoV-2 calculated over

approximately two months was twice higher compared to the one calculated over eight months (until February 2020 and August 2020, respectively) (412).

In addition, beyond the possible problems associated with the estimation of the evolutionary rate, we need to consider that the dataset might not meet all the characteristics for robust phylogenetic and phylogeographic inference, in particular during ongoing outbreaks (137). For example, unevenly sampled data could lead to inadequate geographic representation resulting in misleading conclusions about the geographical source of the outbreak.

During the analysis of the early introductions and spread of SARS-CoV-2 in the northern regions of France we had to deal with situations similar to those mentioned above. First, the data was collected over a short period of time, from 24 January to 23 March 2020. Second, at that moment there were scarce early sequences in many countries in Europe including France. Third, the limited genetic diversity of SARS-CoV-2 detected at that time. Fourth, the uneven sampling of several French regions according to case counts (Fig.1 of the manuscript). And lastly, the nature of the syndromic-only based genomic surveillance for a virus associated with a large proportion of asymptomatic infections.

Because of this, we considered that the dataset was not sufficient for reliable phylogeographic inferences and hence, when and where the virus was introduced to France remained unclear. Nevertheless, with our work we provided important insights about the initial introductions of SARS-CoV-2 in the country and the later spread of the virus at the local level. In particular, we found that the first cases of SARS-CoV-2 infection detected in France, and even in Europe, on January 24, 2020 from travelers from Hubei, China, belonged to the V clade. The following detected cases in February belonged also to the V and the S clade. Our analysis indicated that at least these first introductions did not lead to further local transmission, highlighting the efficacy of preventive measures (i.e. contract tracing and isolation) imposed on the initial imported COVID-19 cases (based on syndromic surveillance). Still, SARS-CoV-2 managed to make its way into the country as in contrast to the first cases, sequences detected later on fell in the diversity of clade G, today better known as the B.1 lineage, according to the PANGO nomenclature. Interestingly, the earliest sequence of the G clade, sampled from an individual with no travel history and no contact with returning travelers, was not the least divergent in the clade, suggesting local circulation in undocumented infections prior to the wave of COVID-19 cases in France.

To conclude this first part, although genomic epidemiology can provide crucial information to support public health outbreak responses, phylogenetic analyses, especially from ongoing outbreaks, needs to be interpreted with caution in light of limitations and assumptions. Similarly, it is important to consider that the viral sequences obtained at any given time, may only represent the tip of the iceberg of the underlying viral genetic diversity. Therefore, the phylogenetic relationships drawn can be challenged as more samples are obtained.

In the second project presented in this chapter, we aimed at characterizing the intra-host dynamics of SARS-CoV-2 population during the long-term infection of an immunocompromised patient who received convalescent plasma therapy. For this purpose, we deep-sequenced samples collected from the lower respiratory tract at regular intervals over the course of the infection.

One of the strengths of our study is the special efforts made to ensure reliable and accurate low variant calling and analysis. Like every other type of measurement, measuring the viral genetic diversity is limited by the signal-to-noise ratio of the experiment. Indeed, one of the biggest challenges when characterizing

within-host viral diversity is distinguishing "true" from false-positive low frequency variants (413). Although deep sequencing techniques can produce enough data to look at these low frequency variants, errors can be introduced at various steps of the sequencing workflow. Starting from the library preparation, cross-contamination during sample processing can lead to the detection of spurious mutations. In addition, during the sample processing viral RNA samples are generally reverse transcribed (RT) to cDNA, which is subsequently amplified by PCR in order to obtain sufficient amounts of material for sequencing. These two steps can be very error-prone. For instance, errors can be introduced during the cDNA synthesis, with current commercial reverse transcriptases having an estimated error rate of 1.8×10^{-4} for Superscript IV and 1.3×10^{-4} for TGIRT (414). Despite newly engineered high-fidelity polymerases, PCR amplification can result in polymerase mistakes generating point mutations in the resulting PCR products; or in population skewing due to unequal amplification. The latter can occur due to primer mismatches against the target sequence altering the efficiency of the PCR (415). Lastly, although the major improvements over the last few years, errors can also be introduced during sequencing with NGS technologies having an associated error rate between 1/100 and 1/1000 per base pairs sequenced (416).

Although errors can occur at these various steps, their impact on the analysis would differ according to which step they appear; the earlier the spurious mutation appears during the process, the more significant the impact. For example, an error introduced during the RT will be dragged in the following steps and may be amplified during the PCR and the library preparation (despite the usual short number of cycles). As a result, errors introduced during the RT may not necessarily be present at very low frequencies.

Beyond these technical errors, other factors can alter the accuracy of low frequency variants detection. In particular, the number of input genomes, with lower input samples being more sensitive to cross-contamination and RT, PCR and sequencing errors leading to an inflation of false positives; and sequencing coverage depth (72). In this regard, a minimum of 1000 virus RNA copies is usually recommended as an input (415) and it is generally accepted that the coverage should be 10 times the reciprocal of a variant's frequency (72). In this way, to reliably detect variants present at 1% of frequency, a minimum coverage of 1000X coverage is needed.

In sum, all these factors can produce erroneous identification of low frequency variants thereby leading to biased measurement of the viral diversity. Indeed, in several works, erroneous conclusions may have been drawn by calling spurious low frequency variants. For instance, as pointed out by Bloom et al. (417), calling spurious mutations led to larger transmission bottlenecks estimates for human influenza virus by Poon et al. (418) compared with other several studies. Similarly, false-positive variants biased the results of SARS-CoV-2 transmission bottleneck sizes by Popa et al. (419). The re-analysis of this work by Martin and Koelle (420) resulted in the estimation of a narrower bottleneck, in agreement with previous studies (421, 422).

Therefore, to characterize the SARS-CoV-2 intra-host evolutionary dynamics including low frequency variants, we aimed to use a robust method to ensure accurate low frequency variant calling and analysis. For this purpose, samples were sequenced using an untargeted metagenomic deep sequencing approach which allowed us to capture the composition of the viral population in an unbiased manner, limiting errors and biases associated to the PCR amplification. In addition, each sample was prepared in duplicates from the extracted RNA, to account for possible errors during the reverse transcription, and sequenced on different batches, in view of possible errors introduced during the sequencing. To perform our analysis, the raw files obtained from the variant calling using iVar were filtered to keep the variants that: 1) passed

the strand bias test; 2) were found across duplicates; 3) had a minimum frequency threshold of 1%; and 4) had a minimum coverage of 1000X. We found that per each sample replicate, an average 60% of the total variants were detected in both replicates, suggesting that a large amount of variants were artifacts, probably introduced during the sequencing or library preparation, in particular during the reverse transcription. Not surprisingly, in the samples in which we found less consistency between replicates, the Ct values were higher. This fact reflects how the input virus concentration may affect the low frequency variants detection. Nevertheless, among the variants found across replicates, there was little divergence in the frequency of the variants between replicates (in average a difference of 5% of frequency).

After having decided the cutoff for the analysis of the low frequency variants, the next step was to look at the evolution of SARS-CoV-2 over the course of the infection in the immunocompromised patient treated with CP therapy. Similarly to a report of Adam Lauring's team (423), we showed that the long-term evolution of SARS-CoV-2 within an immunocompromised patient is not always associated with a strong accumulation of AA changes, especially in the spike protein of the virus. However, this contrasts with other published works. For example, the virus described by Choi et al., (394) after 146 days of evolution in an immunocompromised patient harbored 7 non-synonymous mutations and two deletions only in the spike protein. Similarly, several works have described a strong selection on SARS-CoV-2 during CP therapy and emergence of mutations in the spike believed to affect SARS-CoV-2 affinity to ACE-2 receptors such as Y453F (424) and N501Y (425, 426) or believed to be involved in immune evasion (ΔH69-V70 and E484K) (424, 426). In contrast, after the CP treatment we observed a major shift in the viral population that restored the genotype dominant in the early state of the infection, characterized by the spike protein variant S50 (S:S50) instead of L50 (S:L50). Indeed, the amino acid sequence of both the replaced and novel dominant population only differ at this position. Given that most neutralizing antibodies target the spike protein (427) and after the CP treatment, we observed only one AA change in the viral population, which was located in the spike protein, we hypothesize the following two scenarios. In the first scenario, the population emerged under positive selection: the antibodies in the CP specifically targeted the viruses harboring the variant S:L50, therefore leading to the replacement of a viral population with an escape genotype (S:S50). In the second scenario, the antibodies managed to eliminate the dominant population in the lower respiratory tract, regardless of the genotype of the spike protein (i.e., S:S50 or S:L50), which was subsequently replenished by a viral population less accessible to the administered antibodies, possibly located in another compartment.

Something interesting that could be done in order to explore these two scenarios would be to generate lentivirus pseudoparticles packaging the wild-type spike and spike 50L mutant and perform neutralization assays to measure the neutralization activity of convalescent plasma against these viruses (Figure 7-7). This system has been described by the Bloom laboratory (428) and it has been recently used in our laboratory (429). It has the advantage of introducing the SARS-CoV-2 spike protein, with or without any particular mutations of interest, into a lentivirus non-replicative particle. With such a system we could study the infectivity and sensitivity of the pseudoparticles to neutralization by convalescent plasma samples. In this way, this type of experiment would help us to understand if the S50L substitution on the spike protein was enough to trigger the replacement of the viral population in the lower respiratory tract, leading to the re-emergence of a population carrying the S50 substitution as the dominant viral population.

Another interesting experiment to assess if the spike S50L could increase the sensitivity to antibody neutralization would be to study how this mutation would affect the spike structure. In this regard, such computational modeling has already been done and it has been shown that the substitution S50L may

have stabilizing effects on SARS-CoV-2 spike protein but no evidence of affecting antibody binding in this region (430).

Finally, in line with previous studies (394, 395), we observed that the CP failed to eliminate the infection, highlighting the challenge of treating immunocompromised individuals. However, the effect of CP observed on the disease progression and virus evolution cannot be translated to immune competent hosts who have a better immune system control.

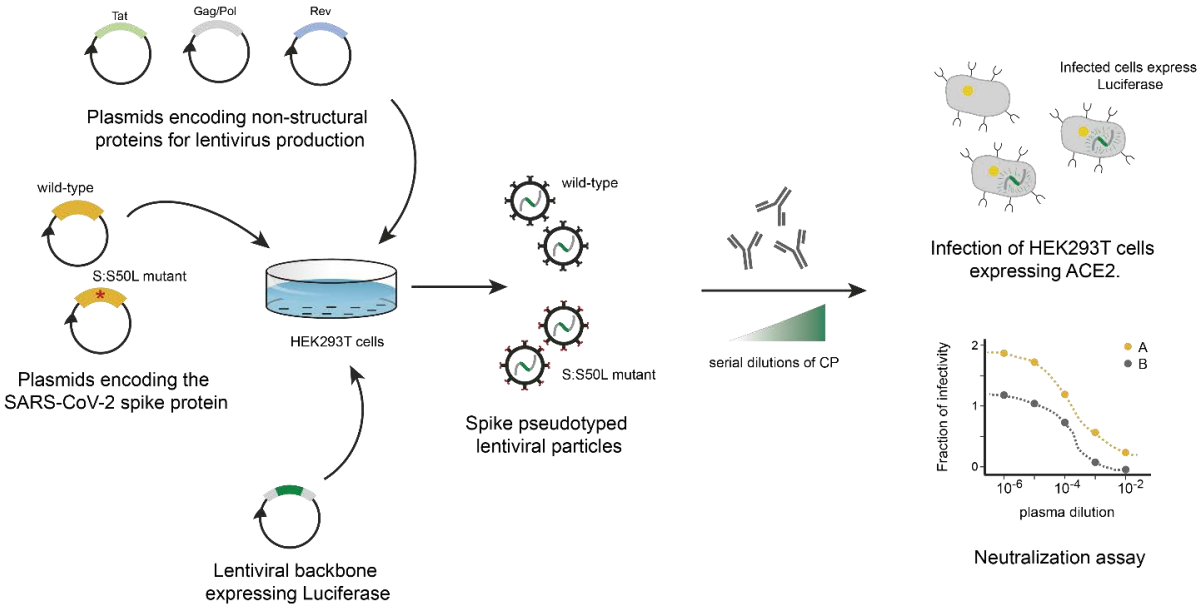


Figure 7-7: *Approach to assess the CP neutralizing potency against the wild-type spike and spike mutant S50L (S: S50L) of SARS-CoV-2. HEK293T cells are transfected with a plasmid encoding a lentiviral backbone expressing Luciferase, a plasmid expressing the wild-type spike or spike mutant S50L and plasmids expressing the Tat, Gag-Pol, and Rev HIV proteins needed for virion formation. The transfected cells will produce lentiviral particles with the spike protein in the surface (the wild-type or the mutant spike version). The neutralization assay would be achieved by first performing serial dilutions of the convalescent plasma and subsequent incubation with the pseudotyped lentiviral particles. Finally, this mix would be added to HEK293T cells expressing ACE2. Infected cells would express the Luciferase, signal that can be measured. To calculate the inhibitory concentrations 50% (IC50s), the fraction of infectivity elicited by the pseudotyped lentiviral particles (for the wild-type and the mutant spike) in the presence of serial dilutions of the CP is studied. The fraction of infectivity is calculated as the luciferase reading for a particular plasma dilution divided the luciferase reading in the absence of plasma also referred to as "maximum infectivity". This lentivirus pseudoparticles packaging a coronavirus spike system for neutralization assays has been described by the Bloom laboratory (428).*

Key points of the Chapter 3

- Phylogenetic analysis of the initial introductions of SARS-CoV-2 shed light on the initial (multiple) introductions in France and the later spread of the virus at the local level, with insights for Europe.

- This work highlights the risk of phylodynamics and phylogeographic inferences based on fractionated data, low SARS-CoV-2 genetic diversity compared to the fast spread, and a syndromic-only genomic surveillance.

- Long term evolution of SARS-CoV-2 in an immunocompromised individual is not always associated with strong accumulation of changes in the viral genome.

- As treatment with convalescent plasma did not result in viral clearance, treated patients should still be considered as potentially infectious, highlighting the challenge of managing such patients in long term care.

8 GENERAL DISCUSSION

Limitations and perspectives in virus genomic epidemiology

This thesis explored the use of genomic epidemiology to dive into the origin, timing, and spread of several viral infectious diseases. In chapter 1 we studied the presence of RNA viruses in CSF samples from patients with aseptic meningitis with unknown etiology. We aimed at reconstructing the dynamics of two CHIKV outbreaks in Cambodia in chapter 2. Finally, in chapter 3, we studied the introductions and early spread of SARS-CoV-2 in France, and we monitored the intrahost evolution of SARS-CoV-2 during a long-term infection in an immunocompromised patient.

Genomic epidemiology together with the field of phylodynamics constitute powerful frameworks to investigate infectious disease. However, both fields face numerous challenges, from technical aspects (i.e., data generation), theoretical challenges (i.e., data analysis and results interpretation), to global trends in data generation and sharing. We have highlighted a few examples of some of these challenges throughout this thesis.

mNGS can be used to identify the potential etiologic agent causing the disease, as we showed in chapter 1 or as it has been demonstrated by many others (174-177). However, several challenges limit its use. First, the low-quality and/or low-concentrated samples can affect the sensitivity for pathogen detection. Indeed, sample stability is essential for sequencing, particularly for RNA, which is labile and susceptible to degradation by RNase enzymes or due to multiple freeze-thaw steps during sample preparation. Overcoming these problems is critical for leveraging mNGS for different applications, including clinical applications. Different laboratory practices and methods have been proposed to overcome the challenges of low-quality and/or low-quantity RNA samples. For example, in order to prevent degradation, RNA samples are generally stored frozen at −20 °C or −80 °C, and the use of RNase inhibitors, RNase-free reagents, and/or protective agents is highly recommended. In addition to this, different methods have been developed to deplete unwanted DNA or RNA, enriching samples with the RNA/DNA of interest (431). For example the method described by Matranga et al. (171), and used through this thesis, which rely in the selective depletion of host ribosomal RNA and contaminating poly(A) carrier, enriching clinical samples with viral content.

Another factor that can influence the sensitivity for pathogen detection is sample contamination. As mentioned in chapter 1, sample contamination can arise at different stages of the mNGS protocol: during the sample extraction (e.g., contamination from skin flora during needle aspiration), aliquoting, nucleic acid extraction, library preparation, or the sequencing run. In addition, contaminants can be detected at the same or even higher level than bona fide pathogens in low-concentrated samples, making it difficult to interpret the data (229). Therefore, it is advisable to follow procedures to ensure that the environment is sterile and RNA and DNA-free and avoid sample cross-contamination. In addition, positive and negative controls (e.g., no-template control, non-infectious control) should be included to guarantee that environmental and sample cross-contamination are not producing false-positive results.

Nevertheless, due to its high utility and cost reduction, mNGS increases in popularity and becomes increasingly used. As a consequence, it is expected to have in the future more extensive sequencing databases of patients with different infectious diseases and with no infection, enriching host-pathogen databases. Having such databases would enhance our ability to discriminate between potential pathogens, commensal microbiota, and contaminants. Additionally, user-friendly tools, bioinformatic pipelines, and computational analysis webservers such as CZ ID (432) are becoming more widely available, helping with the analysis of the data. In a similar way, significant efforts are being put toward updating and increasing the quality of available databases such as FDA- ARGOS (433) or RVDB (233).

As mentioned throughout this thesis, genomes can be used for outbreak investigation, strengthening public health response. One of the major issues for interpreting genomic epidemiological studies is sampling bias (122, 137, 158). The absence or the overrepresentation of specific samples can lead to inadequate representation, impacting the reliability of conclusions made using phylogeographic analysis. As we discussed in chapter 2, during the analysis of the emergence and spread of CHIKV in Cambodia, unbiased sampling is usually hard to achieve, in particular for the following reasons:

1. it demands knowing the extent and the intensity of the outbreak
2. it requires having access to the locations concerned for sampling
3. it requires important sequencing efforts

For this reason, while it is important to extract as much information as possible from the genomic data, it is imperative to remember that the viral diversity sampled represents a snapshot of the outbreak taken at a given time. Therefore the resulting phylogenies are hypotheses that might be challenged if we obtain more data (137).

Several strategies can be implemented in order to limit sampling bias. For example, performing population-based studies, like the one carried out in Iceland to investigate the spread of SARS-CoV-2 (434). Briefly, people tested included symptomatic individuals and asymptomatic individuals between 20 and 70 years old, who were randomly invited to enroll in the study. Nevertheless, carrying out such a study is not always feasible. Furthermore, the inclusion of additional data could help to mitigate sampling bias. For instance, flight and occurrence data (confirmed cases but with no sequences available) can be integrated into phylogeographic reconstructions to estimate the number of introductions to a specific country. Du Plessis et al. implemented such a strategy, which allowed the authors to obtain a more detailed picture of SARS-CoV-2 importations to the UK before the lockdown (435). Additionally, including travel history data in the phylogeographic analysis could also help by adding new locations in the estimations and yielding a more realistic hypothesis of the virus spread (156). This strategy can be very relevant for studying cases of travelers coming from areas with no sequence data available. Such was the case in analyzing SARS-CoV-2 introductions in Rwanda (436). By incorporating travel history information in the phylogeographic inference, the authors included traveler cases from Tanzania, Burundi, and South Sudan, countries for which no SARS-CoV-2 genomes were available at the time of the study. Indeed, their analysis inferred the introduction of SARS-CoV-2 from these countries to Rwanda.

Another major issue for performing and interpreting genomic epidemiological studies is whether our data is sufficiently informative for inferring the viral evolutionary rate and the time scale of the outbreak. There might not be enough nucleotide changes accumulated in the sampled genomes, leading to no temporal signal and thus no usability of the dataset to get any estimates from the molecular clock (136). As discussed

in chapter 3, this problem may be critical at the beginning of a viral outbreak. In the early stages, genomic variation might be very low, an thus insufficient to make reliable inferences even when large numbers of genomes sequences are available (136, 137). For this reason, it is crucial to determine if an outbreak has reached the phylodynamic threshold: the point at which there are enough molecular changes within the available genomes to calculate robust estimates (136).

We were faced with this situation during the analysis of the introductions and early spread of SARS-CoV-2 in the Northern regions of France. Our data was collected over a short period, from 24 January to 23 March 2020, and it had little genomic variation. This, combined with the scarcity of early sequences in many countries in Europe at the time and the uneven sampling of several French provinces, led us to consider that dataset was not sufficient for reliable phylogeographic inferences.

Furthermore, although virus genomic data can by itself provide relevant information, to complement "classic" epidemiological approaches in assisting public health responses to an outbreak, genomic data needs to be accompanied by reliable metadata (109). In order to use viral genomes for outbreak investigation two main pieces of information are required for each clinical or biological sample: the date and location of collection. However, such data might not always be available, for example when doing retrospective analysis. We found ourselves in this situation when studying the genetic diversity of TOSV in Spain. Several sequences available in ViPR did not have a collection date or sampling location. Together with the fractionated genomic data available for TOSV, this situation prevented us from doing extensive phylodynamic analysis.

Additional information, including travel and contact history and ecological and human mobility data, can increase the utility of genomic epidemiology, providing a more comprehensive picture of the outbreak. For example, the typical seasonality of arboviruses in Cambodia is described as increasing cases during the rainy season, peaking during July and August. However, climate change and urbanization might favor continual mosquito circulation. In the future, data about *Aedes* mosquito abundance and distribution in Cambodia would allow knowing, for example, if the circulation of these vectors stops or not during the dry season. Such information would be relevant to understanding whether CHIKV can persist during the inter-epidemic period thanks to mosquito vertical transmission cycles or low and undetected human-mosquito transmission cycles. In addition, this could provide a better understanding of mosquitoes and, therefore, mosquito-borne virus circulation. Such information could ultimately be used for planning vector-control strategies and genomic surveillance systems.

Nevertheless, several open-access databases, namely WorldPop (437), FlowMinder (438), VectorBase (439) and Virion (440) have shown promising progress towards maximizing resources availability for detailed outbreak studies.

Finally, a major challenge for using genomic epidemiology is its implementation.

While more and more countries are using genomic epidemiology as part of their surveillance programs, this might not be true in several regions of the world, particularly in low-income and middle-income countries (441). Despite the overall drop in the cost of NGS over the past few years, purchasing all the necessary equipment and reagents is still a significant barrier to implementing genomic epidemiology, particularly in these countries (109).

Furthermore, to put genomic surveillance in place, it is critical to improve and strengthen the cooperation between partners at various levels (441). First, establishing collaborations between hospitals, public health

agencies, and academic research laboratories to ensure maximum and real-time genomic surveillance at the local or country level. Second, at the international level; a collaborative framework that transcends borders is key to maximizing the benefits of using genomic sequencing for strengthening public health responses. A good example of this was when recently, scientists in South Africa quickly alerted the emergence of the Omicron variant, which enabled other countries to monitor and prevent its circulation (as much as possible).

These international efforts and collaborations are also important due to the global disparities in viral genomic surveillance as we witnessed during the ongoing COVID-19 pandemic. A study in August 2021 indicated that around 77% of low and middle-income countries sequenced less than 0.5% of their cases (442). Due to the potential emergence of new variants of SARS-CoV-2, strengthening international collaborations and improving global surveillance, particularly in low-income and middle-income countries, should be a matter of global concern and global priority (442).

This need for improving global genomic surveillance highlights the importance of creating scientific networks and research collaborations to generate and share data and knowledge in real-time. However, open data sharing might not benefit everyone, particularly researchers working in low-income and middle-income countries, who might not have the possibility to analyze and valorize their data like in other countries. Because of this, the GISAID database (328), has put in place in a user's agreement that demands giving credit to the people who generated the data when using their sequences. However, the equilibrium between real-time generation, sharing of the data and protection of researchers who produced the data is still a critical matter of debate. Hopefully, we will progress towards an agreement favoring both parts while ensuring maximum benefits of genomic epidemiology.

9 CONCLUDING REMARKS

With this thesis, I hope to have contributed to a better understanding of the epidemiology and evolution of several RNA viruses representing important threats to human health. At the same time, I hope to have shown that genomic epidemiology could be a powerful tool to investigate infectious disease outbreaks at various steps.

For instance, upon detecting a disease, the most critical question to be answered is: which is the responsible pathogen?

In chapter 1, we used metagenomic sequencing to attempt identify the potential pathogen causing a series of meningitis cases in Southern Spain, showing how we could answer this question at least in some cases. Indeed, our mNGS analysis detected in multiple cases RNA from TOSV, an arbovirus responsible for an increasing number of infections in countries of the Mediterranean region. mNGS is a promising approach for diagnosing infectious diseases because a broad spectrum of pathogens can be detected in a single assay (viral, bacterial, parasitic, and fungal), including complete new recombinant forms of the virus. Regarding the latter, using this approach, we also identified a novel E13 recombinant form among the EV-positive meningitis cases, an intriguing result that could have remained unnoticed with classic typing methods.

Once the pathogen responsible for the disease is identified, we can determine whether we have the tools for diagnostics (e.g., qPCR or serology tests), investigation (e.g., complete genome sequencing), and prevention (e.g., vaccines). If we do not have them, we can use the information provided by the mNGS to develop them. In this sense, our mNGS findings allowed us to design an amplicon-based sequencing approach. Thanks to this, we successfully obtained several full-length genome sequences of TOSV, providing the first complete genomes circulating in Spain in the last 20 years. In addition, this sequencing protocol will hopefully be useful to colleagues in Spain but also in the Mediterranean region, as it allows to detect and generate complete TOSV genomes, in particular from clinical samples of suboptimal quality and quantity of RNA.

Chapter 2 also combined two sequencing methods (metagenomic and amplicon-based approaches) to obtain complete CHIKV sequences from cases detected in Cambodia during two different outbreaks: 2011-2013 and 2020. To try to get further insights into these outbreaks and their dynamics, we first added temporal data to our genomic data to estimate when CHIKV was introduced into the population. Next, we included geographic data, and we performed subsequent phylogeographic analysis yielding an additional level of detail, shedding light on the likely origin of the outbreak, connections to outbreaks in the same region, and dispersal of the virus within the country. Nevertheless, we are aware of the limitations of our study, particularly for the epidemiological and genomic CHIKV data available for Cambodia and neighboring countries.

Furthermore, when viral genomes sequenced from the same region during different epidemics are available, phylodynamics can provide significant insights into the evolution of the virus during the inter-epidemic period. Such information could be used to answer the following questions: was the virus able to

persist in the population between the two outbreaks? Or, was the virus introduced to the population from a new spillover from an animal reservoir?

This is an exciting question that has already been explored through the different EBOV outbreaks. Most EBOV outbreaks are the result of spillover events from the animal reservoir. In contrast to this, five years after the last outbreak, EBOV re-emerged in Guinea in 2021 from the result of long viral persistence in a human survivor and subsequent flare-up, as the authors suggested (443). In the second chapter, we had the opportunity to slightly explore the inter-epidemic evolution and spread of CHIKV in Cambodia. In light of the available genomic and epidemiological data for the Southeast Asia region, our analysis suggested that the outbreak was not seeded from previously CHIKV circulating in Cambodia but instead from the introduction of the virus from countries where CHIKV seems to circulate almost continuously, such as India.

Viral genomics can also be used to study intra-host evolution. NGS allows the sequencing of viruses at a high depth of coverage, characterizing the complete variant repertoire of the viral population. Studying intra-host viral population diversity can provide insights into how these processes could relate to the virus evolution at a larger scale. This is a major point of discussion on the ongoing COVID-19 pandemic. Indeed, it has been proposed that chronic infections leading to a substantial accumulation of nucleotide changes could be responsible for the emergence of VOCs such as the Omicron or Alpha variants. Chapter 3 superficially explores such an idea by monitoring the evolution of SARS-CoV-2 during a long-term infection in an immunocompromised patient. We found that the long-term evolution of SARS-CoV-2 in an immunocompromised individual is not always associated with a strong accumulation of changes, particularly in the spike protein. Additionally, as the patient was treated with convalescent plasma therapy, we studied the impact of the treatment on the evolution of the viral population. Interestingly, we noted a significant shift in the viral population but no virus clearance after the treatment, which highlights the challenges of treating these vulnerable members of our society.

In summary, I hope to have provided examples of what we can learn from sequencing data, highlighting the utility of genomic epidemiology and phylodynamics for outbreak investigations while adding to the understanding of several RNA viruses.

10 APPENDIX: the initial plan A

I joined the G5- Evolutionary genomics of RNA viruses in October 2018 to work on experimental RNA virus evolution.

My initial project aimed to assess how the environment in which a virus replicates can influence the evolution and composition of the viral population and to evaluate the consequences of such variation in key parameters such as pathogenicity and transmissibility. Here, with the environment, we specifically referred to the host's genetic background and immune history. For this purpose, the project consisted in setting up an *in vitro* evolution system, where we would use cells from the blood of healthy donors from different ancestries (e.g., Asian, African ancestry) to replicate a viral population. After 10 to 15 passages, we would evaluate the composition of the viral population by NGS and its fitness, tropism, and transmissibility. To study the effect of the immune history on the composition and evolution of the viral population, we planned to add other variables, such as antibodies, at non-neutralizing concentration, as it is known that this can modulate the environment and thus constrain the viral population (Figure 10-1-1).

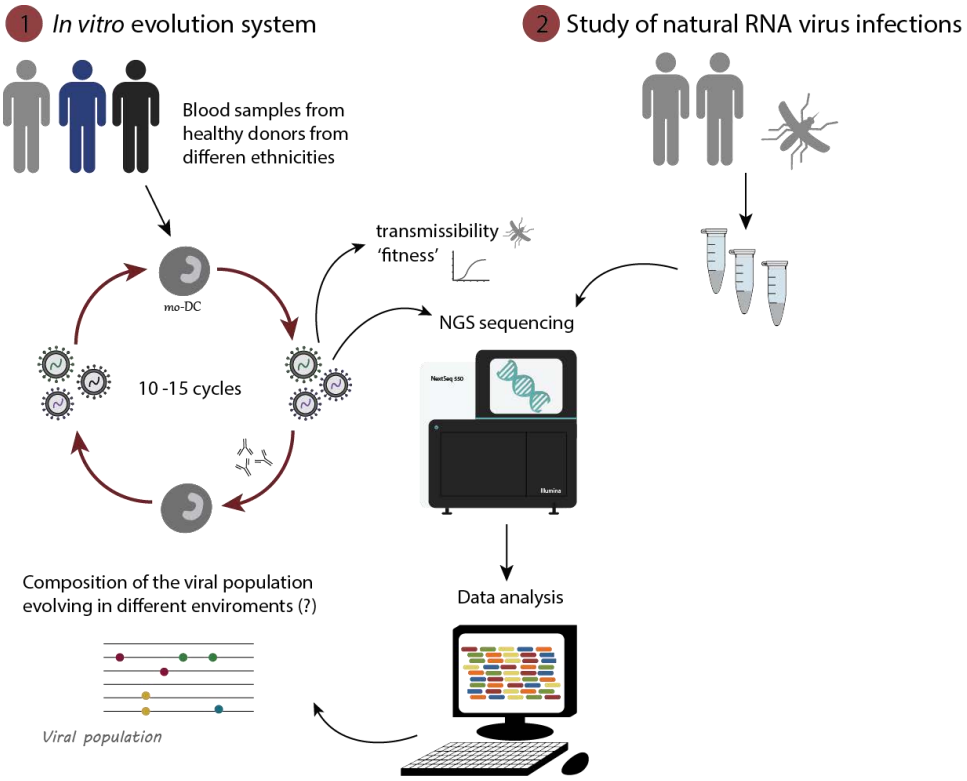


***Figure 10-1**: Initial Ph.D. project with two complementary parts. The first part included an in vitro evolution system to study the impact of the genetic background and immune history of the host in the evolution and composition of a viral population. The second part consists of the study of natural RNA virus infections.*

As a model virus, we planned to use DENV for two main reasons. First, differences in dengue disease symptoms have been reported at the individual level and within human populations (i.e., populations of

different ethnic backgrounds). Second, DENV infection induces long-term immunity against the infecting serotype but, in most cases, not against heterologous serotypes. In line with the antibody-dependent enhancement mechanism, evidence suggests that severe dengue diseases are associated with cases of secondary heterotypic infection. This supported the idea that human populations can constitute very different environments for the virus (e.g., DENV) evolution.

We spent more than one year fine-tuning the *in vitro* evolution system. In particular, we were trying to find the most suitable cell type which we would use to replicate DENV for 10 - 15 passages. In other words, we were looking for a human cell type that we could infect with DENV and that could support enough virus production to continue the cycle of infection. Our first choice was dendritic cells and macrophages, the first primary target cells of DENV. We dedicated a substantial time to learn and optimize the following two critical steps:

1. Purification of CD14+ cells from PBMC (peripheral blood mononuclear cells) that we obtained from healthy donors
2. Differentiation of CD14+ cells into dendritic cells (MO-DC), macrophages type 1 (MO-M1), and macrophages type 2 (MO-M2) using different cocktails of cytokines.

Next, we moved on to perform the infection assays. Briefly, initially we choose to use DENV serotype 1 (DENV-1) to perform the infections, as this low passage isolate grows to high titers compared to all DENV isolates we tested. We used flow cytometry to evaluate the number of infected cells and titration by focus forming assay in Vero cells to evaluate the amount of virus produced post-infection.

The infection of primary human MO-M1 resulted in a low infection rate (less than 1%) and low virus production. On the contrary, the primary human MO-DC were infected by DENV with around a 20% infection rate, and we obtained a higher viral titer; however, it was still not enough virus production to continue the infection cycle. In order to improve this, we tried to optimize the infection experiments by using different MOIs (multiplicity of infection), that is, the ratio of viruses per cell, and by performing DENV growth curves to identify at what point post-infection the virus production was the highest. Despite such efforts, we did not manage to improve the number of infected cells or the virus production.

Subsequently, we tried the infection of human-induced pluripotent stem cells (hiPSCs) with DENV-1. HiPSCs are cells that are generated from adult somatic cells that have been reprogrammed back into an embryonic-like pluripotent cell. Recent evidence suggested that these cells could be infected at a high rate by viruses like DENV or ZIKV. Therefore, using these cells to build our system had promising benefits. First, we could obtain hiPSCs from healthy donors from different ethnicities, which is fundamental to answer our biological question. Second, hiPSCs can be grown and subject to cell-type-specific differentiation, thereby consisting of an unlimited source of dendritic cells and macrophages. At the same time, we recognized a potential drawback of using this cells: the reprogramming process to hiPSCs might induce genomic and genetic changes. Given that the genetic background was critical for the project, to use these cells in our *in vitro* project would have required other steps to ensure the genomic integrity of the cells, for example.

We obtained hiPSCs from three different lineages: European, Asian, and African lineage, and we decided to perform the infections with DENV-1. The infection of hiPSCs resulted in a higher infection rate than with the MO–DC and higher viral production, but not enough to continue the infection cycles.

After a long year of experiments without much success and considering that we had invested a coherent amount of time and effort on this project, we decided, in agreement with my thesis advisory committee to set it aside to focus on more feasible projects. However, we were prepared for that: from the beginning, we anticipated significant technical difficulties (e.g., obtaining enough cells from each donor to perform all the infections or finding a suitable cell type capable of sustaining DENV infection for several rounds replications).

The second part of the project aimed at studying the DENV viral population during human to mosquito transmission. We had biological samples from a Cambodian cohort[3] that included human patients infected with the dengue virus and mosquitoes that were directly fed with the blood of these patients. In some cases, we also had mosquitoes captured from the same environment where the patients lived. We plan to use NGS to characterize DENV intra-host genetic diversity in both patients and matched mosquitoes and look for patterns of transmissions. However, the low quality of samples, probably due to the repeated thawing and freezing processes that these samples had been subjected to, made it very complicated to obtain good quality data from which we could draw significant conclusions. For this reason, after several months of work, we decided to stop working on this project.

Nevertheless, the laboratory had other projects in which I could get involved. The projects that we estimated we could finish in the stipulated time of the Ph.D. were mostly molecular epidemiology studies, such as the study of CHIKV emergence and spread in Cambodia. Although this implied a significant shift in my Ph.D. project from the biological question to the virus model, which required additional know-how, we thought it was the most straightforward solution.

Although I have not produced a publishable body of work from the first and half years of my Ph.D., I appreciate that it still provided me with numerous skills that I could transfer to next projects, including classical virology techniques, library preparation, NGS data analysis, and very importantly, resilience and an adaptable mindset.

Today, the new projects have proved fruitful with two first-author publication and others in preparation. In addition, they allowed me to explore other frameworks to study the evolution of RNA viruses, such as the field of phylodynamics and phylogeography, which I would not have done otherwise. I have, indeed, developed such a great interest in these fields of research that I will continue exploring them during my post-doc.

[3] These samples were obtained in the framework of the European program DENFREE consortium coordinated by Anavaj Sakuntabhai, and in collaboration with Institut Pasteur du Cambodge.

11 REFERENCES

1. Morens DM, Fauci AS. Emerging Infectious Diseases: Threats to Human Health and Global Stability. PLoS Pathogens. 2013;9(7):e1003467.
2. Zhu N, Zhang D, Wang W, Li X, Yang B, Song J, et al. A Novel Coronavirus from Patients with Pneumonia in China, 2019. New England Journal of Medicine. 2020;382(8):727-33.
3. Trust W. From equality to global poverty: the COVID-19 effects on societies and economies 2021 [Available from: https://wellcome.org/news/equality-global-poverty-how-covid-19-affecting-societies-and-economies.
4. Foundation G. COVID-19: A global perspective 2020 [Available from: https://www.gatesfoundation.org/goalkeepers/report/2020-report/#GlobalPerspective.
5. Saladino V, Algeri D, Auriemma V. The psychological and social impact of Covid-19: new perspectives of well-being. Frontiers in psychology. 2020;11:2550.
6. Development OOfEC-oa. Combatting COVID-19's effect on children 2020 [Available from: https://read.oecd-ilibrary.org/view/?ref=132_132643-m91j2scsyh&title=Combatting-COVID-19-s-effect-on-children&_ga=2.115951494.1265123864.1638091357-303687672.1638091357.
7. Pearce-Duvet JM. The origin of human pathogens: evaluating the role of agriculture and domestic animals in the evolution of human disease. Biological Reviews. 2006;81(3):369-82.
8. Wolfe ND, Dunavan CP, Diamond J. Origins of major human infectious diseases. Nature. 2007;447(7142):279-83.
9. Baker RE, Mahmud AS, Miller IF, Rajeev M, Rasambainarivo F, Rice BL, et al. Infectious disease in an era of global change. Nature Reviews Microbiology. 2022;20(4):193-205.
10. Pybus OG, Tatem AJ, Lemey P. Virus evolution and transmission in an ever more connected world. Proceedings of the Royal Society B: Biological Sciences. 2015;282(1821):20142878.
11. Sami L. Understanding Emerging and Re-emerging Infectious Diseases. 2007.
12. Mackey TK, Liang BA, Cuomo R, Hafen R, Brouwer KC, Lee DE. Emerging and Reemerging Neglected Tropical Diseases: a Review of Key Characteristics, Risk Factors, and the Policy and Innovation Environment. Clinical Microbiology Reviews. 2014;27(4):949-79.
13. Woolhouse M, Gaunt E. Ecological Origins of Novel Human Pathogens. Critical Reviews in Microbiology. 2007;33(4):231-42.
14. Jones KE, Patel NG, Levy MA, Storeygard A, Balk D, Gittleman JL, et al. Global trends in emerging infectious diseases. Nature. 2008;451(7181):990-3.
15. Allen T, Murray KA, Zambrana-Torrelio C, Morse SS, Rondinini C, Di Marco M, et al. Global hotspots and correlates of emerging zoonotic diseases. Nature communications. 2017;8(1):1-10.
16. CDC. Zoonotic Diseases 2021 [Available from: https://www.cdc.gov/onehealth/basics/zoonotic-diseases.html.
17. Taylor LH, Latham SM, Woolhouse ME. Risk factors for human disease emergence. Philosophical Transactions of the Royal Society of London Series B: Biological Sciences. 2001;356(1411):983-9.
18. Fooks AR, Cliquet F, Finke S, Freuling C, Hemachudha T, Mani RS, et al. Rabies. Nature Reviews Disease Primers. 2017;3(1):17091.
19. Hampson K, Coudeville L, Lembo T, Sambo M, Kieffer A, Attlan M, et al. Estimating the Global Burden of Endemic Canine Rabies. PLOS Neglected Tropical Diseases. 2015;9(4):e0003709.
20. Davis BM, Rall GF, Schnell MJ. Everything you always wanted to know about rabies virus (but were afraid to ask). Annual review of virology. 2015;2:451-71.

21. Karesh WB, Dobson A, Lloyd-Smith JO, Lubroth J, Dixon MA, Bennett M, et al. Ecology of zoonoses: natural and unnatural histories. The Lancet. 2012;380(9857):1936-45.
22. Morens DM, Folkers GK, Fauci AS. The challenge of emerging and re-emerging infectious diseases. Nature. 2004;430(6996):242-9.
23. Sharp PM, Bailes E, Chaudhuri RR, Rodenburg CM, Santiago MO, Hahn BH. The origins of acquired immune deficiency syndrome viruses: where and when? Philosophical Transactions of the Royal Society of London Series B: Biological Sciences. 2001;356(1410):867-76.
24. Go YY, Balasuriya UB, Lee C-k. Zoonotic encephalitides caused by arboviruses: transmission and epidemiology of alphaviruses and flaviviruses. Clinical and experimental vaccine research. 2014;3(1):58-77.
25. Molaei G, Andreadis TG, Armstrong PM, Anderson JF, Vossbrinck CR. Host feeding patterns of Culex mosquitoes and West Nile virus transmission, northeastern United States. Emerging infectious diseases. 2006;12(3):468.
26. Bowen RA, Nemeth NM. Experimental infections with West Nile virus. Current opinion in infectious diseases. 2007;20(3):293-7.
27. Colpitts TM, Conway MJ, Montgomery RR, Fikrig E. West Nile Virus: biology, transmission, and human infection. Clinical microbiology reviews. 2012;25(4):635-48.
28. Weaver S, Barrett A. Nature reviews. Microbiology. Nat Rev Microbiol. 2004;2(10):789-801.
29. Baker RE, Mahmud AS, Miller IF, Rajeev M, Rasambainarivo F, Rice BL, et al. Infectious disease in an era of global change. Nature Reviews Microbiology. 2021.
30. UNEP. Preventing the next pandemic - Zoonotic diseases and how to break the chain of transmission 2020 [Available from: https://www.unep.org/resources/report/preventing-future-zoonotic-disease-outbreaks-protecting-environment-animals-and.
31. Worldometer. World Population by Year 2021 [Available from: https://www.worldometers.info/world-population/.
32. Lips KR, Brem F, Brenes R, Reeve JD, Alford RA, Voyles J, et al. Emerging infectious disease and the loss of biodiversity in a Neotropical amphibian community. Proceedings of the National Academy of Sciences. 2006;103(9):3165-70.
33. Gottdenker NL, Streicker DG, Faust CL, Carroll C. Anthropogenic land use change and infectious diseases: a review of the evidence. EcoHealth. 2014;11(4):619-32.
34. Levi T, Massey AL, Holt RD, Keesing F, Ostfeld RS, Peres CA. Does biodiversity protect humans against infectious disease? Comment. Ecology. 2016;97(2):536-42.
35. Olivero J, Fa JE, Real R, Márquez AL, Farfán MA, Vargas JM, et al. Recent loss of closed forests is associated with Ebola virus disease outbreaks. Scientific Reports. 2017;7(1).
36. Li Y, Kamara F, Zhou G, Puthiyakunnon S, Li C, Liu Y, et al. Urbanization Increases Aedes albopictus Larval Habitats and Accelerates Mosquito Development and Survivorship. PLoS Neglected Tropical Diseases. 2014;8(11):e3301.
37. Weaver SC. Urbanization and geographic expansion of zoonotic arboviral diseases: mechanisms and potential strategies for prevention. Trends in Microbiology. 2013;21(8):360-3.
38. Gubler DJ. Dengue, urbanization and globalization: the unholy trinity of the 21st century. Tropical medicine and health. 2011;39(4SUPPLEMENT):S3-S11.
39. Simmons CP, Farrar JJ, Van Vinh Chau N, Wills B. Dengue. New England Journal of Medicine. 2012;366(15):1423-32.
40. Powell JR, Gloria-Soria A, Kotsakiozi P. Recent History ofAedes aegypti: Vector Genomics and Epidemiology Records. BioScience. 2018;68(11):854-60.
41. Powell JR, Tabachnick WJ. History of domestication and spread of Aedes aegypti-a review. Memórias do Instituto Oswaldo Cruz. 2013;108:11-7.

42. Jones BA, Grace D, Kock R, Alonso S, Rushton J, Said MY, et al. Zoonosis emergence linked to agricultural intensification and environmental change. Proceedings of the National Academy of Sciences. 2013;110(21):8399-404.
43. Graham JP, Leibler JH, Price LB, Otte JM, Pfeiffer DU, Tiensin T, et al. The Animal-Human Interface and Infectious Disease in Industrial Food Animal Production: Rethinking Biosecurity and Biocontainment. Public Health Reports. 2008;123(3):282-99.
44. Leibler JH, Carone M, Silbergeld EK. Contribution of company affiliation and social contacts to risk estimates of between-farm transmission of avian influenza. PLoS One. 2010;5(3):e9888.
45. Epstein JH, Anthony SJ, Islam A, Kilpatrick AM, Ali Khan S, Balkey MD, et al. Nipah virus dynamics in bats and implications for spillover to humans. Proceedings of the National Academy of Sciences. 2020;117(46):29190-201.
46. Miyamoto M, Parid MM, Aini ZN, Michinaka T. Proximate and underlying causes of forest cover change in Peninsular Malaysia. Forest Policy and Economics. 2014;44:18-25.
47. Daszak P, Plowright R, Epstein JH, Pulliam J, Abdul Rahman S, Field HE, et al. The emergence of Nipah and Hendra virus: pathogen dynamics across a wildlife-livestock-human continuum. Disease Ecology: Community structure and pathogen dynamics. 2006:186-201.
48. Clayton BA. Nipah virus: transmission of a zoonotic paramyxovirus. Current Opinion in Virology. 2017;22:97-104.
49. Singh RK, Dhama K, Chakraborty S, Tiwari R, Natesan S, Khandia R, et al. Nipah virus: epidemiology, pathology, immunobiology and advances in diagnosis, vaccine designing and control strategies–a comprehensive review. Veterinary Quarterly. 2019;39(1):26-55.
50. Luby SP, Hossain MJ, Gurley ES, Ahmed B-N, Banu S, Khan SU, et al. Recurrent zoonotic transmission of Nipah virus into humans, Bangladesh, 2001–2007. Emerging infectious diseases. 2009;15(8):1229.
51. Ching PKG, de Los Reyes VC, Sucaldito MN, Tayag E, Columna-Vingno AB, Malbas Jr FF, et al. Outbreak of henipavirus infection, Philippines, 2014. Emerging infectious diseases. 2015;21(2):328.
52. Kurpiers LA, Schulte-Herbrüggen B, Ejotre I, Reeder DM. Bushmeat and emerging infectious diseases: lessons from Africa. Problematic Wildlife: Springer; 2016. p. 507-51.
53. Nasi R, Taber A, Van Vliet N. Empty forests, empty stomachs? Bushmeat and livelihoods in the Congo and Amazon Basins. International Forestry Review. 2011;13(3):355-68.
54. Coad L, Fa JE, Abernethy K, Van Vliet N, Santamaria C, Wilkie D, et al. Towards a sustainable, participatory and inclusive wild meat sector: CIFOR; 2019.
55. ICAO. The World of Air Transport in 2019. 2019.
56. Russell TW, Wu JT, Clifford S, Edmunds WJ, Kucharski AJ, Jit M. Effect of internationally imported cases on internal spread of COVID-19: a mathematical modelling study. The Lancet Public Health. 2021;6(1):e12-e20.
57. Candido DS, Claro IM, De Jesus JG, Souza WM, Moreira FR, Dellicour S, et al. Evolution and epidemic spread of SARS-CoV-2 in Brazil. Science. 2020;369(6508):1255-60.
58. Deng X, Gu W, Federman S, Du Plessis L, Pybus OG, Faria NR, et al. Genomic surveillance reveals multiple introductions of SARS-CoV-2 into Northern California. Science. 2020;369(6503):582-7.
59. Faria NR, Azevedo RDSDS, Kraemer MUG, Souza R, Cunha MS, Hill SC, et al. Zika virus in the Americas: Early epidemiological and genetic findings. Science. 2016;352(6283):345-9.
60. Semenza JC, Sudre B, Miniota J, Rossi M, Hu W, Kossowsky D, et al. International Dispersal of Dengue through Air Travel: Importation Risk for Europe. PLoS Neglected Tropical Diseases. 2014;8(12):e3278.
61. Kilpatrick AM. Globalization, land use, and the invasion of West Nile virus. Science. 2011;334(6054):323-7.

62. Grantz KH, Meredith HR, Cummings DAT, Metcalf CJE, Grenfell BT, Giles JR, et al. The use of mobile phone data to inform analysis of COVID-19 pandemic epidemiology. Nature Communications. 2020;11(1).
63. Oliver N, Lepri B, Sterly H, Lambiotte R, Deletaille S, De Nadai M, et al. Mobile phone data for informing public health actions across the COVID-19 pandemic life cycle. American Association for the Advancement of Science; 2020. p. eabc0764.
64. Naicker PR. The impact of climate change and other factors on zoonotic diseases. Archives of Clinical Microbiology. 2011;2(2).
65. UNEP. Alarming rise in global temperatures 2021 [Available from: https://www.unep.org/news-and-stories/story/alarming-rise-global-temperatures.
66. Reinhold J, Lazzari C, Lahondère C. Effects of the Environmental Temperature on Aedes aegypti and Aedes albopictus Mosquitoes: A Review. Insects. 2018;9(4):158.
67. Delatte H, Gimonneau G, Triboire A, Fontenille D. Influence of Temperature on Immature Development, Survival, Longevity, Fecundity, and Gonotrophic Cycles ofAedes albopictus, Vector of Chikungunya and Dengue in the Indian Ocean. Journal of Medical Entomology. 2009;46(1):33-41.
68. Reiter P. Climate change and mosquito-borne disease: knowing the horse before hitching the cart. Revue scientifique et technique (International Office of Epizootics). 2008;27(2):383-98.
69. Ryan SJ, Carlson CJ, Mordecai EA, Johnson LR. Global expansion and redistribution of Aedes-borne virus transmission risk with climate change. PLOS Neglected Tropical Diseases. 2019;13(3):e0007213.
70. Kraemer MU, Sinka ME, Duda KA, Mylne AQ, Shearer FM, Barker CM, et al. The global distribution of the arbovirus vectors Aedes aegypti and Ae. albopictus. elife. 2015;4:e08347.
71. Dolan PT, Whitfield ZJ, Andino R. Mechanisms and Concepts in RNA Virus Population Dynamics and Evolution. Annual Review of Virology. 2018;5(1):69-92.
72. Lauring AS. Within-host viral diversity: a window into viral evolution. Annual Review of Virology. 2020;7:63-81.
73. Lemey P, Salemi M, Vandamme A-M. The phylogenetic handbook: a practical approach to phylogenetic analysis and hypothesis testing: Cambridge University Press; 2009.
74. Holmes EC. The evolution and emergence of RNA viruses: Oxford University Press; 2009.
75. Acevedo A, Brodsky L, Andino R. Mutational and fitness landscapes of an RNA virus revealed through population sequencing. Nature. 2014;505(7485):686-90.
76. Sanjuán R, Moya A, Elena SF. The distribution of fitness effects caused by single-nucleotide substitutions in an RNA virus. Proceedings of the National Academy of Sciences. 2004;101(22):8396-401.
77. Sanjuán R. Mutational fitness effects in RNA and single-stranded DNA viruses: common patterns revealed by site-directed mutagenesis studies. Philosophical Transactions of the Royal Society B: Biological Sciences. 2010;365(1548):1975-82.
78. Elena SF, Carrasco P, Daròs JA, Sanjuán R. Mechanisms of genetic robustness in RNA viruses. EMBO reports. 2006;7(2):168-73.
79. Perelson AS, Neumann AU, Markowitz M, Leonard JM, Ho DD. HIV-1 dynamics in vivo: virion clearance rate, infected cell life-span, and viral generation time. Science. 1996;271(5255):1582-6.
80. Tsetsarkin KA, Vanlandingham DL, McGee CE, Higgs S. A Single Mutation in Chikungunya Virus Affects Vector Specificity and Epidemic Potential. PLoS Pathogens. 2007;3(12):e201.
81. Urbanowicz RA, McClure CP, Sakuntabhai A, Sall AA, Kobinger G, Müller MA, et al. Human adaptation of Ebola virus during the West African outbreak. Cell. 2016;167(4):1079-87. e5.
82. Harvey WT, Carabelli AM, Jackson B, Gupta RK, Thomson EC, Harrison EM, et al. SARS-CoV-2 variants, spike mutations and immune escape. Nature Reviews Microbiology. 2021;19(7):409-24.
83. Lauring AS, Hodcroft EB. Genetic Variants of SARS-CoV-2—What Do They Mean? JAMA. 2021;325(6):529.
84. Hou YJ, Chiba S, Halfmann P, Ehre C, Kuroda M, Dinnon KH, et al. SARS-CoV-2 D614G variant exhibits efficient replication ex vivo and transmission in vivo. Science. 2020;370(6523):1464-8.

85. Volz E, Hill V, McCrone JT, Price A, Jorgensen D, O'Toole Á, et al. Evaluating the effects of SARS-CoV-2 spike mutation D614G on transmissibility and pathogenicity. Cell. 2021;184(1):64-75. e11.
86. Plante JA, Liu Y, Liu J, Xia H, Johnson BA, Lokugamage KG, et al. Spike mutation D614G alters SARS-CoV-2 fitness. Nature. 2021;592(7852):116-21.
87. O'Toole Á, Hill V, Pybus OG, Watts A, Bogoch II, Khan K, et al. Tracking the international spread of SARS-CoV-2 lineages B. 1.1. 7 and B. 1.351/501Y-V2 with grinch. Wellcome open research. 2021;6.
88. Starr TN, Greaney AJ, Hilton SK, Ellis D, Crawford KH, Dingens AS, et al. Deep mutational scanning of SARS-CoV-2 receptor binding domain reveals constraints on folding and ACE2 binding. Cell. 2020;182(5):1295-310. e20.
89. Gu H, Chen Q, Yang G, He L, Fan H, Deng Y-Q, et al. Adaptation of SARS-CoV-2 in BALB/c mice for testing vaccine efficacy. Science. 2020;369(6511):1603-7.
90. Montagutelli X, Prot M, Levillayer L, Salazar EB, Jouvion G, Conquet L, et al. The B1.351 and P.1 variants extend SARS-CoV-2 host range to mice. 2021.
91. Gmyl AP, Belousov EV, Maslova SV, Khitrina EV, Chetverin AB, Agol VI. Nonreplicative RNA recombination in poliovirus. Journal of virology. 1999;73(11):8958-65.
92. Simon-Loriere E, Holmes EC. Why do RNA viruses recombine? Nature Reviews Microbiology. 2011;9(8):617-26.
93. Kyriakopoulou Z, Pliaka V, Amoutzias GD, Markoulatos P. Recombination among human non-polio enteroviruses: implications for epidemiology and evolution. Virus Genes. 2015;50(2):177-88.
94. Muslin C, Mac Kain A, Bessaud M, Blondel B, Delpeyroux F. Recombination in enteroviruses, a multi-step modular evolutionary process. Viruses. 2019;11(9):859.
95. Lefkowitz EJ, Dempsey DM, Hendrickson RC, Orton RJ, Siddell SG, Smith DB. Virus taxonomy: the database of the International Committee on Taxonomy of Viruses (ICTV). Nucleic acids research. 2018;46(D1):D708-D17.
96. Oberste MS, Maher K, Pallansch MA. Evidence for frequent recombination within species human enterovirus B based on complete genomic sequences of all thirty-seven serotypes. Journal of virology. 2004;78(2):855-67.
97. Jackson B, Boni MF, Bull MJ, Colleran A, Colquhoun RM, Darby AC, et al. Generation and transmission of interlineage recombinants in the SARS-CoV-2 pandemic. Cell. 2021;184(20):5179-88. e8.
98. Jackwood MW, Boynton TO, Hilt DA, McKinley ET, Kissinger JC, Paterson AH, et al. Emergence of a group 3 coronavirus through recombination. Virology. 2010;398(1):98-108.
99. McDonald SM, Nelson MI, Turner PE, Patton JT. Reassortment in segmented RNA viruses: mechanisms and outcomes. Nature Reviews Microbiology. 2016;14(7):448-60.
100. Gerber M, Isel C, Moules V, Marquet R. Selective packaging of the influenza A genome and consequences for genetic reassortment. Trends in microbiology. 2014;22(8):446-55.
101. Westgeest KB, Russell CA, Lin X, Spronken MI, Bestebroer TM, Bahl J, et al. Genomewide analysis of reassortment and evolution of human influenza A (H3N2) viruses circulating between 1968 and 2011. Journal of virology. 2014;88(5):2844-57.
102. Smith GJ, Bahl J, Vijaykrishna D, Zhang J, Poon LL, Chen H, et al. Dating the emergence of pandemic influenza viruses. Proceedings of the National Academy of Sciences. 2009;106(28):11709-12.
103. Garten RJ, Davis CT, Russell CA, Shu B, Lindstrom S, Balish A, et al. Antigenic and genetic characteristics of swine-origin 2009 A (H1N1) influenza viruses circulating in humans. science. 2009;325(5937):197-201.
104. Ma W, Kahn RE, Richt JA. The pig as a mixing vessel for influenza viruses: human and veterinary implications. Journal of molecular and genetic medicine: an international journal of biomedical research. 2009;3(1):158.
105. Grubaugh ND, Ladner JT, Lemey P, Pybus OG, Rambaut A, Holmes EC, et al. Tracking virus outbreaks in the twenty-first century. Nat Microbiol. 2019;4(1):10-9.

106. Adam D. A guide to R--the pandemic's misunderstood metric. Nature. 2020;583(7816):346-9.
107. Brookmeyer R. Incubation period of infectious diseases. Encyclopedia of biostatistics. 2005;4.
108. Kraemer MUG, Pybus OG, Fraser C, Cauchemez S, Rambaut A, Cowling BJ. Monitoring key epidemiological parameters of SARS-CoV-2 transmission. Nature Medicine. 2021;27(11):1854-5.
109. Pollett S, Fauver J, Maljkovic Berry I, Melendrez M, Morrison A, Gillis L, et al. Genomic epidemiology as a public health tool to combat mosquito-borne virus outbreaks. The Journal of Infectious Diseases. 2020;221(Supplement_3):S308-S18.
110. Hill V, Ruis C, Bajaj S, Pybus OG, Kraemer MU. Progress and challenges in virus genomic epidemiology. Trends in parasitology. 2021;37(12):1038-49.
111. Houldcroft CJ, Beale MA, Breuer J. Clinical and biological insights from viral genome sequencing. Nature Reviews Microbiology. 2017;15(3):183-92.
112. Chiu CY, Miller SA. Clinical metagenomics. Nature Reviews Genetics. 2019;20(6):341-55.
113. Briese T, Paweska JT, McMullan LK, Hutchison SK, Street C, Palacios G, et al. Genetic detection and characterization of Lujo virus, a new hemorrhagic fever–associated arenavirus from southern Africa. PLoS pathogens. 2009;5(5):e1000455.
114. Stelzl E, Haas B, Bauer B, Zhang S, Fiss EH, Hillman G, et al. First identification of a recombinant form of hepatitis C virus in Austrian patients by full-genome next generation sequencing. PLOS ONE. 2017;12(7):e0181273.
115. Gardy JL, Loman NJ. Towards a genomics-informed, real-time, global pathogen surveillance system. Nature Reviews Genetics. 2018;19(1):9-20.
116. Mongan AE, Tuda JSB, Runtuwene LR. Portable sequencer in the fight against infectious disease. Journal of Human Genetics. 2020;65(1):35-40.
117. Hoenen T, Groseth A, Rosenke K, Fischer RJ, Hoenen A, Judson SD, et al. Nanopore sequencing as a rapidly deployable Ebola outbreak tool. Emerging infectious diseases. 2016;22(2):331.
118. Quick J, Loman NJ, Duraffour S, Simpson JT, Severi E, Cowley L, et al. Real-time, portable genome sequencing for Ebola surveillance. Nature. 2016;530(7589):228-32.
119. Faria N, Sabino E, Nunes M, Alcantara L, Loman N, Pybus O. Mobile real-time surveillance of Zika virus in Brazil. Genome Med 8: 97. 2016.
120. Castro-Wallace SL, Chiu CY, John KK, Stahl SE, Rubins KH, McIntyre AB, et al. Nanopore DNA sequencing and genome assembly on the International Space Station. Scientific reports. 2017;7(1):1-12.
121. Grenfell BT, Pybus OG, Gog JR, Wood JLN, Daly JM, Mumford JA, et al. Unifying the Epidemiological and Evolutionary Dynamics of Pathogens. Science. 2004;303(5656):327-32.
122. Baele G, Suchard MA, Rambaut A, Lemey P. Emerging concepts of data integration in pathogen phylodynamics. Systematic biology. 2017;66(1):e47-e65.
123. Kristian, B, Christian, Sealfon R, Aaron, Lina, et al. Clinical Sequencing Uncovers Origins and Evolution of Lassa Virus. Cell. 2015;162(4):738-50.
124. Gire SK, Goba A, Andersen KG, Sealfon RSG, Park DJ, Kanneh L, et al. Genomic surveillance elucidates Ebola virus origin and transmission during the 2014 outbreak. Science. 2014;345(6202):1369-72.
125. Baum D. Reading a phylogenetic tree: the meaning of monophyletic groups. Nature Education. 2008;1(1):190.
126. Kapli P, Yang Z, Telford MJ. Phylogenetic tree building in the genomic age. Nature Reviews Genetics. 2020;21(7):428-44.
127. Huelsenbeck JP, Ronquist F, Nielsen R, Bollback JP. Bayesian inference of phylogeny and its impact on evolutionary biology. science. 2001;294(5550):2310-4.
128. community B. Summarizing Trees 2017 [Available from: https://beast.community/summarizing_trees.

129. Ronquist F, Teslenko M, Van Der Mark P, Ayres DL, Darling A, Höhna S, et al. MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space. Systematic biology. 2012;61(3):539-42.
130. Drummond AJ, Rambaut A. BEAST: Bayesian evolutionary analysis by sampling trees. BMC evolutionary biology. 2007;7(1):1-8.
131. Bouckaert R, Heled J, Kühnert D, Vaughan T, Wu C-H, Xie D, et al. BEAST 2: a software platform for Bayesian evolutionary analysis. PLoS computational biology. 2014;10(4):e1003537.
132. Ho S. The molecular clock and estimating species divergence. Nature Education. 2008;1(1):1-2.
133. Yoder AD, Yang Z. Estimation of primate speciation dates using local molecular clocks. Molecular biology and evolution. 2000;17(7):1081-90.
134. Li WLS, Drummond AJ. Model averaging and Bayes factor calculation of relaxed molecular clocks in Bayesian phylogenetics. Molecular biology and evolution. 2012;29(2):751-61.
135. community B. Molecular Clocks 2017 [Available from: https://beast.community/clocks#fixed-local-clock.
136. Duchene S, Featherstone L, Haritopoulou-Sinanidou M, Rambaut A, Lemey P, Baele G. Temporal signal and the phylodynamic threshold of SARS-CoV-2. Virus evolution. 2020;6(2):veaa061.
137. Villabona-Arenas CJ, Hanage WP, Tully DC. Phylogenetic interpretation during outbreaks requires caution. Nature Microbiology. 2020;5(7):876-7.
138. Schierup MH, Hein J. Recombination and the molecular clock. Molecular biology and evolution. 2000;17(10):1578-9.
139. Volz EM, Koelle K, Bedford T. Viral Phylodynamics. PLoS Computational Biology. 2013;9(3):e1002947.
140. Stadler T, Vaughan TG, Gavryushkin A, Guindon S, Kühnert D, Leventhal GE, et al. How well can the exponential-growth coalescent approximate constant-rate birth–death population dynamics? Proceedings of the Royal Society B: Biological Sciences. 2015;282(1806):20150420.
141. Stadler T, Kouyos R, von Wyl V, Yerly S, Böni J, Bürgisser P, et al. Estimating the basic reproductive number from viral sequence data. Molecular biology and evolution. 2012;29(1):347-57.
142. Volz EM, Kosakovsky Pond SL, Ward MJ, Leigh Brown AJ, Frost SDW. Phylodynamics of Infectious Disease Epidemics. Genetics. 2009;183(4):1421-30.
143. Stadler T, Kuhnert D, Bonhoeffer S, Drummond AJ. Birth-death skyline plot reveals temporal changes of epidemic spread in HIV and hepatitis C virus (HCV). Proceedings of the National Academy of Sciences. 2013;110(1):228-33.
144. Ayoub HH, Chemaitelly H, Kouyoumjian SP, Abu-Raddad LJ. Characterizing the historical role of parenteral antischistosomal therapy in hepatitis C virus transmission in Egypt. International Journal of Epidemiology. 2020;49(3):798-809.
145. Stadler T, Kühnert D, Rasmussen DA, du Plessis L. Insights into the early epidemic spread of Ebola in Sierra Leone provided by viral sequence data. PLoS currents. 2014;6.
146. Faria NR, Suchard MA, Rambaut A, Lemey P. Toward a quantitative understanding of viral phylogeography. Current opinion in virology. 2011;1(5):423-9.
147. Pybus OG, Suchard MA, Lemey P, Bernardin FJ, Rambaut A, Crawford FW, et al. Unifying the spatial epidemiology and molecular evolution of emerging epidemics. Proceedings of the National Academy of Sciences. 2012;109(37):15066-71.
148. Lemey P, Rambaut A, Bedford T, Faria N, Bielejec F, Baele G, et al. Unifying viral genetics and human transportation data to predict the global transmission dynamics of human influenza H3N2. PLoS pathogens. 2014;10(2):e1003932.
149. Ferreira MA, Suchard MA. Bayesian analysis of elapsed times in continuous-time Markov chains. Canadian Journal of Statistics. 2008;36(3):355-68.

150. Lemey P, Rambaut A, Drummond AJ, Suchard MA. Bayesian Phylogeography Finds Its Roots. PLoS Computational Biology. 2009;5(9):e1000520.
151. Lemey P, Rambaut A, Welch JJ, Suchard MA. Phylogeography takes a relaxed random walk in continuous space and time. Molecular biology and evolution. 2010;27(8):1877-85.
152. Dellicour S, Gill MS, Faria NR, Rambaut A, Pybus OG, Suchard MA, et al. Relax, keep walking-a practical guide to continuous phylogeographic inference with BEAST. Mol Biol Evol. 2021.
153. Rezza G, Nicoletti L, Angelini R, Romi R, Finarelli A, Panning M, et al. Infection with chikungunya virus in Italy: an outbreak in a temperate region. The Lancet. 2007;370(9602):1840-6.
154. Faye O, de Lourdes Monteiro M, Vrancken B, Prot M, Lequime S, Diarra M, et al. Genomic epidemiology of 2015–2016 Zika virus outbreak in Cape Verde. Emerging infectious diseases. 2020;26(6):1084.
155. Minin VN, Suchard MA. Fast, accurate and simulation-free stochastic mapping. Philosophical Transactions of the Royal Society B: Biological Sciences. 2008;363(1512):3985-95.
156. Lemey P, Hong SL, Hill V, Baele G, Poletto C, Colizza V, et al. Accommodating individual travel history and unsampled diversity in Bayesian phylogeographic inference of SARS-CoV-2. Nature Communications. 2020;11(1).
157. Faria NR, Kraemer MU, Hill S, De Jesus JG, Aguiar R, Iani FC, et al. Genomic and epidemiological monitoring of yellow fever virus transmission potential. Science. 2018;361(6405):894-9.
158. Kalkauskas A, Perron U, Sun Y, Goldman N, Baele G, Guindon S, et al. Sampling bias and model choice in continuous phylogeography: Getting lost on a random walk. PLOS Computational Biology. 2021;17(1):e1008561.
159. Suchard MA, Lemey P, Baele G, Ayres DL, Drummond AJ, Rambaut A. Bayesian phylogenetic and phylodynamic data integration using BEAST 1.10. Virus Evolution. 2018;4(1).
160. Lemey P, Rambaut A, Bedford T, Faria N, Bielejec F, Baele G, et al. Unifying Viral Genetics and Human Transportation Data to Predict the Global Transmission Dynamics of Human Influenza H3N2. PLoS Pathogens. 2014;10(2):e1003932.
161. Dudas G, Carvalho LM, Bedford T, Tatem AJ, Baele G, Faria NR, et al. Virus genomes reveal factors that spread and sustained the Ebola epidemic. Nature. 2017;544(7650):309-15.
162. Dellicour S, Rose R, Faria NR, Lemey P, Pybus OG. SERAPHIM: studying environmental rasters and phylogenetically informed movements. Bioinformatics. 2016;32(20):3204-6.
163. Dellicour S, Lequime S, Vrancken B, Gill MS, Bastide P, Gangavarapu K, et al. Epidemiological hypothesis testing using a phylogeographic and phylodynamic framework. Nature Communications. 2020;11(1).
164. Ladner JT, Grubaugh ND, Pybus OG, Andersen KG. Precision epidemiology for infectious disease control. Nature Medicine. 2019;25(2):206-11.
165. Mate SE, Kugelman JR, Nyenswah TG, Ladner JT, Wiley MR, Cordier-Lassalle T, et al. Molecular evidence of sexual transmission of Ebola virus. New England Journal of Medicine. 2015;373(25):2448-54.
166. Butler D. What first case of sexually transmitted Ebola means for public health. Nature News. 2015;10.
167. Keita AK, Koundouno FR, Faye M, Düx A, Hinzmann J, Diallo H, et al. Resurgence of Ebola virus in 2021 in Guinea suggests a new paradigm for outbreaks. Nature. 2021;597(7877):539-43.
168. Grubaugh ND, Ladner JT, Kraemer MU, Dudas G, Tan AL, Gangavarapu K, et al. Genomic epidemiology reveals multiple introductions of Zika virus into the United States. Nature. 2017;546(7658):401-5.
169. Ampofo WK, Azziz-Baumgartner E, Bashir U, Cox NJ, Fasce R, Giovanni M, et al. Strengthening the influenza vaccine virus selection and development process: Report of the 3rd WHO Informal Consultation for Improving Influenza Vaccine Virus Selection held at WHO headquarters, Geneva, Switzerland, 1–3 April 2014. Vaccine. 2015;33(36):4368-82.

170. Pfizer. Pfizer and BioNTech Provide Update on Omicron Variant 2021 [Available from: https://www.pfizer.com/news/press-release/press-release-detail/pfizer-and-biontech-provide-update-omicron-variant.
171. Matranga CB, Andersen KG, Winnicki S, Busby M, Gladden AD, Tewhey R, et al. Enhanced methods for unbiased deep sequencing of Lassa and Ebola RNA viruses from clinical and biological samples. Genome biology. 2014;15(11):1-12.
172. Quick J, Grubaugh ND, Pullan ST, Claro IM, Smith AD, Gangavarapu K, et al. Multiplex PCR method for MinION and Illumina sequencing of Zika and other virus genomes directly from clinical samples. Nature Protocols. 2017;12(6):1261-76.
173. Pickett BE, Greer DS, Zhang Y, Stewart L, Zhou L, Sun G, et al. Virus pathogen database and analysis resource (ViPR): a comprehensive bioinformatics database and analysis resource for the coronavirus research community. Viruses. 2012;4(11):3209-26.
174. Wilson MR, Sample HA, Zorn KC, Arevalo S, Yu G, Neuhaus J, et al. Clinical metagenomic sequencing for diagnosis of meningitis and encephalitis. New England Journal of Medicine. 2019;380(24):2327-40.
175. Saha S, Ramesh A, Kalantar K, Malaker R, Hasanuzzaman M, Khan LM, et al. Unbiased metagenomic sequencing for pediatric meningitis in Bangladesh reveals neuroinvasive chikungunya virus outbreak and other unrealized pathogens. MBio. 2019;10(6):e02877-19.
176. Wilson MR, Zimmermann LL, Crawford ED, Sample HA, Soni PR, Baker AN, et al. Acute West Nile Virus Meningoencephalitis Diagnosed Via Metagenomic Deep Sequencing of Cerebrospinal Fluid in a Renal Transplant Patient. American Journal of Transplantation. 2017;17(3):803-8.
177. Wilson MR, O'Donovan BD, Gelfand JM, Sample HA, Chow FC, Betjemann JP, et al. Chronic Meningitis Investigated via Metagenomic Next-Generation Sequencing. JAMA Neurology. 2018;75(8):947.
178. James SL, Abate D, Abate KH, Abay SM, Abbafati C, Abbasi N, et al. Global, regional, and national incidence, prevalence, and years lived with disability for 354 diseases and injuries for 195 countries and territories, 1990–2017: a systematic analysis for the Global Burden of Disease Study 2017. The Lancet. 2018;392(10159):1789-858.
179. McGill F, Griffiths MJ, Bonnett LJ, Geretti AM, Michael BD, Beeching NJ, et al. Incidence, aetiology, and sequelae of viral meningitis in UK adults: a multicentre prospective observational cohort study. The Lancet Infectious Diseases. 2018;18(9):992-1003.
180. Hasbun R, Rosenthal N, Balada-Llasat JM, Chung J, Duff S, Bozzette S, et al. Epidemiology of Meningitis and Encephalitis in the United States, 2011–2014. Clinical Infectious Diseases. 2017;65(3):359-63.
181. Shukla B, Aguilera EA, Salazar L, Wootton SH, Kaewpoowat Q, Hasbun R. Aseptic meningitis in adults and children: Diagnostic and management challenges. Journal of Clinical Virology. 2017;94:110-4.
182. John CC, Carabin H, Montano SM, Bangirana P, Zunt JR, Peterson PK. Global research priorities for infections that affect the nervous system. Nature. 2015;527(7578):S178-S86.
183. McGill F, Griffiths MJ, Solomon T. Viral meningitis: current issues in diagnosis and treatment. Current opinion in infectious diseases. 2017;30(2):248-56.
184. Kohil A, Jemmieh S, Smatti MK, Yassine HM. Viral meningitis: an overview. Archives of Virology. 2021;166(2):335-45.
185. Gailson T, Vohra V, Saini AG, Bhatia V. Mumps infection with meningoencephalitis and cerebellitis. BMJ case reports. 2021;14(11).
186. Ramachandran PS, Wilson MR. Metagenomics for neurological infections—expanding our imagination. Nature Reviews Neurology. 2020;16(10):547-56.
187. Casas-Alba D, De Sevilla MF, Valero-Rello A, Fortuny C, García-García JJ, Ortez C, et al. Outbreak of brainstem encephalitis associated with enterovirus-A71 in Catalonia, Spain (2016): a clinical observational

study in a children's reference centre in Catalonia. Clinical Microbiology and Infection. 2017;23(11):874-81.
188. Leal Barceló AM, Carrascosa García P, Rincón López EM, Herrero M, Navarro ML, editors. Brote de infección por enterovirus causantes de afectación neurológica grave en un hospital terciario. Anales de Pediatría; 2018.
189. Cabrerizo M, García-Iñiguez JP, Munell F, Amado A, Madurga-Revilla P, Rodrigo C, et al. First cases of severe flaccid paralysis associated with enterovirus D68 infection in Spain, 2015–2016. The Pediatric infectious disease journal. 2017;36(12):1214-6.
190. III IdSC. Vigilancia de la Parálisis Flácida Aguda y Vigilancia de Enterovirus, Informe año 2018 2018 [Available from: https://www.isciii.es/QueHacemos/Servicios/VigilanciaSaludPublicaRENAVE/EnfermedadesTransmisibles/Documents/archivos%20A-Z/POLIO/Resultados_Vigilancia_Polio/Informes_Anuales_Polio/Informe_PFA_y_Enterovirus_2018.pdf.
191. III IdSC. Vigilancia de la Parálisis Flácida Aguda y Vigilancia de Enterovirus, Informe año 2020 2020 [Available from: https://www.isciii.es/QueHacemos/Servicios/VigilanciaSaludPublicaRENAVE/EnfermedadesTransmisibles/Documents/archivos%20A-Z/POLIO/Resultados_Vigilancia_Polio/Informes_Anuales_Polio/Informe_PFA_EV_2020_web.pdf.
192. Antón AIN, de Ory Manchón F, Fariñas MPS-S, Narváez LF, Cámara MIG, Mari JMN, et al. Microbiological diagnosis of emerging arboviral and rodent borne diseases. Enfermedades infecciosas y microbiologia clinica. 2015;33(3):197-205.
193. Centro de Coordinación de Alertas y Emergencias sanitarias MdS, Consumo y Bienestar Social. Informe de situación y evaluación del riesgo de enfermedad por flebovirus trasmitidos por flebotomos en España 2019 [Available from: https://www.sanidad.gob.es/ca/profesionales/saludPublica/ccayes/analisisituacion/doc/ER_Flebovirus.pdf.
194. Palacios G, Oberste M. Enteroviruses as agents of emerging infectious diseases. Journal of neurovirology. 2005;11(5):424-33.
195. Simmonds P, Gorbalenya A, Harvala H, Hovi T, Knowles N, Lindberg AM, et al. Recommendations for the nomenclature of enteroviruses and rhinoviruses. Archives of virology. 2020;165(3):793-7.
196. ICTV. Genus: Enterovirus [Available from: https://talk.ictvonline.org/ictv-reports/ictv_online_report/positive-sense-rna-viruses/w/picornaviridae/681/genus-enterovirus.
197. Harvala H, Broberg E, Benschop K, Berginc N, Ladhani S, Susi P, et al. Recommendations for enterovirus diagnostics and characterisation within and beyond Europe. Journal of clinical virology. 2018;101:11-7.
198. Pons-Salort M, Parker EP, Grassly NC. The epidemiology of non-polio enteroviruses: recent advances and outstanding questions. Current opinion in infectious diseases. 2015;28(5):479.
199. Bubba L, Broberg E, Jasir A, Simmonds P, Harvala H. Enterovirus study collaborators. Circulation of non-polio enteroviruses in 24 EU and EEA countries between 2015 and 2017: a retrospective surveillance study. Lancet Infect Dis. 2020;20:350-61.
200. Trallero G, Avellon A, Otero A, De Miguel T, Pérez C, Rabella N, et al. Enteroviruses in Spain over the decade 1998–2007: virological and epidemiological studies. Journal of clinical virology. 2010;47(2):170-6.
201. III IdSC. Polio and non-polio EV surveillance [Available from: https://www.isciii.es/QueHacemos/Servicios/VigilanciaSaludPublicaRENAVE/EnfermedadesTransmisibles/Paginas/Resultados_Vigilancia_Polio.aspx.

202. Midgley CM, Watson JT, Nix WA, Curns AT, Rogers SL, Brown BA, et al. Severe respiratory illness associated with a nationwide outbreak of enterovirus D68 in the USA (2014): a descriptive epidemiological investigation. The Lancet Respiratory medicine. 2015;3(11):879-87.
203. CDC. Non-Polio Enteroviruses: Outbreaks and Surveillance 2020 [Available from: https://www.cdc.gov/non-polio-enterovirus/outbreaks-surveillance.html.
204. Benschop KS, Albert J, Anton A, Andrés C, Aranzamendi M, Armannsdóttir B, et al. Re-emergence of enterovirus D68 in Europe after easing the COVID-19 lockdown, September 2021. Eurosurveillance. 2021;26(45):2100998.
205. Fernandez-Garcia MD, Majumdar M, Kebe O, Ndiaye K, Martin J. Identification and whole-genome characterization of a recombinant Enterovirus B69 isolated from a patient with Acute Flaccid Paralysis in Niger, 2015. Scientific reports. 2018;8(1):1-8.
206. Midgley SE, Nielsen AG, Trebbien R, Poulsen MW, Andersen PH, Fischer TK. Co-circulation of multiple subtypes of enterovirus A71 (EV-A71) genotype C, including novel recombinants characterised by use of whole genome sequencing (WGS), Denmark 2016. Eurosurveillance. 2017;22(26):30565.
207. Midgley SE, Benschop K, Dyrdak R, Mirand A, Bailly J-L, Bierbaum S, et al. Co-circulation of multiple enterovirus D68 subclades, including a novel B3 cluster, across Europe in a season of expected low prevalence, 2019/20. Eurosurveillance. 2020;25(2):1900749.
208. Baertl S, Pietsch C, Maier M, Hönemann M, Bergs S, Liebert UG. Enteroviruses in Respiratory Samples from Paediatric Patients of a Tertiary Care Hospital in Germany. Viruses. 2021;13(5):882.
209. Oprisan G, Combiescu M, Guillot S, Caro V, Combiescu A, Delpeyroux F, et al. Natural genetic recombination between co-circulating heterotypic enteroviruses. Journal of general virology. 2002;83(9):2193-200.
210. Bouslama L, Nasri D, Chollet L, Belguith K, Bourlet T, Aouni M, et al. Natural recombination event within the capsid genomic region leading to a chimeric strain of human enterovirus B. Journal of virology. 2007;81(17):8944-52.
211. Nikolaidis M, Mimouli K, Kyriakopoulou Z, Tsimpidis M, Tsakogiannis D, Markoulatos P, et al. Large-scale genomic analysis reveals recurrent patterns of intertypic recombination in human enteroviruses. Virology. 2019;526:72-80.
212. Lukashev AN, Lashkevich VA, Ivanova OE, Koroleva GA, Hinkkanen AE, Ilonen J. Recombination in circulating Human enterovirus B: independent evolution of structural and non-structural genome regions. Journal of general virology. 2005;86(12):3281-90.
213. Fischer TK, Simmonds P, Harvala H. The importance of enterovirus surveillance in a post-polio world. The Lancet Infectious Diseases. 2021.
214. Bessaud M, Delpeyroux F. Enteroviruses-the famous unknowns. The Lancet Infectious Diseases. 2020;20(3):268-9.
215. Harvala H, Benschop KS, Berginc N, Midgley S, Wolthers K, Simmonds P, et al. European non-polio enterovirus network: introduction of hospital-based surveillance network to understand the true disease burden of non-polio enterovirus and parechovirus infections in Europe. Microorganisms. 2021;9(9):1827.
216. Ayhan N, Charrel RN. An update on Toscana virus distribution, genetics, medical and diagnostic aspects. Clinical Microbiology and Infection. 2020;26(8):1017-23.
217. Charrel RN. Emergence of Toscana virus in the mediterranean area. World Journal of Virology. 2012;1(5):135.
218. Bichaud L, Dachraoui K, Piorkowski G, Chelbi I, Moureau G, Cherni S, et al. Toscana virus isolated from sandflies, Tunisia. Emerging infectious diseases. 2013;19(2):322.
219. Es-Sette N, Nourlil J, Hamdi S, Mellouki F, Lemrani M. First detection of Toscana virus RNA from sand flies in the genus Phlebotomus (Diptera: Phlebotomidae) naturally infected in Morocco. Journal of medical entomology. 2014;49(6):1507-9.

220. Punda-Polić V, Mohar B, Duh D, Bradarić N, Korva M, Fajs L, et al. Evidence of an autochthonous Toscana virus strain in Croatia. Journal of clinical virology. 2012;55(1):4-7.
221. Papa A, Paraforou T, Papakonstantinou I, Pagdatoglou K, Kontana A, Koukoubani T. Severe Encephalitis Caused by Toscana Virus, Greece. Emerging Infectious Diseases. 2014;20(8):1417-9.
222. Papa A, Kontana A, Tsergouli K. Phlebovirus infections in Greece. Journal of Medical Virology. 2015;87(7):1072-6.
223. Baldelli F, Ciufolini MG, Francisci D, Marchi A, Venturi G, Fiorentini C, et al. Unusual presentation of life-threatening Toscana virus meningoencephalitis. Clinical infectious diseases. 2004;38(4):515-20.
224. Bartels S, de Boni L, Kretzschmar HA, Heckmann JG. Lethal encephalitis caused by the Toscana virus in an elderly patient. Journal of neurology. 2012;259(1):175-7.
225. Suresh S, Rawlinson WD, Andrews PI, Stelzer-Braid S. Global epidemiology of nonpolio enteroviruses causing severe neurological complications: A systematic review and meta-analysis. Reviews in medical virology. 2020;30(1):e2082.
226. Puenpa J, Wanlapakorn N, Vongpunsawad S, Poovorawan Y. The history of enterovirus A71 outbreaks and molecular epidemiology in the Asia-Pacific region. Journal of biomedical science. 2019;26(1):1-11.
227. Gu W, Miller S, Chiu CY. Clinical metagenomic next-generation sequencing for pathogen detection. Annual Review of Pathology: Mechanisms of Disease. 2019;14:319-38.
228. Zinter MS, Mayday MY, Ryckman KK, Jelliffe-Pawlowski LL, Derisi JL. Towards precision quantification of contamination in metagenomic sequencing experiments. Microbiome. 2019;7(1).
229. Zinter M, Mayday M, Ryckman K, Jelliffe-Pawlowski L, DeRisi J. Towards precision quantification of contamination in metagenomic sequencing experiments. Microbiome. 2019;7(1):1-5.
230. Illumina. Effects of Index Misassignment on Multiplexing and Downstream Analysis 2018 [Available from: https://www.illumina.com/content/dam/illumina-marketing/documents/products/whitepapers/index-hopping-white-paper-770-2017-004.pdf.
231. Lindon JC, Tranter GE, Koppenaal D. Encyclopedia of spectroscopy and spectrometry: Academic Press; 2016.
232. Verena. VIRION database [Available from: https://viralemergence.github.io/virion/.
233. Bigot T, Temmam S, Pérot P, Eloit M. RVDB-prot, a reference viral protein database and its HMM profiles. F1000Research. 2020;8:530.
234. Saha S, Ramesh A, Kalantar K, Malaker R, Hasanuzzaman M, Khan LM, et al. Unbiased metagenomic sequencing for pediatric meningitis in Bangladesh reveals neuroinvasive Chikungunya virus outbreak and other unrealized pathogens. Mbio. 2019;10(6).
235. Higgs S, Vanlandingham D. Chikungunya virus and its mosquito vectors. Vector-Borne and Zoonotic Diseases. 2015;15(4):231-40.
236. Burt FJ, Chen W, Miner JJ, Lenschow DJ, Merits A, Schnettler E, et al. Chikungunya virus: an update on the biology and pathogenesis of this emerging pathogen. The Lancet Infectious Diseases. 2017;17(4):e107-e17.
237. Weaver SC, Lecuit M. Chikungunya virus and the global spread of a mosquito-borne disease. New England Journal of Medicine. 2015;372(13):1231-9.
238. Voss JE, Vaney M-C, Duquerroy S, Vonrhein C, Girard-Blanc C, Crublet E, et al. Glycoprotein organization of Chikungunya virus particles revealed by X-ray crystallography. Nature. 2010;468(7324):709-12.
239. Schwartz O, Albert ML. Biology and pathogenesis of chikungunya virus. Nature Reviews Microbiology. 2010;8(7):491-500.
240. de Groot RJ, Hardy WR, Shirako Y, Strauss JH. Cleavage-site preferences of Sindbis virus polyproteins containing the non-structural proteinase. Evidence for temporal regulation of polyprotein processing in vivo. The EMBO journal. 1990;9(8):2631-8.

241. Shirako Y, Strauss JH. Regulation of Sindbis virus RNA replication: uncleaved P123 and nsP4 function in minus-strand RNA synthesis, whereas cleaved products from P123 are required for efficient plus-strand RNA synthesis. Journal of virology. 1994;68(3):1874-85.
242. Rupp JC, Sokoloski KJ, Gebhart NN, Hardy RW. Alphavirus RNA synthesis and non-structural protein functions. The Journal of general virology. 2015;96(Pt 9):2483.
243. Robinson MC. An epidemic of virus disease in Southern Province, Tanganyika territory, in 1952–1953. Transactions of the royal society of tropical medicine and hygiene. 1955;49(1):28-32.
244. Weaver SC. Arrival of Chikungunya Virus in the New World: Prospects for Spread and Impact on Public Health. PLoS Neglected Tropical Diseases. 2014;8(6):e2921.
245. Powers AM, Brault AC, Tesh RB, Weaver SC. Re-emergence of Chikungunya and O'nyong-nyong viruses: evidence for distinct geographical lineages and distant evolutionary relationships. Microbiology. 2000;81(2):471-9.
246. Valentine MJ, Murdock CC, Kelly PJ. Sylvatic cycles of arboviruses in non-human primates. Parasites & Vectors. 2019;12(1).
247. Volk SM, Chen R, Tsetsarkin KA, Adams AP, Garcia TI, Sall AA, et al. Genome-scale phylogenetic analyses of chikungunya virus reveal independent emergences of recent epidemics and various evolutionary rates. Journal of virology. 2010;84(13):6497-504.
248. Njenga MK, Nderitu L, Ledermann J, Ndirangu A, Logue C, Kelly C, et al. Tracking epidemic chikungunya virus into the Indian Ocean from East Africa. The Journal of general virology. 2008;89(Pt 11):2754.
249. Mavalankar D, Shastri P, Raman P. Chikungunya epidemic in India: a major public-health disaster. The Lancet infectious diseases. 2007;7(5):306-7.
250. Lanciotti RS, Kosoy OL, Laven JJ, Panella AJ, Velez JO, Lambert AJ, et al. Chikungunya virus in US travelers returning from India, 2006. Emerging infectious diseases. 2007;13(5):764.
251. Hochedez P, Hausfater P, Jaureguiberry S, Gay F, Datry A, Danis M, et al. Cases of chikungunya fever imported from the islands of the South West Indian Ocean to Paris, France. Eurosurveillance. 2007;12(1):13-4.
252. Tomasello D, Schlagenhauf P. Chikungunya and dengue autochthonous cases in Europe, 2007–2012. Travel medicine and infectious disease. 2013;11(5):274-84.
253. Schuffenecker I, Iteman I, Michault A, Murri S, Frangeul L, Vaney M-C, et al. Genome Microevolution of Chikungunya Viruses Causing the Indian Ocean Outbreak. PLoS Medicine. 2006;3(7):e263.
254. Vazeille M, Moutailler S, Coudrier D, Rousseaux C, Khun H, Huerre M, et al. Two Chikungunya isolates from the outbreak of La Reunion (Indian Ocean) exhibit different patterns of infection in the mosquito, Aedes albopictus. PloS one. 2007;2(11):e1168.
255. Tsetsarkin KA, Chen R, Yun R, Rossi SL, Plante KS, Guerbois M, et al. Multi-peaked adaptive landscape for chikungunya virus evolution predicts continued fitness optimization in Aedes albopictu s mosquitoes. Nature communications. 2014;5(1):1-14.
256. Tsetsarkin KA, Weaver SC. Sequential adaptive mutations enhance efficient vector switching by Chikungunya virus and its epidemic emergence. PLoS pathogens. 2011;7(12):e1002412.
257. Caminade C, Medlock JM, Ducheyne E, McIntyre KM, Leach S, Baylis M, et al. Suitability of European climate for the Asian tiger mosquito Aedes albopictus: recent trends and future scenarios. Journal of the Royal Society Interface. 2012;9(75):2708-17.
258. Medlock JM, Hansford KM, Schaffner F, Versteirt V, Hendrickx G, Zeller H, et al. A review of the invasive mosquitoes in Europe: ecology, public health risks, and control options. Vector-borne and zoonotic diseases. 2012;12(6):435-47.
259. Duong V, Andries A-C, Ngan C, Sok T, Richner B, Asgari-Jirhandeh N, et al. Reemergence of Chikungunya virus in Cambodia. Emerging infectious diseases. 2012;18(12):2066.

260. Wangchuk S, Chinnawirotpisan P, Dorji T, Tobgay T, Dorji T, Yoon I-K, et al. Chikungunya fever outbreak, Bhutan, 2012. Emerging infectious diseases. 2013;19(10):1681.
261. Pulmanausahakul R, Roytrakul S, Auewarakul P, Smith DR. Chikungunya in Southeast Asia: understanding the emergence and finding solutions. International Journal of Infectious Diseases. 2011;15(10):e671-e6.
262. Tsetsarkin KA, Chen R, Leal G, Forrester N, Higgs S, Huang J, et al. Chikungunya virus emergence is constrained in Asia by lineage-specific adaptive landscapes. Proceedings of the National Academy of Sciences. 2011;108(19):7872-7.
263. Leparc-Goffart I, Nougairede A, Cassadou S, Prat C, De Lamballerie X. Chikungunya in the Americas. The Lancet. 2014;383(9916):514.
264. PAHO. Geographical extension of Chikungunya in the Americas 2017 [Available from: https://www.paho.org/en/topics/chikungunya.
265. Nunes MRT, Faria NR, de Vasconcelos JM, Golding N, Kraemer MU, de Oliveira LF, et al. Emergence and potential for spread of Chikungunya virus in Brazil. BMC medicine. 2015;13(1):1-11.
266. PAHO. 23 de diciembre de 2021: Dengue, Chikungunya y Zika en el contexto del COVID-19 2021 [Available from: https://www.paho.org/es/documentos/23-diciembre-2021-dengue-chikungunya-zika-contexto-covid-19.
267. Khongwichit S, Chansaenroj J, Thongmee T, Benjamanukul S, Wanlapakorn N, Chirathaworn C, et al. Large-scale outbreak of Chikungunya virus infection in Thailand, 2018–2019. PLOS ONE. 2021;16(3):e0247314.
268. Khongwichit S, Chansaenroj J, Chirathaworn C, Poovorawan Y. Chikungunya virus infection: molecular biology, clinical characteristics, and epidemiology in Asian countries. Journal of Biomedical Science. 2021;28(1).
269. Times M. Chikungunta virus reappears after 10 years 2019 [Available from: https://www.mmtimes.com/news/chikungunya-reappears-after-10-years.html.
270. 24 C. Cambodia: Thousands infected as chikungunya outbreak spreads September 28 2020 [Available from: https://crisis24.garda.com/insights-intelligence/intelligence/risk-alerts/gyjr4d5ngybjrd26o/cambodia-thousands-infected-as-chikungunya-outbreak-spreads-september-28.
271. Borgherini G, Poubeau P, Jossaume A, Gouix A, Cotte L, Michault A, et al. Persistent arthralgia associated with chikungunya virus: a study of 88 adult patients on reunion island. Clinical Infectious Diseases. 2008;47(4):469-75.
272. Economopoulou A, Dominguez M, Helynck B, Sissoko D, Wichmann O, Quenel P, et al. Atypical Chikungunya virus infections: clinical manifestations, mortality and risk factors for severe disease during the 2005–2006 outbreak on Reunion. Epidemiology & Infection. 2009;137(4):534-41.
273. Lima STSd, Souza WMd, Cavalcante JW, Candido DdS, Fumagalli MJ, Carrera J-P, et al. Fatal outcome of chikungunya virus infection in Brazil. 2020.
274. Javelle E, Tiong TH, Leparc-Goffart I, Savini H, Simon F. Inflammation of the external ear in acute chikungunya infection: experience from the outbreak in Johor Bahru, Malaysia, 2008. Journal of Clinical Virology. 2014;59(4):270-3.
275. Contopoulos-Ioannidis D, Newman-Lindsay S, Chow C, LaBeaud AD. Mother-to-child transmission of Chikungunya virus: A systematic review and meta-analysis. PLoS neglected tropical diseases. 2018;12(6):e0006510.
276. Gérardin P, Barau G, Michault A, Bintner M, Randrianaivo H, Choker G, et al. Multidisciplinary prospective study of mother-to-child chikungunya virus infections on the island of La Reunion. PLoS medicine. 2008;5(3):e60.

277. Gérardin P, Sampériz S, Ramful D, Boumahni B, Bintner M, Alessandri J-L, et al. Neurocognitive outcome of children exposed to perinatal mother-to-child Chikungunya virus infection: the CHIMERE cohort study on Reunion Island. PLoS neglected tropical diseases. 2014;8(7):e2996.
278. Ramful D, Carbonnier M, Pasquet M, Bouhmani B, Ghazouani J, Noormahomed T, et al. Mother-to-child transmission of Chikungunya virus infection. The Pediatric infectious disease journal. 2007;26(9):811-5.
279. Quick J, Grubaugh ND, Pullan ST, Claro IM, Smith AD, Gangavarapu K, et al. Multiplex PCR method for MinION and Illumina sequencing of Zika and other virus genomes directly from clinical samples. Nat Protoc. 2017;12(6):1261-76.
280. Nate Matteson NG, Karthik Gangavarapu ,Josh Quick ,Nick Loman ,Kristian Andersen. PrimalSeq: Generation of tiled virus amplicons for MiSeq sequencing 2020 [Available from: https://www.protocols.io/view/primalseq-generation-of-tiled-virus-amplicons-for-bez7jf9n.
281. Phadungsombat J, Imad H, Rahman M, Nakayama EE, Kludkleeb S, Ponam T, et al. A novel sub-lineage of chikungunya virus East/Central/South African genotype Indian Ocean lineage caused sequential outbreaks in Bangladesh and Thailand. Viruses. 2020;12(11):1319.
282. Chen R, Puri V, Fedorova N, Lin D, Hari KL, Jain R, et al. Comprehensive Genome Scale Phylogenetic Study Provides New Insights on the Global Expansion of Chikungunya Virus. Journal of Virology. 2016;90(23):10600-11.
283. Yang C-F, Su C-L, Hsu T-C, Chang S-F, Lin C-C, Huang JC, et al. Imported chikungunya virus strains, taiwan, 2006–2014. Emerging infectious diseases. 2016;22(11):1981.
284. Pongsiri P, Auksornkitti V, Theamboonlers A, Luplertlop N, Rianthavorn P, Poovorawan Y. Entire genome characterization of Chikungunya virus from the 2008–2009 outbreaks in Thailand. Trop Biomed. 2010;27(2):167-76.
285. Bielejec F, Baele G, Vrancken B, Suchard MA, Rambaut A, Lemey P. SpreaD3: Interactive Visualization of Spatiotemporal History and Trait Evolutionary Processes. Molecular Biology and Evolution. 2016;33(8):2167-9.
286. Somlor S, Vongpayloth K, Diancourt L, Buchy P, Duong V, Phonekeo D, et al. Chikungunya virus emergence in the Lao PDR, 2012–2013. PLOS ONE. 2017;12(12):e0189879.
287. Gérardin P, Couderc T, Bintner M, Tournebize P, Renouil M, Lémant J, et al. Chikungunya virus–associated encephalitis: a cohort study on La Réunion Island, 2005–2009. Neurology. 2016;86(1):94-102.
288. Couderc T, Chrétien F, Schilte C, Disson O, Brigitte M, Guivel-Benhassine F, et al. A mouse model for Chikungunya: young age and inefficient type-I interferon signaling are risk factors for severe disease. PLoS Pathog. 2008;4(2):e29.
289. Hijano DR, Brazelton de Cardenas J, Maron G, Garner CD, Ferrolino JA, Dallas RH, et al. Clinical correlation of influenza and respiratory syncytial virus load measured by digital PCR. PloS one. 2019;14(9):e0220908.
290. Fajnzylber J, Regan J, Coxen K, Corry H, Wong C, Rosenthal A, et al. SARS-CoV-2 viral load is associated with increased disease severity and mortality. Nature communications. 2020;11(1):1-9.
291. Westblade LF, Brar G, Pinheiro LC, Paidoussis D, Rajan M, Martin P, et al. SARS-CoV-2 viral load predicts mortality in patients with and without cancer who are hospitalized with COVID-19. Cancer cell. 2020;38(5):661-71. e2.
292. Rianthavorn P, Prianantathavorn K, Wuttirattanakowit N, Theamboonlers A, Poovorawan Y. An outbreak of chikungunya in southern Thailand from 2008 to 2009 caused by African strains with A226V mutation. International Journal of Infectious Diseases. 2010;14:e161-e5.
293. Pulmanausahakul R, Roytrakul S, Auewarakul P, Smith DR. Chikungunya in Southeast Asia: understanding the emergence and finding solutions. Int J Infect Dis. 2011;15(10):e671-6.
294. Walsh J, Ty M. Cambodian migrants in Thailand: Working conditions and issues. Asian Social Science. 2011;7(7):23-9.

295. Lequime S, Lambrechts L. Vertical transmission of arboviruses in mosquitoes: A historical perspective. Infection, Genetics and Evolution. 2014;28:681-90.
296. Grubaugh ND, Gangavarapu K, Quick J, Matteson NL, De Jesus JG, Main BJ, et al. An amplicon-based sequencing framework for accurately measuring intrahost virus diversity using PrimalSeq and iVar. Genome biology. 2019;20(1):1-19.
297. Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Molecular biology and evolution. 2013;30(4):772-80.
298. Nguyen L-T, Schmidt HA, Von Haeseler A, Minh BQ. IQ-TREE: A Fast and Effective Stochastic Algorithm for Estimating Maximum-Likelihood Phylogenies. Molecular Biology and Evolution. 2015;32(1):268-74.
299. Kalyaanamoorthy S, Minh BQ, Wong TKF, Von Haeseler A, Jermiin LS. ModelFinder: fast model selection for accurate phylogenetic estimates. Nature Methods. 2017;14(6):587-9.
300. Hoang DT, Chernomor O, Von Haeseler A, Minh BQ, Vinh LS. UFBoot2: Improving the Ultrafast Bootstrap Approximation. Molecular Biology and Evolution. 2018;35(2):518-22.
301. Baele G, Li WLS, Drummond AJ, Suchard MA, Lemey P. Accurate model selection of relaxed molecular clocks in Bayesian phylogenetics. Molecular biology and evolution. 2012;30(2):239-43.
302. Baele G, Lemey P, Bedford T, Rambaut A, Suchard MA, Alekseyenko AV. Improving the accuracy of demographic and molecular clock model comparison while accommodating phylogenetic uncertainty. Molecular biology and evolution. 2012;29(9):2157-67.
303. Agarwal A, Sharma AK, Sukumaran D, Parida M, Dash PK. Two novel epistatic mutations (E1:K211E and E2:V264A) in structural proteins of Chikungunya virus enhance fitness in Aedes aegypti. Virology. 2016;497:59-68.
304. Huber JH, Childs ML, Caldwell JM, Mordecai EA. Seasonal temperature variation influences climate suitability for dengue, chikungunya, and Zika transmission. PLOS Neglected Tropical Diseases. 2018;12(5):e0006451.
305. Cousien A, Ledien J, Souv K, Leang R, Huy R, Fontenille D, et al. Predicting Dengue Outbreaks in Cambodia. Emerging infectious diseases. 2019;25(12):2281.
306. Faull KJ, Williams CR. Intraspecific variation in desiccation survival time of Aedes aegypti (L.) mosquito eggs of Australian origin. Journal of Vector Ecology. 2015;40(2):292-300.
307. Staples JE, Fischer M. Chikungunya Virus in the Americas — What a Vectorborne Pathogen Can Do. New England Journal of Medicine. 2014;371(10):887-9.
308. Boyer S, Marcombe S, Yean S, Fontenille D. High diversity of mosquito vectors in Cambodian primary schools and consequences for arbovirus transmission. PLOS ONE. 2020;15(6):e0233669.
309. Niyas KP, Abraham R, Unnikrishnan R, Mathew T, Nair S, Manakkadan A, et al. Molecular characterization of Chikungunya virus isolates from clinical samples and adult Aedes albopictus mosquitoes emerged from larvae from Kerala, South India. Virology Journal. 2010;7(1):189.
310. Kumar NP, Sabesan S, Krishnamoorthy K, Jambulingam P. Detection of Chikungunya virus in wild populations of Aedes albopictus in Kerala State, India. Vector-Borne and Zoonotic Diseases. 2012;12(10):907-11.
311. Shrinet J, Jain S, Sharma A, Singh SS, Mathur K, Rana V, et al. Genetic characterization of Chikungunya virus from New Delhi reveal emergence of a new molecular signature in Indian isolates. Virology Journal. 2012;9(1):100.
312. CDC. SARS-CoV Images 2020 [Available from: https://www.cdc.gov/sars/lab/images.html.
313. Hu B, Guo H, Zhou P, Shi Z-L. Characteristics of SARS-CoV-2 and COVID-19. Nature Reviews Microbiology. 2021;19(3):141-54.
314. Wertheim JO, Chu DK, Peiris JS, Kosakovsky Pond SL, Poon LL. A case for the ancient origin of coronaviruses. Journal of Virology. 2013;87(12):7039-45.

315. Cook JKA, Jackwood M, Jones RC. The long view: 40 years of infectious bronchitis research. Avian Pathology. 2012;41(3):239-50.
316. Chen F, Knutson TP, Rossow S, Saif LJ, Marthaler DG. Decline of transmissible gastroenteritis virus and its complex evolutionary relationship with porcine respiratory coronavirus in the United States. Scientific Reports. 2019;9(1).
317. Liu Q, Gerdts V. Transmissible gastroenteritis virus of pigs and porcine epidemic diarrhea virus. Reference Module in Life Sciences. 2019.
318. CDC. Middle East respiratory syndrome coronavirus (MERS-CoV) 2019 [Available from: https://www.who.int/health-topics/middle-east-respiratory-syndrome-coronavirus-mers#tab=tab_1.
319. V'kovski P, Kratzel A, Steiner S, Stalder H, Thiel V. Coronavirus biology and replication: implications for SARS-CoV-2. Nature Reviews Microbiology. 2021;19(3):155-70.
320. Zhu N, Zhang D, Wang W, Li X, Yang B, Song J, et al. A novel coronavirus from patients with pneumonia in China, 2019. New England journal of medicine. 2020.
321. Deng S-Q, Peng H-J. Characteristics of and Public Health Responses to the Coronavirus Disease 2019 Outbreak in China. Journal of Clinical Medicine. 2020;9(2):575.
322. Wu Z, McGoogan JM. Characteristics of and important lessons from the coronavirus disease 2019 (COVID-19) outbreak in China: summary of a report of 72 314 cases from the Chinese Center for Disease Control and Prevention. jama. 2020;323(13):1239-42.
323. Worobey M. Dissecting the early COVID-19 cases in Wuhan. Science. 2021;374(6572):1202-4.
324. WHO. COVID-19 - China 2020 [Available from: https://www.who.int/emergencies/disease-outbreak-news/item/2020-DON229.
325. Wu F, Zhao S, Yu B, Chen Y-M, Wang W, Song Z-G, et al. A new coronavirus associated with human respiratory disease in China. Nature. 2020;579(7798):265-9.
326. Zhou P, Yang X-L, Wang X-G, Hu B, Zhang L, Zhang W, et al. A pneumonia outbreak associated with a new coronavirus of probable bat origin. nature. 2020;579(7798):270-3.
327. Website V. Novel 2019 coronavirus genome 2020 [Available from: https://virological.org/t/novel-2019-coronavirus-genome/319.
328. GISAID. GISAID database [Available from: https://www.gisaid.org/.
329. Chen N, Zhou M, Dong X, Qu J, Gong F, Han Y, et al. Epidemiological and clinical characteristics of 99 cases of 2019 novel coronavirus pneumonia in Wuhan, China: a descriptive study. The lancet. 2020;395(10223):507-13.
330. Chan JF-W, Yuan S, Kok K-H, To KK-W, Chu H, Yang J, et al. A familial cluster of pneumonia associated with the 2019 novel coronavirus indicating person-to-person transmission: a study of a family cluster. The lancet. 2020;395(10223):514-23.
331. WHO. Emergency Committee regarding the outbreak of novel coronavirus (2019-nCoV) 2020 [Available from: https://www.who.int/news/item/30-01-2020-statement-on-the-second-meeting-of-the-international-health-regulations-(2005)-emergency-committee-regarding-the-outbreak-of-novel-coronavirus-(2019-ncov).
332. WHO. Naming the coronavirus disease (COVID-19) and the virus that causes it 2020 [Available from: https://www.who.int/emergencies/diseases/novel-coronavirus-2019/technical-guidance/naming-the-coronavirus-disease-(covid-2019)-and-the-virus-that-causes-it.
333. Spiteri G, Fielding J, Diercke M, Campese C, Enouf V, Gaymard A, et al. First cases of coronavirus disease 2019 (COVID-19) in the WHO European Region, 24 January to 21 February 2020. Eurosurveillance. 2020;25(9).
334. Holmes EC, Goldstein SA, Rasmussen AL, Robertson DL, Crits-Christoph A, Wertheim JO, et al. The origins of SARS-CoV-2: A critical review. Cell. 2021;184(19):4848-56.
335. Andersen KG, Rambaut A, Lipkin WI, Holmes EC, Garry RF. The proximal origin of SARS-CoV-2. Nature medicine. 2020;26(4):450-2.

336. Lu R, Zhao X, Li J, Niu P, Yang B, Wu H, et al. Genomic characterisation and epidemiology of 2019 novel coronavirus: implications for virus origins and receptor binding. The lancet. 2020;395(10224):565-74.
337. Robson F, Khan KS, Le TK, Paris C, Demirbag S, Barfuss P, et al. Coronavirus RNA proofreading: molecular basis and therapeutic targeting. Molecular cell. 2020;79(5):710-27.
338. Temmam S, Vongphayloth K, Salazar EB, Munier S, Bonomi M, Regnault B, et al. Bat coronaviruses related to SARS-CoV-2 and infectious for human cells. Nature. 2022:1-10.
339. Rambaut) VoA. Phylogenetic analysis of nCoV-2019 genomes 2020 [Available from: https://virological.org/t/phylodynamic-analysis-176-genomes-6-mar-2020/356.
340. Boni MF, Lemey P, Jiang X, Lam TT-Y, Perry BW, Castoe TA, et al. Evolutionary origins of the SARS-CoV-2 sarbecovirus lineage responsible for the COVID-19 pandemic. Nature Microbiology. 2020;5(11):1408-17.
341. Zhang T, Wu Q, Zhang Z. Probable Pangolin Origin of SARS-CoV-2 Associated with the COVID-19 Outbreak. Current Biology. 2020;30(7):1346-51.e2.
342. Xiao K, Zhai J, Feng Y, Zhou N, Zhang X, Zou J-J, et al. Isolation of SARS-CoV-2-related coronavirus from Malayan pangolins. Nature. 2020;583(7815):286-9.
343. Li X, Giorgi EE, Marichannegowda MH, Foley B, Xiao C, Kong X-P, et al. Emergence of SARS-CoV-2 through recombination and strong purifying selection. Science advances. 2020;6(27):eabb9153.
344. Lam TT-Y, Jia N, Zhang Y-W, Shum MH-H, Jiang J-F, Zhu H-C, et al. Identifying SARS-CoV-2-related coronaviruses in Malayan pangolins. Nature. 2020;583(7815):282-5.
345. Hoffmann M, Kleine-Weber H, Schroeder S, Krüger N, Herrler T, Erichsen S, et al. SARS-CoV-2 cell entry depends on ACE2 and TMPRSS2 and is blocked by a clinically proven protease inhibitor. cell. 2020;181(2):271-80. e8.
346. Shang J, Ye G, Shi K, Wan Y, Luo C, Aihara H, et al. Structural basis of receptor recognition by SARS-CoV-2. Nature. 2020;581(7807):221-4.
347. Ou X, Liu Y, Lei X, Li P, Mi D, Ren L, et al. Characterization of spike glycoprotein of SARS-CoV-2 on virus entry and its immune cross-reactivity with SARS-CoV. Nature communications. 2020;11(1):1-12.
348. Zhu X, Mannar D, Srivastava SS, Berezuk AM, Demers J-P, Saville JW, et al. Cryo-electron microscopy structures of the N501Y SARS-CoV-2 spike protein in complex with ACE2 and 2 potent neutralizing antibodies. PLoS biology. 2021;19(4):e3001237.
349. Hoffmann M, Kleine-Weber H, Pöhlmann S. A multibasic cleavage site in the spike protein of SARS-CoV-2 is essential for infection of human lung cells. Molecular cell. 2020;78(4):779-84. e5.
350. Zhou H, Chen X, Hu T, Li J, Song H, Liu Y, et al. A novel bat coronavirus closely related to SARS-CoV-2 contains natural insertions at the S1/S2 cleavage site of the spike protein. Current biology. 2020;30(11):2196-203. e3.
351. Hadfield J, Megill C, Bell SM, Huddleston J, Potter B, Callender C, et al. Nextstrain: real-time tracking of pathogen evolution. Bioinformatics. 2018;34(23):4121-3.
352. GISAID. Clade and lineage nomenclature aids in genomic epidemiology studies of active hCoV-19 viruses 2021 [Available from: https://www.gisaid.org/references/statements-clarifications/clade-and-lineage-nomenclature-aids-in-genomic-epidemiology-of-active-hcov-19-viruses/.
353. Rambaut A, Holmes EC, O'Toole Á, Hill V, McCrone JT, Ruis C, et al. A dynamic nomenclature proposal for SARS-CoV-2 lineages to assist genomic epidemiology. Nature microbiology. 2020;5(11):1403-7.
354. O'Toole Á, Scher E, Underwood A, Jackson B, Hill V, McCrone JT, et al. Assignment of epidemiological lineages in an emerging pandemic using the pangolin tool. Virus Evolution. 2021;7(2):veab064.
355. cov-lineages. Lineage List 2022 [Available from: https://cov-lineages.org/lineage_list.html.

356. Trevor Bedford EBH, Richard A Neher. Updated Nextstrain SARS-CoV-2 clade naming strategy 2021 [Available from: https://nextstrain.org/blog/2021-01-06-updated-SARS-CoV-2-clade-naming.
357. WHO. Tracking SARS-CoV-2 variants 2022 [Available from: https://www.who.int/en/activities/tracking-SARS-CoV-2-variants/.
358. WHO. WHO announces simple, easy-to-say labels for SARS-CoV-2 Variants of Interest and Concern 2021 [Available from: https://www.who.int/news/item/31-05-2021-who-announces-simple-easy-to-say-labels-for-sars-cov-2-variants-of-interest-and-concern.
359. Andrew Rambaut NL, Oliver Pybus, Wendy Barclay, Jeff Barrett, Alesandro Carabelli, Tom Connor, Tom Peacock, David L Robertson, Erik Volz. Preliminary genomic characterisation of an emergent SARS-CoV-2 lineage in the UK defined by a novel set of spike mutations 2020 [Available from: https://virological.org/t/preliminary-genomic-characterisation-of-an-emergent-sars-cov-2-lineage-in-the-uk-defined-by-a-novel-set-of-spike-mutations/563.
360. Meng B, Kemp SA, Papa G, Datir R, Ferreira IA, Marelli S, et al. Recurrent emergence of SARS-CoV-2 spike deletion H69/V70 and its role in the Alpha variant B. 1.1. 7. Cell reports. 2021;35(13):109292.
361. Rajah MM, Hubert M, Bishop E, Saunders N, Robinot R, Grzelak L, et al. SARS-CoV-2 Alpha, Beta, and Delta variants display enhanced Spike-mediated syncytia formation. The EMBO journal. 2021;40(24):e108944.
362. Tegally H, Wilkinson E, Giovanetti M, Iranzadeh A, Fonseca V, Giandhari J, et al. Detection of a SARS-CoV-2 variant of concern in South Africa. Nature. 2021;592(7854):438-43.
363. Faria NR, Mellan TA, Whittaker C, Claro IM, Candido DdS, Mishra S, et al. Genomics and epidemiology of the P. 1 SARS-CoV-2 lineage in Manaus, Brazil. Science. 2021;372(6544):815-21.
364. Abdool Karim SS, de Oliveira T. New SARS-CoV-2 variants—clinical, public health, and vaccine implications. New England Journal of Medicine. 2021;384(19):1866-8.
365. Pearson CA, Russell TW, Davies NG, Kucharski AJ, group CC-w, Edmunds WJ, et al. Estimates of severity and transmissibility of novel SARS-CoV-2 variant 501Y. V2 in South Africa. CMMID Repository. 2021.
366. Baum A, Fulton BO, Wloga E, Copin R, Pascal KE, Russo V, et al. Antibody cocktail to SARS-CoV-2 spike protein prevents rapid mutational escape seen with individual antibodies. Science. 2020;369(6506):1014-8.
367. Greaney AJ, Loes AN, Crawford KH, Starr TN, Malone KD, Chu HY, et al. Comprehensive mapping of mutations in the SARS-CoV-2 receptor-binding domain that affect recognition by polyclonal human plasma antibodies. Cell host & microbe. 2021;29(3):463-76. e6.
368. Wang P, Nair MS, Liu L, Iketani S, Luo Y, Guo Y, et al. Antibody resistance of SARS-CoV-2 variants B. 1.351 and B. 1.1. 7. Nature. 2021;593(7857):130-5.
369. cov-lineages. P.1 - cov-lineages 2022 [Available from: https://cov-lineages.org/global_report_P.1.html.
370. Dhar MS, Marwal R, Vs R, Ponnusamy K, Jolly B, Bhoyar RC, et al. Genomic characterization and epidemiology of an emerging SARS-CoV-2 variant in Delhi, India. Science. 2021;374(6570):995-9.
371. Articles N. Coronavirus variants are spreading in India — what scientists know so far 2021 [Available from: https://www.nature.com/articles/d41586-021-01274-7.
372. Bolze A, Luo S, White S, Cirulli ET, Wyman D, Dei Rossi A, et al. SARS-CoV-2 variant Delta rapidly displaced variant Alpha in the United States and led to higher viral loads. 2021.
373. McCrone JT, Hill V, Bajaj S, Pena RE, Lambert BC, Inward R, et al. Context-specific emergence and growth of the SARS-CoV-2 Delta variant. medRxiv. 2021.
374. cov-lineages. B.1.617.2 lineage 2021 [Available from: https://cov-lineages.org/global_report_B.1.617.2.html.

375. Twohig KA, Nyberg T, Zaidi A, Thelwall S, Sinnathamby MA, Aliabadi S, et al. Hospital admission and emergency care attendance risk for SARS-CoV-2 delta (B. 1.617. 2) compared with alpha (B. 1.1. 7) variants of concern: a cohort study. The Lancet Infectious Diseases. 2022;22(1):35-42.
376. Sheikh A, McMenamin J, Taylor B, Robertson C. SARS-CoV-2 Delta VOC in Scotland: demographics, risk of hospital admission, and vaccine effectiveness. The Lancet. 2021;397(10293):2461-2.
377. Bernal JL, Andrews N, Gower C, Gallagher E, Simmons R, Thelwall S, et al. Effectiveness of Covid-19 vaccines against the B. 1.617. 2 (Delta) variant. New England Journal of Medicine. 2021.
378. Lucas C, Vogels CB, Yildirim I, Rothman JE, Lu P, Monteiro V, et al. Impact of circulating SARS-CoV-2 variants on mRNA vaccine-induced immunity. Nature. 2021;600(7889):523-9.
379. Bernal J, Andrews N, Gower C, Gallagher E, Simmons R, Thelwall S, et al. Effectiveness of COVID-19 vaccines against the B. 1.617. 2 variant. medRxiv. Preprint posted online May. 2021;24.
380. Cherian S, Potdar V, Jadhav S, Yadav P, Gupta N, Das M, et al. SARS-CoV-2 spike mutations, L452R, T478K, E484Q and P681R, in the second wave of COVID-19 in Maharashtra, India. Microorganisms. 2021;9(7):1542.
381. Saito A, Nasser H, Uriu K, Kosugi Y, Irie T, Shirakawa K. SARS-CoV-2 spike P681R mutation enhances and accelerates viral fusion. bioRxiv. Preprint. 2021;10(2021.06):17.448820.
382. Zhang J, Xiao T, Cai Y, Lavine CL, Peng H, Zhu H, et al. Membrane fusion and immune evasion by the spike protein of SARS-CoV-2 Delta variant. Science. 2021;374(6573):1353-60.
383. Liu Y, Liu J, Johnson B, Xia H, Ku Z, Schindewolf C, et al. Delta spike P681R mutation enhances SARS-CoV-2 fitness over Alpha variant. bioRxiv 2021. Google Scholar.
384. Stern A, Fleishon S, Kustin T, Mandelboim M, Erster O, Mendelson E, et al. The unique evolutionary dynamics of the SARS-CoV-2 Delta variant. medRxiv. 2021.
385. Viana R, Moyo S, Amoako DG, Tegally H, Scheepers C, Althaus CL, et al. Rapid epidemic expansion of the SARS-CoV-2 Omicron variant in southern Africa. Nature. 2022:1-10.
386. cov-lineages. B.1.1.529 2022 [Available from: https://cov-lineages.org/global_report_B.1.1.529.html.
387. Planas D, Saunders N, Maes P, Benhassine FG, Planchais C, Porrot F, et al. Considerable escape of SARS-CoV-2 variant Omicron to antibody neutralization (preprint).
388. Andrews N, Stowe J, Kirsebom F, Toffa S, Rickeard T, Gallagher E, et al. Effectiveness of COVID-19 vaccines against the Omicron (B. 1.1. 529) variant of concern. MedRxiv. 2021.
389. Flemming A. Omicron, the great escape artist. Nature Reviews Immunology. 2022:1-.
390. Pulliam JR, van Schalkwyk C, Govender N, von Gottberg A, Cohen C, Groome MJ, et al. Increased risk of SARS-CoV-2 reinfection associated with emergence of the Omicron variant in South Africa. MedRxiv. 2021.
391. Andersen KG. Twitter - 12:18 AM · Dec 1, 2021 [Available from: https://twitter.com/K_G_Andersen/status/1465822536629821442.
392. Mallapaty S. Where did Omicron come from? Three key theories. Nature. 2022;602(7895):26-8.
393. Avanzato VA, Matson MJ, Seifert SN, Pryce R, Williamson BN, Anzick SL, et al. Case study: prolonged infectious SARS-CoV-2 shedding from an asymptomatic immunocompromised individual with cancer. Cell. 2020;183(7):1901-12. e9.
394. Choi B, Choudhary MC, Regan J, Sparks JA, Padera RF, Qiu X, et al. Persistence and evolution of SARS-CoV-2 in an immunocompromised host. New England Journal of Medicine. 2020;383(23):2291-3.
395. Kemp SA, Collier DA, Datir RP, Ferreira IA, Gayed S, Jahun A, et al. SARS-CoV-2 evolution during treatment of chronic infection. Nature. 2021;592(7853):277-82.
396. Oude Munnink BB, Sikkema RS, Nieuwenhuijse DF, Molenaar RJ, Munger E, Molenkamp R, et al. Transmission of SARS-CoV-2 on mink farms between humans and mink and back to humans. Science. 2021;371(6525):172-7.

397. Garry RF. Mutations arising in SARS-CoV-2 spike on sustained human-to-human transmission and human-to-animal passage. Image. 2021;908(1148):292.
398. Montagutelli X, van der Werf S, Rey FA, Simon-Loriere E. SARS-CoV-2 Omicron emergence urges for reinforced One-Health surveillance. EMBO Molecular Medicine.e15558.
399. Bedford T. Twitter - 5:48 PM · Jan 28, 2022 2022 [Available from: https://twitter.com/trvrb/status/1487105396879679488.
400. NIH-COVID19. Clinical Spectrum of SARS-CoV-2 Infection 2021 [Available from: https://www.covid19treatmentguidelines.nih.gov/overview/clinical-spectrum/.
401. CDC. People with Certain Medical Conditions 2021 [Available from: https://www.cdc.gov/coronavirus/2019-ncov/need-extra-precautions/people-with-medical-conditions.html.
402. Sah P, Fitzpatrick MC, Zimmer CF, Abdollahi E, Juden-Kelly L, Moghadas SM, et al. Asymptomatic SARS-CoV-2 infection: A systematic review and meta-analysis. Proceedings of the National Academy of Sciences. 2021;118(34).
403. Ma Q, Liu J, Liu Q, Kang L, Liu R, Jing W, et al. Global percentage of asymptomatic SARS-CoV-2 infections among the tested population and individuals with confirmed COVID-19 diagnosis: a systematic review and meta-analysis. JAMA network open. 2021;4(12):e2137257-e.
404. articles N. What the data say about asymptomatic COVID infections 2020 [Available from: https://www.nature.com/articles/d41586-020-03141-3.
405. Nonaka CK, Franco MM, Gräf T, de Lorenzo Barcia CA, de Ávila Mendonça RN, De Sousa KAF, et al. Genomic evidence of SARS-CoV-2 reinfection involving E484K spike mutation, Brazil. Emerging infectious diseases. 2021;27(5):1522.
406. CDC. The Possibility of COVID-19 after Vaccination: Breakthrough Infections 2021 [Available from: https://www.cdc.gov/coronavirus/2019-ncov/vaccines/effectiveness/why-measure-effectiveness/breakthrough-cases.html.
407. Netzl A, Tureli S, Legresley E, Mühlemann B, Wilks SH, Smith DJ. Analysis of SARS-CoV-2 Omicron Neutralization Data up to 2021-12-22. 2022.
408. Levin EG, Lustig Y, Cohen C, Fluss R, Indenbaum V, Amit S, et al. Waning immune humoral response to BNT162b2 Covid-19 vaccine over 6 months. New England Journal of Medicine. 2021;385(24):e84.
409. Chemaitelly H, Tang P, Hasan MR, AlMukdad S, Yassine HM, Benslimane FM, et al. Waning of BNT162b2 vaccine protection against SARS-CoV-2 infection in Qatar. New England Journal of Medicine. 2021;385(24):e83.
410. Duchêne S, Holmes EC, Ho SY. Analyses of evolutionary dynamics in viruses are hindered by a time-dependent bias in rate estimates. Proceedings of the Royal Society B: Biological Sciences. 2014;281(1786):20140732.
411. Holmes EC, Dudas G, Rambaut A, Andersen KG. The evolution of Ebola virus: Insights from the 2013–2016 epidemic. Nature. 2016;538(7624):193-200.
412. Mahan Ghafari LdP, Oliver Pybus and Aris Katzourakis. Time dependence of SARS-CoV-2 substitution rates 2020 [Available from: https://virological.org/t/time-dependence-of-sars-cov-2-substitution-rates/542.
413. McCrone JT, Lauring AS. Measurements of Intrahost Viral Diversity Are Extremely Sensitive to Systematic Errors in Variant Calling. Journal of Virology. 2016;90(15):6884-95.
414. Zhao C, Liu F, Pyle AM. An ultraprocessive, accurate reverse transcriptase encoded by a metazoan group II intron. Rna. 2018;24(2):183-95.
415. Grubaugh ND, Gangavarapu K, Quick J, Matteson NL, De Jesus JG, Main BJ, et al. An amplicon-based sequencing framework for accurately measuring intrahost virus diversity using PrimalSeq and iVar. Genome Biology. 2019;20(1).

416. Salk JJ, Schmitt MW, Loeb LA. Enhancing the accuracy of next-generation sequencing for detecting rare and subclonal mutations. Nature Reviews Genetics. 2018;19(5):269-85.
417. Xue KS, Bloom JD. Reconciling disparate estimates of viral genetic diversity during human influenza infections. Nature genetics. 2019;51(9):1298-301.
418. Poon LL, Song T, Rosenfeld R, Lin X, Rogers MB, Zhou B, et al. Quantifying influenza virus diversity and transmission in humans. Nature genetics. 2016;48(2):195-200.
419. Popa A, Genger J-W, Nicholson MD, Penz T, Schmid D, Aberle SW, et al. Genomic epidemiology of superspreading events in Austria reveals mutational dynamics and transmission properties of SARS-CoV-2. Science Translational Medicine. 2020;12(573):eabe2555.
420. Martin MA, Koelle K. Comment on "Genomic epidemiology of superspreading events in Austria reveals mutational dynamics and transmission properties of SARS-CoV-2". Science translational medicine. 2021;13(617):eabh1803.
421. Braun KM, Moreno GK, Wagner C, Accola MA, Rehrauer WM, Baker DA, et al. Acute SARS-CoV-2 infections harbor limited within-host diversity and transmit via tight transmission bottlenecks. PLoS pathogens. 2021;17(8):e1009849.
422. Lythgoe KA, Hall M, Ferretti L, de Cesare M, MacIntyre-Cockett G, Trebes A, et al. SARS-CoV-2 within-host diversity and transmission. Science. 2021;372(6539):eabg0821.
423. Baang JH, Smith C, Mirabelli C, Valesano AL, Manthei DM, Bachman MA, et al. Prolonged Severe Acute Respiratory Syndrome Coronavirus 2 Replication in an Immunocompromised Patient. The Journal of Infectious Diseases. 2021;223(1):23-7.
424. Bazykin GA SO, Danilenko D, Fadeev A, Komissarova K, Ivanova A, Sergeeva M, Safina K, Nabieva E, Klink G, Garushyants S, Zabutova J, Kholodnaia A, Skorokhod I, Ryabchikova VV, Komissarov A, Lioznov D. Emergence of Y453F and Δ69-70HV mutations in a lymphoma patient with long-term COVID-19 2021 [Available from: https://virological.org/t/emergence-of-y453f-and-69-70hv-mutations-in-a-lymphoma-patient-with-long-term-covid-19/580.
425. Choi B, Choudhary MC, Regan J, Sparks JA, Padera RF, Qiu X, et al. Persistence and Evolution of SARS-CoV-2 in an Immunocompromised Host. New England Journal of Medicine. 2020;383(23):2291-3.
426. Kemp SA, Collier DA, Datir RP, Ferreira IATM, Gayed S, Jahun A, et al. SARS-CoV-2 evolution during treatment of chronic infection. Nature. 2021.
427. Pang NY-L, Pang AS-R, Chow VT, Wang D-Y. Understanding neutralising antibodies against SARS-CoV-2 and their implications in clinical practice. Military Medical Research. 2021;8(1).
428. Crawford KHD, Eguia R, Dingens AS, Loes AN, Malone KD, Wolf CR, et al. Protocol and Reagents for Pseudotyping Lentiviral Particles with SARS-CoV-2 Spike Protein for Neutralization Assays. Viruses. 2020;12(5):513.
429. Delaune D, Hul V, Karlsson EA, Hassanin A, Ou TP, Baidaliuk A, et al. A novel SARS-CoV-2 related coronavirus in bats from Cambodia. Nature communications. 2021;12(1):1-7.
430. Teng S, Sobitan A, Rhoades R, Liu D, Tang Q. Systemic effects of missense mutations on SARS-CoV-2 spike glycoprotein stability and receptor-binding affinity. Briefings in Bioinformatics. 2021;22(2):1239-53.
431. Adiconis X, Borges-Rivera D, Satija R, DeLuca DS, Busby MA, Berlin AM, et al. Comparative analysis of RNA sequencing methods for degraded or low-input samples. Nature methods. 2013;10(7):623-9.
432. ID C. CZ ID [Available from: https://czid.org/.
433. Sichtig H, Minogue T, Yan Y, Stefan C, Hall A, Tallon L, et al. FDA-ARGOS is a database with public quality-controlled reference genomes for diagnostic use and regulatory science. Nature Communications. 2019;10(1).
434. Gudbjartsson DF, Helgason A, Jonsson H, Magnusson OT, Melsted P, Norddahl GL, et al. Spread of SARS-CoV-2 in the Icelandic population. New England Journal of Medicine. 2020;382(24):2302-15.

435. Du Plessis L, McCrone JT, Zarebski AE, Hill V, Ruis C, Gutierrez B, et al. Establishment and lineage dynamics of the SARS-CoV-2 epidemic in the UK. Science. 2021;371(6530):708-12.
436. Butera Y, Mukantwari E, Artesi M, O'Toole ÁN, Hill V, Rooke S, et al. Genomic sequencing of SARS-CoV-2 in Rwanda reveals the importance of incoming travelers on lineage diversity. Nature communications. 2021;12(1):1-11.
437. WorldPop. WorldPop [Available from: https://www.worldpop.org/.
438. Flowminder. Flowminder [Available from: https://www.flowminder.org.
439. VectorBase. VectorBase [Available from: https://vectorbase.org.
440. Carlson CJ, Gibb RJ, Albery GF, Brierley L, Connor R, Dallas T, et al. The Global Virome in One Network (VIRION): an atlas of vertebrate-virus associations. bioRxiv. 2021.
441. Hill V, Ruis C, Bajaj S, Pybus OG, Kraemer MUG. Progress and challenges in virus genomic epidemiology. Trends in Parasitology. 2021;37(12):1038-49.
442. Brito AF, Semenova E, Dudas G, Hassler GW, Kalinich CC, Kraemer MU, et al. Global disparities in SARS-CoV-2 genomic surveillance. medRxiv. 2021.
443. Keita AK, Koundouno FR, Faye M, Düx A, Hinzmann J, Diallo H, et al. Resurgence of Ebola virus in 2021 in Guinea suggests a new paradigm for outbreaks. Nature. 2021;597(7877):539-43.