From viral origins to global dispersion : exploring factors that drive and sustain viral diffusion over time and space

Andrew Mark Holtz

Université Paris Cité

Frontiére de l'Innovation en Recherche et Education

Ecole doctorale 474

Institut Pasteur

Lyssavirus, épidémiologie et neuropathologie

From Viral Origins to Global Dispersion: Exploring Factors That Drive and Sustain Viral Diffusion Over Time and Space

Par: Andrew Holtz

Thèse de doctorat de GENE, OMIQUES, BIOINFORMATIQUE ET BIOLOGIE DES SYSTEMES

Dirigée par:
Hervé BOURHY et Guy BAELE,
Encadrante de thèse : Anna ZHUKOVA

Présentée et soutenue publiquement le 7. December 2023

Devant un jury composé de:

Professor Dr. Katie Hampson, Rapportrice, University of Glasgow

Group Leader Dr. Denise Kühnert, Rapportrice, Max Planck Inst Geoanthropology

Associate Professor Dr. Sebastian Lequime, Examinateur, University of Groningen

Group Leader Dr. Sandie Munier, Présidente, Institut Pasteur

Professor Dr. Hervé Bourhy, Directeur de thése, Institut Pasteur

Associate Professor Dr. Guy Baele, Co-directeur de thése, KU Leuven

Research Engineer Dr. Anna Zhukova, Encadrante de thèse, Institut Pasteur

2

Andrew Holtz
PhD Evolutionary Biology / 2023
Frontiére de l'Innovation en Recherche et Education
Université de Paris Cité

Abstract

From Viral Origins to Global Dispersion: Exploring Factors That Drive and Sustain Viral Diffusion Over Time and Space

As shifting migration patterns increase the arrival of viruses in new territories, our capacity to understand the factors that enable their emergence must evolve in parallel. Technological leaps in genomic sequencing and global collaboration provide an immense wealth of data, enabling researchers to decode the dynamics of disease emergence and dissemination. Phylogenetics and phylogeographic tools which map evolutionary relationships between viruses and how they have spread over time and space have revolutionized the way we understand viral emergence. These tools are instrumental in testing the association of viral movement with environmental, epidemiological, and social factors. They have been used successfully to reveal, for example, migration patterns and drivers of the 2013–2016 Ebola epidemic in West Africa and in the 2019–2023 SARS-CoV-2 pandemic.

This thesis aims to develop and apply novel methods in phylogeography to reveal factors in human migration involved in current and past disease outbreaks. Chapter one focuses on rabies virus (RABV), a neglected tropical disease responsible for an estimated 59,000 annual deaths, circulating mostly among canines. Our investigation, using Maximum Likelihood methods, evaluates the role of human migration in the global dispersion of rabies by harnessing the plethora of available genetic sequence data. We introduce a newly-developed method of alignment concatenation including over 14,000 sequences from over 120 nations across a 40-year period. From this analysis, we estimate introduction patterns of canine rabies globally since its estimated origin between 1301 and 1403 and infer how human migration over long-distance and early European colonization may have influenced its spread. Our analysis highlights a relationship between historical human migration and rabies spread.

Chapter two builds on the concept of factor evaluation in disease spread by using a more rigorous approach using the generalized linear model in a Bayesian framework. We were able to investigate the emergence and origin of the novel SARS-CoV-2 variant B.1.214.2, which was first identified

Andrew Holtz
PhD Evolutionary Biology / 2023
Frontiére de l'Innovation en Recherche et Education
Université de Paris Cité

in Belgium in 2021. Using travel history-aware phylogeographic methods, we were able to integrate both patient travel history from interviews and global air passenger data to reconstruct the transmission path of this variant. We estimated that the emergence of the variant in Europe originates in Central Africa, most likely the Republic of Congo and was transmitted with minimal detection during the early months of 2021.

In both chapters, we unveil new perspectives on how human migration influences the spread of infectious diseases, underscoring the significance of integrating genomic data with epidemiological, social, and geographic contexts. In the first chapter, we deploy efficient methods to process the 14,000 RABV sequences in our dataset, shedding light on its geographic spread and the impact of human movement. In the second chapter, we delve deeper by employing more rigorous predictor testing, leveraging the more compact dataset of SARS-CoV-2 sequences. In the final chapter, we advocate for a synthesis of these approaches. By employing the generalized linear model within a maximum likelihood framework, predictor evaluation could be possible on large sequence datasets, ultimately providing statistically relevant approaches to disease control strategies.

Keywords :
Phylogenetics, phylogeography, rabies, SARS-CoV-2, zoonosis, molecular epidemiology, outbreak, infectious disease dynamics, viral evolution

Andrew Holtz
PhD Evolutionary Biology / 2023
Frontiére de l'Innovation en Recherche et Education
Université de Paris Cité

Résumé

Des Origines Virales à la Dispersion Mondiale : Exploration des Facteurs qui Stimulent et Maintiennent la Diffusion Virale à Travers le Temps et l'Espace.

Les changements migratoires augmentent l'arrivée de virus sur de nouveaux territoires. Notre capacité à comprendre les facteurs favorisant leur émergence doit évoluer en parallèle. Les avancées technologiques en séquençage génomique et la collaboration mondiale fournissent une immense richesse de données, permettant aux chercheurs de décoder la dynamique d'émergence et de propagation des maladies. Les outils de phylogénétique et de phylogéographie qui cartographient les relations évolutives entre les virus et leur propagation dans le temps et l'espace ont révolutionné notre compréhension de l'émergence virale. Ils ont été utilisés avec succès, par exemple, pour révéler les schémas de migration et les moteurs de l'épidémie d'Ebola de 2013-2015 en Afrique de l'Ouest et de la pandémie de SARS-CoV-2 de 2019-2023.

Cette thèse vise à développer et appliquer de nouvelles méthodes en phylogéographie pour révéler les facteurs de la migration humaine impliqués dans les épidémies actuelles et passées. Le premier chapitre se concentre sur le virus de la rage (RABV), une maladie tropicale négligée responsable d'une estimation de 59 000 décès annuels. Notre enquête, utilisant des méthodes de Maximum de Vraisemblance, évalue le rôle de la migration humaine dans la dispersion mondiale de la rage en exploitant la multitude de données de séquences génétiques disponibles. Nous introduisons une nouvelle méthode de concaténation d'alignements incluant des séquences de plus de 120 pays sur une période de 40 ans. De cette analyse, nous estimons les schémas d'introduction de la rage canine à l'échelle mondiale depuis son origine estimée entre 1301 et 1403 et déduisons comment la migration humaine sur de longues distances et la colonisation européenne précoce ont pu influencer sa propagation. Notre analyse souligne une relation entre la migration humaine historique et la propagation de la rage.

Le deuxième chapitre élabore sur le concept d'évaluation des facteurs dans la propagation des maladies en utilisant une approche plus rigoureuse avec le modèle linéaire généralisé dans un cadre bayésien. Nous avons pu étudier l'émergence et l'origine du nouveau variant SARS-CoV-2

Andrew Holtz
PhD Evolutionary Biology / 2023
Frontiére de l'Innovation en Recherche et Education
Université de Paris Cité

B.1.214.2, qui a été identifié pour la première fois en Belgique en 2021. En utilisant des méthodes phylogéographiques intégrant l'historique des voyages, nous avons pu combiner à la fois l'historique de voyage des patients issu des entretiens et les données mondiales des passagers aériens pour reconstruire le chemin de transmission de ce variant. Nous avons estimé que l'émergence du variant en Europe trouve son origine en Afrique centrale, très probablement la République du Congo, et qu'il a été transmis avec une détection minimale durant les premiers mois de 2021.

Dans les deux chapitres, nous dévoilons de nouvelles perspectives sur la manière dont la migration humaine influence la propagation des maladies infectieuses, soulignant l'importance d'intégrer les données génomiques avec les contextes épidémiologiques, sociaux et géographiques. Dans le premier chapitre, nous déployons des méthodes efficaces pour traiter les 14 000 séquences RABV de notre ensemble de données, éclairant sa propagation géographique et l'impact des mouvements humains. Dans le deuxième chapitre, nous approfondissons en employant des tests prédictifs plus rigoureux, en exploitant l'ensemble de données plus compact de séquences SARS-CoV-2. Dans le dernier chapitre, nous plaidons pour une synthèse de ces approches. En employant le modèle linéaire généralisé dans un cadre de Maximum de Vraisemblance, l'évaluation des prédicteurs pourrait être possible sur de grands ensembles de données de séquences, fournissant finalement des approches statistiquement pertinentes aux stratégies de contrôle des maladies.

Mots clefs :
Phylogénétique, Phylogéographie, Rage, SARS-CoV-2, Zoonose, Épidémiologie moléculaire, Éclosion, Dynamique des maladies infectieuses, Évolution virale.

Andrew Holtz
PhD Evolutionary Biology / 2023
Frontiére de l'Innovation en Recherche et Education
Université de Paris Cité

Résumé Substantiel en Français

Introduction

Dans cette thèse de doctorat, je me lance dans une exploration approfondie des maladies zoonotiques, qui sont des infections transmises des animaux aux humains et constituent une part significative de toutes les maladies infectieuses humaines. On estime que 60 % de ces maladies sont des zoonoses, acquises par divers moyens tels que le contact direct avec des animaux, la consommation d'aliments et d'eau contaminés ou l'exposition à des environnements abritant des agents infectieux. La nature cyclique de ces maladies leur permet de persister dans des réservoirs animaux et de réapparaître périodiquement dans les populations humaines, présentant une menace persistante pour la santé publique.

L'objectif de ma recherche est double : la rage et le SARS-CoV-2, tous deux des exemples de maladies zoonotiques capables de franchir les barrières d'espèces. La rage, presque toujours mortelle chez l'homme une fois les symptômes apparus, ne se transmet pas entre humains, tandis que le SARS-CoV-2, le virus responsable de la pandémie de COVID-19, est très contagieux parmi les humains. L'émergence de ces zoonoses est influencée par une confluence de facteurs, y compris les pratiques d'assainissement, la proximité avec la faune et les schémas de migration et d'urbanisation humaines. Pour disséquer la dynamique de la propagation des maladies et identifier les facteurs contribuant à l'émergence zoonotique, j'emploie des méthodes statistiques telles que les rapports de cotes et les études phylogéographiques.

Les activités humaines, notamment la déforestation, l'urbanisation et le commerce de la faune, ont été impliquées dans le débordement des maladies zoonotiques des animaux aux humains. L'adaptabilité des virus zoonotiques à de multiples hôtes et leur capacité de transmission par des vecteurs, illustrée par le rôle du moustique Aedes dans la propagation de la fièvre Dengue, augmentent considérablement leur potentiel de transmission. De plus, la migration humaine a été identifiée comme un facteur pivot dans la propagation des maladies zoonotiques, avec des camps de réfugiés émergeant comme des points chauds pour la transmission de maladies évitables par la vaccination. Des études ont lié la densité de la population humaine à la propagation de maladies comme la dengue et la rage, soulignant le rôle critique du mouvement humain et de l'urbanisation dans l'émergence des zoonoses.

Le changement climatique complique davantage ce paysage, permettant aux vecteurs tels que les moustiques et les tiques d'envahir de nouvelles régions, comme l'Europe du Sud et le nord-est des États-Unis, conduisant à une augmentation marquée des maladies à transmission vectorielle. Ce phénomène défie les systèmes de santé publique de s'adapter continuellement à ces schémas changeants. L'interconnexion de notre société mondiale signifie qu'une épidémie dans une région peut avoir des implications mondiales en quelques jours, mettant en évidence l'importance de l'approche « Une seule santé », qui intègre la santé humaine, animale et environnementale pour renforcer la sécurité sanitaire mondiale grâce à des actions et une gouvernance coordonnées.

Cette thèse aborde également les maladies tropicales négligées (MTN), telles que la rage, qui touchent principalement les pays à revenu faible et intermédiaire et passent souvent inaperçues malgré leur impact significatif. La nature cyclique de ces maladies, en particulier dans les régions où les populations de chiens ne sont pas vaccinées, peut conduire à des épidémies locales. Des efforts soutenus en matière de vaccination et l'adhésion à l'approche Une seule santé sont essentiels pour comprendre et atténuer la propagation de ces maladies.

Pour comprendre les processus complexes de l'émergence et de la propagation des maladies infectieuses, les modèles épidémiologiques mathématiques sont devenus indispensables. Ces modèles analysent des facteurs tels que la densité de population, les schémas de mouvement et l'impact des mesures préventives comme les vaccinations, permettant aux experts de prédire le cours d'une épidémie et l'efficacité des interventions potentielles. L'avènement de la surveillance génomique a révolutionné notre compréhension des maladies infectieuses, permettant le séquençage de nombreux organismes et améliorant notre compréhension de la phylogénétique. Depuis que Walter Fiers a séquencé le premier génome viral dans les années 1970, les technologies de séquençage ont évolué, devenant plus accessibles et rentables, cruciales pendant des épidémies comme Ebola.

Les mutations dans le matériel génétique des virus sont centrales pour la surveillance génomique. Elles peuvent se produire sous forme de mutations ponctuelles, d'insertions, de délétions ou par des mécanismes comme la recombinaison et le réassortiment. Ces mutations sont essentielles pour comprendre les voies évolutives des virus, car elles peuvent conférer des avantages tels

que la résistance aux traitements ou la capacité d'infecter de nouveaux hôtes. Par exemple, le taux de mutation rapide du VIH-1 lui a permis de développer une résistance aux monothérapies. Les taux de mutation parmi les virus varient, les virus à ARN mutent généralement plus rapidement que les virus à ADN.

L'alignement de séquences multiples est une étape fondamentale de l'analyse phylogénétique, impliquant la comparaison de séquences d'ADN, d'ARN ou d'acides aminés pour identifier les sites homologues et les distances évolutives. Cet alignement est crucial pour comprendre l'histoire évolutive complète d'un organisme. Les séquences génomiques complètes fournissent une image complète, englobant les régions codantes et non codantes du génome. Les mutations synonymes, qui n'altèrent pas la séquence protéique, sont contrastées avec les mutations non synonymes qui le font. Les régions non codantes, bien qu'elles ne soient pas directement impliquées dans le codage des protéines, peuvent influencer d'autres processus génétiques.

Les arbres phylogénétiques sont essentiels pour comprendre les relations évolutives entre les séquences, nous permettant d'inférer leur histoire évolutive. La construction d'arbres phylogénétiques est un processus complexe et intensif en calcul, utilisant des méthodes heuristiques telles que le voisinage pour les méthodes basées sur la distance et la parcimonie maximale, la vraisemblance maximale (ML) et les approches bayésiennes pour les méthodes basées sur les caractères. L'estimation ML recherche l'arbre qui maximise la vraisemblance des données observées sous un modèle d'évolution spécifique. Les modèles de substitution nucléotidique, comme le modèle général réversible dans le temps (GTR), sont centraux pour les calculs ML.

La datation d'un arbre phylogénétique implique de calibrer les longueurs des branches avec le temps, ce qui nécessite une horloge moléculaire. Les méthodes bayésiennes offrent une alternative à la ML, intégrant des connaissances antérieures et mettant à jour les croyances sur les paramètres de l'arbre en fonction des données observées. L'échantillonnage par chaîne de Markov Monte Carlo (MCMC) est une technique clé en phylogénétique bayésienne, utilisée par des logiciels comme BEAST, qui permettent des modèles évolutifs complexes et des tests d'hypothèses.

Andrew Holtz
PhD Evolutionary Biology / 2023
Frontiére de l'Innovation en Recherche et Education
Université de Paris Cité

Le virus de la rage (RABV) est un pathogène zoonotique responsable de la rage, une maladie avec un long historique et un défi actuel pour la santé publique mondiale. Malgré la disponibilité de vaccins efficaces, la rage reste endémique dans de nombreuses régions, avec des cas transmis par les canidés représentant plus de 99 % des infections humaines. Le génome à ARN simple brin du virus code pour cinq protéines et se propage à travers le système nerveux, s'avérant invariablement fatal une fois les symptômes manifestés.

Le RABV infecte principalement deux catégories d'hôtes : les Chiroptères (chauves-souris) et les Carnivores (canidés et autres mammifères comme les ratons laveurs et les mouffettes). Bien que les ongulés puissent être infectés, ils ne transmettent généralement pas le virus. Les infections humaines sont considérées comme des événements sans issue, mais la transmission interhumaine est théoriquement possible. La vaccination des chiens est la stratégie de prévention la plus efficace, agissant comme un tampon contre la transmission humaine. Pour les humains, la prophylaxie pré-exposition (PrEP) est recommandée pour les groupes à haut risque, et la prophylaxie post-exposition (PEP) est cruciale après une exposition potentielle.

Cette thèse se concentre sur les méthodes de contrôle épidémiologique éclairées par des études phylogénétiques et phylodynamiques, qui fournissent des aperçus sur la propagation du RABV et informent des campagnes de vaccination animale efficaces. La surveillance génomique a révélé deux grands groupes de RABV, avec le RABV maintenu par les canidés montrant une expansion mondiale significative. Les analyses phylogéographiques ont été instrumentales pour tracer les origines et la propagation des maladies, aidant dans les stratégies de vaccination ciblées. Cependant, de nombreuses études ont des limitations en raison de leur focalisation étroite, et cette thèse vise à surmonter ces limites en intégrant une gamme plus large de données génomiques et en examinant divers prédicteurs de la propagation du RABV.

En contraste, la thèse aborde également le SARS-CoV-2, le coronavirus responsable de la pandémie de COVID-19. Originaire de Wuhan, en Chine, le SARS-CoV-2 est un virus à ARN simple brin enveloppé. Les chauves-souris sont considérées comme le réservoir probable, avec un hôte intermédiaire inconnu facilitant le transfert zoonotique. Le virus a évolué, résultant en plusieurs Variants Préoccupants (VOC) et Variants d'Intérêt (VOI) qui ont augmenté la transmissibilité, la virulence ou la résistance aux vaccins.

Andrew Holtz
PhD Evolutionary Biology / 2023
Frontiére de l'Innovation en Recherche et Education
Université de Paris Cité

L'analyse phylogénétique a été cruciale pour comprendre l'émergence et la propagation des variantes du SARS-CoV-2, avec des données d'historique de voyage aidant à tracer les origines et les voies de transmission. Cette thèse vise à améliorer les méthodes d'inférence phylogénétique et phylogéographique en intégrant divers types de données, y compris l'historique de voyage des patients, pour mieux comprendre la propagation de maladies comme le RABV et le SARS-CoV-2.

La thèse propose d'utiliser des méthodes de vraisemblance maximale pour étudier les modèles de transmission historiques du RABV canin et l'analyse bayésienne pour enquêter sur le SARS-CoV-2 contemporain, visant à fournir une compréhension complète de la propagation des maladies et de l'influence de la migration humaine. L'objectif ultime est d'améliorer les stratégies de contrôle des maladies grâce à des informations basées sur les données et de développer une nouvelle méthode qui peut évaluer efficacement les interventions de santé publique.

Chapitre 1 : Élucider la Dispersion Historique du Virus de la Rage

L'objectif principal du premier projet était de comprendre la dispersion historique du virus de la rage (RABV). J'ai cherché à déterminer quand et où la souche actuelle du RABV adaptée aux canidés a émergé pour la première fois et à estimer sa propagation ultérieure à travers le monde. De plus, j'ai cherché à fournir un aperçu de la manière dont les différentes clades, les emplacements géographiques et les espèces hôtes sont représentés dans la chronologie évolutive du virus. Historiquement, les études phylogéographiques du RABV s'appuyaient sur des sous-ensembles de gènes spécifiques, tels que les gènes N et G, introduisant ainsi involontairement un biais. Pour remédier à ce problème, j'ai introduit une nouvelle méthode impliquant la concaténation de séquences au sein d'alignements de séquences multiples. Cette approche nous permet d'utiliser tous les gènes du génome du RABV, garantissant une analyse plus complète. J'ai validé la méthode de concaténation de séquences en comparant les résultats avec des études antérieures et en réalisant des sous-échantillonnages et des répétitions d'analyses. Les résultats ont présenté une nouvelle méthode d'inférence phylogéographique capable d'analyser d'importants ensembles de données de séquences à travers diverses régions de gènes, pays, dates et espèces hôtes. En utilisant cette méthode, l'étude a estimé le moment le plus précis du plus récent ancêtre commun (tMRCA) du RABV et les dates d'émergence dans

44 clades, pays et régions du monde entier. L'origine de cette souche de RABV a été estimée entre 1301 et 1403, et l'étude a dressé une carte des routes de transmission remontant aux années 1300. Cette analyse a également révélé des cas de migration à longue distance par les humains et suggéré que la colonisation européenne aux XVIIe et XVIIIe siècles expliquait partiellement la transmission du RABV chez les canidés. Le succès de ce projet résulte des efforts de collaboration dans la surveillance du séquençage du virus de la rage et de la validation par comparaison avec des études phylogéographiques antérieures sur le virus de la rage chez les canidés.

Chapitre 2 : Étude de l'Émergence de la Variante B.1.214.2 du SARS-CoV-2

Le deuxième projet visait à comprendre l'émergence d'une variante préoccupante du virus SARS-CoV-2, connue sous le nom de B.1.214.2. Cette variante a été initialement identifiée en Belgique et a suscité des inquiétudes internationales en raison de sa propagation rapide. La variante a été observée pour la première fois dans des zones métropolitaines de Belgique, de France et de Suisse, avec des preuves de voyages internationaux et de travailleurs étrangers contribuant à sa dissémination. Les entretiens avec les patients ont révélé que certains cas étaient dus à des voyages dans d'autres pays européens ou en Afrique centrale, en particulier en République du Congo. De plus, la B.1.214.2 s'est avérée relativement courante en Afrique centrale au début de l'année 2021, ce qui indique un nombre significatif de cas dans la région. Pour estimer l'origine de la variante, j'ai utilisé une analyse phylogéographique prenant en compte l'historique des voyages, en intégrant les entretiens avec les patients et les données des passagers aériens. La recherche a estimé que l'origine probable de la B.1.214.2 était en République du Congo vers juin 2020. Elle a ensuite été introduite en Belgique et en France entre août et novembre 2020, très probablement par le biais des voyages aériens, avant de se propager à d'autres pays européens. En collaboration avec un laboratoire d'immunologie de l'Université de Louvain (KU Leuven), l'étude a réalisé des tests immunologiques sur la variante. Étonnamment, ils ont constaté une immunité adaptative accrue chez les individus infectés par la B.1.214.2 par rapport aux autres variants du SARS-CoV-2 circulant à la même époque. Ce projet est en phase de finalisation et sera bientôt soumis aux revues scientifiques pour examen.

Andrew Holtz
PhD Evolutionary Biology / 2023
Frontiére de l'Innovation en Recherche et Education
Université de Paris Cité

Perspective

Cette thèse a démontré le potentiel de l'exploitation de la richesse des données de séquençage disponibles pour l'inférence phylogéographique et des facteurs prédictifs, en particulier en ce qui concerne l'influence de la migration humaine sur la dispersion virale. Le premier chapitre a souligné l'importance de traiter le biais des séquences et de la géographie dans les projets phylogéographiques. La nouvelle méthode de concaténation de séquences introduite dans le chapitre 1 a permis une analyse plus complète en incluant tous les gènes disponibles dans le génome, réduisant ainsi le biais. Cette approche pourrait être appliquée à diverses bases de données de séquences virales, atténuant à la fois le biais des séquences et de la géographie dans les reconstructions phylogénétiques et phylogéographiques. Le chapitre 3 a souligné que le biais géographique dans le séquençage n'était pas limité aux études historiques, mais avait également un impact sur les enquêtes contemporaines, comme le montre l'étude de la variante B.1.214.2 du SARS-CoV-2. L'analyse phylogéographique prenant en compte l'historique des voyages est apparue comme un outil précieux pour résoudre ce biais, avec des applications s'étendant à divers virus pouvant facilement traverser les frontières internationales.

La thèse a proposé un modèle linéaire généralisé (GLM) dans un cadre de vraisemblance maximale comme une avancée méthodologique significative. Cette approche pourrait faciliter l'analyse phylogéographique à grande échelle, permettant de tester de nombreux prédicteurs pour la dispersion virale. L'approche du GLM avait le potentiel de restructurer la manière dont les prédicteurs sont évalués dans la dynamique de propagation virale.

Limitations

Bien que ces études aient fourni des informations précieuses sur la propagation des virus zoonotiques, il convient de tenir compte de certaines limitations. La définition des frontières nationales, en particulier dans les études historiques, peut ne pas représenter avec précision la nature fluide des frontières au fil du temps. De plus, se fier à l'historique des voyages pour l'inférence phylogéographique pourrait introduire des inexactitudes, attribuant potentiellement l'historique des voyages comme l'origine géographique d'une souche de virus. Malgré ces limitations, la thèse a souligné la nécessité d'une meilleure collaboration entre la science et la politique pour mettre en œuvre des stratégies efficaces de lutte contre les maladies. Elle a souligné l'importance de comprendre les différents facteurs qui influencent la propagation des

maladies et le potentiel de nouvelles méthodologies pour améliorer notre compréhension et le contrôle des maladies infectieuses.

Conclusions

En conclusion, cette thèse a montré comment les avancées dans les données génomiques et les méthodologies phylogéographiques innovantes peuvent éclairer la propagation des virus zoonotiques. Grâce à des efforts de collaboration et à des approches interdisciplinaires, les chercheurs peuvent mieux comprendre l'impact de la migration humaine sur l'émergence et la dissémination des maladies infectieuses. Ces informations ont le potentiel d'informer les stratégies de lutte contre les maladies et de contribuer à la prévention des épidémies futures, favorisant ainsi la santé et la sécurité mondiales.

Andrew Holtz
PhD Evolutionary Biology / 2023
Frontiére de l'Innovation en Recherche et Education
Université de Paris Cité

Acknowledgements

To start, I would like to thank all the members of my jury for taking the time to listen to the research I have done over the last three years. Thank you to Katie Hampson and Denise Kühnert for being reviewers for this thesis and providing your expert feedback. I would also like to thank Sebastian Lequime for being not only a member of my jury, but also for being a member of my thesis advisory committee and providing very valuable insights over the course of my project. I would also like to strongly thank Eddie Holmes for providing invaluable support during a difficult time, providing edits and comments to prepare this thesis manuscript in addition to being a member of my thesis advisory committee.

I would thank the directors of my thesis Hervé Bourhy and Guy Baele for giving me the opportunity to work with them during the duration of my PhD. It was under difficult circumstances, and I am very thankful that they took me as their student. I want to express my special gratitude to Hervé for his unwavering support during some difficult times both at the beginning and conclusion of my thesis. You consistently ensured that I had the assistance I required, and I appreciate that.

I am incredibly grateful to Anna Zhukova for her patience, supervision, and kindness throughout the various bumps of my PhD, including the first difficult months. If it was not for Anna, I'm not sure where my PhD would be since she provided me with not only the computational guidance but also encouraged me to pursue different aspects of the research and motivated me throughout Although I arrived with very little computational skills, you allowed me to discover and develop. Thank you for being there for me over the course of the three years.

I would like to thank the members of the previous Evolutionary Bioinformatic group, who adopted me at the beginning of my PhD. Thank you for providing me with a sense of community- Marie, Luc, Jakub, Anna, and Fred. Thank you as well to Ruopeng Xie and Maylis Layan, and to the members of the Lyssavirus Epidemiology and Neuropathology group. I would also like to thank all the members of Guy's group in Leuven for welcoming me with open arms: Kanika, Sam, Barney, Nena

Andrew Holtz
PhD Evolutionary Biology / 2023
Frontiére de l'Innovation en Recherche et Education
Université de Paris Cité

Next, I would like to extend gratitude to the other PhDs/Post-docs who supported me throughout this experience- Maria Lopopolo, Katharina Kloppenborg, Vanessa Kramer, Mariatou Dramé, and Elizabeth Meloney, and Jamie Sugrue. Furthermore, I would like to thank my friends in Paris for their support- Chris, Jimmy, Joël, Andrew, Rafa and my friends beyond- Amir, Aube, Jake, Kathy, Ane, Rachel, Leire, Al, Tessa, Dana, and Tiffany.

Finally, I would like to thank my sister and especially my father and Joy for supporting me in every part of my education and emotional journey. Your constant encouragement and emphasis on science and curiosity has helped define who I am today. Thank you to my Sierra Leonean Mama for introducing me to foreign culture and language and sparking my interest to live abroad.

Lastly, I'd like to dedicate this work to my late mother. As a fellow scientist and academic, seeing photos of you presenting your research projects fills me with pride, and I'm honored to follow in your footsteps. I wish you could have been here to see this work.

Andrew Holtz
PhD Evolutionary Biology / 2023
Frontiére de l'Innovation en Recherche et Education
Université de Paris Cité

Scientific Output

Articles

Holtz, A., Baele, G., Bourhy, H. et al. Integrating full and partial genome sequences to decipher the global spread of canine rabies virus. Nat Commun 14, 4247 (2023). https://doi.org/10.1038/s41467-023-39847-x

Holtz, A., Hong, S.L., Van Weyenbergh, J. et al. Origin and geographic spread of the B.1.214.2 SAR-CoV-2 lineage with omicron-like 3AA spike protein insertion. PrePrint (2023).

Posters & Oral Presentations

A phylogenetic study on the emergence of rabies virus in Europe using Bayesian and maximum-likelihood inference. Epidemics. December 2021. Online poster. https://www.dropbox.com/scl/fi/r1q298xofmtmdkfq42n30/epidemics_poster_holtz.pdf?rlkey=koo5ypxdswrvsdcy4nheprthq&dl=0

Deciphering the global spread of canine rabies virus in the modern era. International Dynamics & Evolution of Human Viruses. April 2021. In person poster. https://www.dropbox.com/scl/fi/heggeyc500hhr6jlhipib/EvoVirus_Poster_Better.pdf?rlkey=sic3kdfk071spywdwldx9ia2n&dl=0

Integrating full and partial genome sequences to decipher the global spread of canine rabies virus. Epidemics. December 2023. In person poster

Teaching

PhiNDaccess: M3.1 Evolution and Phylogenetics. IP Tunis. Bayesian Phylogeographic. Lectures and Practical Session. 14 hours.

Invited Jury Member for Masters thesis
Duret, Loréna. Exploring Dengue Virus Genetic Diversity and Fitness in Southeast Asia. ENS. Oral Defense. 31 August 2021.

List of Abbreviations

ACE2	Angiotensin-Converting Enzyme 2
ACR	Ancestral Character Reconstruction
AIC	Akaike Information Criterion
BIC	Bayesian Information Criterion
BSSVS	Bayesian Stochastic Search Variable Selection
CI	Confidence Interval
DEIC	Dutch East India Company
DNA	Deoxyribonucleic Acid
DRC	The Democratic Republic of the Congo
ESS	Effective Sample Size
GLM	Generalized Linear Model
GTR	General Time Reversible
HIV	Human Immunodeficiency Virus
LSD	Least-Squares Dating
MAP	Maximum a Posteriori
MCMC	Markov Chain Monte Carlo
ML	Maximum Likelihood
MPPA	Marginal Posterior Probabilities Approximation
MSA	Multiple Sequence Alignment
NNI	Nearest Neighbor Interchange
NJ	Neighbor Joining
PCR	Polymerase Chain Reaction
PEP	Post-Exposure Prophylaxis
PrEP	Pre-Exposure Prophylaxis
RABV	Rabies Virus
RBD	Receptor Binding Domain
RC	The Republic of the Congo
RBR	RNA-Dependent RNA Polymerase
RIR1	Recurrent Insertion Region 1
RNA	Ribonucleic Acid
SARS-CoV-2	Severe Acute Respiratory Syndrome Coronavirus 2
SNPs	Single Nucleotide Polymorphisms
SPR	Subtree Pruning and Regrafting
TBR	Tree Bisection and Reconnection
tMRCA	Time to the Most Recent Common Ancestor
VOC	Variant of Concern

Andrew Holtz

PhD Evolutionary Biology / 2023

Frontiére de l'Innovation en Recherche et Education

Université de Paris Cité

VOI	Variant of Interest
WHO	World Health Organization
WT	Wild Type

Software

BEAST v1.10.5 [1]

FastTree v2.1.11 [2]

HyPhy v2.5.45 [3]

IQ-TREE2 v2.2.2.2 [4]

iTol v5 [5]

LSD2 v1.8.8 [6]

MAFFT v7.505 [7]

NextClade v1.7.1 [8]

Pangolin v1.2.81 [9]

PastML v1.9.34 [10]

RABV-GLUE 1.1.107 [11]

TempEST v1.5.3 [12]

TreeTime v0.8.6 [13]

Andrew Holtz

PhD Evolutionary Biology / 2023

Frontiére de l'Innovation en Recherche et Education

Université de Paris Cité

Table of Contents

Andrew Holtz

PhD Evolutionary Biology / 2023

Frontiére de l'Innovation en Recherche et Education

Université de Paris Cité

Andrew Holtz

PhD Evolutionary Biology / 2023

Frontiére de l'Innovation en Recherche et Education

Université de Paris Cité

1 General Introduction

1.1 Thesis organization

This thesis contains two main chapters. The largest of the two is an investigation of the historical spread of canine rabies virus using fast methods in maximum likelihood. The later chapter involves the estimation of the origin and spread of a variant of SARS-CoV-2 which spread in Europe and beyond in 2021 using rigorous methods in Bayesian statistics. The theme that connects these two topics is the use of large data sets including sequence, migration, and geographic data to build strong models to understand the impact of factors in the spread of human disease. The two projects use different methods: two different viruses, over different time frames, using different mathematical frameworks. I then brings these approaches together as a perspective within the discussion section, suggesting a method to incorporate a predictor evaluation to a maximum likelihood framework to allow for rigorous covariate testing on large sequence data sets. I first introduce the concepts that are fundamental to this thesis.

1.2 Background

1.2.1 Infectious Disease Overview and Control

1.2.1.1.1 Definition of Infectious Disease

Humans do not live alone. We are intertwined with a rich tapestry of visible and invisible life forms. These intricate communities of yeast, bacteria, and viruses envelop our world and permeate our environment, including our own bodies, and we rely on these microorganisms for the fundamental workings of life on Earth, from the nitrogen cycle to human digestion. Yet, driven by their natural biology and ecology, these communities can sometimes adversely affect our well-being, giving rise to diseases. Pathogenic microorganisms may emerge due to disruptions in our own resident microbial communities, as seen in skin infections like *Staphylococcus aureus*, or from more foreign invaders like influenza, Ebola, HIV, or *Clostridium botulinum*. Their spread hinges on expansive networks spanning homes, schools, towns, and nations, occasionally involving additional species and co-infecting agents such as HIV and Mycobacterium tuberculosis[14,15], Staphylococcus aureus (pneumonia) and RSV [16].

Andrew Holtz
PhD Evolutionary Biology / 2023
Frontiére de l'Innovation en Recherche et Education
Université de Paris Cité

These microorganisms can disrupt the body's normal function and trigger a range of symptoms, varying from mild to severe, depending on the pathogen and the individual's immune response. By the necessity to replicate and transmit, they have adapted direct and indirect modes of transmission, being either between an infected individual and a susceptible person, or through exposure to contaminated surfaces, food, water, or animal vectors like insects. The course of an infectious disease may involve an incubation period during which the pathogen multiplies and progress and in the host body, followed by the onset of symptoms and potential progression to more severe stages if left untreated. Prevention and control strategies have been established to combat infectious diseases such as hygiene practices, antimicrobial treatments, and public health measures to mitigate the spread of these diseases within populations [17].

Infectious diseases can be endemic in a population by being consistently present at a baseline level in a specific geographic area or population group. They might not always cause outbreaks, but they maintain a steady presence within the population. Examples of endemic diseases include dengue fever in parts of Southeast Asia and South America, and Chagas disease in parts of Latin America [17]. Seasonal diseases such as influenza also fall into this category. An epidemic is the occurrence of a significantly higher number of cases of a particular disease within a specific population, geographic area, or community than what is normally expected, potentially indicating a rapid and widespread outbreak. Epidemics can be localized to a certain region or population, or they can affect larger areas, even crossing national or international boundaries. They can be caused by various factors, such as the introduction of a new infectious agent, changes in the behavior of the disease-causing organism, alterations in human susceptibility to the disease, or shifts in environmental conditions that promote the spread of the disease [17].

1.2.1.1.2 What is a Zoonosis

Introduction to Zoonoses

Zoonoses are infectious diseases transmitted from non-human animals to humans, caused by various pathogens including bacteria, viruses, and parasites. These diseases can be acquired through direct contact with animals, consumption of contaminated food and water, or exposure to a tainted environment [18,19]. Such diseases pose a significant global health concern due to humans' extensive interactions with animals in various domains like livestock, pet ownership, and wildlife. It is estimated that 60% of human infectious disease are zoonoses [19–21] .

Andrew Holtz
PhD Evolutionary Biology / 2023
Frontiére de l'Innovation en Recherche et Education
Université de Paris Cité

Origin and Persistence of Zoonotic Diseases

Zoonoses have a cyclical nature, often cycling through human, wildlife, and livestock populations, constantly reemerging and disappearing. This poses significant threats, as zoonoses can persist in non-human hosts even after they have been eliminated from human communities. Consequently, these diseases can resurge within human populations years after apparent elimination. This is in stark contrast to diseases that exclusively affect humans, such as smallpox and polio, which are the only infectious diseases of humans that have been successfully eradicated [20]. Although zoonoses can result from spillovers in domestic animals and livestock, 71.8% of zoonosis are the result from spillover events from wildlife species [22]

In this thesis I focus on two main zoonoses: rabies virus and SARS-CoV-2. These both jump between species and cause infections. In humans, rabies is terminal, meaning the disease is always fatal and there are no human-to-human transmissions. SARS-CoV-2, on the other hand, is a zoonosis that can spread easily from person to person.

Factors that amplify zoonotic emergence

The study of zoonotic emergence investigates certain factors that can stimulate the emergence of infectious diseases. There are several different statistical methods to test the association of predictors to the emergence of a disease. Most traditionally in epidemiology, odds ratios are used to understand the likelihood of a factor being associated to another [17]. The first study including odds ratios was John Snow reviewing cases of Cholera in London, where he determined that households using certain water pumps were more likely to contain cholera cases [23]. This type of study has been extrapolated in modern epidemiology to understand infectious disease spread such as SARS-CoV-2. Odds-ratios in retrospective studies are very powerful to understand disease dynamics and are used to measure the effects of certain sanitation practices on the emergence and control of SARS-CoV-2 or other respiratory viruses [24,25]. We can now study disease emergence using phylogeographic studies by reviewing factors that help explain the geographic dispersal and evolutionary history of infectious disease (the specifics of these test are explain in later chapters) [26].

In addition to sanitation practices, proximity to wild animal markets, farms, local sanitation or water sources, and the forest/wild borders have been associated to zoonotic disease emergence 24

Andrew Holtz
PhD Evolutionary Biology / 2023
Frontiére de l'Innovation en Recherche et Education
Université de Paris Cité

[20,21]. Zoonotic spillover from wildlife can also be mediated by biodiversity loss and habitat degradation [27,28].

More specifically, studies have linked ecological and environmental data such as landscape, precipitation, temperature, and humidity to the emergence of infectious diseases. As an example, Kiltting et al. were able to build a relationship between temperature, precipitation, and the presence of pastures and the spread of Lassa virus in West Africa. Using these data, they then built a projection for the regions of Africa that are potentially susceptible to the virus [29].

Human migration (travel, displacement) and urbanization have been strongly linked to zoonotic infectious disease emergence. In Lemey et al. the influence of air passenger flow and human mobility was highlighted in a phylogeographic study, where they used these data alongside that of virus evolution and the spread of influenza virus [26]. Refugee camps that provide needed refuge from war and natural disasters have become central cogs in infectious disease transmission, most often diseases that are vaccine-preventable [30]. A scoping review in 2021 found that migrant populations in the European Union and European Economic Area were often victim to outbreaks of measles, varicella, hepatitis, rubella and mumps [30]. Similarly, a study from 2019 found an increase in diarrheal disease in refugee camps due to crowding and lack of proper sanitary conditions [31].

In Rabaa et al. human population density was linked to the spread of dengue 2 virus in Vietnam using 168 full length DENV-2 genome sequences, with the influence of population density estimated using a Bayesian approach (i.e., the BEAST software). They found that there were higher rates of viral exchange across a gradient of human population density [32,33]. Dellicour et al. obtained similar results on Rabies virus using the generalize linear model in BEAST [26], highlighting how viral lineages seem to be introduced and maintained in human accessible areas associated to high human population density. This shows how human migration can influence the introduction and transmission of canine rabies in wildlife [34].

Wild animal markets, farms, areas proximate to wilderness, and regions witnessing rapid urbanization and habitat destruction are all potential hotspots for the emergence of these diseases [20,21]. Furthermore, the adaptability of zoonotic viruses across multiple hosts enhances

their transmission potential, especially in immunologically naïve populations. For instance, Dengue fever can spread wherever its vector, mosquitoes of the genus *Aedes*, can establish. Given a mosquito's ability to covertly travel in international shipments, introduction of the mosquito and the diseases it can transmit becomes effortless [35]. Wild animal movements are important for the spread of certain diseases such wild board in Belgium and the spread of African Swine Fever [36] and foxes and racoons in the spread of RABV [37,38].

As international travel and trade increases, so does the threat of emergent infectious disease. In recent years, there have been several examples of epidemics that were able to emerge partially due to changes in human behavior. The recent SARS-CoV-2 pandemic and the Ebola Epidemic in West Africa of 2013-2015 are both hypothesized (although still debated) to have emerged from interactions between humans and wild-life through the trade in wild-life and live animal markets and/or bushmeat. Lassa virus, Ebola virus, rabies virus and other related lyssaviruses, as well as SARS-like viruses, are just a small sample of the large group of viruses that are more likely to emerge given more interactions with wildlife populations. These vector-borne zoonoses are affected by the ecology of the vector, the range of the host and complex environmental factors which make emergence prediction challenging.

Climate change and human activity will only continue to support the spread of infectious disease. As many arthropod vectors such as mosquitos and ticks are ectotherms, they are positively affected by temperature increases, allowing them to inhabit warming areas such as southern Europe, and parts of Northeast United States [19,35,39,40]. The US CDC has reported a 200% increase in the number of tick-borne diseases in the US in the last decade, and an increase in the geographic spread of illness [41]. Similarly, Europe has seen a more than double increase in the number of cases of Lyme borreliosis, tick-borne encephalitis, and Crimean-Congo fever since 2009 [42,43]. Between 2010 and 2021, there have been autochthonous outbreaks of dengue virus in Croatia, France, Italy, and Spain [44], and it is widely hypothesized that these outbreaks will continue to increase in occurrence and size [35] . The increase of tropical storms and consequent flooding can also impact the spread and density of infectious diseases. One study found that after flood events, there are increases cases of leptospirosis, campylobacter, and cryptosporidiosis [45] and an increase in habitats for mosquito reproduction.

Andrew Holtz
PhD Evolutionary Biology / 2023
Frontiére de l'Innovation en Recherche et Education
Université de Paris Cité

Lastly, the increase in global trade and tourism facilitates long-distance disease transmission between animals, humans, and vectors. As people and goods move across borders, they can unknowingly carry pathogens with them, allowing diseases to spread to new regions and to new host populations, increasing vulnerability. The high rate of mutation in some viruses enables them to overcome existing immune defenses and infect individuals who may not have encountered them before. This adaptability poses a significant challenge for public health systems, as they need to constantly monitor and respond to emerging infectious diseases. Furthermore, the interconnectedness of our world means that a disease outbreak in one part of the world can now have global repercussions within a matter of days.

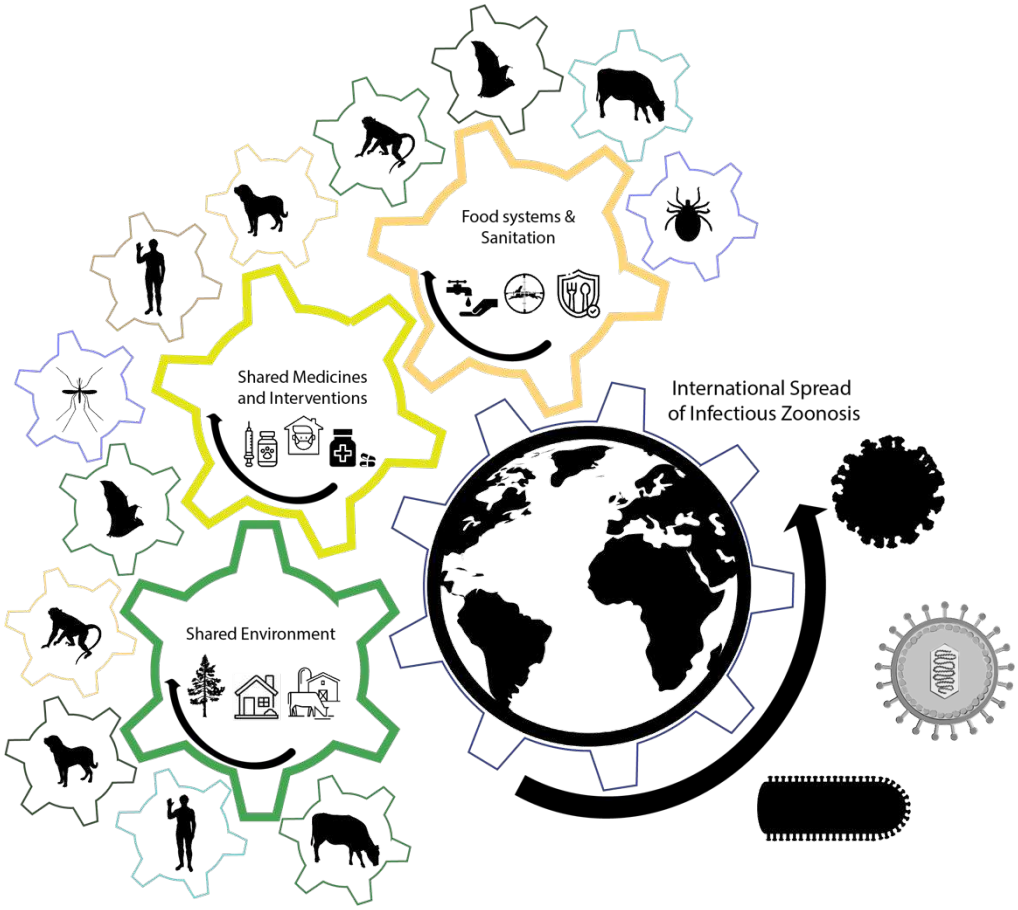


Figure 1. One Health influence on the spread and maintenance of infectious disease*. Humans, livestock, companion animals, and wildlife connect through their environments. Their interactions*

Andrew Holtz
PhD Evolutionary Biology / 2023
Frontiére de l'Innovation en Recherche et Education
Université de Paris Cité

influence how environment, interventions, and food systems function and these together can result in how infectious diseases are able to emerge and sustained in communities, environments, and in the world.

The "One Health" Approach

In the face of such complex challenges, the "One Health" approach has emerged over the past two decades as a comprehensive strategy to understand and hopefully help prevent disease emergence. It underscores the deep interconnections between human, animal, and environmental health. By recognizing these interplays, "One Health" provides a holistic framework from prevention to response, bolstering global health security through integrated governance, open communication, and coordinated actions [46,47] (Figure 1).

1.2.1.1.3 What is a Neglected Tropical Disease?

In addition to being a zoonosis, rabies is also categorized as a neglected tropical disease (NTD). These diseases primarily affect populations in lower-middle income countries, and they often emerge as silent epidemics, drawing limited attention despite their significant impact on the affected communities. NTD epidemics reveal ongoing disparities in healthcare access and underscore the need for targeted efforts to address these pervasive yet often overlooked health issues. Despite efforts to tackle the burden of NTD, 1.65 billion people were reported to require mass or individual treatment and care for NTDs [48]. Although many NTDs can be considered endemic, a pattern of emergence, transmission, and recovery is a common cycle seen in many regions, especially with diseases such as RABV. In some regions, particularly areas with free roaming and/or unvaccinated dog populations, rabies can lead to periodic local epidemics. These epidemics can occur when an infected animal, often a dog, transmits the virus to other animals or humans, creating a chain of transmission. During such outbreaks, the disease can spread relatively quickly within a free-roaming community. Due to the nature of rabies and its transmission dynamics, these local epidemics eventually subside once the infected animals die off and if preventive measures like vaccination are put in place, such that the transmission cycle can be broken. This can lead to periods of lower rabies activity or even temporary absence of new cases. The cyclic nature of rabies within and across different species underscores the

importance of sustained efforts in One Health to understand its spread and emergence dynamics (Figure 1) [49,50].

1.2.1.1.4 Disease and genomic surveillance

Many of the studies listed above studied factors that sustain viral emergence using mathematical epidemiological models. By analyzing data on disease transmission, researchers can identify patterns and factors that contribute to the emergence and propagation of infectious diseases. Mathematical models provide insights into the dynamics of outbreaks, such as the reproduction number, infectious period, and serial interval, allowing experts to simulate different scenarios and assess the potential effectiveness of interventions. These models consider variables such as population density, movement patterns, and the effectiveness of preventive measures like vaccination and quarantine. By understanding the intricate interplay between these factors, public health officials can develop targeted strategies to mitigate the impact of emerging infectious diseases, ultimately safeguarding communities and preventing widespread outbreaks.

Starting with the HIV pandemic, the addition of genomic surveillance has played a pivotal role in enhancing our understanding of infectious disease emergence. By analyzing the genetic material of pathogens, such as viruses, researchers can uncover crucial insights into their origins, evolution, and patterns of transmission. This involves studying the genetic sequences of pathogens sampled from various cases across different geographic locations and time periods. Phylogenetics and phylogeography are powerful tools within genomic surveillance that help trace the origins and patterns of spread of an outbreak. Phylogenetics involves estimating evolutionary trees to trace genetic relationships between different strains of a pathogen. Phylogeography, on the other hand, uses these phylogenetic relationships along with geographic information to map the spread of a pathogen across spatial regions. Together, these approaches enable scientists to identify the initial source of an outbreak, track its geographic expansion, and even discern how it adapts over time. Many new efforts in disease surveillance include the isolation of DNA or RNA of infectious agents to understand how the disease is evolving and to generate real-time epidemiological information that can be used by public health officials in actionable and meaningful ways.

Andrew Holtz
PhD Evolutionary Biology / 2023
Frontiére de l'Innovation en Recherche et Education
Université de Paris Cité

1.2.2 Genomic Surveillance

1.2.2.1.1 Nucleic Acid Sequencing

Walter Fiers, a Belgian molecular biologist at the University of Ghent was the first to ever sequence a complete nucleotide sequence of a gene in 1972 called Bacteriophage MS2 Coat Protein [51] and he continued on to sequence the entire RNA viral genome in 1976 [52]. Since then, researchers have applied and expanded these methods greatly, allowing for the sequencing of genes and genomes of many living organisms. This was a trigger point for modern research, including phylogenetics. In a later chapter, I will explore how sequences are specifically employed in this field. The use of sequencing to understand evolutionary relationships between organisms has surged.

Since the inception of sequencing technology, there have been significant improvements in accessibility, cost, and quality. Before the Ebola epidemic of 2013-2016, viral sequencing primarily focused on specific virus genes only, resulting in extensive databases of multiple gene sequences per virus. Partial sequences were often deemed adequate due to their cost-effectiveness and speed, proving particularly useful in field studies during ongoing epidemics. They also sufficed for diagnostic purposes. Comparative genetic studies often relied on a single, evolutionarily diverse gene to gain insights into viral evolution: while powerful, it is not possible to compare evolutionary history across genes. Whole-genome sequencing solves this problem, paving the way for phylogenetic analyses based on complete genome. However, even with technical developments, such as novel sequencers including the MinION sequencer from Oxford Nanopore Technologies [53,54], which enables on-site sequencing in various environments for real-time analysis, whole-genome sequencing for many viruses still lags the submission of partial sequences. Although whole-genome sequencing is gradually increasing, the number of WGS compared to partial sequences for most several neglected tropical disease viruses, is small (Figure 2). Importantly, by selecting just WGS, the number of sequences included in a study is substantially reduced and could introduce a geographic bias, especially if certain genetic regions are predominantly sequenced in specific geographic locations compared to others.

Andrew Holtz
PhD Evolutionary Biology / 2023
Frontiére de l'Innovation en Recherche et Education
Université de Paris Cité

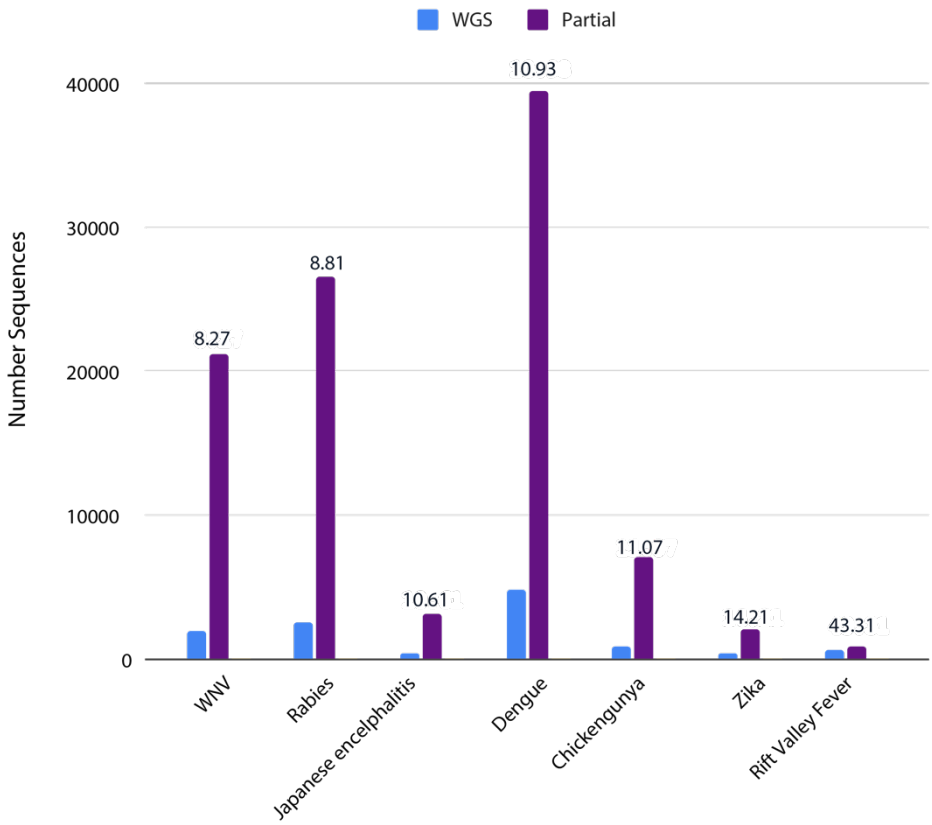


Figure 2. *Neglected tropical disease viruses explained by the number of whole-genome sequence and partial sequence submissions. Data is according to sequences submitted to NCBI GenBank and by NCBI Virus as of September 2023*

Andrew Holtz

PhD Evolutionary Biology / 2023

Frontiére de l'Innovation en Recherche et Education

Université de Paris Cité

1.2.2.1.2 Mutations in DNA and RNA

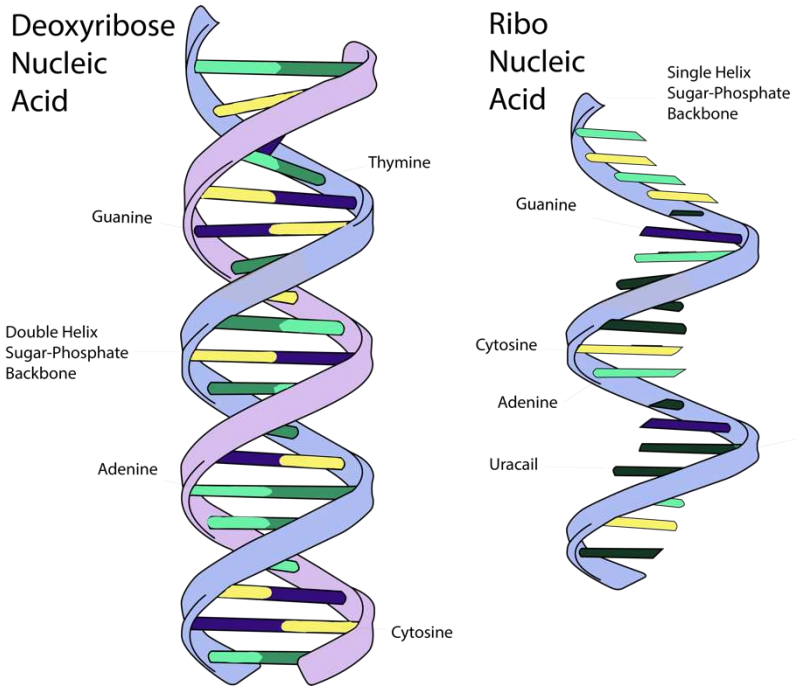


Figure 3. Deoxyribose Nucleic Acid (DNA) and Ribonucleic Acid

Codons are a pattern of three nucleotides of RNA that are translated into protein. They are defined by the originating DNA sequence (Figure 3) (RNA viruses are an exemption). When a mutation occurs in DNA there is the risk for a subsequent mutation in the codon and protein [55]. To limit structural mutations, several three nucleotide long codons bring rise to the same amino acid. This means that mutations, called synonymous mutations, are less likely to cause an amino acid change. Synonymous mutations are less likely to cause consequent effects in function for the organism, and can, therefore, remain in the genome silently. Mutations in the first or second codon position are more likely to cause an amino acid change and consequently cause a structural and functional change, which is more likely to affect the organism. These mutations can be advantageous, deleterious, or neutral.

Mutation Type	**Description**
Point Mutations	Single nucleotide substitutions that can lead to missense

Insertions/Deletions	Addition or removal of nucleotides
Repeat Expansions	Repeated DNA sequences expand
Frameshift Mutations	Indels alter reading frame
Reassortment	Exchange of genetic segments between two co-infecting viruses
Recombination	Exchange of genetic material between different organisms

Table 1. Mutation Types and Descriptions for Viral Sequences

In phylogenetics, we are focused on all mutations that arise in the DNA/RNA sequences, even if they have no structural and function effect. In viruses, the four most common type of mutation types are point mutations, insertions or deletions, repeat expansions, and frameshift mutations (Table 1). Many viruses also undergo recombination or reassortment, which are mechanisms by which genetic material can be exchanged and reorganized. Reassortment is specific to viruses with segmented genomes and occurs when two viruses infect a cell simultaneously and their genetic segments mix and match during the assembly of new virions. In contrast, recombination involves the exchange of genetic material between two similar or identical nucleotide sequences, leading to a hybrid sequence that contains elements of both original sequences. HIV serves as a prime example of recombination. When a cell is infected with two different strains of HIV, the reverse transcriptase enzyme can switch between the two RNA templates during replication, leading to the production of a recombinant viral genome. These mutations can arise from several types of events but occur most often during replication [56].

Different functions such as viral polymerase fidelity, cellular microenvironment, replication mechanisms, proof reading capacity, all can introduce mutations. As in all organisms, most mutations in viruses are deleterious, greatly reducing fitness. However, some mutations may be advantageous, allowing the organism to escape immune responses, colonize new ecological niches and evolve to become more infectious, pathogenetic, transmissible. For example, HIV-1 is a fast-mutating virus that successfully evolved resistance against monotherapies [57,58]. Mutation rates among viruses vary (Table 2). In general, RNA viruses mutate faster than DNA viruses due to the presence of the complimentary strain which allows for some proofreading mechanism [57,59]. Each mutation that remains on the genome can be passed down to future generations, which gives scientists a method to track the genealogy and lineages of these viruses. For example, the unique mutational profile in the HIV genome of an individual (the unique sequence) has been

Andrew Holtz
PhD Evolutionary Biology / 2023
Frontiére de l'Innovation en Recherche et Education
Université de Paris Cité

used in the past to rebuild transmission chains for HIV [60,61]. We can take advantage of the mutations that arise in genomes to dig deep into their past. By aggregating these mutations together and comparing them across other similar sequences, we can better understand where, when, and sometimes by what factors did viruses spread.

In zoonoses, host adaptations can arise and disappear in the viral genome depending on the host species, and the underlying mechanisms are not completely understood. Troupin et. al highlights host-adaptation in the rabies virus genome among ferret-badgers and mongoose populations. They report amino acid mutations associated with emergence of rabies virus in Asia and differences in evolutionary rates among ferret-badgers and mongoose [62]. Influenza also undergoes host-adaptation. Swine and avian flu lineages normally remain circulating among their targeted host, although with occasional spillover into humans, giving rise to potential influenza epidemics (i.e. 2009 Swine Flu-H1N1, Avian Flu H5N1 spillover events) [63,64].

Class	Virus	Genome size (kb)	Average mutation rate
ss(+)RNA			
	SARS-CoV-2	29.9	2.0 × 10-6[65]
	Human norovirus G1	7.65	1.5 × 10−4
	Hepatitis C virus	9.65	3.8 × 10−5
ss(−)RNA			
	Influenza A virus	13.6	2.5 × 10−5
	Rabies Virus	11.9	2.0 × 10−4[66]
	Measles virusd	15.9	3.5 × 10−5
Retrovirus			
	HIV-1 (free virions)	9.18	6.3 × 10−5
	HIV-1 (cellular DNA)	9.18	4.4 × 10−3
	Foamy virus	13.2	2.1 × 10−5
dsDNA			
	Herpes simplex virus	152	5.9 × 10−8

Table 2. Mutation rates of different virus types (Adapted from Sanjuán et. al) [57]

Andrew Holtz
PhD Evolutionary Biology / 2023
Frontiére de l'Innovation en Recherche et Education
Université de Paris Cité

1.2.2.1.3 Multiple Sequence Alignment

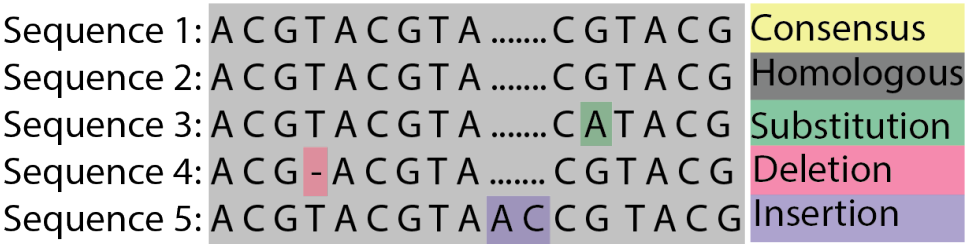


Figure 4. Example sequence alignment with common mutation types shown.

The first step in phylogenetics is multiple sequence alignment. One of the most powerful tools in comparing evolutionary history is a multiple sequence alignment (MSA). DNA, RNA or amino acid sequences can be compared by visually or computationally analyzing similarities and differences. Homologous sites, which refer to positions in RNA or DNA that exhibit the same nucleotide identity, can be stacked one on top of the other forming columns. This stacked aggregation of sequences allows for evolutionary distance to be observed by the number of mutations across the sites for the sequences. The sites in a sequence can only be compared to other sequences which also contain those same sites. Mutations such as those discussed in the previous section (Figure 2) then become visible in the MSA (Figure 3). It is, therefore, important to determine your target sequence for an analysis. Gaps can be introduced into sequences to account for deletions or missing data.

In the context of whole genome sequences (WGS), the entire genome is analyzed, encompassing both non-coding and coding regions. Coding regions generally accumulate fewer mutations because they are pivotal for the organism's function. Within coding areas, two primary types of mutations arise based on their influence on protein sequence: synonymous and nonsynonymous mutations. Synonymous mutations are alterations in the DNA sequence that do not change the amino acid sequence of the protein, often due to the higher tolerance at the third position of a codon. This position can mutate more frequently without altering the resulting amino acid. In contrast, nonsynonymous mutations lead to a modification in the amino acid sequence of the protein. In noncoding regions, while mutations may not directly impact protein translation, they can affect processes such as transcription factor binding. Given these differences across the genome, accurate alignment is vital for any phylogenetic study, as it reveals the complete history of mutations and, therefore, evolutionary history. MSAs stack these mutations along the entire

genome, but since sequences contain hundreds or thousands of nulceotides, this is a computationally demanding task to determine the alignment with the highest likelihood. Different MSA approaches weigh sequence evolution factors differently, some optimizing gap penalties to better account for insertions/deletions, while others focus on conserved motifs or regions to anchor alignments. Therefore, sophisticated software like MAFFT [7], T-Coffee [67], and MUSCLE [68] exist to address this task, each employing unique algorithms and heuristics to optimize sequence alignment.

1.2.2.1.4 Phylogenetic Inference

Multiple sequence alignments provide a blueprint for the evolutionary connections between sequences of interest. These connections allow for phylogenetic inference, which is the reconstruction of evolutionary history and relationships between different organisms, species, and subspecies, depicting their common ancestors and the branching that leads to their diversification. Although this was originally only based on morphological traits [69,70], we can now make use of nucleic acid sequences that define these morphological traits to redraw evolutionary relationships. These evolutionary histories are best presented by phylogenetic trees.

The evolutionary distance between two taxa can be calculated by summing branch lengths on the shortest path connecting them in a phylogenetic tree. Branch lengths normally represent the number of substitutions per site in the nucleic acid sequence. A dated tree, however, is a type of phylogenetic tree where branch lengths correspond to chronological time rather than genetic change alone. On a dated tree, branch lengths often represent time following a molecular clock principle. The molecular clock is the hypothesis in molecular evolution that proposes a relatively constant rate of genetic change or mutation over time [71]. Unrooted phylogenetic trees simply show evolutionary relationships between taxa, while a rooted tree contains the most common ancestor at the root position, or the most ancestral node of the tree. In phylogenetic inference, we often use rooted bifurcating (binary) trees. These are trees made of nodes that have exactly two daughter branches, meaning the tree splits into two distinct lineages from each node. A multifurcating tree, on the other hand, is made up of nodes with more than two branching. This is also called a polytomy and is often the result of unresolved relationships in the tree or instances

Andrew Holtz
PhD Evolutionary Biology / 2023
Frontiére de l'Innovation en Recherche et Education
Université de Paris Cité

where there is insufficient data in the tree to make a confident relationship approximation [72]. The inference of a phylogenetic tree by sequence data can be computationally intensive even for a small number of taxa. As the number of taxa increases the computing time increases significantly.

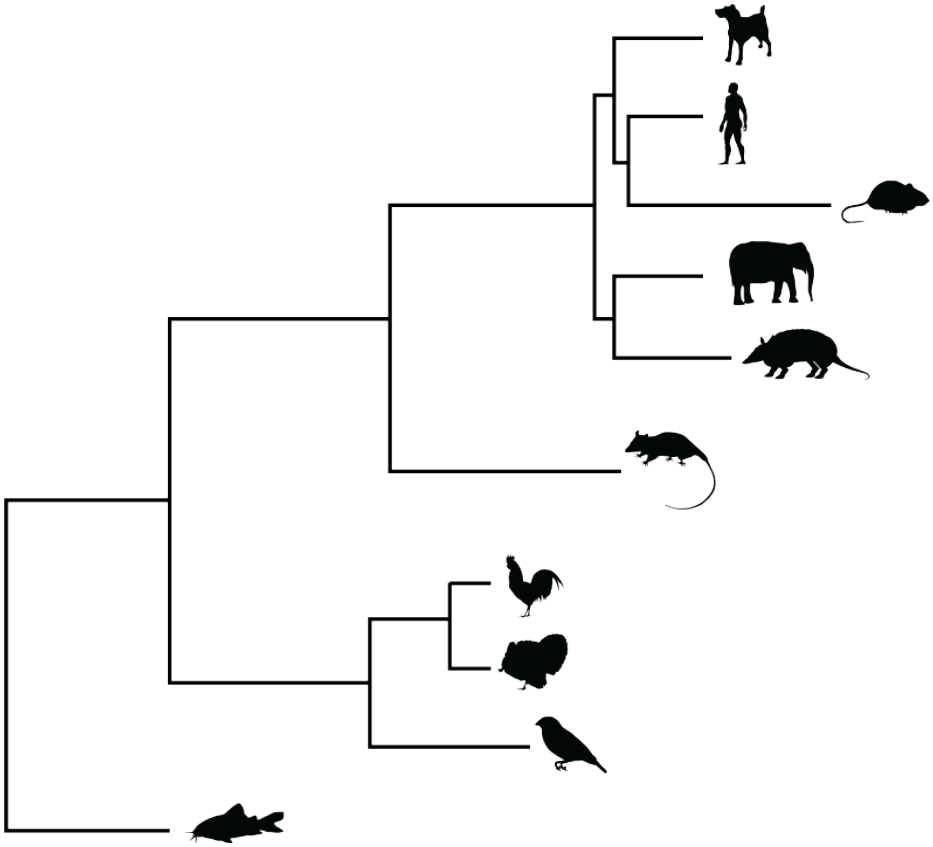


Figure 5. *Phylogenetic tree example showing mammals, birds, and fish. Adapted from Amemiya et al. Silhouettes are from phylopic.org under their CC0 1.0 Universal Public Domain Dedication license.*

There are $(2n-3)!/(2n-2(n-2))!$ number of possible bifurcating trees for n number of taxa. To demonstrate the exponentiality of this estimate, for a tree with just 5 taxa, there are 210 possible tree configurations [73]. Multiply this by ten, so 50 taxa, and there are ~4×10^{150} possible tree formations, which is greater than the number of atoms in the universe. Evaluating all potential trees, therefore, is impractical, and tree-construction methods have been developed based on heuristics to search the tree space. These methods encompass distance-based methods such as neighbor-joining and character-based methods, such as maximum parsimony, maximum likelihood, and Bayesian approaches, with our focus in this thesis being on the latter two.

Andrew Holtz
PhD Evolutionary Biology / 2023
Frontiére de l'Innovation en Recherche et Education
Université de Paris Cité

1.2.3 **Phylogenetics ML Methods Overview**

1.2.3.1.1 ML Tree Estimation Theory

Maximum Likelihood (ML) estimation of a phylogenetic tree is a character-based method that aims to find the tree with the highest likelihood within the tree space, considering a specific statistical model of evolution. The idea of ML is to test many trees in the tree space to find the tree topology that maximizes the probability of observing the character states (e.g., DNA bases) at its leaves given a model of evolution.

A likelihood function (L) represents the conditional probability of the observed data (i.e., aligned sequences) given parameters θ, which constitute the specific hypothesis for the relationship of the sequences:

$$L(\theta) = P(\text{Data} \mid \theta)$$

Where θ consists of the tree including branch lengths, the substitution model (m), and additional parameters [72,73].

Nucleotide Substitution Model

A central component of the maximum likelihood calculation (as well as Bayesian algorithms), is the evolutionary model which outlines the substitutions that occur in our data, or in DNA sequences (or RNA). Commonly used models are continuous-time Markov models [74] of the GTR (general …) family. They can be represented by the Q matrix, which defines the relative rates of change of each nucleotide in a nucleic acid sequence and it consists of a 4x4 matrix, each nucleic acid of DNA represented on the columns and rows (Figure X). There are three central assumptions: (1) each site evolves independently, (2) sites are time reversible, meaning that given i and j as nucleic acids, the probability of evolving from i to j over time t weighted by the equilibrium frequency of i (πi), is equal to the probability of evolving from j to i over the same time weighted by πj: $\forall i,j \in \{A,C,T,G\}: \pi_i P_{ij}(t) = \pi_j P_{ji}(t)$, and (3) evolution is site-independent, i.e., consistent across the entire sequence (or genome).

Andrew Holtz
PhD Evolutionary Biology / 2023
Frontiére de l'Innovation en Recherche et Education
Université de Paris Cité

In the equation (Figure X), we can estimate the rate of each transition. For example, for DNA the rate of the transition from C to A can be estimated by $a\pi_c$, where *a* describes the occurrence of the substitution of A to C [72]. Each entry represents the rate among all substitutions from *i* to *j*. It is then possible to calculate the probabilities of change between bases over time t by P(t) = e^{Qt} or $e^{\Lambda t}$, where Λ represents the generator matrix of the process.

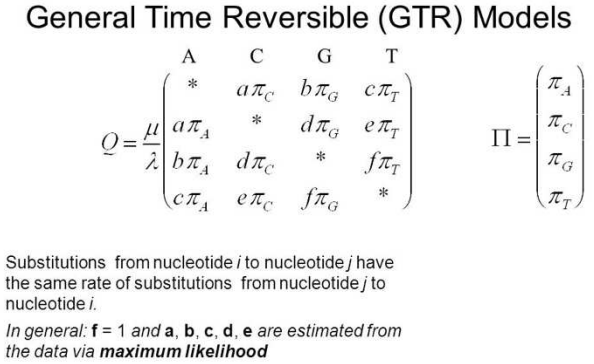


Figure 6. Markov Model, GTR, showing Q matrix. Nucleic acids of DNA are shown as columns, and the instantaneous substitution rate is shown. From: Computational Molecular Evolution

P(t) is used in tree likelihood calculations, as described in the next subsection.

There are numerous substitution models which can be used to model the evolution of the observed data (aligned sequences). They all describe mutations over time and consist of evolutionary rates, *R* and equilibrium frequencies π. At the core is the general time reversible model (GTR) [75], which is considered more general, and realistic compared to simpler models, such as the Jukes-Cantor [76](assuming equal rates of substitution for all types of nucleic acid changes), because it allows for different substitution rates and base frequencies at each site of the DNA sequence. However, in situations where the available data is limited, such as with a small data set or a short sequence length, a simpler model like Jukes-Cantor can be preferable. The flexibly, therefore, represents DNA sequences which feature more complex rates. Figure X shows the GTR and portrays six different rates of change (R(A→T), R(A→C), R(A→G), R(T→C), R(T→G), R(C→G),

Andrew Holtz
PhD Evolutionary Biology / 2023
Frontiére de l'Innovation en Recherche et Education
Université de Paris Cité

and the equilibrium frequencies, which represent the proportions of each nucleotide in the sequence.

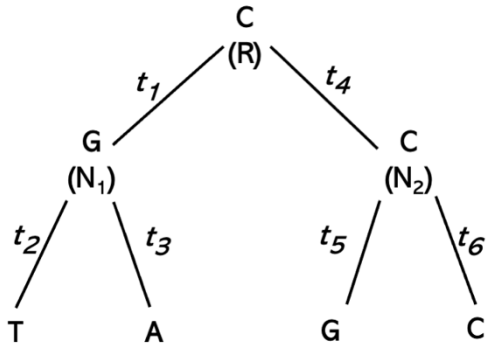
There are additional parameters that can be added to the overall model for each tree inference. For example, gamma rate heterogeneity [77] accounts for the variation in substitution rates across different sites in the DNA sequence. In real biological data, different sites may evolve at different rates due to various factors such as functional constraints or variation in selective pressures. To accommodate this variation, the gamma distribution can be introduced to the GTR model. This distribution assigns different rates to different sites in the sequence, allowing for rate heterogeneity. The gamma distribution is characterized by a shape parameter (alpha), which determines the level of rate heterogeneity. In addition, there is the Free Rate model, which is an addition rate heterogeneity which allows different branches of the tree to have their own independent rates of substitution [78,79]. Each branch is assigned an individual rate multiplier and is then used to modify the substitution rates for those branches [72,73]. Lastly, a site model that assumes a proportion of sites are invariable can be set as a parameter [80].

1.2.3.1.2 ML Tree Estimation

The inference of the tree by maximum likelihood is time demanding. As said, it is impossible to test every possible tree. We therefore rely on computational algorithms that use heuristics to search the tree space. It is important to describe the mathematic framework under these two methods.

Typically, the tools require first a test tree which is created by a faster and simpler method, such as maximum parsimony (IQTREE2), or neighbor-joining (FastTree) method. With this optimized tree, the topology is iteratively optimized until a tree with the best maximum likelihood is found.

In a basic view, we will consider a tree topology with just one site.



To estimate the likelihood for this tree with this one site s_i, we can multiply the probability of each of the sites mutating at each node of the tree.

$$L(\theta \mid s_i) = P(s_i \mid \theta) = \pi_{c} * P_{C \to G}(t_1) * P_{G \to T}(t_2) * P_{G \to A}(t_3) * P_{C \to C}(t_4) * P_{C \to G}(t_5) * P_{C \to C}(t_6)$$

The site of the tips of the trees are known, but the sites of the internal nodes are unknown. Therefore, to calculate the probability of the internal nodes, all nucleotides must be evaluated for each internal node and the root [72,73].

The expression for $L(\theta \mid s_1)$ is given by the summation over all possible nucleotide combinations for the ancestral nodes:

$$L(\theta \mid s_i) = P(s_i \mid \theta) = \sum_R \sum_{N1} \sum_{N2} \pi_R * [P_{R \to N1}(t_1) * P_{N1 \to T}(t_2) * P_{N1 \to T}(t_3) * P_{R \to N2}(t_4) * P_{N2 \to G}(t_5) * P_{N2 \to C}(t_6)]$$

This can be efficiently calculated using pruning algorithms [73].

Next, the topology of the trees is optimized during calculation to maximize the likelihood of the observed data by improving the branch pattern of the tree. Commonly used methods for topology optimization include Subtree Pruning and Regrafting (SPR), Nearest Neighbor Interchange (NNI), and Tree Bisection and Reconnection (TBR) (see Fig. 6) [81].

Andrew Holtz
PhD Evolutionary Biology / 2023
Frontiére de l'Innovation en Recherche et Education
Université de Paris Cité

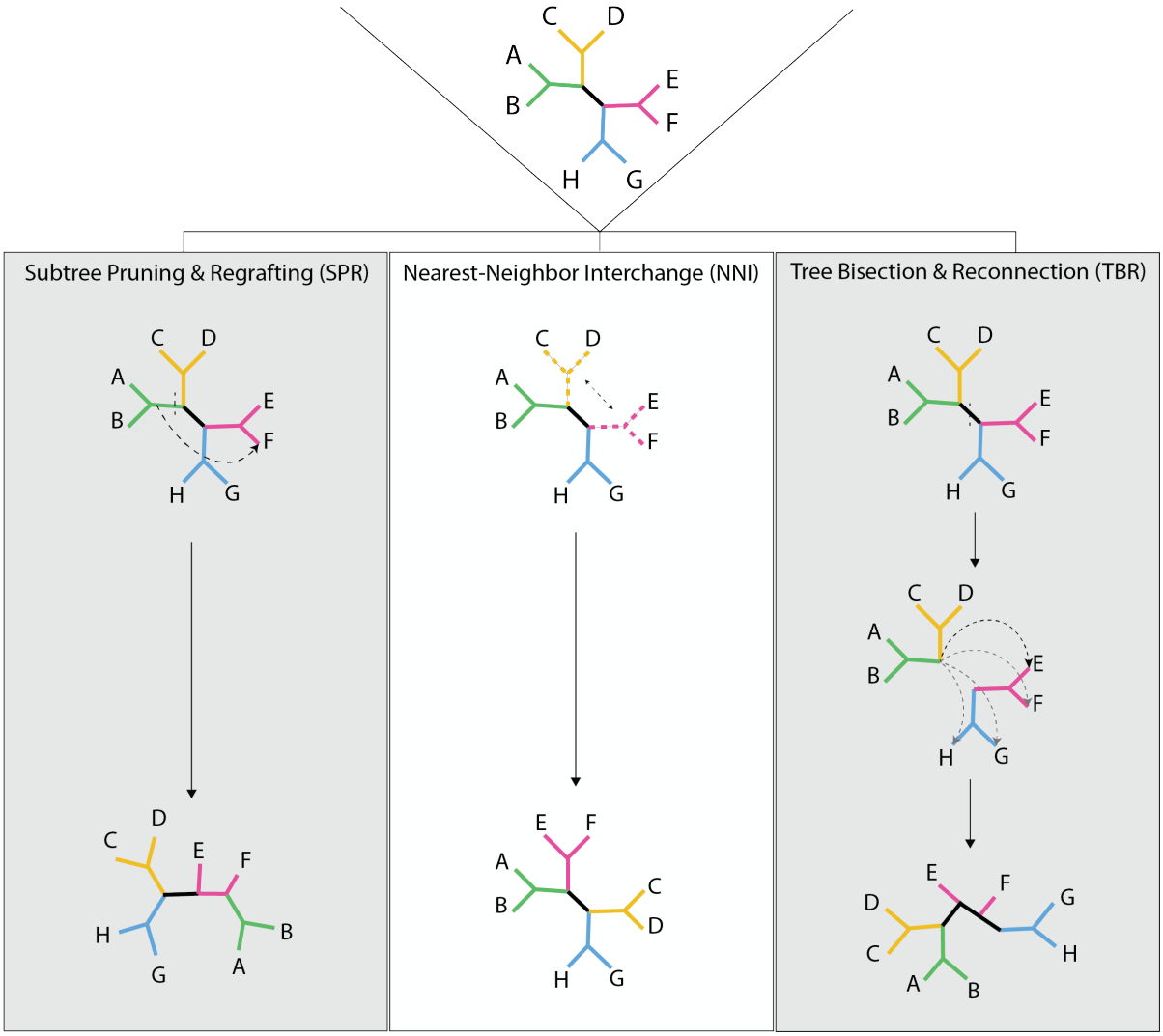


Figure 7. Three basic topology optimizations (SPR, NNI, TBR). Adapted from Lemey et al. 2009

In Subtree Pruning and Regrafting (SPR), a subtree is selected and temporarily removed from the main tree (pruned). It is then reattached (regrafted) onto another branch of the tree. This process is performed iteratively, exploring different arrangements to find a better topology that maximizes the likelihood. Nearest Neighbor Interchange (NNI) involves detaching four subtrees connected by an internal branch and rearranging these subtrees to explore alternative tree topologies. The best arrangement that optimizes the likelihood is selected. Tree Bisection and Reconnection (TBR) starts by removing a branch from the tree, resulting in two subtrees. These subtrees are then

Andrew Holtz
PhD Evolutionary Biology / 2023
Frontiére de l'Innovation en Recherche et Education
Université de Paris Cité

reconnected by inserting a new branch to explore various tree topologies. The best arrangement that maximizes the likelihood is determined [72,82].

There is very efficient software with various added functionality (model finder, boot-strapping, gene partitioning) that can estimate maximum likelihood phylogenetic trees.

IQ-TREE is a popular phylogenetic tree inference tool that uses a hill-climbing method with random perturbations to search for the best tree topology by iteratively searching the tree space based on its current estimation of the best trees [83]. It generates an initial pool of candidate trees using parsimony-based starting trees and performs NNI hill-climbing to find the top 5 trees. At each iteration, a randomly chosen candidate tree is perturbed with random NNIs, and if the resulting tree is better, it replaces the best tree [4,72,82,83]. IQ-TREE2, released in 2020, builds on IQ-TREE and improves the algorithm and offers advanced models.

FastTree works well as a preliminary tool and in instances with an abundance of data, where more complex methods become computationally infeasible, as it is designed to reduce the time and space complexity of ML phylogenetic inference. In return it explores less options in the tree space and it lacks branch length optimization. It builds an initial tree using a modified neighbor-joining algorithm with heuristics to lower memory consumption [82,84]. FastTree includes minimum evolution steps, aiming to minimize branch lengths, and then performs NNI and SPR rearrangements to improve the tree topology. It stops the NNI rounds based on heuristics and executes a predefined number of rounds to optimize the tree. FastTree has been updated to FastTree2 in 2010 to improve accuracy, speed and includes additional features for handling larger and more complex datasets [2].

Although these are efficient methods, maximum likelihood is, unfortunately, restrained by its hill-climbing nature, where it incrementally modifies the tree based on a neighboring tree within the tree space that has a higher likelihood score. This approach can be problematic as it may become trapped in a local optimum, preventing it from exploring other potentially better tree topologies with higher likelihoods [82]. One method to tackle this hurdle is my starting the search with several different starting trees. This way, each search starts from a different point and are more likely to escape local optimums. Alternatively, Bayesian approaches exist which allow for exploring the

entire tree space instead of only focusing on the local optimum. I will explore this approach as well as the software BEAST in chapter 8.2.6.

Phylogenetic tree inference is often only the beginning of the way we explore evolution of infectious diseases. Understanding the evolutionary relationships provides the base. When combined with dating and geographic context, the historical spread of a disease begins to emerge. In this thesis, I utilize Least-Squares Dating [44] and PastML [10] for geographic reconstruction.

1.2.3.1.3 ML Tree Dating Methods

The dating of an ML tree as well as any additional features (such as Ancestral Character Reconstruction (ACR)) occurs after the tree has been fixed. Dating a phylogenetic tree involves pairing the tree with an estimated substitution rate over time. The rate itself can often be estimated from the dates and sequences given (e.g. for rapidly evolving organisms, such as many viruses). The dating process converts the scale of the branch lengths of the tree from number of mutations to time (years) transforming the phylogeny into a time-scaled tree. Before any dating can be done, however, temporal signal of the sequence data set must be evaluated [82]. Temporal signal involves testing if the mutations accumulated in sequences have occurred in a relatively similar pattern and rate across all sequences (i.e., in a clock-like manner). This can be visually assessed in software such as TempEST [12], where the dates of the sequences are plotted against the root-to-tip divergence, which is the accumulated genetic changes along all branches of the tree from the root to tips.

A central part of tree dating is the calibration with a molecular clock. The molecular clock is an assumption that mutations in sequences accumulate over a constant rate over time [72]. A strict molecular clock assumes a constant rate of evolution over time across the whole tree [71,82]. However, strict molecular clocks rarely accurately represent the true evolution due to numerous factors, such as selective pressures, population sizes, and host species effects. Therefore, so-called “relaxed” molecular clocks were developed to allow for variation in the time-calibration of a tree. This allows for different rates for the branches of a tree, but there is still an overall statistical distribution which defines these rates [72,82,85]. There are complex methods to estimate the best fitting clock for a data set. In general, a strict molecular clock will be demonstrated by

Andrew Holtz
PhD Evolutionary Biology / 2023
Frontiére de l’Innovation en Recherche et Education
Université de Paris Cité

sequences with dates with a small R^2 value, meaning that the points (time vs. root-to-tip distance) on the graph described above are close to the estimated evolutionary rate which is demonstrated by the linear line. In contrast, if the points are divergent from the line, it could be a reason to test the relaxed clock.

Two popular pieces of software that tackle the dating of a phylogenetic tree (given temporal signal) are TreeTime and Least-Squares Dating **(**LSD**)**. TreeTime uses an iterative method for maximizing the likelihood of time-scaled phylogenies by estimating ancestral sequences based on the branch lengths of the input tree [13] . Least-Squares Dating calculates branch lengths in nucleic acid substitutions by the following equation:

$$b_i = y_i \cdot \omega + \varepsilon_i$$

where for the branch *i,* y represents the length of the branch in years, ω is the rate of substitution, and ε is the error determined from the normal distribution [82]. This means that LSD assumes a strict clock, but the noise error term allows for uncorrelated non-normal deviations. Magnitude of the noise variable is set by the user. The larger the value the more relaxed the clock will be. LSD dating includes a method to reduce error using the weighted least squares criterion. Confidence intervals can be provided using Poisson distributions [6,82].

Finally, a tree can be rooted by dates by searching for the point in the tree that minimizes the objective function when that point is used as the root. It searches for the point in the tree where the branch lengths correlate best to the molecular clock [6].

1.2.3.1.4 Ancestral Character Reconstruction in ML

Ancestral Character Reconstruction (ACR) is a complex term to describe how rates between additional variables can be traced in a phylogenetic tree, the most common of which being discrete geographic states. These states can be cities, states, regions, countries, and even continents. The idea is parallel to the reconstruction of the phylogenetic tree and is typically completed using maximum likelihood. However, instead of calculating the probabilities of transitioning from one nucleotide to another (A to G), we are now estimating the probability of a virus (for example) transitioning from one state to another- or from one country to another. These

Andrew Holtz
PhD Evolutionary Biology / 2023
Frontiére de l'Innovation en Recherche et Education
Université de Paris Cité

probabilities are calculated and estimated by maximum likelihood methods while considering a model of substitution (e.g. GTR-like) and additional factors (Figure X). This, however, requires optimization and becomes complex with additional characters and tips of the tree [73].

PastML is one of the pieces of software designed to help tackle this computation problem. It implements both parsimony and Maximum Likelihood (ML) methods to predict the ancestral character states. Ancestral character states under ML can be predicted in three different ways. Maximum a posteriori (MAP) [82] computes the marginal posterior probabilities of every state for each internal node of the tree and then simply chooses the state with the highest probability for each node. The Joint method picks the ancestral character scenario with the highest likelihood. PastML offers a novel method called marginal posterior probabilities approximation (MPPA)[10] , which curates a subset of likely states for each node which minimizes the prediction error measured by the Brier score. Therefore, in the region of the tree with high uncertainty (typically close to the root), MPPA keeps several states with similar probabilities, while it predicts one state for the nodes with low uncertainty (typically close to the leaves). This method reduces bias since it allows the user to understand the other states that have similar values [10,82]. For example, in the MAP method if a state (S1) was predicted to have a probability of 32% for node X and this was the highest probability for this node, it would be chosen and presented as the estimated ancestral character. In MPPA, however, this character is presented along with the other characters with similar probabilities (for example, S2 at 29.5%, S3 at 19%, but not S4 at a much lower probability of 5%). This can have large consequences on the confidence of your result and is an important caveat in ACR.

1.2.4 Overview of Bayesian Phylogenetics Methods

1.2.4.1.1 Bayesian Statistics Overview

Bayesian statistics is a powerful approach to statistical inference that is often used in epidemiology and phylogenetics. It offers a flexible framework which can be adapted for very complex models. Unlike maximum likelihood estimation, which aims to find the parameter values that maximize the likelihood of the observed data given a specific model, Bayesian statistics introduces a fundamentally different perspective. In Bayesian analysis, probability is used to

Andrew Holtz
PhD Evolutionary Biology / 2023
Frontiére de l'Innovation en Recherche et Education
Université de Paris Cité

express uncertainty about parameters, treating them as random variables. This means that prior information about the parameters can be combined with the likelihood of the data to form the posterior distribution using Bayes' theorem. This posterior distribution represents our updated knowledge about the parameters after observing the data [72].

Core to Bayesian statistics is the Bayes' theorem, which calculates the probability of a hypothesis being true, given observed data, by combining two components: the prior probability distribution of the hypothesis (our initial belief about it) and the likelihood of observing the data under that hypothesis. This combination results in the posterior probability of the hypothesis, representing our updated belief after considering the data. Mathematically, we can represent this with an equation.

$$\text{Posterior} = [\text{Likelihood} * \text{Prior}] / \text{Evidence}$$

--

$$P(A|B) = [P(B|A) * P(A)]/ P(B)$$

--

$$(Pr(\theta|e) = [Pr(e|\theta)*Pr(\theta)] / [(Pr(e|\theta)*Pr(\theta) + Pr(e|\theta')]$$

$$\mathbf{(Pr(\theta|e) = [Pr(e|\theta)*Pr(\theta)] / [\int (Pr(e|\theta')*Pr(\theta')d\theta']}$$

Let's say for a given model representing a hypothesis (θ) with parameters (*e*), the posterior probability (Pr(θ|*e*) is equal to the Pr(*e*|θ), which represents the compatibility of the observed data with the model and the prior Pr(θ), which reflects the initial beliefs and information about the parameters before observing any data. This is then divided by all evidence, which is the likelihood over all parameter possibilities. This can be written as, $\int Pr(e|\theta')p(\theta')d\theta'$, which represents the integral product of the likelihood and the prior over all possible parameter values. This calculation is performed to obtain the total evidence or marginal likelihood, which is used as a normalizing factor in Bayes' theorem to convert the product of the prior and likelihood into the posterior distribution [72,73,86].

To translate this into a phylogenetic context, the analysis is initiated by defining the prior assumptions and information regarding the relationship of the sequences. Then we use

Andrew Holtz
PhD Evolutionary Biology / 2023
Frontiére de l'Innovation en Recherche et Education
Université de Paris Cité

observational data and employ a probabilistic model of evolution and the Bayes' theorem to revise the initial assumptions into the posterior.

The derivation of the posterior probability distribution is a computational complex problem. It is impossible to derive the analytical form of the total evidence. Random sampling is not sufficient since the posterior probability is concentrated in a small part of the parameter space, and sampling effort is not sufficient since it is unlikely we sample enough of the specific parameter space with the maximized posterior probability. To solve this problem, Markov chain Monte Carlo sampling (MCMC) has been developed which employs a Markov chain to sample from a target distribution that's hard to sample directly. It constructs a chain of parameter values, where each value depends on the previous one according to a specific transition rule. The key idea is to design this transition rule so that, over time, the chain's distribution converges to the desired target distribution. The Markov chain is, therefore, a sequence of random variables where the probability distribution of each variable depends on the previous one. An algorithm is used for the process of searching, testing, accepting or rejecting and reiterating [72,87]. Commonly, the Metropolis-Hastings algorithm is used. There are six steps: (1) initialized with parametric value, (2) exploration of the parameter space using a random walk, (3) propose a new parametric value at each iteration (3) accept or reject the proposed value based on the likelihood of the data and prior probability, (4) move to the proposed value or continue with the current value, and (5) reiterate a set number of times[73]. This process is stochastic by changing and exploring the entire parameter space, but it is time-consuming depending on the number of parameters [72,73].

MCMC methods within the Bayesian framework extends beyond statistical inference, finding particular use in phylogenetics. MCMC facilitates the exploration of complex parameter spaces, allowing for the estimation of evolutionary relationships and divergence times while considering the uncertainty inherent in the data. BEAST 1.10 (Bayesian Evolutionary Analysis Sampling Trees) [1] and BEAST 2.0 [88] are widely-used software package for conducting Bayesian phylogenetic analysis and molecular evolution studies employs Markov Chain Monte Carlo (MCMC) methods to sample from the posterior distribution of evolutionary parameters, allowing researchers to make probabilistic inferences about the evolutionary history of sequences.

Andrew Holtz
PhD Evolutionary Biology / 2023
Frontiére de l'Innovation en Recherche et Education
Université de Paris Cité

1.2.4.1.2 BEAST

Several software packages are available for phylogenetic inference, but BEAST 1.10 brings together the MCMC methods in the Bayesian framework along with complex models for phylogenetic reconstruction. In particular, BEAST 1.10 allowed for the inclusion of several models to build one analysis for tree reconstruction, time-tree calibration, and ancestral character reconstruction in one platform. BEAST 1.10 and future iterations have built on this to allow for more complicated epidemiological models, phylogeographic reconstruction, and predictor testing via the generalized linear model. BEAST 2.0 similarly offers several models, but it is more flexible and allows for the integration of customized models. The process of using BEAST 1.0 depends on the objective, but, in general, the users input an XML file detailing data, models, and MCMC parameters facilitated by the auxiliary program, BEAUti. The output is a collection of tab-delimited plain text files summarizing estimated parameter values and trees, reflecting the posterior distribution. Tracer then enables the analysis of MCMC outputs by visually assessing convergence, which in phylodynamics is considered when the effective sample size (ESS) has reached 200. In this thesis, I focus on two relatively recent additions to the BEAST arsenal: (1) the generalized linear model to test potential covariates in the spread of a disease and therefore the relationships between sequences on the tree, and (2) travel-aware phylogeography which allows the user to complement the geography of a sequence with known travel information that is tied to that sequence.

1.2.4.1.3 Generalized Linear Model

The generalized linear model (GLM) is a method to test transition rates between discrete locations and covariates that are presented as matrices. It is an alternative method in phylogeography for reconstructing the state history for transitions.

As discussed in Section 8.2.4.1.4. phylogeography aims to calculate the transition rates between states to estimate the most likely ancestral location. These transition rates are estimated by the generalize linear model. For each predictor there is a matrix which is made of all the countries in the analysis and their relation to one another. For example, if you are testing geographic distance, the matrix will be a nxn matrix, where n represents the number of countries, and each entry in the matrix the geographic distance between each country. For binary variables such as sharing a

border, 1 and 0 can be used. From this matrix, a substitution matrix is created and a weight is calculated. All the predictors are then summed together for each relationship between two countries to year a final transition rate between two countries.

	A	B	C	…
A	-	d_{AB}	d_{AC}	…
B	d_{BA}	-	d_{BC}	…
C	d_{CA}	d_{CB}	-	…
…	…	…	…	-

--

$$\text{Log}(\Lambda_{AB}) = \boldsymbol{\delta}_1 \boldsymbol{\beta}_1 \mathbf{P}_{1AB} + \boldsymbol{\delta}_2 \boldsymbol{\beta}_2 \mathbf{P}_{2AB} + \boldsymbol{\delta}_3 \boldsymbol{\beta}_3 \mathbf{P}_{3AB} + \boldsymbol{\delta}_n \boldsymbol{\beta}_n \mathbf{P}_{nAB}$$

Transition rate from A to B =**Σ**(*n* Predictors for AB* coefficient weight *binary support)

--

Transition rates are calculated for each *A->B (or i -> j)* relation in the data set. These transition rates are components of the rate matrix Λ of discrete location states. The matrix is then used in BEAST to estimate the ancestral character reconstruction. The coefficient weight (**δ**) is estimated in BEAST as a measurement for the overall contribution of this predictor relative. The binary indicator (**β**) is a support value estimated by how often this predictor is included in the model for the observed data. If the predictor is estimated to contribute, it could be used in all instances (1) or in none (0) [89] .

The GLM model has been used extensively in the past to understand how epidemiological, climate, and human-travel history can effect and mediate the spread of disease. During the 2014-2016 Ebola epidemic in West Africa, the GLM in BEAST 1.10 was used to illustrate that areas with high population density in close geographic proximity exhibited a role in the increased transmission of the virus [90]. In addition, Lemey *et al.* used the generalized linear model to highlight the influence of human travel patterns on the global transmission dynamics of H3N2

Andrew Holtz
PhD Evolutionary Biology / 2023
Frontiére de l'Innovation en Recherche et Education
Université de Paris Cité

influenza. Using air transportation data between 2004 and 2006 as a potential predictor to spatial context of the spread of the disease, they found that air passenger flow data can predict disease spread and that coupling human mobility with phylogenetic data builds confidence and accuracy in inference [26].

1.2.4.1.4 Travel History-Aware Phylogeography

As a response to low diversity and sampling bias during the 2019-2023 SARS-CoV-2 Pandemic, a novel method was developed to include additional source of information to inform the BEAST model for phylogeography. The pandemic was an exceptional demonstration on the worlds whole-genomic sequencing capacity. Already six-months into the pandemic, there were over 100,000 genomes generated globally. Despite the sheer volume of sequencing, there were still issues in phylogeographic reconstruction due to sampling bias and uneven distribution of sequencing efforts. Lemey *et al* presents travel history-aware phylogeography, which incorporates travel data with transportation data (as GLM) to reconstruct the geographic spread of SARS-CoV-2 [91]. This analysis requires the careful collection of epidemiological data from patients who have had the virus sequenced. This data includes their recent travel locations, the length of their trips, and the time since they returned. In the phylogeographic analysis, this travel data offers potential new locations for tracing the virus's transition. Consequently, an artificial ancestral sequence is generated in the analysis based on that location. Using this method, the potential to include countries that have limited sequencing capacity is recovered. For example, a passenger is diagnosed with SARS-CoV-2 infection in France. Superficially, a sample from this patient would indicate in the metadata a French origin. However, it is discovered upon interviewing the patient that they have just returned from a trip to Sierra Leone, a country with limiting SARS-CoV-2 sequence capacity. Therefore, the sequence from this patient now potentially represents a version of SARS-CoV-2 circulating in Sierra Leone (or the general region). This information can now be used in the travel history-aware phylogeographic analysis since a Sierra Leonean artificial sequence as an ancestor of the French sequence can be included in the analysis, potentially explaining previous unexplained transmissions.

This model presents opportunities beyond SARS-CoV-2 since human-mediated migration of viruses is common in many zoonoses, as explained in previous sections. Rabies virus in

Andrew Holtz
PhD Evolutionary Biology / 2023
Frontiére de l'Innovation en Recherche et Education
Université de Paris Cité

particularly is a zoonotic infectious disease, with a long history of the relevance of human migration in its emergence and maintenance around the world [92].

1.2.5 Overview of Rabies Virus

Rabies virus (RABV) is a zoonotic single-stranded negative-sense RNA virus that is the main responsible agent for Rabies disease. It is one of the oldest documented human diseases, with many references to rabies like disease in ancient Rome, Greece and Egypt [92–94] . Despite effective vaccination for both animals and humans, endemic rabies in animals is still an issue in the 21st century causing localized outbreaks around the world. Canine-transmitted rabies is responsible for more than 99% of human rabies disease [95]. However, as a zoonotic disease, a One Health perspective is vital for the global control and understanding of its spread. Even though human cases are only caused by spill over from animal populations, there are still a projected 59,000 cases of RABV a year, and due to the 100% fatality rate of RABV it is still a serious public health issue [96].

1.2.5.1.1 Genome and Virology

Rabies virus is a member of the genus *Lyssavirus* of the family *Rhabdoviridae* of enveloped RNA viruses. The virion is comprised of five proteins: (1) **N**ucleoprotein, (2) **P**hosphoprotein, (3) **M**atrix protein, (4) **G**lycoprotein, (5) **L**arge RNA polymerase protein (Figure 7). These proteins are coded in the five genes of the virus which make up the 11,900 nucleotides long genome. The genome is relatively stable with a nucleotide substitution (i.e. evolutionary) rate of around $2.00 \cdot 10^{-4}$ substitutions per site per year [66].

Andrew Holtz
PhD Evolutionary Biology / 2023
Frontiére de l'Innovation en Recherche et Education
Université de Paris Cité

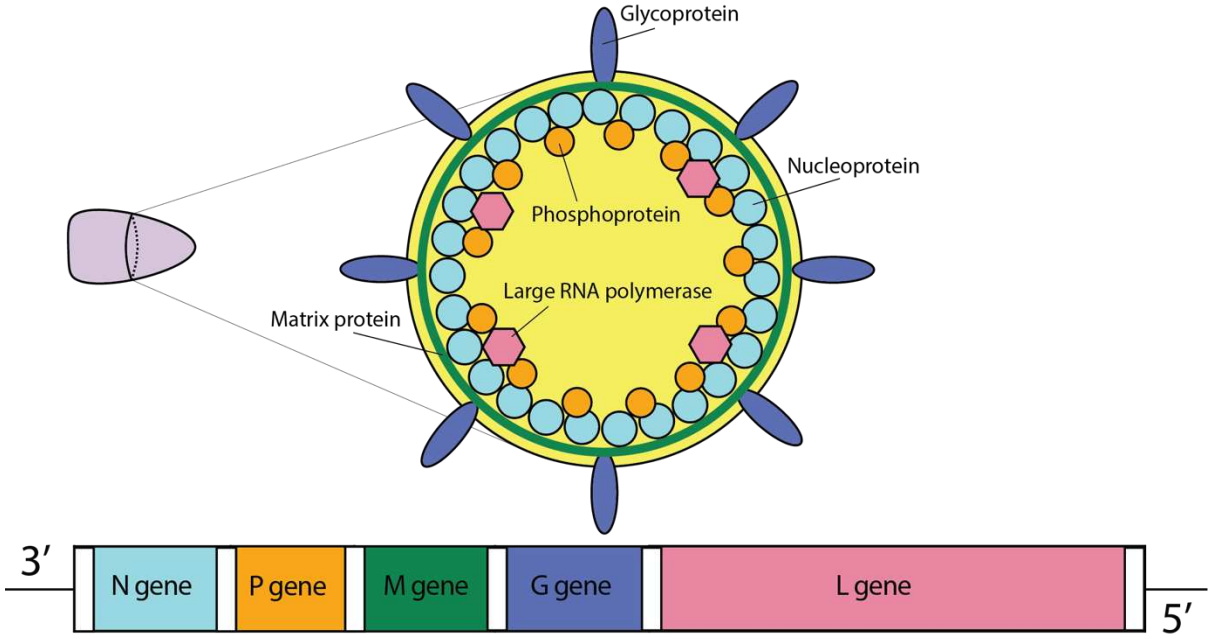


Figure 8. Cartoon of Rabies virus showing each gene and the schematic localization of its representative protein in the virus structure

RABV transmission occurs after broken skin or mucosa is exposed to infectious material (saliva in most of the case) from a host animal. This accounts for the vast majority of virus transmissions. The virus enters the host cells through receptor-mediated endocytosis, facilitated by the viral glycoprotein (G) binding to several cell receptors including the nicotinic acetylcholine receptors at neuromuscular junctions, entering peripheral nervous system (PNS) neurons. Once inside the cell, the virus uncoats, and the RNA genome is transcribed and replicated in the cytoplasm of the infected cell. Viral proteins are synthetized by the cell machinery. The nascent genomic RNA are encapsidated by the viral nucleoprotein to form the nucleocapsid. This is then surrounded by the matrix protein. The new viral particles are assembled in the cytoplasm and are progressively surrounded by the envelope containing the viral glycoprotein, a mechanism that leads to the budding from the cell membrane. The newly formed virions are then released from the cell and can infect new cells. RABV is a neurotropic virus, meaning it has an affinity for the nervous system and is capable of infecting nerve cells. Using retrograde axonal transport, RABV travels along the PNS neurons to the central nervous system (CNS), where it can cause encephalitis, leading to the clinical symptoms of rabies. Rabies initially manifests as fever, headache, and general weakness, but as it progresses, neurological symptoms develop, including insomnia, anxiety, confusion,

Andrew Holtz
PhD Evolutionary Biology / 2023
Frontiére de l'Innovation en Recherche et Education
Université de Paris Cité

hallucinations, hydrophobia (fear of water), and paralysis. Rabies is always fatal once clinical symptoms appear [97].

The virus is able to spread via the retrograde axoplasmic flow to the salivary glands and is excreted in the saliva. Once in the saliva of the host, the transmission cycle can repeat with a fresh bite of a new host.

1.2.5.1.2 Epidemiology

There are two main categories of primary hosts of RABV – *Chiroptera* (which includes all bats), and canine carnivores (sub order Carniformia) including dogs, foxes, raccoons, wolfs, raccoon dogs, and skunks. Additional mammals like ungulates experience infections through spill over events through these primary hosts, but they normally do not sustain further transmission. Humans are considered as dead-end infections, although interhuman transmission is theoretically possible as the virus is also present in the saliva of infected humans.

In canine populations, rabies is a vaccine-preventable disease and the vaccination of dogs is the most effective measure to prevent its spread [98]. Regular vaccination not only protects the animal but also acts as a buffer, preventing transmission to humans. In regions with high incidences of rabies, mass vaccination campaigns have proven effective in drastically reducing cases [99]. For humans, pre-exposure prophylaxis (PrEP) is recommended for those at high risk, such as those with contact with wild animals and people travelling to rabies endemic countries. Although PrEP is recommended, it is not required for protection and alone does not enough provide full protection. After a bite from a rabid animal is suspected, it is best to rapidly begin the post-exposure prophylaxis (PEP) regimen. For previously unvaccinated individuals, PEP typically involves a series of three vaccine doses administered over a one-week period (days 0, 3, and 7) and eventually in the most severe cases of exposure a dose of rabies immune globulin (RIG). For those previously vaccinated (either by PrEP or previous PEP), the treatment typically involves only two doses of the rabies vaccine (days 0 and 3) without the need for RIG [95,100].

Despite the known efficacy and cost-efficiency in vaccination of canines and the fact that 99% of human cases are transmitted by common-canines, canine rabies still circulates in much of the Middle East, Africa, and Asia[96,101]. Endemic disease constantly poses the potential to be

Andrew Holtz
PhD Evolutionary Biology / 2023
Frontiére de l'Innovation en Recherche et Education
Université de Paris Cité

reintroduced into regions of the world that have previous eliminated the disease. Pet importation, human travel to rabies-endemic areas, and wildlife transmission, emphasizes the global need for a One Health, multidisciplinary and multi-national effort in rabies control [102,103].

In this thesis, I will focus on epidemiological control methods that are influenced by phylogenetic and phylodynamic studies. These studies help provide tactical information on how rabies virus spreads and where animal vaccination campaigns could be effective.

1.2.5.1.3 RABV phylogenetics

Genomic surveillance of RABV has been ongoing since the 1970s. Before whole-genome sequencing, studies focused on either the G or N gene and from these sequences the first phylogenetic studies on RABV were performed. An analysis of a global sample of RABV reveals two major groups, *Chiroptera* (bats) and *Carnivora* (canine-maintained and North Americans raccoons and skunks) [104]. The canine-maintained group has led to the largest expansion of RABV recorded around the world, accounting for documented cases on every content (besides Antarctica). This group can be subdivided into more than 44 sublineages which most of them showing a strong geographical clustering [62,105].

Phylogenetics, especially phylogeographic analyses, have been instrumental in tracing disease origins and spread, aiding in targeted vaccination strategies. For instance, an analysis of fox rabies resurgence in Northern Italy (2008-2011) traced its origin to the Western Balkans and Slovenia, revealing that certain areas, fortified by oral fox vaccination campaigns, prevented further lineage spread, while unvaccinated neighboring regions were more susceptible. This study also underscored the cryptic transmission of RABV in wildlife, which can remain under the radar of national surveillance for years [37]. Additional studies studying how raccoons, skunks, and bats can spread RABV [38], have also been done, further highlighting the importance of wild animal transmission in the spread of RABV.

This study is just one of many that used phylogenetics and phylogeography to understand a localized RABV epidemic or the general endemic situation in a specific country [106–111]. However, a common limitation of these studies is their narrow focus: many focus on particular genes, exclusively analyze sequences from their own regions or countries, and are constrained to specific

time frames. Moreover, there has been a lack of exploration into predictors that could influence the maintenance or sustainability of RABV epidemics, such as country borders, geographic distances, human migration patterns, and colonial histories. This kind of isolated approach can limit our broader understanding of RABV's global spread. A core aspect of this thesis seeks to bridge these gaps. By integrating sequences from various genomic regions, irrespective of their sequencing time, host organism, or geographical origin, and investigating the predictors, I aim to provide a comprehensive reconstruction of the global spread of RABV. This holistic approach is geared towards enhancing our epidemiological campaigns, aligning with the broader One Health objective of integrated health solutions.

1.2.6 SARS-CoV-2

Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2) is the causative agent of the COVID-19 pandemic that emerged in late 2019. Originating in Wuhan, China, this novel virus belongs to the coronavirus family, which also includes the viruses responsible for the SARS outbreak in 2002 and MERS in 2012. With a varying degree of symptoms from asymptomatic to severe respiratory distress, the virus has put significant strain on global health infrastructures.

1.2.6.1.1 Genome and Virology

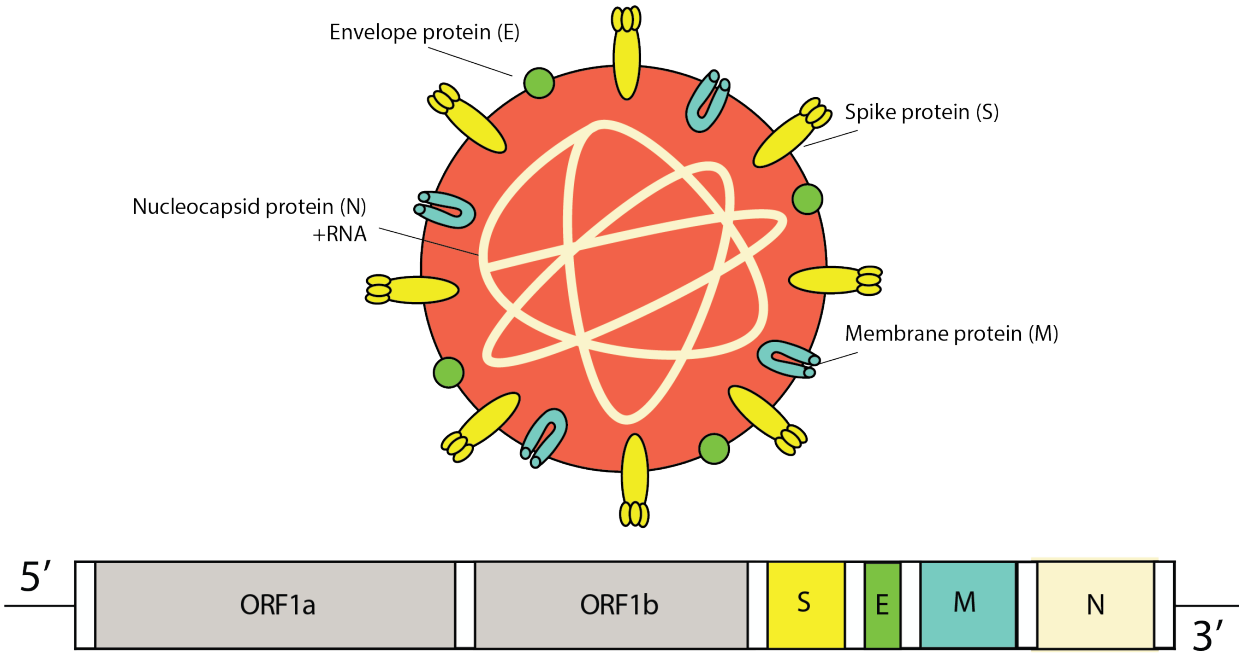


Figure 9. Cartoon of SARS-CoV-2 virus showing each gene and its representative protein in the virus structure

SARS-CoV-2 is an enveloped, positive-sense single-stranded RNA virus. Its genome consists of approximately 29,900 nucleotides, encoding multiple structural proteins: Spike (S), Envelope (E), Membrane (M), and Nucleocapsid (N) [112]. The Spike protein, in particular, has garnered attention for its role in mediating viral entry into host cells by binding to the human ACE2 receptor. It is also the location of lineage defining mutations, perhaps as a result of vaccination epitopes and antibody host immune pressure [113].

1.2.6.1.2 Epidemiology

Bats are considered the probable primary reservoir for SARS-CoV-2, with a potential intermediary host of currently unknown species playing a role in zoonotic transfer [114]. The virus was first detected at a live animal market in Wuhan China, from which human-to-human transmission rapidly ensued transmitting to communities all around the world.

Although it has a relatively low mutation rate of 1.0-2.0 × 10^{-6} mutations per genome replication compared to other viruses [65], it has transmitting to over 640 million people as of December 57

Andrew Holtz
PhD Evolutionary Biology / 2023
Frontiére de l'Innovation en Recherche et Education
Université de Paris Cité

2022 [115] enabling rapid virus evolution. As the virus spread it generated new lineages including Variants of Concern (VOC) that seem to have an increase transmissibility, virulence of change in clinical presentation, or decreased effect of vaccination. Variants of Interest (VOI) on the other hand are defined as variants that have been able to cause small transmission clusters, causing multiple community transitions and have been detected in several regions of countries, in addition to also having a potential increase in transmissibility or virulence. The World Health Organization (WHO) is responsible for defining and maintain these definitions and the latest can be found on their technical document: https://www.who.int/publications/m/item/updated-working-definitions-and-primary-actions-for--sars-cov-2-variants [116].

1.2.6.1.3 Variants of Concern and Phylogenetics

Each dominate variant showed some transmission advantage and replace previously circulating variants (Figure 10). As of Summer 2023, there are multiple BA sublineages that co-exist circulating around the world, and there is very transmission of pre-Omicron lineages (including Alpha, Beta, Delta, Gamma). Below is a brief overview of the history of VOCs.

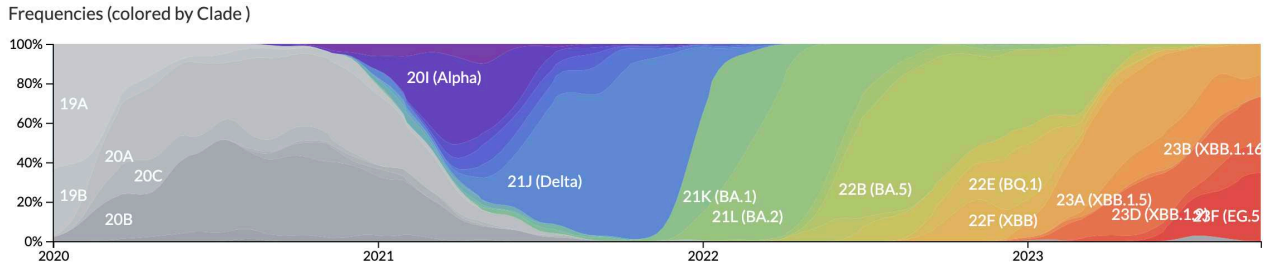


Figure 10. Nextstrain Variants frequency over time since the start of the pandemic

Alpha (B.1.1.7) [65,115]:

UK/Europe/US

Winter/Spring 2021

Spike mutations: N501Y, P681H, T716I…

Hypothesized advantages: Increased transmissibility and possibly increased virulence.

Beta (B.1.351) [65,115]:

South Africa/ Global

58

Andrew Holtz

PhD Evolutionary Biology / 2023

Frontiére de l'Innovation en Recherche et Education

Université de Paris Cité

Winter/Spring 2021
Spike mutations: N501Y, E484K, K417N…
Hypothesized advantages: Potential resistance to neutralization by some antibodies.

Gamma (P.1) [65,115]:
Brazil/Global
Spring 2021
Spike mutations: N501Y, E484K, K417T…
Hypothesized advantages: Potential resistance to neutralization and increased transmissibility.

Delta (B.1.617.2) [117]:
Global
Spring/Summer/Fall 2021
Spike mutations: L452R, P681R, T478K
Hypothesized advantages: Increased transmissibility and potential resistance to neutralization by some antibodies.

Omicron (B.1.1.529 & BA and XBB sublineage[118 119]):
Global
Late 2021 - Present (Summer 2023)
Omicron is defined by more 30 amino acid substitutions in the Spike protein, 15 of which are in the receptor binding domain (RBD). These include three new clusters: S371L, S373P and S375F, N440K, G446S, Q493, G496, Q498, and Y505H, as well as other accumulated mutations: K417N, S477N, T478K, E484A, and N501Y and 3-residue insertion sequence
Hypothesized advantage: Greater resistance to neutralizing antibodies and vaccine-induced humoral immunity

Phylogeographic and phylodynamic analyses have been used to understand the origin and spread of SARS-CoV-2 variants (examples using BEAST 1.0 are A.27 and B.1.620). Addressing the emergence of new lineages is crucial in disease control, as it prompts questions about the origins, transmission, and mutations of the variant. The BEAST 1.0 pipeline, incorporating travel history-aware phylogeographic data, has been pivotal in decoding the origins of various lineages. For

instance, lineage A.27, detected predominantly in south-west Germany between December 2020 and June 2021, exhibited spike gene mutations shared with other variants of concern. With the integration of travel-history data, researchers were able to pinpoint its origins to West Africa [120]. In a parallel manner, lineage B.1.620, initially identified in Lithuania, was traced back to Central Africa when analyzed with air passenger and patient travel histories [121]. These lineages not only underscore the dynamic nature of the virus but also highlight the importance of global interconnectedness and the need for vigilant monitoring in preventing the spread of future variants.

Both studies highlight an important pattern: genomic surveillance in countries with strong sequencing programs often first identity variants that might have been cryptically transmitting elsewhere. This could partially be explained by the lack of early control measures such as quarantine and lockdowns in many low-middle income countries, which allowed the virus to continue to transmit and acquire new mutations [122].

Further studies like these can continue to provide interesting angles for transmission dynamics of not just SARS-CoV-2 but other infectious diseases which can transmit over long distances, aided by human migration.

1.3 **Aims of this thesis**

In this thesis, I aim to enhance methodologies through the incorporation of diverse data types to bolster phylogenetic and phylogeographic inference. Two projects are undertaken: the first investigates historical transmission patterns of canine RABV using maximum likelihood methods, and the second focuses on contemporary SARS-CoV-2 using Bayesian analysis, integrating patient travel data.

The core message is that employing diverse approaches for phylogeographic analysis yields nuanced insights, improving our understanding of disease spread. In the RABV project, the focus is to integrate a large historical dataset of RABV sequences irrespective of which area of the genome the sequence represents. The aim is to integrate these sequences to increase the phylogenetic and phylogeographic diversity, which in response allows us to test the impact of

human migration in its spread, in addition to uncovering transmission patterns of canine RABV spread.

Although using a different approach in a Bayesian framework, I aim to reveal the origin and spread of the SARS-CoV-2 variant, B.1.214.2. By incorporating not only sequence and geographic data but also integrating patient travel history data, I am able to make accurate estimations of origin and introduction highlighting the importance of human-migration in the spread of the variant.

Overall, this adaptability is crucial for accommodating evolving infectious diseases and human migration influences, enabling more effective, data-driven disease control strategies. Lastly, I aim in this thesis to combine the methods, insights, and models in the first two chapters to introduce a novel GLM method in a maximum likelihood and envisions its application in evaluating public health interventions.

Andrew Holtz
PhD Evolutionary Biology / 2023
Frontiére de l'Innovation en Recherche et Education
Université de Paris Cité

2 Global Origins of Canine-mediated RABV Virus

The primary objective for this chapter was to understand the historical dispersion of rabies virus. We wanted to know when and where the current strain of rabies virus (RABV) adapted to canids first emerged and estimate the locations and times of consequent spread across the world. In addition, we wanted to have a large picture overview highlighting how the different clades, geographic locations, and host-species are represented on the evolutionary history of the virus.

It became very clear that if we wanted an accurate representation of sequences, we would want to include the total amount of sequences available to download in order to keep the time and geographic diversity. This was a first for any type of phylogenetic analysis. In the past, either whole-genome sequences or specific gene subsets were used. For RABV this was historically the N and G genes. It was evident, that previous phylogeographic studies had unknowingly introduced bias by using only sequences from certain genes. Therefore, we wanted to introduce a new method that could make use of all genes in the genome.

In this chapter we introduce a sequence concatenation method in the multiple sequence alignment. The fear with this method is that in a multiple sequence alignment, evolutionary relationships are read by observing shared mutations. This is not possible if there is no overlap among two sequences in the genome. Therefore, we needed to ensure there were enough whole-genome sequences that would allow us to make this comparison. We validated this method of sequence concatenation by comparing our results with previous studies and by subsampling our data and repeating the analysis. Our results present a method for phylogeographic inferencing for large-scale sequence datasets, across many gene regions, countries, dates, and even host-species.

Using this method, we estimate the most precise tMRCA of RABV and dates of consequent emergence into the 44 clades of RABV and into countries and regions around the world. We estimate the origin emergence of this strain of RABV to be between 1301 and 1403, and estimate the origin of many global introductions, painting a map of all transmission routes between from the 1300s and now. This enabled us to identity across the phylogenetic tree incidences of long-distance migration by humans. In addition, we used ancestral character

Andrew Holtz
PhD Evolutionary Biology / 2023
Frontiére de l'Innovation en Recherche et Education
Université de Paris Cité

reconstruction to highlight where and when canine RABV transmission could be partially explained by European colonization from the 17th and 18th centuries.

This work would not have been possible without collaborative efforts in sequencing surveillance for rabies virus. In addition, our validation relies on the previous phylogeographic studies of canine rabies virus. In the study, we review several studies and their phylogeographic estimates and how those relate to our own estimates.

The article in this chapter was published in *Nature Communications* on July 17, 2023

Andrew Holtz
PhD Evolutionary Biology / 2023
Frontiére de l'Innovation en Recherche et Education
Université de Paris Cité

nature communications

Article

https://doi.org/10.1038/s41467-023-39847-x

Integrating full and partial genome sequences to decipher the global spread of canine rabies virus

Received: 2 March 2023

Accepted: 30 June 2023

Andrew Holtz[1] ✉, Guy Baele[2], Hervé Bourhy[1,3] & Anna Zhukova[4] ✉

Integrating full and partial genome sequences to decipher the global spread of canine rabies virus

Nature Communications volume 14, Article number: 4247 (2023)

https://doi.org/10.1038/s41467-023-39847-x

AUTHORS

Andrew Holtz[1], Guy Baele[2], Hervé Bourhy[2,3] & Anna Zhukova[4]

AFFILIATIONS

1. Institut Pasteur, Université Paris Cité, Lyssavirus Epidemiology and Neuropathology Unit, F-75015, Paris, France
2. Department of Microbiology, Immunology and Transplantation, Rega Institute, KU Leuven, Leuven, Belgium
3. World Health Organization Collaborating Center for Reference and Research on Rabies, Institut Pasteur, Paris, France
4. Institut Pasteur, Université Paris Cité, Bioinformatics and Biostatistics Hub, F-75015, Paris, France

CO-CORRESPONDING AUTHORS

Andrew.holtz@pasteur.fr (AH), anna.zhukova@pasteur.fr (AZ)
64

Andrew Holtz

PhD Evolutionary Biology / 2023

Frontiére de l'Innovation en Recherche et Education

Université de Paris Cité

2.1 Abstract

Despite the rapid growth in viral genome sequencing, statistical methods face challenges in handling historical viral endemic diseases with large amounts of underutilized partial sequence data. We propose a phylogenetic pipeline that harnesses both full and partial viral genome sequences to investigate historical pathogen spread between countries. Its application to rabies virus (RABV) yields precise dating and confident estimates of its geographic dispersal. By using full genomes and partial sequences, we reduce both geographic and genetic biases that often hinder studies that focus on specific genes. Our pipeline reveals an emergence of the present canine-mediated RABV between years 1301 and 1403 and reveals regional introductions over a 700-year period. This geographic reconstruction enables us to locate episodes of human-mediated introductions of RABV and examine the role that European colonization played in its spread. Our approach enables phylogeographic analysis of large and genetically diverse data sets for many viral pathogens.

2.2 Introduction

Studies that investigate the dynamics of viral emergence are vital to inform epidemiological decision making [123–125]. Whole-genome sequencing has become the norm for phylogenetic studies focused on modern epidemics/pandemics, such as the recent SARS-CoV-2 pandemic [126]. Alternatively, many zoonotic pathogens have decades of partial genome submissions, greatly outnumbering the number of whole-genome sequences (WGS) [127]. Such neglected zoonotic diseases—such as those caused by West Nile, rabies, and Lassa viruses—receive inadequate attention from public health officials, leading to limited funding [102] and a focus on sequencing only certain parts of the viral genome, particularly in cases involving non-human hosts. WGS contain more mutations and offer greater phylogenetic resolution compared to single genes [59]. The development of novel models and methods that harness the plethora of genetic data, WGS and partial alike, could lead to stronger insights in pathogen control such as epidemic origins [128], geographic spread [26,129] and cryptic transmission [130,131].

Among these pathogens is rabies virus (RABV), one of the most well-documented zoonotic viruses, which is responsible for many local epidemics and an estimated 59,000 human deaths
65

Andrew Holtz
PhD Evolutionary Biology / 2023
Frontiére de l'Innovation en Recherche et Education
Université de Paris Cité

annually [96]. RABV is perhaps unique in how it has circulated in a large diversity of mammals over thousands of years, although 99% of human cases are caused by household dog transmission [132]. Despite much of Western Europe being rabies-free, 109 countries still battle the deadly disease [95,103]. There is a constant risk of RABV re-emergence by pet importation, travel to rabies-endemic areas, and wildlife transmission, which emphasizes the global need for a One Health, multidisciplinary and multi-national effort in rabies control [102,133]. Programming for epidemic control and prevention often relies on phylogenetic analysis for canine vaccination [37,134], which estimate RABV genealogies to infer spatio-temporal characteristics of rabies spread. This relies on global efforts in sample collection and sequencing [110].

Despite the global objective of zero human RABV cases by 2030 [135], rabies remains a neglected disease primarily affecting low and middle-income countries, and there is suboptimal funding, disease investigation, and programming dedicated towards elimination [96]. This has stalled the reallocation of sequencing efforts towards WGS and leaves phylogenetic and phylogeographic analyses still reliant on subgenomic fragments. These analyses have been used to study the origin and spread of canine-maintained RABV in various regions, but often require reducing sample sizes for computational ease (choosing specific genes [136–138] or whole genomes [62,139]) which can introduce bias and reduce geographic and genetic diversity. This trend in partial sequence availability is not unique to RABV, as several other viral diseases also have a historical diversity of partial sequences, including West Nile virus, rotavirus species, and dengue virus.

Past and present phylogenetic analyses of RABV reveal a major divide in virus sequences associated with Chiroptera (bats) and Carnivora (canine-maintained and North Americans raccoons and skunks) [104], which infers a host switch between these two groups of animals. Within this division three phylogenetic groups of sequences have been identified: bat-maintained RABV and raccoon-skunk-maintained RABV and canine-maintained RABV within the Carnivora [104,140] (Supplementary Fig. 3). It is believed that canine-maintained RABV most likely rose in Europe and Asia during a period of dog domestication [104,140]. From this point, canine RABV was able to flourish across the globe [105], generating to two major phylogenetic groupings, the old world and cosmopolitan clades, that can themselves be subdivided into 44 subclades [11,62,105,141]. Despite the potential presence of bat-maintained RABV in the Americas pre-European colonization,

canine-maintained RABV is believed to have emerged in Europe and then spread to the Americas following European colonization between 1642 and 1782 [62,104,105,140].

This study proposes a novel gene concatenation method that takes advantage of the historical diversity of partial gene submissions to analyze the spread of canine-mediated RABV. We use 14,752 sequences to create a concatenated alignment, estimate a phylogenetic tree with 10,044 canine-mediated sequences and, from this, infer the ancestral dispersal around the world. Our findings are not only precise but, for the first time, provide ancestral geographic reconstruction on a global level across clades, regions, and countries. This further enables us to identify episodes of human-mediated introductions of RABV around the globe and inspect how European colonization starting in the fifteenth century impacted the spread of canine rabies. This not only provides valuable historical information for reconstructed transmission paths of RABV globally, but provides, for the first time, a useful phylogenetic framework for pathogens with heterogeneous sequencing data and to help inform introduction control policy.

2.3 Results

RABV sequence and metadata acquisition and composition

All 25,787 available sequences for the five genes of the RABV genome were downloaded from the NCBI Virus database. After quality control, the RABV data set was reduced to 14,752 sequences that spanned 121 countries and were extracted from 192 different host species from 1972 to 2020. The multiple sequence alignment (MSA) comprised a concatenation of the five genes of the RABV genome (Fig. 1a and Supplementary Table 2). The maximum-likelihood (ML) tree from this MSA (Supplementary Fig. 3) revealed three major phylogenetic groups with bootstrap values greater than 0.95 corresponding to bat-, skunk-/raccoon-, and canine-related RABV, consistent with previous global RABV analyses [62,105]. The tree topology is spatially structured with clades, while gene fragments had a very low impact on sequence clustering (Supplementary Fig 3: color tips for gene fragments and simplified clade), indicating that sequences are not clustering by genetic region. All 10,209 sequences in the canine-related cluster were extracted and used for the remainder of the study.

Canis familiaris and genus Vulpes (i.e., foxes) were the most common sources of canine-clustering sequences, accounting for 51.9% and 8.4%, respectively. Other families, including the Bovidae, Hominidae, Mustelidae, Felidae, Mephitidae, and Herpestidae, contributed the remaining sequences. The location of the sampling impacted the predominant source of sequences, with most European sequences coming from the genus Vulpes and those in Asia and Africa primarily coming from household dogs. In Asia, the Hominidae and Mustelidae families, particularly the Chinese ferret badger, were major sources of sequences (Fig. 1b, d). Since the 1970s nearly all continents have experienced an exponential increase in RABV sequence submissions to the NCBI Virus database. This increase is most notable for the N gene and the G gene, while WGS have experienced a slower rate in submissions, especially for Europe and Asia (Fig. 1d). WGS still only represents 13% of sequences and 62% of countries. This demonstrates the quantity of partial sequence diversity available and the potential loss of phylogenetic signal and resolution that occurs when only using WGS for phylogenetic studies. In addition, studies favor certain genes depending on which country/continent the study is located in. For example, proportionally, Europe sequences mostly the N gene, while Asia sequences many samples for the G, M, and P genes. As a result, only using the N gene or WGS introduces systematic sampling bias into any phylogeographic analysis by ignoring sequences that exist for other regions of the genome.

Using all five gene fragments increases phylogenetic signal and tree-time calibration precision

Our ML tree of the 'canine' data set of 10,209 sequences containing all five genes confirms the clustering of canine rabies cases previously detected by Bayesian phylogenetic tree reconstruction using WGS (Fig. 3) [62]. The tree topology is spatially structured by countries of origin, while gene fragment and host species seem to have a very low impact on sequence clustering, providing evidence that rabies transmissions are more defined by geography than host species.

To better investigate the spatio-temporal spread and history of rabies across the world, we dated the phylogenetic tree, comparing evolutionary rates across the entire genome and for individual genes. The estimated evolutionary rate for whole-genome sequences was $2.00 \cdot 10^{-4}$ substitutions [95% CI: $1.95 \cdot 10^{-4}$; $2.22 \cdot 10^{-4}$] per site per year, while the rate for individual genes ranged from $1.9 \cdot 10^{-4}$ to $2.4 \cdot 10^{-4}$ (Supplementary Fig. 5), indicating little difference in evolutionary rate. This is consistent with previous estimates [62,140]. By applying the

Andrew Holtz
PhD Evolutionary Biology / 2023
Frontiére de l'Innovation en Recherche et Education
Université de Paris Cité

WGS evolutionary rate to the tree using LSD2 (where 163 sequences were removed by outlier removal), we dated the tMRCA of canine-related RABV to 1356 (95% CI: 1301; 1403). This estimate narrows previously published estimates in Troupin et al. (1308–1510) by 101 years [62] and in Velasco-Villa et al. (1273–1562) by 187 years [140] (Fig. 2). In general, our methods resulted in older age estimates for the tree compared to the findings of the other two studies.

The improvement in precision of dating is evident in previously established major clades [11,105] compared to Troupin et al.[62] and Velasco-Villa et al. [140] (Fig. 2). Our mean divergence time estimates were also consistently older, which can be explained by including more and older sequences ancestral to the clade roots compared to the other two studies. To name a few, we estimated the emergence of the Asian clade back to 1561 (95% CI: 1524–1594), the cosmopolitan clade back to 1656 (95% CI: 1627–1683), the Africa-2 back to 1799 (95% CI: 1761–1832), and Africa-3 back to 1723 (95% CI: 1697 and 1752). Notably, all point estimates of major clade emergence, with the exception of the cosmopolitan clade, fall within the 95% confidence intervals reported in Troupin et al. [62] (321 whole-genome sequences) and Velasco-Villa et al. [140] (266N gene sequences) as shown in Fig. 2 and Supplementary Fig. 5. The date difference for emergence of the cosmopolitan clade can be partially explained by the inclusion of more partial sequences in this clade clustering closer to the clade root.

Subsampling method provides confidence in phylogenetic analysis and dating

To test the confidence in phylogenetic estimation, we performed country-aware subsampling on the original canine tree for five unique replicate sequence data sets of 5500 sequences each. To ensure the validity of phylogenetic analysis using the efficient but less thorough method of FastTree for the full-canine tree, we performed a phylogenetic analysis using the more accurate IQ-TREE2 method (using an evolutionary model with gene-fragment partitioning) and compared the trees with the non-subsampled FastTree topology by triplet distance (where 0 corresponds to identical tree topologies and 1 to no triplet in common). Between the 5 subsamples and the full-tree, only 0.60–0.64% of tip triplets (Supplementary Table 3) varied between IQ-TREE2 with gene partitioning and the FastTree analysis (normalized triplet distance). This validates the fast method with concatenated gene sequences. Molecular dating of subsamples also further validates the tMRCA found for the larger tree. We found all tMRCAs for subsamples (ranging from 1368

Andrew Holtz
PhD Evolutionary Biology / 2023
Frontiére de l'Innovation en Recherche et Education
Université de Paris Cité

to 1375) were very precise and all are located within the 95% CI of the full-canine tree tMRCA estimation, 1301–1403 (Supplementary Table 3, tMRCA).

Purification and diversifying selection does not have a substantial impact on RABV tree dating

To investigate the impact of selection pressure on dating, we used the aBSREL, FEL, and MEME methods in HyPhy and found evidence of purifying selection on 2793 sites out of 3490 variable sites tested on a WGS subsampled tree with 236 sequences from Troupin et al. [62]. We also found evidence of diversifying selection on 24 sites (p value < 0.01): (101, 436, 506, 587, 605, 633, 639, 716, 748, 839, 891, 969, 971, 1150, 1164, 1428, 1438, 1567, 1823, 1907, 3511, 3571, 3594, 3612). Notably, however, branch length-estimation for purifying and diversifying selection did not remarkably impact tree branch lengths and tMRCA estimates when compared to the original tree (intersecting confidence intervals) (Supplementary Fig. 4).

700 Years of canine-maintained RABV spread

To investigate the geographic transmission of canine RABV, we inferred the ancestral locations at the country level on the full and subsampled trees with PastML (https://github.com/amholtz/GlobalRabies/tree/main/data/ACR_Results) [142]. We created a consensus tree with 6,096 sequences by taking the union of all sequences in subsampled trees and reducing the full-canine tree through pruning. The clade ancestral estimates that were consistent between the aggregated subsamples and the full tree are shown in Fig. 3. For the first time, we estimated the country origin of 34 out of the 44 identified canine-mediated clades in the consensus tree, with ancestral estimates as well for their internal nodes, revealing historical transmission between and across the clades (Supplementary Table 4 and Fig. 3).

To resolve location uncertainty that remained from the country level ancestral character reconstruction (ACR) we reconstructed ancestral states again, by grouping countries into 23 geographic regions (18 of which are represented in our data set) previously defined by the World Bank grouping [143,144]. Out of 3114 internal nodes, the PastML MPPA method estimated a unique region for 3091 (99.26%) of the internal nodes. By grouping, we were able to estimate regional origins for nodes that are unresolved in the country reconstruction visualization (Fig. 4a). We estimated the regional origin for 100% of the previously defined canine-mediated clades [62] and the regional origin for 38/44 (86.7%) of their parental nodes (Supplementary Table 4).

Andrew Holtz
PhD Evolutionary Biology / 2023
Frontiére de l'Innovation en Recherche et Education
Université de Paris Cité

Consistent with previous studies, we estimate the presence of canine rabies in Eastern Asia by 1561 (1524–1594, light brown node "1561 Eastern Asia" in Fig. 4a) and in Southern Asia by 1760 (1716–1799, light blue node "1760 Southern Asia", Fig. 4a). From Eastern Asia, we estimate that there have been multiple introductions to South-East Asia (SEA, pink nodes in Fig. 4a), revealing a continuous reintroduction pattern in the region (Fig. 4b). Independent from the SEA clusters are cases of RABV from China and South Korea (Arctic-AL2, light brown node "1866 Eastern Asia" in Fig. 4a) that cluster with arctic sequences. Its ancestral origin was estimated to be China, dating back to 1866 (1843-1885) (Supplementary Table 4 and Fig. 3).

By regional grouping, we estimated the emergence and maintenance of RABV in West Africa by 1799 (1761–1832; 95% CI, salad-green node "1799 West Africa" in Fig. 4a; Supplementary Table 4 and Fig. 3). We can further define this emergence on the country level, where we see three diverging events within this West African clade associated with sequences from Central African Republic, Senegal, and Nigeria. The major node estimated to be Nigeria emerging in 1933 (1924–1943) contains 97% of the Africa-2 sequences.

We estimated the date of emergence and maintenance of the cosmopolitan clade slightly older than previous studies, 1656 (95% CI: 1627–1683) compared to 1720 (95% CI: 1642–1782) and 1730 (95% CI:1687–1773 [62]) (Fig. 2). This is the first study that estimates the ancestral geographic origin of the downstream subclades. At the regional level, our method estimates the ancestral origin as Northern America, dating to 1656 (Fig. 4, yellow node "1656 Northern America"). Northern America most likely represents sequences from early European colonization of the Americas. For subsequent subclades, we are able to confidently infer the country origin of 23 out of 25, and the regional origin of 25 out of 25 of the previously defined cosmopolitan subclades (Supplementary Table 4).

We can further visualize the subsequent emergence from Northern America (USA) to Central America in 1851 (1836–1864, pale green node "1851 Central America" in Fig. 4a), leading to cases in Cuba in 1905 (1898–1911, maroon node "1905 Caribbean", see also Figs. 4b and 3). We also infer that the AM2a subclade emerged via dissemination from the USA to Mexico in 1851 (1836–1864). Additional introductions in Southern America can be seen throughout the

Andrew Holtz
PhD Evolutionary Biology / 2023
Frontiére de l'Innovation en Recherche et Education
Université de Paris Cité

cosmopolitan clade during different periods but do not seem to be widespread or well sampled (Supplementary Table 4, Figs. 3 and 4).

The transmission of cases of canine rabies in Eastern and Southern Africa can be traced back to an introduction from Northern America (early European colonies) (Fig. 4) to an intermediary position (unresolved between Western Asia or Eastern Africa) in 1805 (1791–1818), and eventually to Eastern Africa in 1826 (1812–1837). We can see repetitive transmissions between Namibia, South Africa, and Zambia, representing Southern and Eastern Africa, supporting previous studies [108].

From the same intermediary region in 1805 (1898–1911), canine rabies further spread to Western Asia in 1809 (1794–1823), and eventually to Europe leading to the emergence of European fox rabies in 1873 (1863–1881, green node "1873 Eastern Europe" in Fig. 4). This is a major node of interest since it is a reflection point that led to non-imported European cases of Rabies. The CA1 subclade which contains sequences from Russia, Ukraine, Latvia, Tajikistan and even shows transmission to parts of China and Mongolia [106,107] is estimated to have emerged in 1924 (1913–1935) via Russia (Fig. 3).

Canine rabies present in the Western Europe and Central Europe subclades most likely originated from Germany in 1948 (1940–1954) and from either Germany or Poland in 1968 (1962–1973). Interestingly, these two subclades share an ancestral node dating back to 1931 (1921–1938) estimated as either Germany or Poland. Poland is also the inferred ancestral origin of the NEE subclade in 1887 (1873–1899) leading to cases in Lithuania, Latvia, Estonia, and Poland (Supplementary Table 4 and Fig. 3). Additional patterns of rabies emergence can be seen on the PastML compressed visualization (https://github.com/amholtz/GlobalRabies/tree/main/data/ACR_Results)[142] and the table of significant clades (Supplementary Table 4).

European colonization likely contributed to RABV transmission

For the first time, we reveal how historical colonization could have shaped canine rabies spread around the world by reconstructing ancestral scenarios on the colonization history of countries of isolation. Major nodes closest to the root are unresolved; however, the colony with the highest

Andrew Holtz
PhD Evolutionary Biology / 2023
Frontiére de l'Innovation en Recherche et Education
Université de Paris Cité

probability (35.9–38.5%) for these unresolved nodes is the British Empire (Fig. 5). This indicates that colonization could have played an instrumental role in the expansion of this strain of RABV. We also find that the French Empire is the inferred ancestral location (marginal probability of 84.1%) for the Africa-2 clade by 1799 (Fig. 5, "**") and the inferred ancestral location (marginal probability of 99.8%) for the majority of the SEA3 clade by 1942 (Fig. 5, "*"). Further, we estimate the Spanish Empire as the ancestral location (marginal probability of 85.5%) for the cosmopolitan clade (Fig. 5, "***") and subsequent AM2a subclade in Latin America. It is important to note that Mexico is grouped as the Spanish Empire in this analysis and is grouped as Northern America in the regional analysis. Although this shows how colonization could have sustained RABV spread in the world, there are many variables and factors not accounted for, such as early global trading companies[104,109,140].

Identification of human-mediated transmissions of canine RABV

To identify human-mediated events of rabies transmission, we consolidated transmissions in the full-canine tree across long distances (greater than 2000 km) and rate of spread (faster than 100 km/year) and/or over large bodies of water. Of 14,640 parent-to-child node transmissions observed in the tree, 1131 transmissions were identified between different countries, and 232 of these were to non-neighboring countries. Of these 232, we identified 43 transmissions of interest according to their speed of transmission (distance over time) and whether the transmission crossed a water barrier (Fig. 6 and Supplementary Table 5). The majority of the transmissions identified have not been previously reported and indicate cryptic transmission by human mobility. It is important to note that nodes without ancestral country estimates are ignored in this analysis.

2.4 Discussion

We present a phylogenetic pipeline that harnesses the information from partial and whole-genome sequences and their metadata to investigate the spatio-temporal dispersal of historic epidemics and the role of humans in their spread. Using this analysis, we were able to reveal the patterns and timing of the global spread of RABV.

The challenge in analyzing sequence data from pathogens that have existed for a long time is the variability in the quality and completeness of the data, such as the presence of partial genes versus complete genomes, the accuracy of the sampling date and location information, and the unequal representation of different geographic areas. When performing phylogenetic analyses on such data, a choice needs to be made between discarding some data by selecting a more homogeneous and higher-quality subsample (e.g., only the whole genomes or a particular gene) to avoid potential noise and keeping more or even all the data to increase the inference power. The former approach allows using more complex and hence time-consuming methods, while the latter requires faster methods as a result of many more sequences being available for inclusion. Using a larger data collection for phylogenetic and phylogeographic inference will result in the consideration of a wider range of sampling times, as well as data from additional countries or regions, allowing for a more detailed analysis and conclusion. Previous RABV studies [62,105,109,145] have opted for the former approach, focusing on smaller (hundreds of sequences) data sets representing specific partial genes or WGS, which were analyzed with Bayesian or ML inference methods. We investigate the opposite approach and shows its advantage in terms of inference power.

Our analysis is based on more than 10,000 RABV partial and whole-genome sequences available in the NCBI Virus database. Our pipeline employs a concatenated alignment of the five RABV gene fragments, with gaps for the regions absent in partial sequences (Supplementary Fig. 2). The phylogenetic analyses were performed with time-efficient inference methods suited for large data sets: an approximate maximum-likelihood phylogenetic analysis using FastTree, time-scaling with the least-square dating method of LSD2, and ancestral geographic character inference using a maximum-likelihood method implemented in PastML. We validated our pipeline by comparing its results to those from previous studies, to phylogenetic analyses using subsampled data sets and to those utilizing more complex evolutionary models accounting for potential evolutionary rate changes between genes and for potential selection pressure:[62,105,109,145] Using a larger data set, we obtained a compatible tree topology and compatible (between different trees and evolutionary models) but more precise dating (Fig. 2).

Additional challenges in pathogen spread analyses are posed by sampling bias in dating and country representation, which can influence phylogeographic reconstructions. The fact that with

Andrew Holtz
PhD Evolutionary Biology / 2023
Frontiére de l'Innovation en Recherche et Education
Université de Paris Cité

long-lasting epidemics (e.g. for centuries) even country definitions might change over time (e.g. from the British Empire to the United Kingdom; UK) further complicates the story. Our method attempts to remove sampling bias as a possible factor in our ancestral character reconstruction in three ways. First, we established a subsampling protocol which removes sequences from oversampled countries, since it is well known that ACR is heavily influenced by the number of sequences per character state [146]. In addition, we report ACR on both regional and country levels since there are some regions with less representation by countries. Finally, we were able to include subgenomic fragments independent of which gene is sequenced by representing the five genes of the RABV genome.

Employing time-efficient methods allowed us to harness the plethora of information of a very large data set to achieve a precise estimation of tMRCA of modern canine RABV and of dates of emergence for each of its 45 clades (Supplementary Table 4 and Fig. 3). Our estimates are compatible with previous studies by Troupin et al. (whole-genome data) [62] and Velasco-Villa et al. [140] (N gene) (Supplementary Table 4) (Fig. 2). However, our confidence intervals are almost two times narrower. For instance, we found the tMRCA to be 1356 (95% CI: 1301–1403), while Troupin et al. found the tMRCA to be 1404 (95% CI:1308–1510) [62] and Velasco-Villa et al. estimated 1435 (95% CI:1273–1562) [140] (Fig. 2). In addition, we also validated that, although present, purification selection [147,148] does not impact time-tree calibration of RABV. Interestingly, there are historical records of rabies predating our date of emergence in ancient Rome, Greece, and Egypt [92–94]. We believe that over history, there have been many different clades of RABV in circulation, and these versions were outcompeted by more infectious renditions. Our analysis is relevant for the current RABV that is in circulation today, which spread and replaced all other circulating RABV strains, aided by human urbanization, global trade, and dog introduction to the human home [105].

Despite the long history of canine-mediated RABV, it does not seem to be strongly impacted by recombination. This is due to the fast fatality rate and infection/transmission dynamics, which make coinfection of different virus types rare [149]. In contrast, bat communities can experience non-fatal RABV infections, allowing for multiple strain coinfections. Our study also shows a distinct geographic distribution of canine RABV genetic variants with minimal local variation.

Therefore, any recombination would likely occur among closely related viruses and remain concealed in large-scale phylogenetic analyses, like those in our analysis.

Considering these factors, we performed for the first time geographic ancestral character reconstruction on internal nodes of the global canine RABV phylogenetic tree. These estimates, however, are reliant on sequence availability and must be represented under this light. For example, we do not have records for historical rabies sequences that could have been circulating around the time of canine rabies expansion. Previous colonizing powers, like the UK, eradicated rabies before RABV sampling was possible [92]. The only representation we have of UK sequences are those left behind in their early colonies, such as the USA and Canada. In light of our analysis, we believe there is a cryptic node before the spread of canine rabies to Northern America in 1627–1683, which represents main European colonizers or global trade organizations. The cosmopolitan clade would originate from this node, which represents old, pre-eradication, European sequences. The node that defines Northern America would only lead to subsequent introductions to Central and South America.

The role of historical events in shaping RABV dispersal patterns is a central question that remains largely unanswered by measurable, quantitative approaches. To test the role of European colonization on historical RABV dispersal, we reconstructed the ancestral dispersal of canine RABV by the British, French, Spanish, Portuguese, and Russian colonial powers (Fig. 5). We observe some clear trends: the French Empire is associated with the spread of canine RABV in West Africa, the Spanish Empire with the spread in Central and South America, and the Russian Empire with the spread in Eastern Europe and Central Asia. The impact of the British Empire is, however, the most dominant throughout the phylogeny (Fig. 5) and is the most likely location for most ancestral nodes. This suggests that the British Empire played the most important role in the spread of canine RABV across the world. It is also the most likely location for the parental node shared by the Cosmopolitan and Arctic clades (Fig. 5) as well as defining the parental node that defines Africa-2. Further investigation could include additional international players of the time, such as trade networks by the Dutch East India Company (DEIC; 16th and 17th centuries) and the Chinese Ming Dynasty (fifteenth century) [62,140]. Indeed, the DEIC could be associated with spread of the Cosmopolitan clade, since the parental node (1526–1598) to the North American introduction has two branches: (1) hypothesized North American introduction 1627–1683, and

Andrew Holtz
PhD Evolutionary Biology / 2023
Frontiére de l'Innovation en Recherche et Education
Université de Paris Cité

(2) hypothesized introduction in South Africa in 1697–1752. The DEIC established trading routes with the USA in 1607 and with South Africa in 1652. Dogs were gifted for purposes of hunting, dog fighting, and companionship, and could have exacerbated the transmission of rabies around the world [150]. Similarly, the Zheng fleet of the Ming Dynasty (fifteenth century) could help explain introductions of RABV from an early unresolved ancestral node between 1341–1443 to Eastern (1524–1594), South-Eastern (1553–1623), and Southern (1716–1799) Asian nodes (Fig. 4a and Supplementary Table 4). The Zheng fleet established trade routes, often transporting animals, in Southern and Eastern Asia and the Arabian Sea during the 1400 s potentially introducing RABV to Eastern, Southern, and South-Eastern Asia [151]. However, we did not find evidence of early rabies transmissions from Eastern Asia to the Arabian Peninsula and Eastern Africa (Fig. 4a) despite the establishment of Zheng trade routes between 1405 and 1433.

The human-mediated importation of RABV was not unique to colonial times. Each year, there are instances of human-mediated introduction of RABV via immigration, travel, or dog importation [133]. We detected human-mediated transmissions on a time-calibrated tree with country annotations by searching for transmissions over branches representing long geographic distances in a short time frame [152]. Detectable instances must have a sequenced case in both the arrival (descendent) and departure (ancestor) country or some additional information on travel history [91]. Due to a lack of sequencing, the direction of importation cannot be reliably inferred. For the importation between France and Mexico (Fig. 6, superscript 2), we identified a known case of RABV which was sequenced in France on a sample obtained from a traveler who had returned from a trip to Mexico [153]. Therefore, the sequence labeled as 'French' actually represents the imported case from Mexico rather than a lineage currently circulating in France. This pattern could be present for other samples as well, such as cases between Spain and Morocco. Future work could incorporate travel history into the analysis to inform patterns of transmission [91]. Additional importations in our analysis (superscripts 1, 3, 4) correspond well to documented cases from historical records [153]. This approach is, therefore, not comprehensive but does provide a time and geographic range for possible human-mediated transmissions. In addition, using distance between most-populated cities in each country is an oversimplification of travel between two countries. Certain transmissions could be identified as significant if two cases might be from bordering areas (for example, a transmission between Tehran and Moscow might appear as significant (>3000 km) even if both cases were collected next to the Azerbaijani border, in Russia

Andrew Holtz
PhD Evolutionary Biology / 2023
Frontiére de l'Innovation en Recherche et Education
Université de Paris Cité

and Iran respectively). We do observe a trend that long-distance transmissions have increased over time. It is tempting to suggest, therefore, that long-distance canine RABV transmissions increase with globalization, although this is highly biased by the fact that we have greater resolution closest to the tips of the tree as a result of more sequencing.

Although we aimed to reduce sampling bias by subsampling our data set by country, utilizing all genes in NCBI, and grouping by region to reduce focus, our study still contains biases. A recent study sheds light on the bias in lineage definitions for RABV and introduces a proposal for new lineage definitions using dynamic nomenclature. Furthermore, a recent review, addresses the geographic bias resulting from the under-representation of sequences in countries with high numbers RABV-related deaths, such as India, Afghanistan, Russia, and middle/west Africa [111]. This bias requires the use of statistical methods to account for underrepresented countries. If all countries with RABV cases submitted sequences uniformly, the need to subsample would be less, and overall, the confidence of RABV phylogeographic studies would be stronger. In addition, it is important to note that while the number of RABV sequences shared on NCBI has increased since 1970, there has been a significant decrease in submissions in recent years. There are many factors that could explain this decrease including the COVID-19 pandemic, political issues within countries or decreased global funding for neglected tropical diseases. To address this issue, continued global and One-Health efforts to increase RABV surveillance and sequencing are necessary [102], as is the open sharing of that data to enable effective genomic surveillance of RABV.

By concatenating the five genes of the RABV genome and using time-efficient methods, we achieve the most precise date and geographic estimates for the emergence of canine-maintained RABV yet reported. This is the first investigation of this type on the global spread of canine RABV and provides valuable resources for epidemiological investigation of RABV introduction events and epidemic control. This concatenation method and dispersal history reconstruction will not only allow for a more precise understanding of global trends for rabies but also can be applied to other pathogens with a large deposit of partial sequences for phylogeographic investigation and dating purposes.

Andrew Holtz
PhD Evolutionary Biology / 2023
Frontiére de l'Innovation en Recherche et Education
Université de Paris Cité

2.5 Methods

Data set compilation

A total of 25,787 sequences of rabies virus (RABV) were obtained from the NCBI Virus database [127] in September 2021. A quality check was conducted to remove sequences that were (1) missing date and country information, (2) older than 1972 (for sequencing quality), (3) identified as vaccine or laboratory strains, (4) or shorter than 200 nucleotides in length. As a result, 14,752 sequences were retained for this study. Sequence filtering is demonstrated in Supplementary Fig. 1, and a complete list of sequences with criteria for exclusion can be found in Supplementary Table 1 ('exclusion' column).

Sequence alignment

An initial global alignment using a custom script (https://github.com/amholtz/GlobalRabies/blob/main/R/clean_RABV.R) [142] was used to define the sequences based on which gene their nucleotides correspond to. Sequences with more than 200 nucleotides in a gene-coding region, were sorted into gene-specific fasta files. For example, WGS were included in the N, P, M, G, and L files. All sequences in these independent files were then cut at the start codons (according to the reference genome) of the corresponding gene (top, Supplementary Fig. 2).

Each newly cut gene-specific fasta file (void of non-coding regions) was then aligned against the (also cut) corresponding reference sequence, using MAFFT (v7.505) with the FFT-NS-1 strategy, addfragments, and keep-length method [7]. This resulted in five multiple sequence alignments for each gene-coding region: N gene (1353 nt), P gene (894 nt), M gene (609 nt), G gene (1575 nt), and L gene (6429 nt). Once all genes were realigned, they were concatenated by ID using Goalign (v0.3.5) [154] column-wise in original order, introducing gaps (-) where gene fragments were missing (non-sequenced locations for each ID). The final product is a global multiple sequence alignment, called concat_seq_genes.fasta with a total length of 10,860 nucleotides (bottom right, Supplementary Fig. 2).

Phylogenetic estimation, dating, and subsampling

All RABV sequences (bats, skunks, canines)

Andrew Holtz
PhD Evolutionary Biology / 2023
Frontiére de l'Innovation en Recherche et Education
Université de Paris Cité

FastTree (v2.1.11)[2] was used for the initial phylogenetic analysis of the 14,752 sequences to confirm the separation of the previously identified clades by Troupin et al. [62]: (1) bats, (2) skunks and raccoons, and (3) canines and their myriad subclades (ran on 32 threads for approximately 1 h:15 min). To this end, the GTR nucleotide substitution model [75] and a discretized gamma distribution to accommodate among-site rate heterogeneity were employed, and bootstrap support values were estimated using default settings (Shimodaira-Hasegawa test [12,155]) (Supplementary Fig. 3).

Canine sequences

The canine-mediated RABV clade was identified and a subset alignment from the original was created (n = 10209). A canine-mediated tree was reconstructed from this alignment using FastTree (v2.1.11) [140] (ran on 32 threads for approximately 47 min). The evolutionary rate was estimated using whole-genome sequences exclusively to mitigate potential sequencing errors from gene-specific projects. To estimate the evolutionary rate, the least-squares dating (LSD2) method version 1.8.8 [6] was used on the whole-genome-only canine tree without replacing the root and with default outlier removal. The resulting evolutionary rate was then provided as a fixed parameter in LSD2 [6] to yield a time-calibrated tree for the full phylogenetic tree with 10,209 sequences with outlier removal and with confidence intervals based on 1000 replicates. A total of 165 sequences were detected as outliers and removed, according to a Z-score of 3 (listed in Supplementary Table 1), resulting in a full-dated canine-derived phylogenetic tree with 10,044 sequences, referred to as the full-canine tree henceforth.

To visualize differences in rates of evolution according to each gene, phylogenetic trees using FastTree (v2.1.11) [140] were estimated using the previously reported settings, according to five gene-specific multiple sequence alignments. Gene-specific evolutionary rates were then determined using LSD2 [6] following outlier removal (Supplementary Fig. 5).

Subsampling

The full-canine tree was subsampled to reduce geographic bias towards countries with a disproportionally high number of sequences by employing a custom script (https://github.com/amholtz/GlobalRabies/blob/main/python/py_subsampling.py)[142] which removes sequences from oversampled countries by phylogenetic diversity [156] given an input tree.

Andrew Holtz
PhD Evolutionary Biology / 2023
Frontiére de l'Innovation en Recherche et Education
Université de Paris Cité

At each step of the algorithm, the tree tip with the shortest branch among those sequences corresponding to over-represented countries is removed. The process is repeated until a desired target number of sequences is reached for all the countries. As several candidates for removal might exist at each step (i.e. tips with the same, shortest, branch length), the tip to be removed is chosen randomly in such cases. The subsampling was repeated five times, generating slightly different (due to stochasticity) trees with 5500 sequences each. Phylogenies from these subsamples were estimated using IQTREE-2 (v2.2.2.2) [4] employing the GTR + I + G4 model (determined by ModelFinder [157]) with gene partitioning to account for potential variation in substitution rates by gene and evaluated with 1000 ultrafast bootstraps [4,158,159] (ran on 8 threads for 164 h:28 min). The resulting tree topologies were compared with the full-canine tree by measuring triplet distance [160,161]. The subsampled phylogenies were time-calibrated in the same way as the full-canine phylogeny and the tMRCA estimates were compared. Finally, the ancestral country reconstructions on the subsampled and full timetrees were compared (see below).

Purifying selection model

To quantify the effect of purifying and diversifying selection on the time calibration, we used the software package HyPhy [3] on the smaller data set of 236 whole-genome sequence subset from Troupin et al. [62] to lighten the computational load. This subset is made of only WGS and its dating has been previously analyzed, such that it is useful to assess model variations. Site-specific rates of the numbers of nonsynonymous and synonymous (dN and dS) substitutions per set were inferred in each codon using the Mixed Effects Model of Evolution (MEME) [162], Fixed Effects Likelihood (FEL) [163] and the adaptive Branch-Site Random Effects model (aBSREL) [164] in HyPhy. The dN/dS ratio was then used by each model to optimize and re-estimate branch lengths of the tree with a fixed topology (Supplementary Methods 1). The branch lengths of the initial tree topology were re-estimated under each model, and the resulting trees were then dated using LSD2. The root dates were compared to those of the initial tree (estimated under the GTR model) (Supplementary Fig. 4).

Phylogeographic reconstruction and subsampling consensus

Ancestral geographic characters of canine RABV transmission were investigated using the fast likelihood method available in PastML version 1.9.34 [10] to reconstruct the historical spread of canine RABV. Marginal probabilities of locations of ancestral nodes were estimated given the

Andrew Holtz
PhD Evolutionary Biology / 2023
Frontiére de l'Innovation en Recherche et Education
Université de Paris Cité

geographic location of sampled sequences in a rooted time-calibrated tree. To assess the confidence in state estimates of internal nodes with respect to potential geographic sampling bias, we repeated PastML on each of the five independent and uniquely subsampled reconstructed and time-calibrated trees and compared the ancestral state estimates. Only ancestral state estimates with marginal posterior probabilities of 50% or greater were considered. A custom script (https://github.com/amholtz/GlobalRabies/blob/main/R/ACR_Sub_comparison.R) [142] was used to identify and compare nodes with identical cluster compositions between the subsampled and full-canine trees to create an aggregated consensus country-annotation (Supplementary Fig. 6 and Supplementary Methods 2). The consensus tree was created by pruning the full tree to keep the union of all sequences in subsampled trees, yielding a final tree with 6096 sequences (Fig. 3), and the aggregated ACR result was used as a new annotation for the consensus tree using PastML.

Regional and colonization history ancestral character reconstruction

ACR using PastML version 1.9.34_[10] was repeated on the 6096-sequence aggregated tree for two geographic characters. The first defined the character using a World Bank regional grouping variable using R package countrycode [143] (Fig. 4) which reflects cultural and geographic similarities. This resulted in the grouping of 120 countries in the original data set into 18 regions. The second defined the character according to the colonial history of the sequence country of isolation (Fig. 5). Colonial powers were considered from 1600 to 1950, resulting in five states: British, Spanish, French, Portuguese, and Russian Empires (not including USSR satellite states). Countries with more than one colonial history were considered a member of the empire that maintained the longest and largest control (for example, India is grouped by the British Empire, even though a small part was also a French colony for several hundred years). Countries that have never been colonized were grouped as 'non-colonial' (Supplementary Table 1). The aggregated tree was used since it has been subsampled to limit bias from geographic origin. For both character analyses, the marginal posterior probabilities approximation (MPPA) method [10], was used for ancestral character estimation. MPPA uses decision-theory concepts to estimate a unique state in tree regions with low uncertainty, and several states in uncertain ones (typically around the root).

Andrew Holtz
PhD Evolutionary Biology / 2023
Frontiére de l'Innovation en Recherche et Education
Université de Paris Cité

Introduction event exploration

To visualize human-mediated transmissions of RABV, long geographic distances over short time intervals were located on the phylogeny using a custom script (https://github.com/amholtz/GlobalRabies/blob/main/R/introduction_events_FullTree.R) [142]. The package cepiigeodist [165] was used to calculate geographic distances between each most-populated city in the country. Character states from the phylogeographic analysis were considered only if there was a consensus between the aggregated subsamples and the full-canine tree. A branch of length Y with a resolved consistent parent node state A and a resolved consistent child node state B was considered as a transmission from country A to country B over Y years. Transmissions within countries and between neighboring countries were discarded, resulting in 232 long-range transmissions. Of these transmissions, those over a distance of more than 2000 km or that were across a water barrier and had a transmission dispersal interval faster than 200 km/year were conserved (Supplementary Table 3), resulting in 43 transmissions.

2.6 Acknowledgements

The authors thank S.L. Kosakovsky Pond for his help running HyPhy and interpreting the results. We would also like to thank Edward Holmes, Jake Faber and Joël Brehin for their helpful reading and critical comments in the review of the manuscript. G.B. acknowledges support from the Research Foundation - Flanders (Fonds voor Wetenschappelijk Onderzoek - Vlaanderen, FWO, Belgium; grant n° G098321N, G0E1420N) and from the Internal Funds KU Leuven (Grant No. C14/18/094). A.H. acknowledges École doctorale Frontières de l'Innovation en Recherche et Education-Programme Bettencourt. A.H is funded by the INCEPTION programme (Investissements d'Avenir grant ANR-16-CONV-0005). This work was funded by Institut Pasteur.

AUTHOR CONTRIBUTIONS

A.H. and A.Z. conceived the idea and developed the theory. H.B. provided insight to rabies epidemiology. A.H. carried out the computational, phylogenetic and phylogeographic analysis with the supervision of A.Z. A.H. wrote the manuscript in consultation with A.Z. All authors provided critical feedback and helped shape the research and analysis. H.B. and G.B provided scientific and technical guidance and review of the paper.

COMPETING INTERESTS

83

Andrew Holtz
PhD Evolutionary Biology / 2023
Frontiére de l'Innovation en Recherche et Education
Université de Paris Cité

The authors declare no competing interests.

PEER REVIEW INFORMATION

Nature Communications thanks Joseph Fauver, Andrés Velasco-Villa and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. A peer review file is available.

2.7 Figures

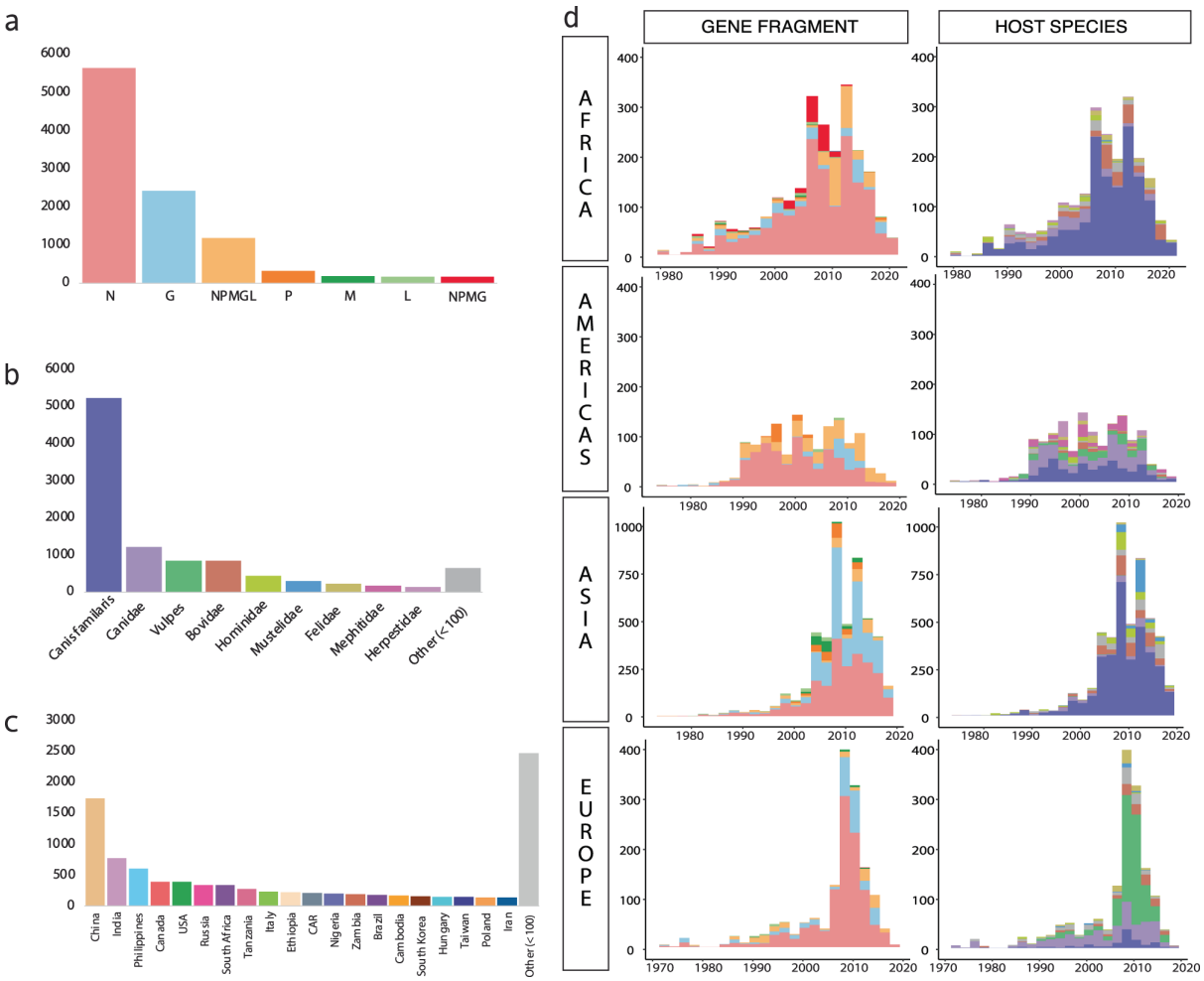


***Figure 1. Metadata exploration of the canine RABV sequences**. Composition by (a) gene fragment, (b) host species, (c) country of origin, (d) evolution of sequence deposition over time by fragment and host sorted by continent represented by stacked bar plots showing accumulated totals per year. Color definitions of host species and gene fragment are seen in plots a and b.*

Andrew Holtz

PhD Evolutionary Biology / 2023

Frontiére de l'Innovation en Recherche et Education

Université de Paris Cité

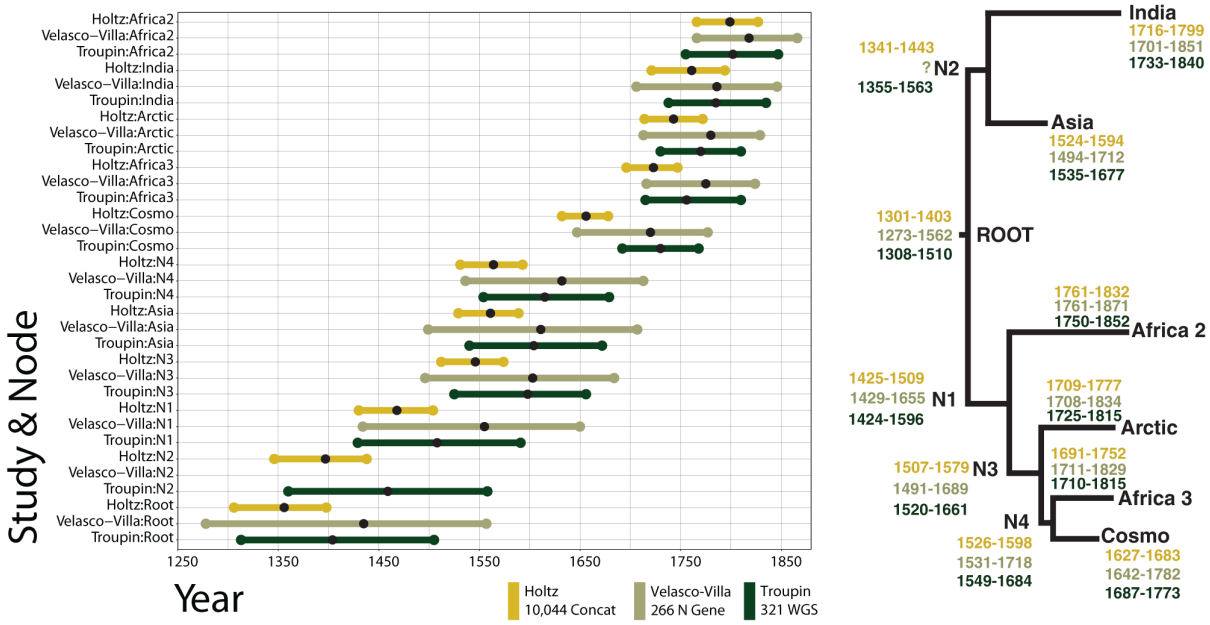


Figure 2. Comparison of previously published global time-calibrated trees of canine-mediated RABV. *The present study is shown in comparison with two previously BEAST reconstructed timetrees, Velasco-Villa et al.[140] used 266 nucleoprotein (N) gene sequences and Troupin et al.[62] used 321 WGS. 95% Confidence intervals of all node defining-major clades and ancestral nodes to the root are plotted. The bullseye on each line represents the reported point estimates. Locations of each node (presented by N) on the tree are shown on the right, as well as the numerical confidence intervals for the reported date estimate. Velasco-Villa et al.[140] does not report the date estimation or confidence interval for N2.*

Andrew Holtz

PhD Evolutionary Biology / 2023

Frontiére de l'Innovation en Recherche et Education

Université de Paris Cité

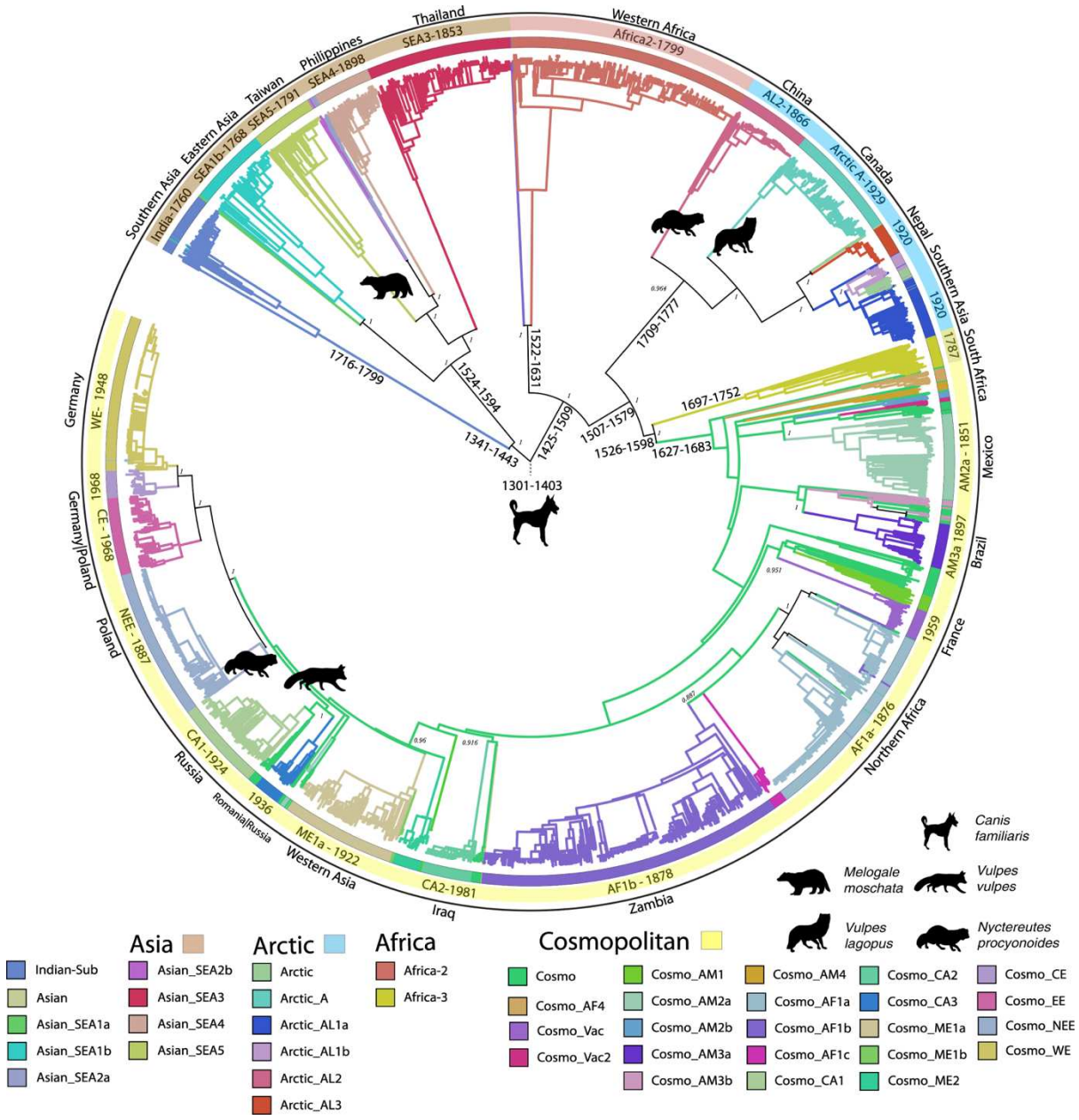


Figure 3. ML consensus phylogenetic tree of canine RABV sequences (n = 6096). *Confidence intervals of LSD2 dating are provided on the most internal nodes. Branch colors and colorstrip present estimated ancestral clade by ancestral character reconstruction (ACR). Unresolved branches are black (the legend is shown below the tree). Names of clades and date of clade tMRCA are provided if space is sufficient. The outer label presents the country or regional ancestral estimate for that clade. Canis familiaris is the most dominant host species in the tree. Clades with a majority other than Canis familiaris are shown on the clade-defining branch. Bootstrap support from 1000 replicates is shown on informative clade-defining nodes (silhouette images of animals are sourced from http://phylopic.org/. These are available for reuse under the CC0 1.0 Universal Public Domain Dedication license).*

Andrew Holtz

PhD Evolutionary Biology / 2023

Frontiére de l'Innovation en Recherche et Education

Université de Paris Cité

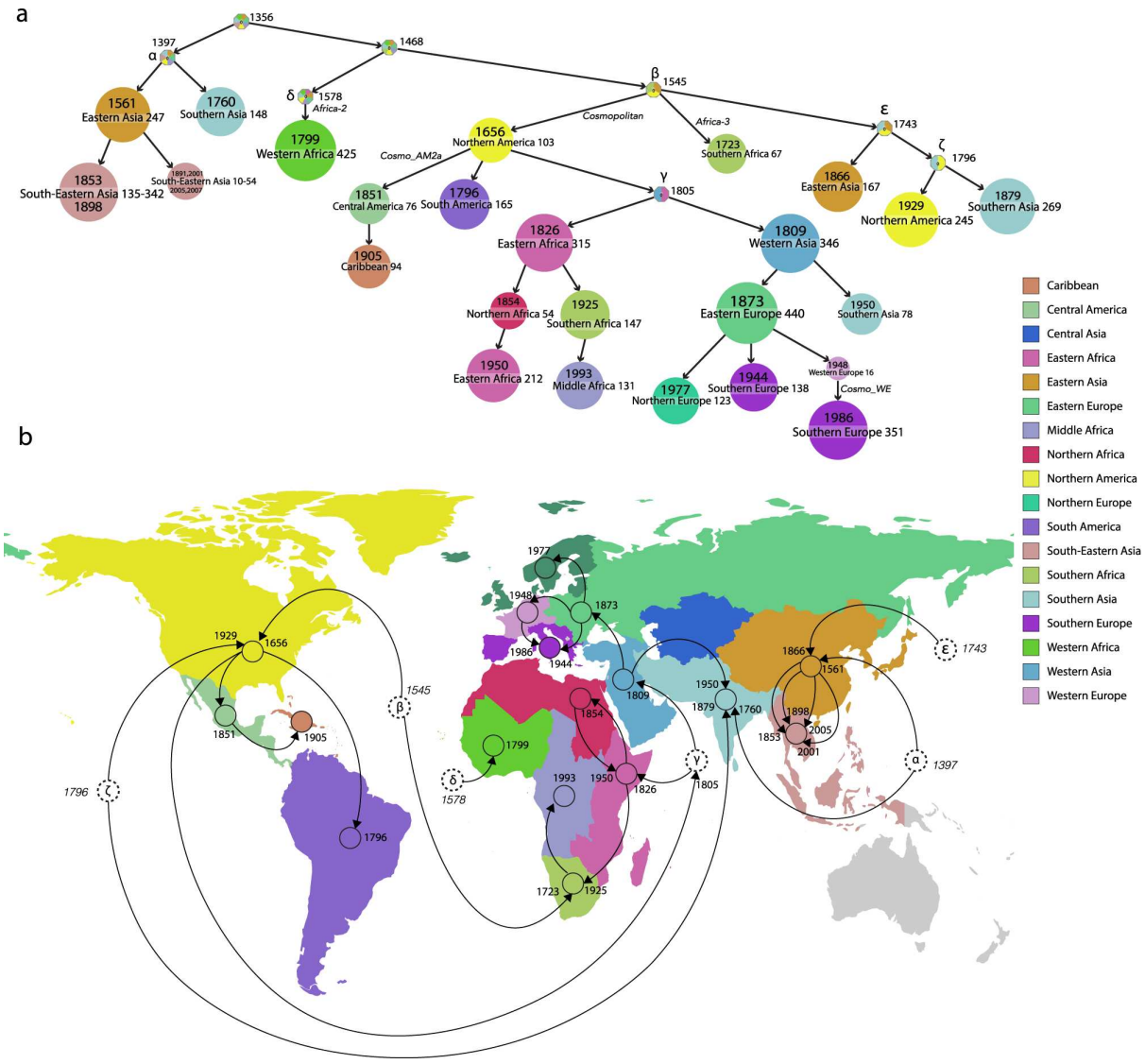


Figure 4. Regional ancestral character reconstruction for canine RABV tree from PastML MPPA. ***a*** *Compressed visualization of consensus tree (of 6096 samples), where the parts of the tree without regional changes are clustered. For each cluster, the date of the most ancestral node, its region and the number of samples represented are shown. The clusters are colored by region (the color legend is shown on the right).* ***b*** *RABV regional spread represented on the world map. Dates represent the end of the branch, while italicized dates represent the dates of unresolved nodes (dotted circles with greek letters corresponding to* ***a****). Dotted circles without a prior path are unresolved up until the root, which is dated at 1356. Confidence intervals can be seen in Supplementary Table 4. Circle locations represent the region rather than country of origin. For example, 138 of the sequences in South-Eastern Asia are isolated from the Philippines, even though the circle is located in Thailand. Grayed regions represent regions with no samples in our data set.*

Andrew Holtz

PhD Evolutionary Biology / 2023

Frontiére de l'Innovation en Recherche et Education

Université de Paris Cité

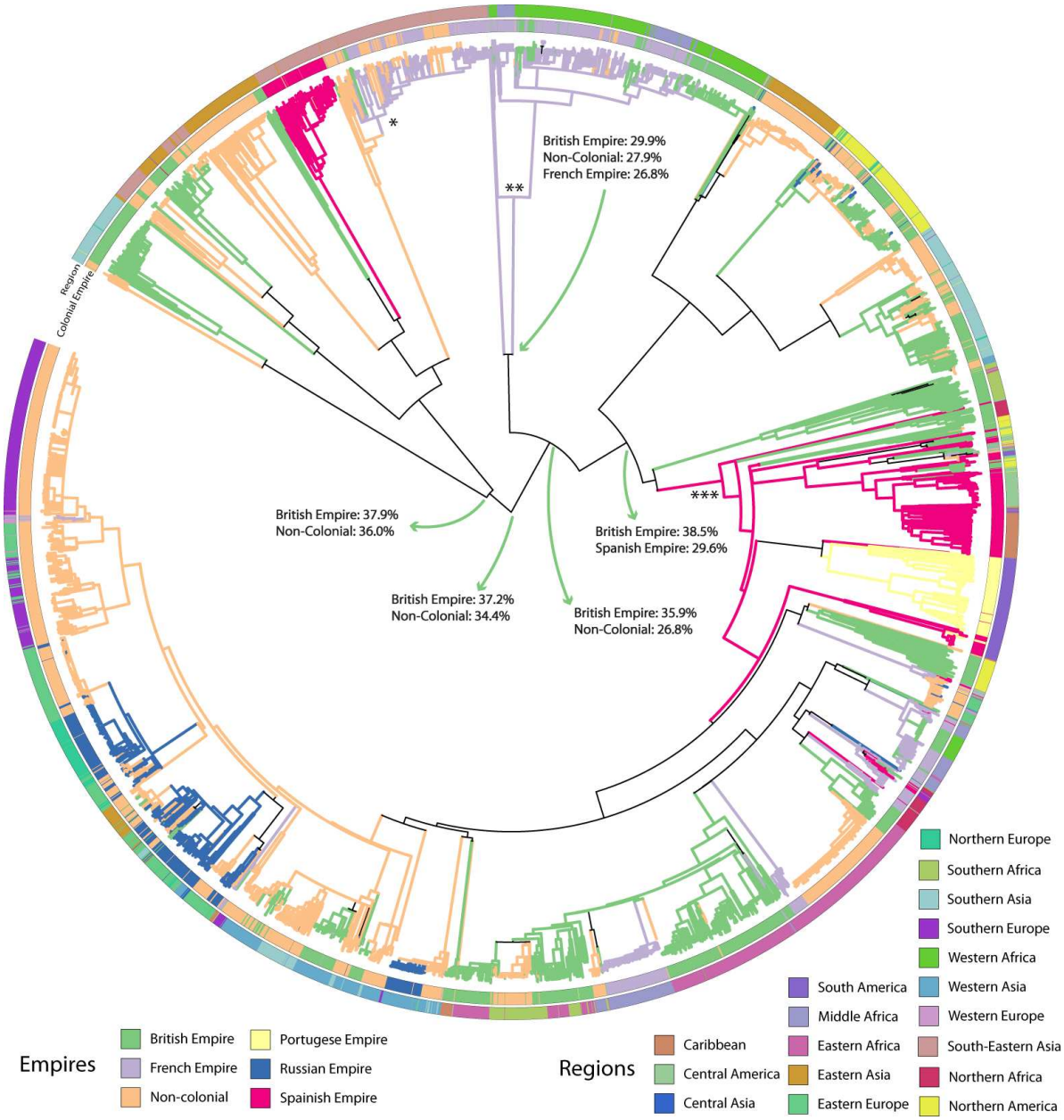


Figure 5. RABV transmission history organized by colonial powers from 1600 to 1950s. *The consensus tree represents 6,096 sequences, colored according to the inferred colonial empires at the nodes (the color code is explained in the legend on the bottom). The color strips around the tree represent historical empires (inner) and regional grouping (outer). Nodes of interest indicated: (*) 99.8% French Empire- 1942, (**) 84.1% French Empire-1799 (***) 85.5% Spanish Empire 1656. The visualization is performed with iTOL [5] on a PastML [10]-annotated tree.*

Andrew Holtz

PhD Evolutionary Biology / 2023

Frontiére de l'Innovation en Recherche et Education

Université de Paris Cité

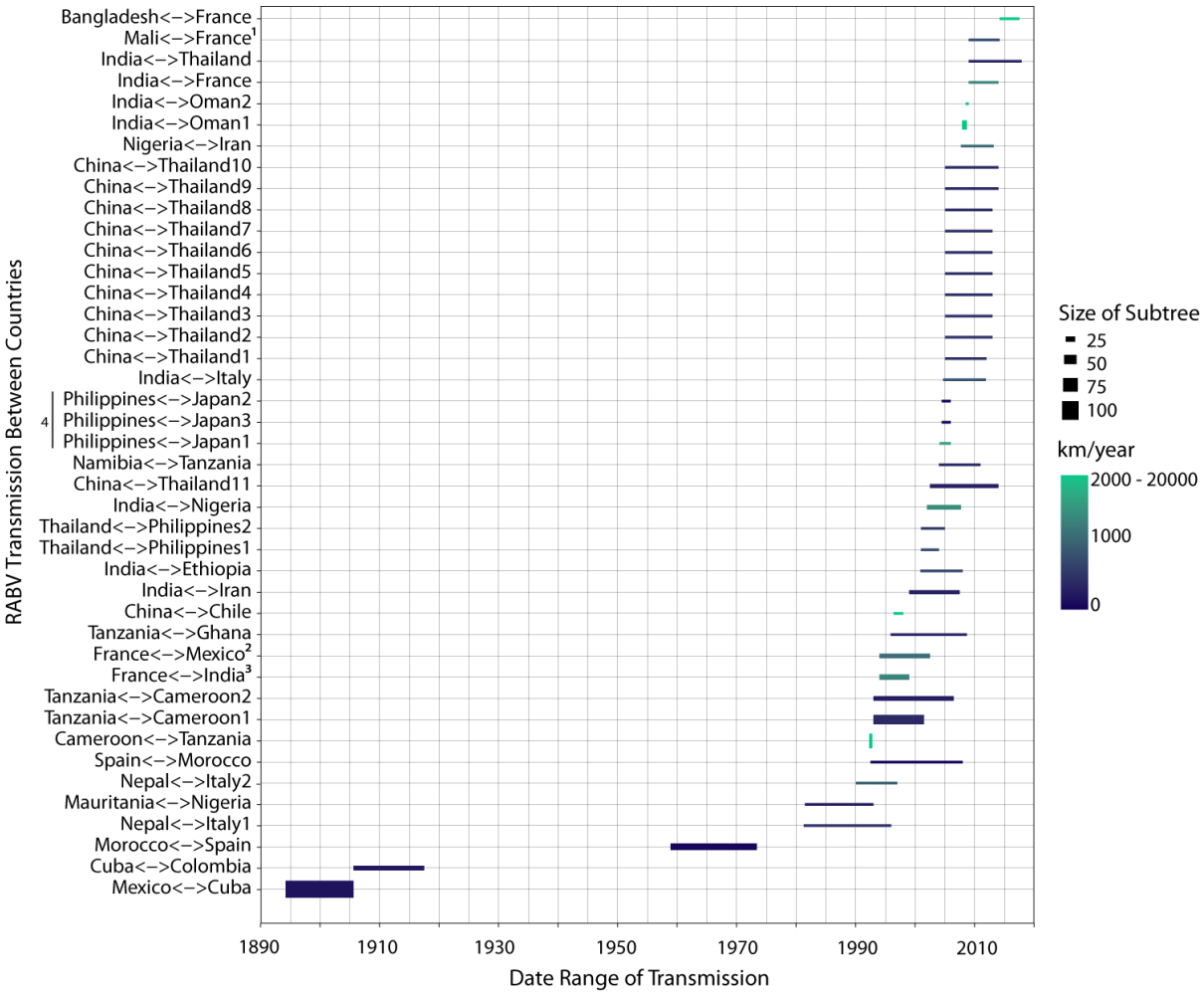
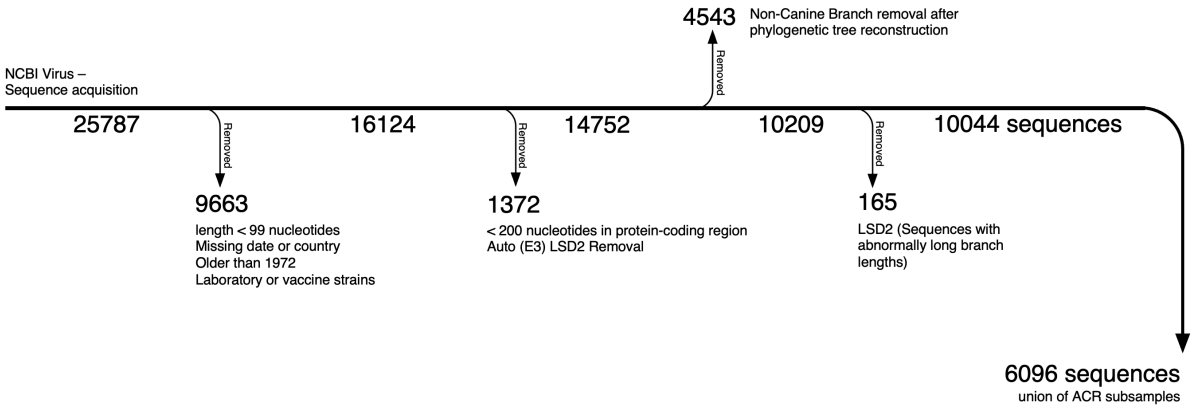


Figure 6. Inferred human-mediated introductions identified on the phylogenetic tree by date. *Human-mediated transmissions are presented by time of transmission between estimated parent and child dates. Transmissions between two countries are shown on the y axis. In some cases, there are multiple transmissions between two countries. The size of the subtree (descendants) from the transmission is shown by line width. For example, the child node for Mexico-Cuba has 109 descendants. Transmission speed is represented by the geographic distance (km) over one year and is calculated by the distance between the most-populated city between parent and child and the branch length. Human-mediated introductions are defined as transmissions with speed faster than 200 km/year and between two countries that are more than 2000 km apart or separated by a body of water (e.g. seas/oceans). Certain transmissions (identified by superscripts) can be linked to documented introductions (see Discussion). This figure depicts importations between two countries rather than indicating the direction of importation due to limited sequence availability.*

2.8 **Supplemental Information**

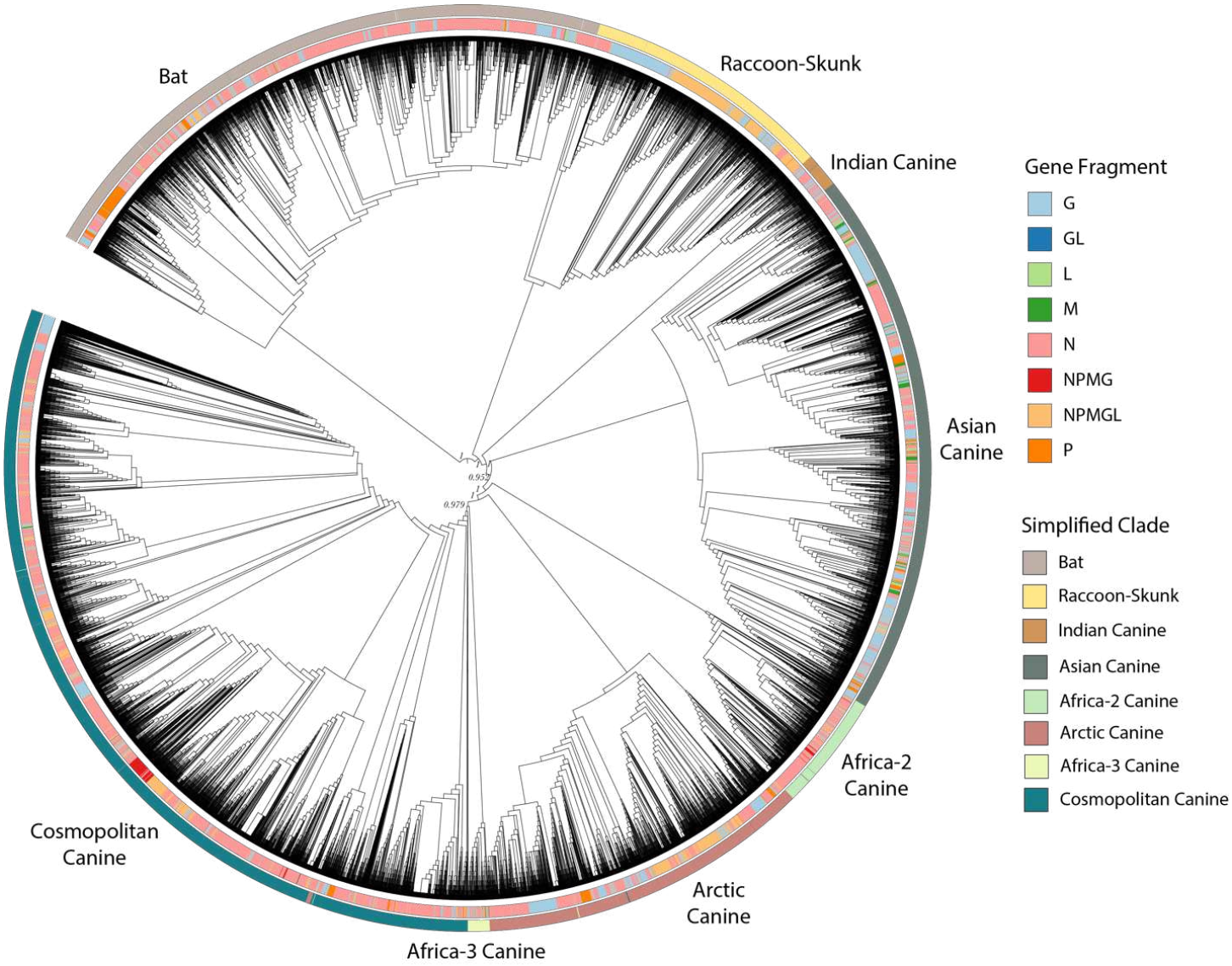


Supplementary Figure 1. *Flowchart of inclusion and exclusion of RABV sequences downloaded from the NCBI Virus database. The number of sequences excluded and the reason for their removal is shown. The number of sequences present at each step is visible under the horizontal line.*

Andrew Holtz
PhD Evolutionary Biology / 2023
Frontiére de l'Innovation en Recherche et Education
Université de Paris Cité

Supplementary Figure 2. Cartoon demonstration on the method of partial and WGS sequence concatenation. *Sequences were defined by an initial global alignment using a custom script, and then sorted into gene-specific fasta files with sequences cut at the start codons of the corresponding gene. The gene-specific fasta files were aligned and concatenated to form a global multiple sequence alignment with a total length of 10,860 nucleotides. The illustration of the RABV genome is shown at the top. Numbers inside each gene area define the positions from the reference genome where the gene-specific sequences were cut. The final resulting multiple sequence alignment is shown in the box on the bottom right. Corresponding file names on the GitHub page (https://github.com/amholtz/GlobalRabies/tree/main/data/sequence_alignments/gene_specific_analysis) are given.*

Andrew Holtz
PhD Evolutionary Biology / 2023
Frontiére de l'Innovation en Recherche et Education
Université de Paris Cité



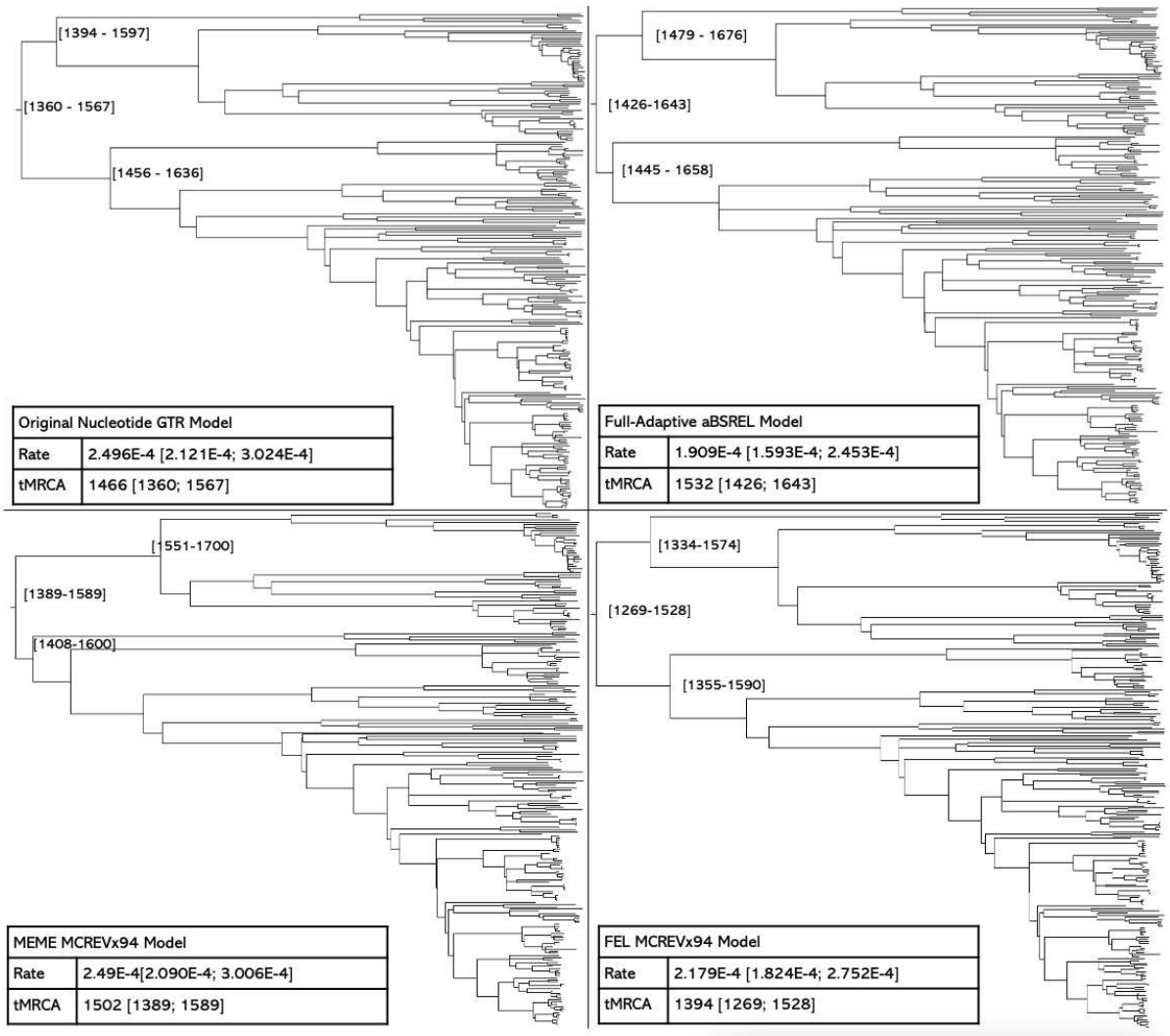
Supplementary Figure 3. Phylogenetic analysis of 14,752 RABV sequences. *Sequences are labeled by simplified major clades (outer circle) and gene fragment (inner circle). Phylogenetic grouping by clade is clear, while there is a scattered distribution by gene fragment. Bootstrap support values (FastTrees Shimodaira-Hasegawa test) of simplified clade defining nodes are displayed.*

Andrew Holtz

PhD Evolutionary Biology / 2023

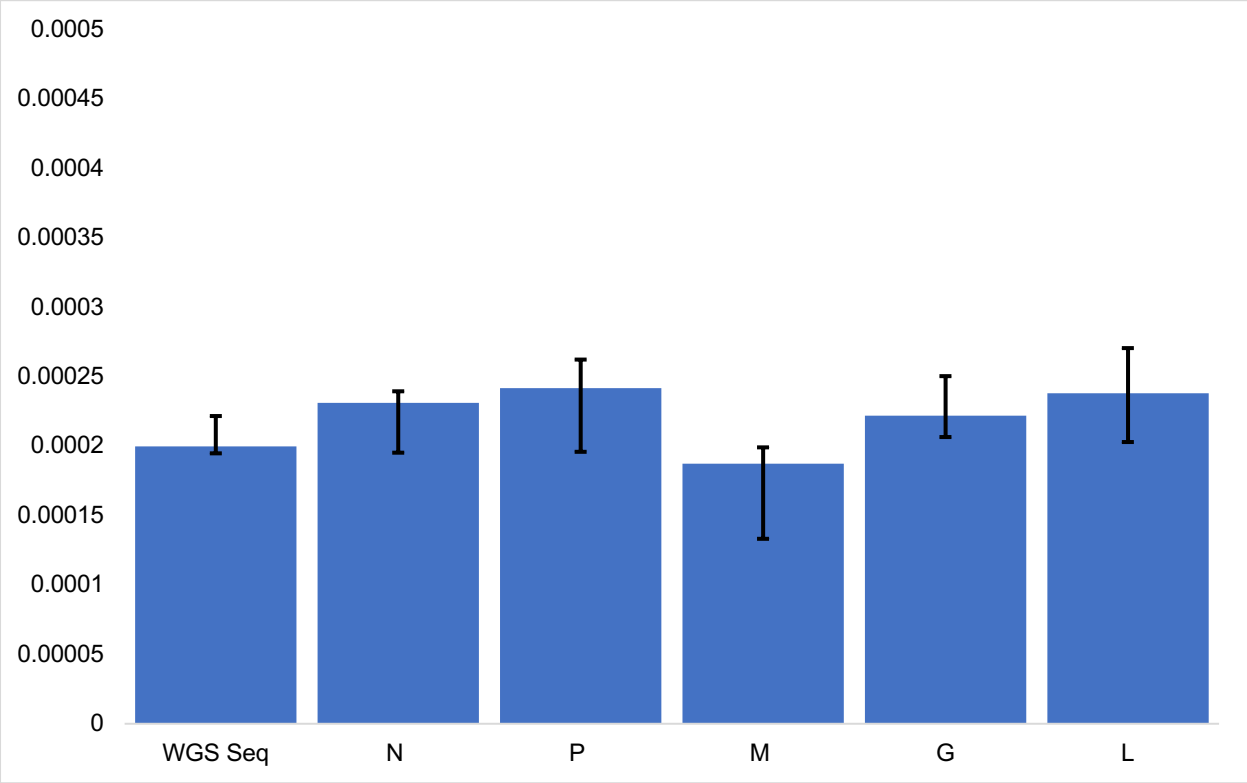
Frontiére de l'Innovation en Recherche et Education

Université de Paris Cité

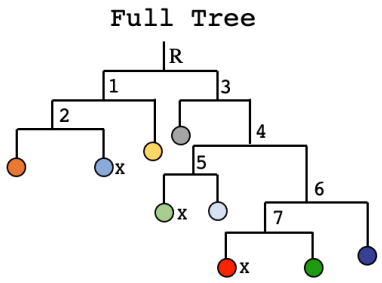


Supplementary Figure 4. Time-calibrated phylogenetic tree after branch-length optimization. *Branch lengths were re-estimated by positive and purifying selection models in HyPhy by aBSREL, FEL, and MEME.*

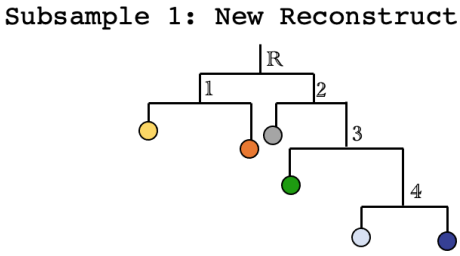
Andrew Holtz
PhD Evolutionary Biology / 2023
Frontiére de l'Innovation en Recherche et Education
Université de Paris Cité



Supplementary Figure 5. Evolutionary rate of RABV genes in the canine-maintained cluster. *Evolutionary rates were estimated by LSD2 from gene-specific MSAs with non-coding regions removed. Non-coding regions were not removed from WGS sequences. Regions tested include whole-genome sequences (WGS), nucleoprotein (N), phosphoprotein (P), matrix (M), glycoprotein (G) and polymerase (L). The point rate estimates are shown as vertical bars and are as follows: WGS (2.08E-4), N(2.17E-4), P(2.29E-4), M(1.66E-4), G(2.28E-4), L(2.37E-4). Error bars present 95% confidence intervals (CIs) from 1000 replicates. The LSD2 CIs are generated by parametric bootstrap: (i) A set of phylogenetic trees is generated by keeping the same topology as in the input tree and pooling each branch length from a Poisson distribution with the mean estimated by LSD2; (ii) The evolutionary rate is estimated on each simulated tree; (iii) The rate CIs are obtained from the 95% quantile of these rates.*



FT:N	FT: ACR	FT: Tips
R	?	
1	France	
2	Germany	
3	Spain	
4	Netherlands	
5	UK	
6	Austria	
7	Belgium	



Sl:N	Sl: ACR	Sl: Tips	FT:N	FT:ACR
R	?		R	?
1	France		1	France
2	Spain		3	Spain
3	UK		4	Netherlands
4	Belgium		4	Netherlands

Supplementary Figure 6. Mock example of sub node comparison. Full Tree contains nine tips. *Tips removed during subsampling are marked with an X. Subsample 1 contains 5 tips. Subsample 1 has been completely newly reconstructed using IQ-TREE2 starting from the subsampled sequence alignment. Notice that the topologies between the Full Tree (FT) and Subsample 1 are not identical (the light blue and green nodes have swapped). The goal is to find the node in Full Tree which represents the same tips in each node of Subsample 1. To do this, we want to find the node in Full Tree that contains all the tips found under a given node in Subsample 1. We pick the most specific node (i.e., with the least tips) among such nodes in the Full tree (or else each node in the subsample tree could be compared to the root since all tips are found under the root). In this example, node 4 in Subsample 1, can be compared to node 4 in Full Tree. Node 3 can be compared to node 4 as well. Node 2 in Subsample 1 can be compared to node 3 in Full Tree. Node 1 in Subsample 1 can be compared to Node 1 in Full Tree. If you now compare the ACR, you can see that the full tree nodes that share the same ACR are 1 and 3. This is the idea of the node comparison script which was applied on a much larger scale.*

Andrew Holtz

PhD Evolutionary Biology / 2023

Frontiére de l'Innovation en Recherche et Education

Université de Paris Cité

Supplementary Table 1. Full RABV Metadata with added columns and exclusion reason

https://github.com/amholtz/GlobalRabies/blob/main/data/metadata_edited_exclusion.tab

https://amholtz.github.io/GlobalRabies/

Supplementary Table 2. Subgenomic fragments by sequence and country. *Number of sequences and countries represented by each sequence fragment.*

Fragment	Sequences	Countries
N	8136	120
G	3649	85
L	254	77
P	572	78
M	196	77
WGS	1972	75
Total	14779	121

Andrew Holtz
PhD Evolutionary Biology / 2023
Frontiére de l'Innovation en Recherche et Education
Université de Paris Cité

Supplementary Table 3. Full Canine RABV tree compared to subsampled trees. *Subsamples are the result of a subsampling protocol which produces a unique set of sequences for a more equal representation of countries. Internal nodes and their geographic estimates were compared to the full tree. Nodes with estimated character state probabilities greater than 50% were considered resolved. Internal nodes of two trees are compatible if their subtrees have the same tips. Of the compatible nodes in a subsample and the full tree, the estimated character (if resolved in both) was compared. 'Agg' column presents the full tree that has been pruned by the aggregation of sequences represented by subsamples 1-5. Triplet distance represents the fraction of triplets that differ across the trees normalized with all triplet possibilities in the compared trees.*

	Full Tree	**Sub 1**	**Sub 2**	**Sub 3**	**Sub 4**	**Sub 5**	**Agg**
tMRCA	1356 [1301-1403]	1375 [1310-1429]	1375 [1327-1416]	1369 [1310-1416]	1374 [1317-1423]	1368 [1311-1417]	1356 [1301-1403]
num. tips	10044	5371	5367	5367	5372	5371	6096
num. internal nodes	4779	2729	2724	2751	2713	2715	3114
normalized triplet distance to Full Tree	-	0.0062	0.0064	0.0062	0.0063	0.0060	-
# reso. internal nodes (%)	4706 (98.5%)	2673 (97.9%)	2665 (97.8%)	2687 (97.7%)	2646 (97.5%)	2656 (97.8%)	-
inter. nodes Full Tree compatible (%)	-	60.16%	60.84%	59.03%	59.94%	59.65%	-
Full Tree shared ACR	-	1770 (94.40%)	1741 (94.47%)	1817 (94.24%)	1778 (94.32%)	1787 (94.10%)	2169 (93.61%)

Andrew Holtz

PhD Evolutionary Biology / 2023

Frontiére de l'Innovation en Recherche et Education

Université de Paris Cité

(of compatible)							

Supplementary Table 4. Ancestral character reconstruction (ACR) by region and country of both parental and child node for 44 canine subclades. *95% confidence intervals are provided for the tMRCA for each clade. tMRCA of previous Troupin et al. [26] are displayed. Rows with more than one country show the conflict between the confidence analysis between the full tree ACR result and the subsamples (Full Tree | Subsampled aggregation).*

Clade Definition	Regional Origin	Country Origin	TMCRA	Troupin et al. [26] TMRCA	Parent Region	Parent Country	Parent - TMRCA
Africa-2	Western Africa		1799 (1761 - 1832)	1802 (1750 - 1852)			1578 (1522 - 163)
Africa-3	Southern Africa	South Africa	1723 (1691 - 1752)	1756 (1710 - 1815)			1564 (1526 - 159)
Arctic_A	Northern America	Canada	1929 (1918 - 1939)	1942 (1929 - 1954)			1796 (1766 - 1824)
Arctic_AL1a	Southern Asia		1920 (1905 - 1931)	1940 (1927 - 1953)	Southern Asia		1879 (1859 - 1898)
Arctic_AL1b	Southern Asia		1948 (1940 - 1954)	1936 (1919 - 1953)	Southern Asia		1938 (1931 - 1945)
Arctic_AL2	Eastern Asia	China	1866 (1843 - 1885)	1886 (1852 - 1921)			1743 (1709 - 1777)
Arctic_AL3	Southern Asia	Nepal	1977 (1973 - 1981)	1881 (1856 - 1906)	Southern Asia		1907 (1892 - 1921)
Asian_SEA1a	Eastern Asia	China	1904 (1880 - 1926)	1973 (1967 - 1978)	Eastern Asia		1711 (1675 - 1747)
Asian_SEA1b	Eastern Asia		1768 (1737 - 1797)	1830 (1789 - 1873)	Eastern Asia		1711 (1675 - 1747)
Asian_SEA2a	Eastern Asia	China	1826 (1799 - 1850)	1956 (1945 - 1967)	Eastern Asia		1740 (1701 - 1773)
Asian_SEA2b	Eastern Asia	China	1975 (1966 - 1982)	1951 (1937 - 1968)	Eastern Asia	China	1857 (1831 - 1881)

Andrew Holtz
PhD Evolutionary Biology / 2023
Frontiére de l'Innovation en Recherche et Education
Université de Paris Cité

Asian_SEA3	South-Eastern Asia	Thailand	1853 (1827 - 1876)	1898 (1876 - 1920)	Eastern Asia		1589 (1553 - 1623)
Asian_SEA4	South-Eastern Asia	Philippines	1898 (1885 - 1910)	1925 (1904 - 1946)	Eastern Asia		1671 (1635 - 1701)
Asian_SEA5	Eastern Asia	Taiwan	1791 (1762 - 1819)	1957 (1939 - 1973)	Eastern Asia		1651 (1616 - 1680)
Cosmopolitan	Northern America		1656 (1627 - 1683)	1730 (1687 - 1773)			1564 1526 - 1598)
Cosmo_AF1a	Northern Africa		1876 (1862 - 1889)	1872 (1851 - 1895)	Northern Africa		1854 (1842 - 1866)
Cosmo_AF1b	Eastern Africa	Zambia	1878 (1863 - 1891)	1907 (1890 - 1925)	Eastern Africa		1826 (1813 - 1838)
Cosmo_AF1c	Eastern Africa	Madagascar	1983 (1980 - 1985)	1983 (1980 - 1985)	Eastern Africa		1826 (1813 - 1838)
Cosmo_AF4	Northern Africa	Egypt	1910 (1894 - 1927)	1932 (1923 - 1940)	Northern America	USA	1679 (1656 - 1701)
Cosmo_AM1	Northern America	USA	1867 (1850 - 1883)	1908 (1886 - 1928)	Northern America	USA	1810 (1794 - 1825)
Cosmo_AM2a	Central America	Mexico	1851 (1836 - 1864)	1890 (1861 - 1920)	Northern America	USA	1772 (1750 - 1791)
Cosmo_AM2b	Northern America	USA	1939 (1928 - 1949)	1830 (1781 - 1852)	Northern America	USA	1791 (1768 - 1811)
Cosmo_AM3a	South America	Brazil	1897 (1880 - 1912)	1912 (1890 - 1936)	South America	Brazil	1808 (1785 - 1827)
Cosmo_AM3b	South America	Brazil	1861 (1847 - 1872)	1890 (1863 - 1916)	South America	Brazil	1808 (1785 - 1827)
Cosmo_AM4	Northern America	USA	1821 (1800 - 1840)	1846 (1811 - 1883)	Northern America	USA	1710 (1688 - 1732)
Cosmo_CA1	Eastern Europe	Russia	1924 (1913 - 1935)	1946 (1936 - 1957)	Eastern Europe		1898 (1889 - 1908)
Cosmo_CA2	Western Asia	Iraq	1981 (1977 - 1985)	1944 (1927 - 1960)	Western Asia		1817 (1798 - 1833)

Andrew Holtz

PhD Evolutionary Biology / 2023

Frontiére de l'Innovation en Recherche et Education

Université de Paris Cité

Cosmo_CA3	Eastern Europe	Romania|Russia	1936 (1925 - 1946)	1942 (1926 - 1956)	Eastern Europe		1880 (1870 - 1889)
Cosmo_CE	Eastern Europe	Germany|Poland	1968 (1962 - 1973)	1967 (1961 - 1974)	Eastern Europe	Germany|Poland	1931 (1921 - 1938)
Cosmo_EE	Southern Europe	Serbia	1944 (1936 - 1950)	1943 (1934 - 1954)	Eastern Europe	Poland|Serbia	1907 (1897 - 1916)
Cosmo_ME1a	Western Asia		1922 (1911 - 1933)	1938 (1927 - 1948)	Western Asia		1861 (1850 - 1872)
Cosmo_ME1b	Western Asia	Israel	1990 (1987 - 1991)	1987 (1984 - 1990)	Western Asia		1922 (1911 - 1933)
Cosmo_ME2	Western Asia	Turkey	1960 (1952 - 1968)	1986 (1984 - 1989)	Western Asia		1859 (1847 - 1870)
Cosmo_NEE	Eastern Europe		1887 (1873 - 1899)	1954 (1926 - 1964)	Eastern Europe	Poland|Serbia	1882 (1870 - 1892)
Cosmo_Vac	Western Europe	France	1959 (1950 - 1967)	not in study	Northern America	USA	1788 (1771 - 1805)
Cosmo_Vac2	Central America	Mexico	1942 (1926 - 1954)	not in study	Northern America	USA	1864 (1845 - 1880)
Cosmo_WE	Western Europe	Germany	1948 (1940 - 1954)	1949 (1942 - 1958)	Eastern Europe	Germany|Poland	1931 (1921 - 1938)
Cosmo_YUGCOW	Southern Europe	Montenegro	1977 (1977 - 1979)	not in study	Western Asia		1819 (1804 - 1833)
Cosmo_YUGFOX	Southern Europe	Serbia	1967 (1954 - 1977)	not in study	Southern Europe	Serbia	1909 (1894 - 1923)
Indian-Sub	Southern Asia		1760 (1716 - 1799)	1785 (1733 - 1840)			1397 (1341 - 1443)

Andrew Holtz
PhD Evolutionary Biology / 2023
Frontiére de l'Innovation en Recherche et Education
Université de Paris Cité

Supplementary Table 5. Inferred human-mediated introductions identified on the phylogenetic tree. *Using the full-canine tree consisting of 10,044 sequences, 14,640 transmissions were identified, 232 of which were to non-neighboring countries, and 43 of which fit the criteria as a human-mediated introduction. Human-mediated introductions are defined as transmissions that occurred within 16 years and between two countries that are more than 2000 km apart or separated by a body of water.*

Parent Country	Child Country	Parent Date	Child Date	Branch Length (years)	Subtree Size (num. of tips)	Distance (km)	(km/year)
Cameroon	Tanzania	1992.8	1993.0	0.2	73.0	3308.4	19477.4
China	Chile	1996.4	1998.0	1.6	1.0	19079.9	11754.8
Bangladesh	France	2014.6	2016.5	1.9	1.0	7916.8	4131.3
India	Oman	2008.5	2009.0	0.5	1.0	1936.2	3556.2
India	Oman	2007.9	2008.7	0.8	13.0	1936.2	2442.4
Philippines	Japan	2004.5	2006.0	1.5	1.0	2999.5	2003.3
Philippines	Japan	2004.5	2006.0	1.5	1.0	2999.5	2003.3
Philippines	Japan	2004.1	2006.0	1.9	1.0	2999.5	1605.9
India	Nigeria	2002.0	2007.7	5.7	3.0	8088.0	1430.3
France	India	1994.0	1999.0	5.0	5.0	6594.2	1318.8
India	France	2009.0	2014.0	5.0	1.0	6594.2	1316.7
France	Mexico	1994.0	2002.5	8.5	3.0	9206.8	1079.8
Nigeria	Iran	2007.7	2013.2	5.5	1.0	5865.2	1065.1
Nepal	Italy	1990.0	1997.0	7.0	1.0	6640.5	948.6
India	Italy	2004.7	2011.9	7.1	1.0	5922.2	831.8

Andrew Holtz

PhD Evolutionary Biology / 2023

Frontiére de l'Innovation en Recherche et Education

Université de Paris Cité

Mali	France	2009.0	2014.2	5.3	1.0	4139.8	785.9
Thailand	Philippines	2001.0	2004.0	3.0	1.0	2210.0	736.0
India	Ethiopia	2000.9	2008.0	7.1	1.0	4561.5	644.2
Thailand	Philippines	2001.0	2005.0	4.0	1.0	2210.0	552.5
China	Thailand	2005.0	2012.0	7.0	1.0	3303.9	471.8
Nepal	Italy	1981.3	1996.0	14.7	1.0	6640.5	452.2
Namibia	Tanzania	2004.0	2011.0	7.0	1.0	2953.5	421.9
China	Thailand	2005.0	2013.0	8.0	1.0	3303.9	412.8
China	Thailand	2005.0	2013.0	8.0	1.0	3303.9	412.8
China	Thailand	2005.0	2013.0	8.0	1.0	3303.9	412.8
China	Thailand	2005.0	2013.0	8.0	1.0	3303.9	412.8
China	Thailand	2005.0	2013.0	8.0	1.0	3303.9	412.8
China	Thailand	2005.0	2013.0	8.0	1.0	3303.9	412.8
China	Thailand	2005.0	2013.0	8.0	1.0	3303.9	412.8
Tanzania	Cameroon	1993.0	2001.5	8.5	24.0	3308.4	388.3
China	Thailand	2005.0	2014.0	9.0	1.0	3303.9	367.0
China	Thailand	2005.0	2014.0	9.0	1.0	3303.9	367.0
Tanzania	Ghana	1995.9	2008.7	12.9	1.0	4604.6	357.4
India	Thailand	2009.0	2017.9	8.9	1.0	2919.7	327.1
India	Iran	1999.0	2007.5	8.5	2.0	2544.6	297.6

China	Thailand	2002.5	2014.0	11.5	2.0	3303.9	286.3
Tanzania	Cameroon	1993.0	2006.5	13.5	3.0	3308.4	244.5
Mauritania	Nigeria	1981.5	1993.0	11.5	1.0	2478.3	216.1
Cuba	Colombia	1905.6	1917.5	12.0	3.0	2203.7	184.4
Mexico	Cuba	1894.2	1905.6	11.3	109.0	1784.2	157.7
Morocco	Spain	1958.9	1973.4	14.5	9.0	762.7	52.8
Spain	Morocco	1992.5	2008.0	15.5	1.0	762.7	49.3

Supplementary Methods 1. Purifying Selection Models used from HyPhy

FEL identifies instances of purifying selection in codons and identifies the internal branches where this has occurred. MEME identifies instances of episodic and pervasive positive selection on internal branches of the tree. aBSREL uses an adaptive model to re-estimate dN/dS ratios for each branch of the tree, yielding a more specific branch-length estimation.

Supplementary Methods 2. ACR Comparison Between Full-Canine Tree and Subsamples

A custom script (https://github.com/amholtz/GlobalRabies/blob/main/R/ACR_Sub_comparison.R) was used to compare the ACR node estimations for each subsample and full-canine tree (see Supp. Figure 6). We considered nodes with ACR marginal probability > 50% as "resolved" and the others as "unresolved". 98% of the tree nodes were resolved in the full-canine tree and in the subsampled trees (see Supp. Table 3 for more details). The state estimates were aggregated across each subsample and compared to the full-canine tree PastML node estimates. Only state (country) estimates consistent between the aggregated subsample and full-canine tree were retained. Of 4706 nodes that had resolved ancestral state (country) estimates from the full-tree, 1103 were found across the 5 subsamples (23%) and could hence be compared. Most nodes that were not comparable were peripheral on the tree. A total of 90% of the comparable nodes share the same ancestral state estimates as the full-canine tree estimate. This result strongly validates the original ancestral character reconstruction (ACR) for the full-canine tree, indicating how only about 10% of the nodes alter after subsampling.

Andrew Holtz
PhD Evolutionary Biology / 2023
Frontiére de l'Innovation en Recherche et Education
Université de Paris Cité

3 Geographic Origin of SARS-CoV-2 Variant, B.1.214.2

Since the emergence of SARS-CoV-2 several variants have been identified. A variant of concern by WHO was originally defined as a sublineage of the virus that spread in a community of individuals. These variants could contain mutations that provide some level of increased transmissibility, immune invasion, or increased pathogenicity. Many contain key mutations in the spike protein, which is a key protein involved in both humoral and cell-mediated immunity.

In early 2021 an increase in cases of a new variant, B.1.214.2, was identified in Belgium and set off international alarms. Quickly, the variant was identified in over 14 countries via genomic surveillance efforts, and efforts to understand its dynamics and phylogenetic history were initiated. Initially, it was found in metropolitan areas of Belgium, France, and Switzerland with international travel and foreign workers. Patient interviews also revealed that several cases were the result of recent travel to either other countries in Europe or to Central Africa. Specifically, the Republic of Congo was a common travel destination among returning travelers. In addition, the relative frequency of B.1.214.2 was as high as 30% in Central Africa in early 2021 suggesting a large number of cases among the population. This information led us to use statistical models to estimate the origin of the variant.

In our travel history-aware phylogeographic analysis, we incorporated data from patient interviews and air passenger records during the epidemic period. This methodology has been previously applied to multiple SARS-CoV-2 variant investigations and has proven valuable in recognizing countries with limited genomic sequencing capabilities. We estimate that the likely origin of B.1.214.2 is in the Republic of the Congo around June 2020. Between August and November 2020 it was introduced multiple times, most likely through air transit, to Belgium and France. From these two countries, we estimate the variant transmitted to the rest of Europe, causing a significant cluster in Basel, Switzerland. The presence of inbound flights from the Republic of the Congo or the Democratic Republic of the Congo in countries with B.1.214.2 sequences further supports the notion that the Republic of the Congo is the likely origin of this variant.

Andrew Holtz
PhD Evolutionary Biology / 2023
Frontiére de l'Innovation en Recherche et Education
Université de Paris Cité

This project was done in collaboration with an immunology lab at KU Leuven, where the variant was tested for immunological markers. These tests were particularly interesting due to a localized outbreak of the variant in a nursing home. This allowed for sample collection and molecular testing of the virus and how its immunological profile. Surprisingly, they found increased adaptive immunity among individuals infected with B.1.214.2 compared to other SARS-CoV-2 variants circulating around the same time.

This project is currently preprint ready and will be submitted to journals for review in the following month.

Andrew Holtz
PhD Evolutionary Biology / 2023
Frontiére de l'Innovation en Recherche et Education
Université de Paris Cité

Emergence of the B.1.214.2 SAR-CoV-2 lineage with Omicron-like spike insertion and a unique upper airway immune signature

Andrew Holtz*, Johan Van Weyenbergh*, Samuel L. Hong, Lize Cuypers, Áine O'Toole, Gytis Dudas, Barney I. Potter, Marco Gerdol, Francine Ntoumi, Claujens Chastel Mfoutou Mapanguy, Bert Vanmechelen, Tony Wawina-Bokalanga, Bram Van Holm, Soraya Maria Menezes, Katja Soubotko, Gijs Van Pottelbergh, Elke Wollants, Pieter Vermeersch, Ann-Sophie Jacob, Brigitte Maes, Dagmar Obbels, Veerle Matheeussen, Geert Martens, Jérémie Gras, Bruno Verhasselt, Wim Laffut, Carl Vael, Truus Goegebuer, COVID-19 Genomics Belgium Consortium, Rob van der Kant, Frederic Rousseau, Joost Schymkowitz, Luis Serrano, Javier Delgado, Tom Wenseleers, Vincent Bours, Emmanuel André, Marc A. Suchard, Andrew Rambaut, Simon Dellicour, Piet Maes*, Keith Durkin*, Guy Baele*

Final preparations for submission

Andrew Holtz
PhD Evolutionary Biology / 2023
Frontiére de l'Innovation en Recherche et Education
Université de Paris Cité

3.1 Abstract

Here we explore the emergence, mutational profile, and spread of the SARS-CoV-2 variant B.1.214.2. First identified in early January 2021 in Belgium, this unique variant exhibited a 3-amino acid insertion in the spike protein, similarly to what was later found in the Omicron variant. This insertion was speculated to confer an evolutionary advantage, possibly enhancing viral transmissibility or evading immune response. First detected in returning international travelers, this strain exhibited a significant, uneven presence across Central Africa, Belgium, Switzerland and France peaking during March-April 2021. Our travel history-aware phylogeographic analysis estimated its origin to be the Republic of the Congo, with a main introduction to Europe via France and Belgium and several smaller introductions across the epidemic period. We link the variant spread to human travel patterns between European countries and air passenger data. Our nation-wide study of SARS-CoV-2 nursing home outbreaks revealed a moderately severe clinical phenotype (8.7% case fatality ratio) but a distinct upper airway immune signature in B.1.214.2-infected high-risk elderly, with higher B- and T-cell activation, type I IFN signaling but lower NK, Th17, and complement system activation, which may help explain its peculiar epidemiological pattern.

3.2 Introduction

Since February 2020, SARS-CoV-2 has rapidly accumulated mutations that have accelerated viral spread. Although its evolutionary rate has been slower than its rate of transmission, SARS-CoV-2 has experienced numerous events of divergent evolution, leading to the emergence of numerous lineages. Some of these lineages and the mutations that define them have been identified as variants of interest (VOI) or variants of concern (VOC). Although the majority of the single nucleotide polymorphisms (SNPs) that characterize these lineages do not result in significant changes in infectivity and virulence, there are key mutations that are associated with an increase in transmissibility or increased host immune escape. These key mutations have led to a rapid viral population replacement in regions where these strains have emerged.

Andrew Holtz
PhD Evolutionary Biology / 2023
Frontiére de l'Innovation en Recherche et Education
Université de Paris Cité

Several mutations have appeared independently in separate VOCs, providing evidence of convergent evolution of mutations that increase fitness of infection [166]. The D614G mutation in the spike protein, which was absent from the ancestral strain but rapidly became dominant among circulating SARS-CoV-2 lineages [167] has been found to increase virion spike density and infectivity [168]. Similarly, mutation N501Y has been found to lead to an increased affinity to ACE2, the receptor to which the spike protein binds during host invasion [169]. E484K is a spike substitution that has been associated with circulating variants such as B.1.1.7, B.1.351, P.1 and Omicron lineages [170]. These along with K417N, Y505H and L452X have been associated with an increased affinity for the virus spike protein to the host ACE2 receptor, which increases the variants transmissibility and pathogenicity [113,171,172] . Moreover, mutations L452R, E484K, K417N, and N501Y [170,171,173] are associated with humoral and cell-mediated immunity escape. Interestingly all these mutations are predicted not to destabilize significantly (+-0.8 cal/mol) the Spike protein (K417N -0.8kcal/mol; L452R -0.8kcal/mol; E484R -0.8kcal/mol; N501Y 0.0 kcal/mol and Y505H 0.49 kcal/mol [174], while the majority of mutations to other aa are quite destabilizing. With respect to binding all these mutations are neutral except N501Y which is quite negative suggesting a conformational change [174]. This suggests that the virus in general selects neutral mutations in terms of stability and binding but that can affect the interaction with the antibodies generated by vaccination.

The majority of these mutations of concern have been documented primarily as either deletions or substitutions, but small spike deletions, such as Δ69/70 and Δ144 have also played a significant role in SARS-CoV-2 evolution. Comparatively, much less attention has been placed on small insertions, even though the spike protein of BA.1, the first omicron sublineage to spread globally, was characterized by the insertion of the tripeptide Glu-Pro-Glu in position 214. Building on this, a recent variant of concern, BA.2.86, which has raised an international alarm in August 2023 for its unnatural number of spike gene mutations and its unprecedented immune invasion contains a four amino acid insertion in a different location in the spike protein. Position 214, found within the N-terminal domain, has been described as an hotspot for recurrent insertions, which occurred both prior and after the advent of omicron, was thereby named RIR1 [175]. Despite the identification of several dozen independent insertions at RIR1, only two SARS-CoV-2 variants besides BA.1 reached a significant international spread, i.e. A.2.5 and B.1.214.2.

Andrew Holtz
PhD Evolutionary Biology / 2023
Frontiére de l'Innovation en Recherche et Education
Université de Paris Cité

The latter, first identified in Europe in late 2020, circulated for 7 months with a two-month peak in April and May of 2021 in Belgium, Switzerland, and France.

The emergence of the B.1.214.2 variant is a part of the larger unfolding of the B.1.214 lineage. The earliest B.1.214 sequences were identified in DR Congo in April 2020. Subsequent sub-lineages were detected: B.1.214.2 in Switzerland by late November 2020; B.1.214.1 in the Republic of the Congo in early December 2020; B.1.214.3 by mid-December 2020 in Luxembourg; and B.1.214.4 back in Switzerland by January 2021 [176–178]. While the pangolin classifications of these variants largely at the time relied on geographic clustering there are notable mutations that define these lineages [9] . All B.1.214 lineages, including its sub-lineages, carry the mutations I1398V, T1881I, and A4016V in ORF1a. The D614G mutation in the spike gene is not only common across the B.1.214 family but also found in B.1.17, B.1.351, and B.1.617.2. Uniquely, B.1.214.2, B.1.214.3, and B.1.214.4 share the T716I substitution, while B.1.214.3 further carries T95I and T478K mutations (neutral in terms of stability, -0.21 kcal/mol[174]). In contrast, B.1.214.2 is characterized by Q414K, N450K mutations (neutral in terms of stability, -0.23 kcal/mol and slightly destabilizing, 0.9kcal/mol, respectively [174]) and the inclusion of the RIR1 insertion sequence (Supp. Figure 2)[9,176,178].

Despite Delta replacing all other circulating strains of SARS-CoV-2 in early summer 2021, our group recently demonstrated that co-circulating non-dominant variants of concern (Gamma) and even variants of interest (Mu) were able to cause high-fatality outbreaks in high-risk elderly, even when fully vaccinated In addition, Delta, Gamma, and Mu nursing home outbreaks revealed a fatal immune signature, characterized by an increase in Th17 activation and high *IFNB1* transcript levels but no difference in type I IFN signalisation[179]. Likewise, the B.1.214.2 strain could have become a significant strain in Europe, as evidenced by its rapid expansion in Belgium. B.1.214.2 was first detected in Belgium, where an initial growth in cases was observed, but its origin remains unknown. Numerous news media outlets in Belgium, Germany, France, and the Democratic Republic of the Congo quickly began warning that, due to this variant's speed of replacement and its unique tripeptide insertion sequence, it could potentially have a high impact on convalescent and vaccine-induced antibodies [180,181]. Several news organizations falsely reported on the variant's origin due to confusion with parental lineage, B.1.214, which circulated

Andrew Holtz
PhD Evolutionary Biology / 2023
Frontiére de l'Innovation en Recherche et Education
Université de Paris Cité

in Democratic Republic of the Congo [182] . It was not until February 2021 that sequences of B.1.214.2 were found in either the Republic of the Congo or Democratic Republic of the Congo.

By late February 2021 as this variant grew in Europe, so did the number of B.1.214.2 sequences in Central Africa where it constituted a higher relative frequency among sequenced cases than in Europe. Between December 2020 and July 2021, B.1.214.2 was identified in 26% of all sequences in the Republic of the Congo, and in March 2021 it accounted for over 50% of the sequenced cases [177]. Therefore, a thorough investigation on the phylogeographic origin and immune characteristics of this variant was warranted. With limited access to sequencing technology and programs in Central Africa, circulating variants can evade detection. Often, these variants are not detected until they emerge in other countries with more financial resources for significant sequencing efforts, as exemplified by B.1.620 which was first identified in Lithuania but estimated to have originated in Central Africa [10].

Here, we investigate the emergence of variant B.1.214.2 using travel history-aware phylogeographic inference to investigate whether cryptic transmission of the variant in Central Africa occurred before its emergence in Europe, or whether B.1.214.2 instead emerged from a circulating sibling strain within Europe. In addition, we perform the first in-depth clinical and immunological characteristics of this variant by studying a large nursing home outbreak with moderately high (8.7%) case fatality ratio.

3.3 Results

Detection and Genomic Surveillance of SARS-CoV-2 lineage B.1.214.2

Between January 3rd and January 19th 2021, seven patients were identified in, Belgium who had recently returned from trips to Republic of the Congo (EPI_ISL_890291, EPI_ISL_890294, EPI_ISL_833185), Kinshasa, Democratic Republic of the Congo (EPI_ISL_912424, EPI_ISL_1123370), and Andalusia, Spain (EPI_ISL_894200, EPI_ISL_894201). These seven samples were not unique in their clinical presentation, but they all tested positive for a unique strain of SARS-CoV-2, which had a characteristic set of mutations—most notably a three amino

acid (AA) insertion sequence at position R214 in the spike (S) protein. Instances of this unique variant rose in the following weeks, ultimately leading to the Pangolin definition of the new lineage, B.1.214.2, on March 02, 2021 [183].

Upon definition, the first case of B.1.214.2 in GISAID was identified in Switzerland collected on November 11, 2020 (EPI_ISL_1296843) (Figure 1). The lineage quickly rose and declined between November 2020 and July 2021 with a peak in March and April 2021 (Figure 1). During this period, 1587 B.1.214.2 genomes were deposited to GISAID from a small set of countries- the top five most represented being Belgium (742, 46.8%), Switzerland (266, 16.8%), France (231, 14.6%), USA (126, 7.9%), Republic of the Congo (49, 3.1%), and Indonesia (29, 1.8%). Of the twenty countries presented, six of them are from Central Africa (RC, DRC, Angola, Rwanda, and Gabon) (Figure 1).

The variant's prevalence was far from negligible, signifying a substantial portion of newly sequenced cases. The variant reached its peak in Central Africa on the week of February 2, 2021, accounting for 28.85% of cases, in Belgium on the week of February 28, 2021 accounting for 5.54% of cases, in Switzerland on the week of March 07, 2021 constituted 2.73% of cases, and in France on the week of March 14, 2021 representing 1.11% of cases (Fig. 2). Although sequences were first collected in Switzerland, the duration of the isolation of B.1.214.2 across the three countries is comparable (Figure 1), with the last being collected in Belgium on June 15, 2021. In the three countries, the beginning of the transmission of B.1.214.2 is observed in the first two months of 2021, and quickly expands in March and April of 2021 (Fig. 2). By the end of May 2021, however, B.1.214.2 is almost no longer detected, and is almost completely replaced by other variants including the rapidly expanding B.1.617.2, commonly known as Delta (Fig. 2 and Supp. Fig 3).

Besides national similarities in B.1.214.2 expansion, we also observed regional patterns of prevalence for the variant. More specifically, when we standardized the number of cases based on the population, we observed a higher prevalence of the B.1.214.2 variant in cosmopolitan and densely populated areas (Figure 3). Notably, this variant was more prevalent in regions such as Brussels in Belgium, Île-de-France in France, and Basel in Switzerland, independent of the total number of cases sequenced. We found 33.6% of sequenced Belgian B.1.214.2 cases originated

Andrew Holtz
PhD Evolutionary Biology / 2023
Frontiére de l'Innovation en Recherche et Education
Université de Paris Cité

in Brussels, 48.1% of sequenced French B.1.214.2 cases originated in Île-de-France, and 69.2% of sequenced Swiss B.1.214.2 cases originated in Basel (Stadt/Landschaft). During this same period, Brussels accounted for 13.8% of total sequences in Belgium, Île-de-France accounted for 24.9% of total sequences in France, and Basel accounted for 17.0% of total sequences in Switzerland. None of these regions stand out as having particularly high sequencing output compared to other regions (Supp. Fig 4), indicating that spatial distribution of B.1.214.2 in the country was due to other factors and not based on more sampling by sequencing capacity alone. The regions with high prevalence of B.1.214.2 are known for attracting a significant number of visitors as well as European and non-European foreign workers and residents. Zürich, which is relatively far from bordering countries, only accounts for 1.5% of B.1.214.2 cases in Switzerland, supporting a cross-border effect in transmission. In all three countries and regions, the incidence of B.1.214.2 almost entirely disappears after May 2021.

B.1.214.2 Mutation Profile, spike protein carries a novel 3AA insertion sequence and numerous substitutions also found in variants of concern

The unique 3AA insertion sequence of lineage B.1.214.2 first caught our attention and led us to investigate the lineage and its introduction to Belgium further. At the time it was the first variant to have an insertion sequence mutation in the spike protein, and before official Pangolin classification, many sequencing labs did not use technology that was insertion aware and insertions were removed as possible error. Quite substantially, we found 174 B.1.214.2 sequences from Switzerland that lacked the 3AA insertion sequence. These samples originated from a lab at ETH Zurich which which used a V-pipe configuration which was not insertion-aware at the time (Marco- personal communication with Sarah Nadeau).

The pangolin definition of B.1.214.2 is characterised by the insertion of 9bp (ACAGATCGA) into the spike protein at position 22204 (adds the amino acids TDR downstream of R214), as well as the spike amino acid changes Q414K, N450K and T716I. The lineage also carries a 30bp deletion in ORF3a (25448-25478), with approximately half the genomes identified in Belgium (or 75% of all B.1.214.2) carry a 9bp deletion (11288-11297) in ORF1a (Figure 4a). The strongest indicators of a sequence clustering with B.1.214.2 sequences are the Q414K and N450K substitutions (Figure 4b, α). In fact, there are four B.1.214.1 sequences from the Republic of the Congo in February 2021, EPI_ISL_1654212, EPI_ISL_1654213, EPI_ISL_1654214, and

Andrew Holtz
PhD Evolutionary Biology / 2023
Frontiére de l'Innovation en Recherche et Education
Université de Paris Cité

EPI_ISL_1671927 (Figure 4b, β), which contain the TDR insertion and the T716I substitution, but lack the critical defining mutations Q414K and N450K. In contrast, there are a number of B.1.214.2 sequences that contain Q414K and N450K but lack the TDR insertion such as EPI_ISL_1096215 (Ireland), EPI_ISL_1915536 (Indonesia), EPI_ISL_1215914 (Germany) (Figure 4b, γ). Lastly, there are sequences characterized as B.1.214 which contain the Q414K, N450K, and the TDR Insertion, EPI_ISL_4096556 (Figure 4a, δ), EPI_ISL_1661248, EPI_ISL_1854777, which cluster with B.1.214.2 on the phylogenetic tree suggesting a discrepancy with Pangolin. The presence of T716I and absence of Q414K, N450K in an earlier B.1.214 sequence from the Republic of the Congo (EPI_ISL_1654214) suggests that the insertion sequence could have been found in ancestors of B.1.214.2, and brought stability to the spike protein to allow for more virulent mutations such as first T716I, and then Q414K and N450K, bringing rise to the B.1.214.2 clade. Interestingly, the ORF1a deletion has occurred independently in the VOC lineages B.1.1.7, B.1.351, and P.1 as well as lineages of note B.1.525 and B.1.526 suggesting a selective advantage for lineages carrying this deletion. The presence of the ORF3a deletion in almost all the B.1.214.2 sequences indicates a vital role in clade definition, in contrast to ORF1a, which is only present in less than half of the B.1.214.2 genomes. This suggests the deletion rose later on the epidemic as it's specific to European clades of the tree. In fact, several sequences from the Democratic Republic of the Congo (such as EPI_ISL_3133642) have the ORF1a deletion and are estimated in the phylogenetic tree (Fig. 6a) as descendants of the European clades indicating that transmissions from Europe to Central Africa also occur.

B.1.214.2 mutations likely to have substantial structural effects on viral function

The N450K mutation is located on the exterior of the Spike protein at a location where immune evading mutations are often found [113]. (Figure 5, A-C). This mutation has also been found to be a key antibody escape mutation [184,185]. This mutation is predicted to destabilize the spike protein slightly, which is compensated by the Q414K mutation, which is located on the interior of the Spike protein and stabilizes the structure thermodynamically (Figure 5, A). Studies have found that N450K confers a mild increase in ACE2 binding affinity [186].

The N450K is often found at the interface between neutralizing antibodies and the Spike protein (Figure 5, C). The mutation is predicted to decrease the affinity of neutralizing antibodies to Spike (Figure 5, D). Mutating N450K is predicted by FoldX to destabilize the spike protein by more

Andrew Holtz
PhD Evolutionary Biology / 2023
Frontiére de l'Innovation en Recherche et Education
Université de Paris Cité

than 0.5 kcal/mol, which is considered destabilizing, for both the closed and open/ACE2-bound conformation (FoldX ΔΔG), while not affecting the interaction energy with ACE2 (FoldX Interaction ΔΔG). Mutating N450K in an example structure of the Spike protein bound to a neutralizing antibody was predicted to severely destabilize the interaction energy, suggesting immune evasion for that particular neutralizing antibody. Several other neutralizing antibodies were predicted to lose their neutralizing effect upon mutating N450K (results not shown). Additionally, the Q414K mutation is exposed in the open conformation of the Spike protein, when it is bound to the ACE2 receptor (Figure 5, A). This could also have an effect on binding to neutralizing antibodies but is less often found at the neutralizing antibody binding sites.

The three amino acid insertion encompassed the sequence TDR between R214 and D215 (Figure 5, E&F). We modeled the conformation of the insertion to the exterior of the spike protein in both the open and closed conformation, which is not the interaction site with ACE2 (Figure 5, E&F). This amino acid location could be important for variant fitness, since the South African variant (501Y.V2, B.1.351, known as Beta) harbors the mutation D215G. Even though the connection between this loop and neutralizing antibodies remains unclear, there's a prevailing theory that insertions at R214 might counterbalance the effects of relatively harmful mutations in the RBD, such as Q414K and N450K. These mutations may reduce antibody affinity, but they could also lessen the efficiency of protein folding [10,20,21].

Phylogeographic analyses reveal the Republic of the Congo as the likely origin of B.1.214.2

We conducted maximum-likelihood (ML) phylogenetic and travel history-aware Bayesian phylogeographic analyses with 14 identified travel cases to elucidate the global diffusion of B.1.214.2 and estimate its ancestral origin (Supp. Table 1). The initial ML phylogenetic tree was reconstructed using IQTREE2 v2.2.2.2 [30,31] with 1660 B.1.214 and descendant sequences (B.1.214, B.1.214.1, B.1.214.2, B.1.214.3, B.1.214.4). A root-to-tip regression analysis showed sufficient temporal signal for a phylogeographic reconstruction. We generated a phylogenetic tree with a calibrated time scale using TreeTime v.0.8.6 [33], dating its most recent common ancestor (tMRCA) to February 2020, indicating the time of divergence of this clade and its derivatives. We were interested specifically in the origin of the B.1.214.2 clade and continued with its subtree for the remainder of the analysis.

Andrew Holtz
PhD Evolutionary Biology / 2023
Frontiére de l'Innovation en Recherche et Education
Université de Paris Cité

We conducted a travel history-aware Bayesian phylogeographic analysis of the B.1.214.2 subtree using travel data from 14 B.1.214.2 infected patients (Supp. Table 1) using BEAST v1.10.5 [1] (pre_thorney_0.1.2). Our analysis estimated the Republic of the Congo as the likely origin of the B.1.214.2 clade receiving posterior support of 57.5%, following France with 29.84% and Belgium with 7.22%. We predicted the tMRCA to mid June 2020 (2020.443) (95% HPD interval: early Mar 2020 (2020.164); early Sep 2020 (2020.681)). We observe two main branches of the B.1.214.2 clade, leading to several independent avenues into Europe (Fig. 6a) originating from the Republic of the Congo. The first likely introduction (88.39% posterior support for the Republic of the Congo) is estimated to have occurred in early August (2020.646, 95% HPD: 2020.49, 2020.777) to France and led to the widespread expansion in Europe (Fig. 6b & Supp. Table 2): Belgium in mid November 2020 (2020.868, 95% HPD: 2020.793,2020.933), Belgium at the end of November 2020 (2020.911, 95% HPD: 2020.853, 2020.976) and Switzerland in mid November 2020 (2020.8454, 95% HPD: 2020.791, 2020.888), and a reintroduction into the Republic of the Congo in mid Dec 2020 (2020.958, 95% HPD: 2020.880, 2021.024). The analysis demonstrates a second branch, estimated to originate from the Republic of the Congo origin (99.89% posterior support) seems to lead to four separate Belgian clusters led to expansion in mostly Belgium, with evidence of transmissions to other Central African and European countries (Figure 6 & Supp. Table 2). We observe multiple clusters in this second main branch of localized transmissions occurring in France, Germany, Belgium, and the United Kingdom from likely the Republic of the Congo (Figure 6). These findings strongly suggest the existence of cryptic transmission of the variant in these countries prior to its detection through genomic surveillance in Jan 2021 (Fig. 1).

The B.1.214.2 variant of the SARS-CoV-2 virus was detected in every region of Belgium and from as far north as Brittany in France to Occitanie in the south, suggesting widespread prevalence throughout both countries. However, the variant was predominantly found in areas with frequent travelers and foreign residents, such as Brussels (Belgium), Île de France (France), and Basel (Switzerland). While this trend could be attributed to increased sequencing in larger cities, it is interesting to note the low prevalence of B.1.214.2 in Zürich and Geneva, despite similar sequencing output. This trend is also observed in Belgium between Brussels and Antwerp (Supplementary Figure 3). Out of 230 sequences from Switzerland, only 5 originated from Zürich

Andrew Holtz
PhD Evolutionary Biology / 2023
Frontiére de l'Innovation en Recherche et Education
Université de Paris Cité

and 10 from Geneva. Phylogeographic analysis revealed that all but one B.1.214.2 case in Switzerland originated from France (Figure 6a) rather than Germany, likely due to the higher number of French border workers in Basel. German cases seem to originate from either Belgium (n = 10) or the Republic of the Congo (n = 8). Three out of the four B.1.214.2 cases in Grand Est, France (Figure 2) cluster within the Basel clade in the phylogenetic tree, showing transmission between the two areas. However, this does not inherently explain the initial introduction of the variant from France to Switzerland. The majority of cases in France are in Paris, and it is worth noting the well-traveled train routes between Paris and Geneva. This suggests that these train routes may have facilitated the spread of the virus from France to Basel.

We do not observe other Central African countries to have played a role in the European expansion, but this could be due to sampling bias, as we have more sequences from the Republic of the Congo than any other Central African country in our dataset (Figure 1). There are only two countries that experienced B.1.214.2 expansion outside of Europe or Central Africa. We estimate an outbreak in Indonesia (n=25) that likely originated from Belgium (and France before that) in early December 2020 (2020.929, 95% HPD: 2020.838, 2021.005), and an outbreak in the United States that likely originated from the United Kingdom (and France before that) in early January 2021 (2021.014, 95% HPD: 2020.961, 2021.058).

To understand the influence of long-distance transmissions, we analyzed air passenger data from December 2019 to July 2021, which revealed 27 countries with inbound flights from either the Republic of the Congo (RC) or the Democratic Republic of the Congo (DRC). Due to the proximity of the two cities and their airports (4km), we included both in the review. Out of these 27 countries, 17 were found to have B.1.214.2 sequences, and all countries with submitted B.1.214.2 sequences had air connections to either RC or DRC (Fig. 6b and Fig. 7). Only 9 out of the 27 countries with inbound flights from RC or DRC did not report B.1.214.2 cases: Uganda, Tanzania, South Korea, New Zealand, Nigeria, Denmark, Chile, and Guinea-Bissau. We found that four of these countries, namely Nigeria, Chile, and Guinea-Bissau, appear in our travel history data, suggesting potential undetected transmission of B.1.214.2 in these countries. These data in conjunction with the phylogeographic analysis which reflect the influence of the Republic of the Congo on the spread of B.1.214.2, seemingly suggests the Republic of the Congo as the origin of the variant.

Andrew Holtz
PhD Evolutionary Biology / 2023
Frontiére de l'Innovation en Recherche et Education
Université de Paris Cité

A large Belgian nursing home outbreak reveals a moderately severe clinical phenotype and a unique B.1.214.2 upper airway immune signature

A B.1.214.2 outbreak was identified through our nation-wide surveillance of SARS-CoV-2 in nursing homes, representing the first large post-vaccination outbreak in nursing homes of the country. Between January 24th and March 3rd, a total of 952 PCR tests were performed on nasal swabs obtained from a total of 86 residents and 114 staff members, interns and volunteers present during the outbreak. Of those, 54 were positive (46 residents and 8 staff members), resulting in a high vs. low attack rate in residents (46/86, 53.5%) and staff members (8/114, 7%), respectively. The outbreak lasted for more than 4 weeks, with the first PCR-positive cases detected on January 24th, and the last ones on February 26th. Although no residents or staff members were hospitalized, four fatal cases were observed, 2 male (aged 85 and 88 years) and 2 female residents (aged 89 and 97 years), who died within 3 to 21 days after their PCR-positive diagnosis, resulting in a moderately high case fatality ratio (4/46, 8.7%). For 16 PCR-positive individuals for which sufficient leftover diagnostic sample was available and viral load was sufficiently high, whole genome sequencing was successful for 14 samples, while two samples were not typable. All 14 successful samples were identified as B.1.214.2, confirming the outbreak was due to a single variant. Of interest, all resident samples are evolutionary related and clustered together on the phylogenetic tree , whereas a single staff member sample was present in a separate cluster (Figure 6a).

The timing of the outbreak occurred in the middle of the national vaccination campaign, with more than 95% of residents receiving their first dose on January 15th and second dose on February 5th. Thus, 28 positive cases were observed after a median of 13 days following the first dose, and 16 positive cases were observed after a median of 10 days following the second dose. Of interest, no differences were observed in anti-S antibodies (not shown) or neutralizing antibodies (Fig. 8A) between residents infected after the first or second vaccination dose. Neither anti-S nor neutralizing antibodies against wild type (WT) or any VOC circulating worldwide in the same year (Alpha, Beta, Gamma, Delta) were correlated to protection against infection ($p>0.1$, data not shown), although the number of serum samples ($n=16$) was too small to draw firm conclusions. Unfortunately, the four fatal cases passed before sample collection was possible. However, all serum samples were obtained at >60 and >15 days after the first and second

Andrew Holtz
PhD Evolutionary Biology / 2023
Frontiére de l'Innovation en Recherche et Education
Université de Paris Cité

vaccination dose, respectively, allowing unbiased comparison across vaccine-derived (PCR-negative) and hybrid immunity (PCR-positive).

Leveraging our nation-wide surveillance effort, we were able to compare neutralizing antibody levels between residents of this outbreak with a similar-sized post-vaccination outbreak, which was also caused by a non-VOC, namely the Mu variant [179]. Of note, first and second dose vaccination occurred simultaneously in both nursing homes, with the same vaccine (Comirnaty). As shown in Fig. 8A, both vaccine-induced (PCR-negative cases) and hybrid (PCR-positive cases) humoral immunity was highly similar in age- and sex-matched residents from both outbreaks, arguing against a major effect of antibody-mediated immune escape of B.1.214.2. In addition, moderate (50% inhibition of ACE2 binding) to high (90% of inhibition) levels of cross-neutralizing antibodies against the major VOC circulating in the same time period were observed in both B.1.214.2-infected and Mu-infected high-risk elderly, with significant decreases respecting the same order of antigenic distance (WT>Alpha>Delta>Gamma>Beta), corroborating previous research [187–190]. A strong individual variation was observed in vaccine-derived as well as hybrid immunity between residents in both outbreaks, without significant differences between the two (Fig. 8A). Therefore, we proceeded to an in-depth immune profiling of the upper respiratory tract in matched PCR-positive vs. PCR-negative (but highly exposed) residents, using digital transcriptomic analysis of nasopharyngeal swabs by nCounter technology (Nanostring). This also allowed us to compare the B.1.214.2 upper airway immune response in matched PCR-positive and PCR-negative exposed individuals from other SARS-CoV-2 nursing home outbreaks, each with a single variant (Gamma, Delta, Mu). Since vaccination status and SARS-CoV-2 variants differed for each outbreak, our strategy was to compare matched age-, sex- and vaccination status-matched PCR-positive vs. PCR-negative exposed individuals for each separate data set, in order to obtain immune signatures that represent specific correlates of protection for B.1.214.2 as compared to other variants. Figure 8B shows a Volcano plot of differentially expressed genes in the upper airway of age- and sex-matched B.1.214.2-infected (PCR-positive) and uninfected (PCR-negative, highly exposed) nursing home residents. We then investigated a possible overlap between this novel B.1.214.2 immune signature and our recently published upper airway signatures in fatal vs. non-fatal COVID-19 cases in other post-vaccine outbreaks [179]. Fig. 8C shows a minimal overlap (11/78 genes) in upregulated immune genes shared between B.1.214.2-infected (PCR+) upper airway samples and matched samples of fatal cases from

Andrew Holtz
PhD Evolutionary Biology / 2023
Frontiére de l'Innovation en Recherche et Education
Université de Paris Cité

Gamma/Delta/Mu nursing home outbreaks, and a slightly higher overlap with immune genes upregulated in mild/moderate (PCR+) matched residents (12/25 genes). Of note, 91 out of 124 upregulated genes were unique to B.1.214.2 upper airway infection (Supp. Table 3), including *IRF3*, which we identified as a "protective" transcriptomic biomarker in previous outbreaks, as well as both type I IFN receptors, *IFNAR1* and *IFNAR2,* the latter was also associated with critical COVID-19 by GWAS [191].

Translating the biological significance of this novel immune signature into signaling pathways, we found significantly increased adaptive immunity (Fig. 8C-D, $p<0.001$), which included both B-cell and T-cell signaling and effector genes such as *LAG3, TAP1, LEF1*, *TNFSF13B* (also known as BAFF, B-cell Activating factor), several proteasome genes (*PSMB7/PSMC2/PSMB9/PSMD7*), and the prototype memory T-cell marker *CD45RO*. Innate and antiviral type I IFN signaling was also significantly increased (Fig. 8C-D, $p<0.05$), represented by *IFNAR2/STAT1/STAT2* signaling and downstream antiviral genes such as *MX1, BST2*, and *IFIT2*. On the other hand, we observed decreased innate "Natural Killer" (NK) CD56"dim" cells (Fig. 8C-D, $p<0.01$) in B.1.214.2 infected residents, as characterized by higher expression of several KIR genes (*KIR2DS1/KIR3DL2/KIR3DL3*, together with genes encoding cell surface markers *NCAM1, KLRG2*, *CD247* and chemokine *XCL1.* Th17 differentiation (characterized by *IL17F/IL22/IL23A/IL23R*) and genes of the complement system (*C5/C6/C7/C8G/C9*) were also decreased in B.1.214.2 PCR+ vs. PCR-negative highly exposed residents. Of note, there was no downregulation of MHC class I antigen presentation (Fig. 8D, $p=0.41$) in B.1.214.2 infected residents, which is strikingly different from our previous findings in other nursing home outbreaks [179], as well as several COVID-19 cohorts worldwide [192–197].

Taken together, cross-comparison of large nursing home outbreaks reveals a moderately severe clinical phenotype of B.1.214.2, with no major difference in cross-neutralizing antibodies but linked to a unique upper airway signature, characterized by heightened B- and T-cell activation, type I IFN signaling but lower NK, Th17, and complement system activation.

Andrew Holtz
PhD Evolutionary Biology / 2023
Frontiére de l'Innovation en Recherche et Education
Université de Paris Cité

3.4 Discussion

In this comprehensive study, we present an analysis of the B.1.214.2 variant, encompassing its epidemiological prevalence, mutational profile, immune signature, and origin. Through a travel history-aware phylogeographic analysis, we estimate that the European expansion of B.1.214.2 may have originated from the Republic of the Congo, leading to localized country-clusters. We observe several episodes of introduction and reintroduction between European countries and Central Africa, revealing a bi-directional transmission route between the two. The high prevalence of B.1.214.2 in the Republic of the Congo during the same period supports early cryptic transmission previous to European expansion. The strong correlation between B.1.214.2 prevalence and air travel to the Republic of the Congo suggests that air passengers played a major role in the spread of this variant to Europe.

We were originally drawn to the unique nine nucleotide insertion sequence at recurrent insertion region (RIR1). This pattern, although novel at the time, has been identified in several other SARS-CoV-2 strains, the most prominent being the Omicron BA.1 sublineage, which has achieved global prevalence. Before Omicron's rise, only 0.3% of all SAR-CoV-2 genomes contained S gene insertions [175]. Previous phylogenetic work revealed independent tripeptide RIR1 insertions, which suggest convergent evolution at this loci. Many of the variants with this mutation also contain non-synonymous spike substitutions and deletions, which hint at the potential of RIR1 insertion in stabilizing deleterious mutations. In Gerdol et. al, which coined and investigated the RIR1 insertion, the authors suggest that the consistent independent emergence of RIR1 insertions in various viral strains, seen in conjunction with Omicron's emergence, points to the likelihood that RIR1 insertions may offer an evolutionary advantage, although the extent of this advantage remains uncertain. Notably, there is a connection between the occurrence of RIR1 insertions, RNA-Dependent RNA polymerase (RDR) deletions, and various non-synonymous mutations in the Receptor Binding Domain (RBD) that are known to affect immune evasion, enhance ACE2 binding, and increase transmissibility. Given these correlations and the predicted impact of RIR1 on the spike protein's structure, it's plausible that RIR1 could serve a permissive role, potentially offsetting the minor disadvantages of certain non-synonymous spike RBD mutations. This clearly convergent insertion mutation deserves further investigation and shows that surveillance and characterization of non-VOC lineages may help us understand the emergence and advantages of novel pandemic lineages.

Andrew Holtz
PhD Evolutionary Biology / 2023
Frontiére de l'Innovation en Recherche et Education
Université de Paris Cité

Our analysis on air passenger data revealed the influence of human migration between continental Europe and Central Africa in the spread of B.1.214.2. The presence of B.1.214.2 cases only in countries with connections to the two Congolese airports suggests a possible pathway for transmission, with three additional potential routes identified through travel history data.The only five countries with air connections to the Republic of the Congo or the Democratic Republic of the Congo without any evidence of B.1.214.2 in circulation are South Korea, New Zealand, Uganda, Tanzania, and Denmark. Tanzania has reported zero sequences of SARS-CoV-2 due to deprioritization of the pandemic. Uganda, in contrast, has reported 2,031 sequences in total (937 over the B.1.214.2 epidemic period), ten of which are B.1.214 [198–200]. South Korea and New Zealand had very strong travel restrictions and arrival testing protocols during the Pandemic, and most likely were able to curb the introduction of B.1.214.2 into their countries. Denmark during this period also had low passenger volume. This could provide some evidence for the success of South Korea and New Zealand in their COVID-19 prevention methods in air travel.

Multiple studies have estimated Central or Western Africa origin of SARS-CoV-2 variants, likely due to a combination of factors including importations from Europe, limited early control measures, and ongoing transmission- mostly notably, B.1.620 and A.27 [120,121,177] . A recent study explains this trend by arguing that African epidemics are the results of importations from Europe, where early control measures were quickly put in place. In Africa, however, transmission mostly progressed throughout the pandemic with the opportunity for new variants of concern to develop. Although trailing earlier on in the pandemic, several African countries have been able to increase their sequencing capacity, which has greatly allowed for phylogeographic studies. This has been greatly in part due to international investment in genomic surveillance, within Africa collaborations, and progress in reagent an equipment allocation [198]. It is important to mention that, especially due to non-uniform sequencing, case identified in a country do not limit the existence of the variant to that country, but rather countries that sequence function as a window into the region. The Republic of the Congo, with a stronger sequencing capacity, is perhaps a representation of the greater Central Africa region [199,201]. We suggest further studies from within Central Africa to investigate the transmission of these variants in the region.

Andrew Holtz
PhD Evolutionary Biology / 2023
Frontiére de l'Innovation en Recherche et Education
Université de Paris Cité

Finally, taking advantage of our nation-wide surveillance of SARS-CoV-2 nursing home outbreaks, we unveil a moderately severe clinical phenotype (in high-risk elderly with an 8.7% case fatality ratio. This clinical phenotype might be explained by systemic broad cross-neutralizing antibodies combined with a novel and distinct upper respiratory tract immune signature in B.1.214.2-infected nursing home residents, as compared to other high-fatality (>10%) post-vaccine outbreaks with Delta, Gamma and Mu variants. Noteworthy limitations of this study include the absence of SARS-CoV-2 aerosol detection during the outbreak, which we recently demonstrated as a useful marker of long-during exposure in other nursing home outbreaks [179]. However, the high attack rate (53.5%), long duration of the outbreak (>4 weeks), and the detection of PCR-positive residents on all three floors of the nursing home strongly suggest high exposure in all residents, including the PCR-negatives. Second, no baseline serum samples were available before the outbreak, nor from fatal cases, to compare the levels of pre-existing and/or vaccine-elicited SARS-CoV-2 neutralizing antibodies. A third limitation is the absence of data on staff pandemic preparedness and population incidence of COVID-19 in the surrounding population, which Suñer et al.[202] identified as major predictors of (pre-vaccine) COVID-19 mortality in a large retrospective study of Spanish nursing homes. Major strengths of this study include the comprehensive testing of the complete nursing home, both staff and residents (>950 PCR tests) for the entire duration of the outbreak, complete metadata, detailed clinical follow-up, and in-depth immunological profiling (systemic anti-S and neutralizing antibodies, upper airway digital transcriptomics). Thus, we found a moderately severe (8.7% case fatality ratio) clinical phenotype of B.1.214.2, with no major difference in cross-neutralizing antibodies across all major VOC (Alpha, Beta, Gamma, Delta) circulating in the same time period. This clinical phenotype of B.1.214.2 was linked to a unique upper airway immune signature. First, it was characterized by higher adaptive (B- and T-cell mediated) immunity, arguing against immunodepression or vaccine failure in these infected high-risk elderly, as well as increased and type I IFN signaling, also associated with protection from critical COVID-19 [203–208]. Moreover, we also observed lower KIR expression and NK CD56 "dim" cells, as well as lower Th17 and complement activation, which are opposed to our previous findings in similar nursing home outbreaks [179]. Taken together, we propose that this unique upper airway immune signature might explain, at least in part, the peculiar epidemiological history of B.1.214.2.

3.5 Methods

123

Andrew Holtz
PhD Evolutionary Biology / 2023
Frontiére de l'Innovation en Recherche et Education
Université de Paris Cité

Study design

This investigation was prompted by the rapid increase in SARS-CoV-2 detections in Belgium, which contained a novel nine nucleotide insertion sequence in the spike protein. By March 2021, already 211 similar sequences were documented in Belgium, and the Pangolin COVID-19 Lineage Assigner [9] assigned the majority of the sequences to lineage B.1.214 which was most commonly found in the Democratic Republic of the Congo. This lineage was then split into B.1.214.1 and B.1.214.2, the latter being the presently described variant of interest. B.1.214.2 is characterized by the insertion of 9bp (ACAGATCGA) into the spike protein at position 22204 (adds the amino acids TDR downstream of R214), as well as the spike amino acid changes Q414K, N450K and T716I. The lineage also carries a 30bp deletion in ORF3a (25448-25478), with approximately half the genomes identified in Belgium also carrying a 9bp deletion (11288-11297) in nsp6, a deletion also observed in the lineages B.1.1.7, B.1.351, B.1.525 and P.1. B.1.214.2 genomes have been deposited in the GISAID [209] database with origins in five additional European countries, North America and Africa.

Data collection

All SARS-CoV-2 genomes used in this study and corresponding metadata were obtained from GISAID in November 2021 filtering for B.1.214 and derivative sequences from November 2020 to August 2021. Travel history was retrieved when available through the GISAID metadata. Additional metadata was collected by contacting sequencing and test laboratories that documented travel data through by contacting the GPs who managed patient care. . We collected 14 travel itineraries for their associated sequences. This information can be found in Supp. Table 1. We have a higher number of documented travel history cases for Belgium, which can be attributed to its robust documentation system and our close collaboration with the hospital systems. Nevertheless, it is important to note that the absence of reported travel history for France and Switzerland does not necessarily mean that such travel cases do not exist. We contacted laboratories in France and Switzerland, aiming to obtain supplemental travel history data that might not be readily available in the GISAID metadata but were unsuccessful due to the strict patient data protections surrounding this information. This work was framed within the role of the NRC respiratory pathogens UZ/KU Leuven (as defined by the Royal Decree of

Andrew Holtz
PhD Evolutionary Biology / 2023
Frontiére de l'Innovation en Recherche et Education
Université de Paris Cité

09/02/2011), as approved by the UZ/KU Leuven Ethical committee for research (S66037 and S65371).

Mutation Analysis and Protein Modelling

Modeling of the insertion variants was done over the 6VXX pdb structure by fragment grafting using the *Bridging* command of the ModelX tool suite [210]. Residues V213 and L219 were the input arguments for the search of length seven \ fragments that resulted in an insertion of 3 residues. The command grafted all geometrically compatible fragments in the ModelX database and obtained models were renumbered to accommodate the insertions corresponding both for the Belgium and for the Costa Rica variants. Fragment search was blind to the sequence so the side chains of the variants were modeled on all the results obtained using the *BuildModel* command of the FoldX package [211], post selection of the energetically more favorable states was done according to total stability energy calculated with FoldX *Stability* command.

Proteins lacking a crystal structure were modeled using I-TASSER [212]. Models were visualized using YASARA [213], schematic representations on proteins were generated using Protter [214]. FoldX version 3.0 beta 6 [215] was used to predict the effect of the mutations on the thermodynamic stability and the interaction energy. First the crystal structures were repaired using the *RepairPDB* command, mutations were modeled using the *BuildModel* command, interaction energy was analyzed using the *AnalyseComplex* command.

Phylogenetic analysis

We downloaded all B.1.214.2 (n = 1587) and B.1.214-derived sequences (B.1.214.*) available on GISAID, resulting in a total of 1986 sequences. From this preliminary dataset, we used NextClade v1.7.1 [8] to identify sequences of poor quality and Pangolin v1.2.81 [9] to characterize the sequence lineage, which filtered 324 sequences resulting in a 1727 high quality sequences. The sequences removed were well spread over the range of countries represented, and no countries were eliminated due to sequence quality. We then added these sequences to a NextStrain[216] build including 1,000 Central African Sequences and 500 global sequences between 28 December 2019 and 08 July 2021, which is the last date of sampling for a B.1.214. divergent sequences. The final dataset contained 3507 sequences and the phylogenetic tree was
125

Andrew Holtz
PhD Evolutionary Biology / 2023
Frontiére de l'Innovation en Recherche et Education
Université de Paris Cité

analyzed using NextStrain and visualized by Auspice [216]. We time-calibrated and rooted the tree using TreeTime v.0.8.6 [13] and "hCoV-10/Wuhan/Hu-1/2019" (GenBank accession MN908947) as the outgroup. We detected and removed 63 outliers. We then extracted the subtree containing all 1662 sequences clustering under the B.1.214* node for further analysis.

We performed maximum likelihood phylogenetic tree reconstruction using IQTREE2 v2.2.2.2 [4,159] on the sequences found within this B.1.214* cluster. The reconstruction was based on a GTR model with empirical frequencies and a three-category FreeRate model of site heterogeneity. This model was selected as the best fitting model using IQTREE's ModelTest functionality. To ensure thorough exploration, we conducted a robust search by considering all possible NNIs (Nearest Neighbor Interchange). Additionally, we optimized the number of initial parsimony trees with an ML (Maximum Likelihood) nearest neighbor interchange (NNI) search, setting the value to 100. We evaluated the temporal signal of our dataset using TempEst [12], and used TreeTime v0.8.6 [13] to root the tree on divergent B.1.214 sequences and use it as a starting tree for our Bayesian analyses, setting the clock rate to 0.0008. No outliers were removed since we had already performed outlier removal in the previous dating step.

To determine the molecular clock that best fits the data, we employed the Bayesian Evaluation of Temporal Signal (BETS) [217] analysis in BEASTv1.10.5 [1]. This utilizes Bayes factors to objectively assess the presence of temporal signal in a dataset and determine the feasibility of calibrating a molecular clock (strict vs relaxed) using the associated sampling dates of genetic sequences. The evaluation is done by comparing the ratio of marginal likelihoods between competing models: a heterochronous model that uses the sequences' sampling dates, and an isochronous model, where all the sampling dates have been removed. We performed four parallel analyses using a skygrid coalescent prior and a HKY+Γ nucleotide substitution model, 1) strict-clock without dates, 2) strict-clock with dates, 3) relaxed clock without dates, 4) relaxed click with dates. We ran the MCMC chains for 10^8 states and used generalized stepping-stone sampling [218] with 100 path steps of 10^6 iterations to compute the marginal likelihoods. The relaxed clock with dates yielded the highest log marginal likelihood and was therefore selected for subsequent analyses.

Andrew Holtz
PhD Evolutionary Biology / 2023
Frontiére de l'Innovation en Recherche et Education
Université de Paris Cité

Travel history-aware phylogeographic reconstruction

As our main goal was to estimate the origin and understand the historical transmission of B.1.214.2, we selected a subtree containing sequences that clustered with B.1.214.2, yielding 1346 sequences including five pangolin defined B.1.214 sequences that cluster with B.1.214.2 (bold are shown in figure 4a: EPI_ISL_1524721, **EPI_ISL_1661248**, **EPI_ISL_1854777**, EPI_ISL_2788222, **EPI_ISL_4096556**). Although 48% of the sequences were collected in Belgium, we decided not to subsample the dataset in order to not add an additional intervention that could be biased. In the event that Belgium seems to over represent the phylogeographic result, we could subsample. This was, however, not the case.

To perform the travel history-aware discrete phylogeographic analysis, we needed to prepare the travel history data. Just three out of 14 sequences with acquired travel history were not collected in Belgium. Each country of origin and country of travel was noted, as well as the sampling date. The days between trip departure and sample collection was fixed if known as to estimate the potential infection time. For sequences without date of travel information, we used a random time calculated from a normal prior distribution with a mean of 10 days before sampling date and a standard deviation of 3 days. This is dictated in the travel history-aware phylogeographic analysis protocol [219,220].

We performed the travel history-aware Bayesian analysis using BEAST v1.10.5 [1] (pre_thorney_0.1.2). We used the general time-reversible substitution model with estimated base frequencies, gamma site heterogeneity model and 4 gamma categories. To infer ancestral locations, we used the Generalized Linear Model [221], using three different covariates- binary neighbour sharing (1 - two countries share a border, 0- two countries do not share a border), geographic distance (km) between capital cities, and finally the number of flights between two countries between 12-2019 to 07-2021 provided by Bluedot [222]. Since Liechtenstein does not have their own airport, these sequences were relabeled as Switzerland. We also estimate state change counts by calculating the number of transitions between two states (countries) called Markov jumps [74] in the dataset. We used an uncorrelated relaxed clock log normal relaxed distribution and a non-parametric skygrid coalescent model with the population size as 50 and cut off of 1.7. Skygrid's parameters were inferred by Hamiltonian Monte Carlo sampling [223] .

Andrew Holtz
PhD Evolutionary Biology / 2023
Frontiére de l'Innovation en Recherche et Education
Université de Paris Cité

Each Markov chain ran for 2 x 10^8 states and sampled every 100,000th state. Convergence was confirmed in Tracer v1.7.1[224] by reviewing effective sample sizes (ESSs). All chains were combined using Logcombiner giving a final MCMC length of 1.5x10^12 states.

To reach convergence, the MCMC was run for 1.5x10^12 states, sampling every one million states. Convergence was confirmed in Tracer v1.7.1[224] once effective sample sizes (ESSs) reached 200 for every parameter. We summarized all trees by constructing a maximum clade credibility (MCC) tree using TeeAnnotator. Branch heights were summarized by the common-ancestor model, which summarizes branch heights of clades across all posterior trees and not only the values for subset of trees that have that clade [225].

ACE2 binding pseudo-neutralization assays

Pseudoneutralization assays, consisting of competitive binding between antibodies and soluble ACE2 receptor to plate-coated Spike proteins of 10 different VOC (MSD), was performed and cross-VOC neutralization calculated as previously described [226].

Digital transcriptomics analysis of upper airway samples

Comprehensive immune profiling of upper airway samples was performed by 600-plex targeted analysis by digital nCounter transcriptomics (NanoString) in a subset of residents with sufficient leftover diagnostic sample ($n = 13$). RNA was extracted from nasopharyngeal swabs as described above and used for hybridization to pre-specified Human Immunology V2 and customized SARS-CoV-2 panels, as described previously [227–229]. Pathway score analyses and cell type deconvolution were performed using nSolver software (NanoString Technologies Ltd.).

Whole genome sequencing

Samples with a sufficiently high viral load (>1000 RNA copies/ml) were analyzed using whole-genome sequencing. An automatic RNA extraction was performed using the MagMAx Viral / Pathogen kit II (MVPII) (Thermo Fisher Scientific, A48383) with 200 µl sample input. The genomes were amplified following the ARTIC network protocol V3 or V4 [230] or as described by Freed *et al.* [231]. After clean-up of the amplicons, libraries were prepared using the SQK-LSK109+EXP-NBD196 ligation sequencing kit from Oxford Nanopore Technologies. Subsequently, the libraries were quantified, and sequencing was performed on a GridION platform using MinKNOW's built-
128

Andrew Holtz
PhD Evolutionary Biology / 2023
Frontiére de l'Innovation en Recherche et Education
Université de Paris Cité

in basecalling, demultiplexing and adapter trimming. Sequencing runs were processed using the ARTIC analysis pipeline and custom scripts.

3.6 Acknowledgements

We are grateful to the diagnostics and sequencing laboratories in all countries includes in this study, but particularly the Republic of the Congo, France, Switzerland, and Belgium for their efforts in screening, sequencing, and publicizing their genomic data via GISAID. We would also like thank GISAID for their efforts in curating the global database of SARS-CoV-2 sequences. GD is supported by the European Molecular Biology Organization (EMBO) Installation Grant (IG) EMBO-IG-5305-2023. A.H. acknowledges École doctorale Frontières de l'Innovation en Recherche et Education-Programme Bettencourt. A.H is funded by the INCEPTION programme (Investissements d'Avenir grant ANR-16-CONV-0005).

3.7 Figures

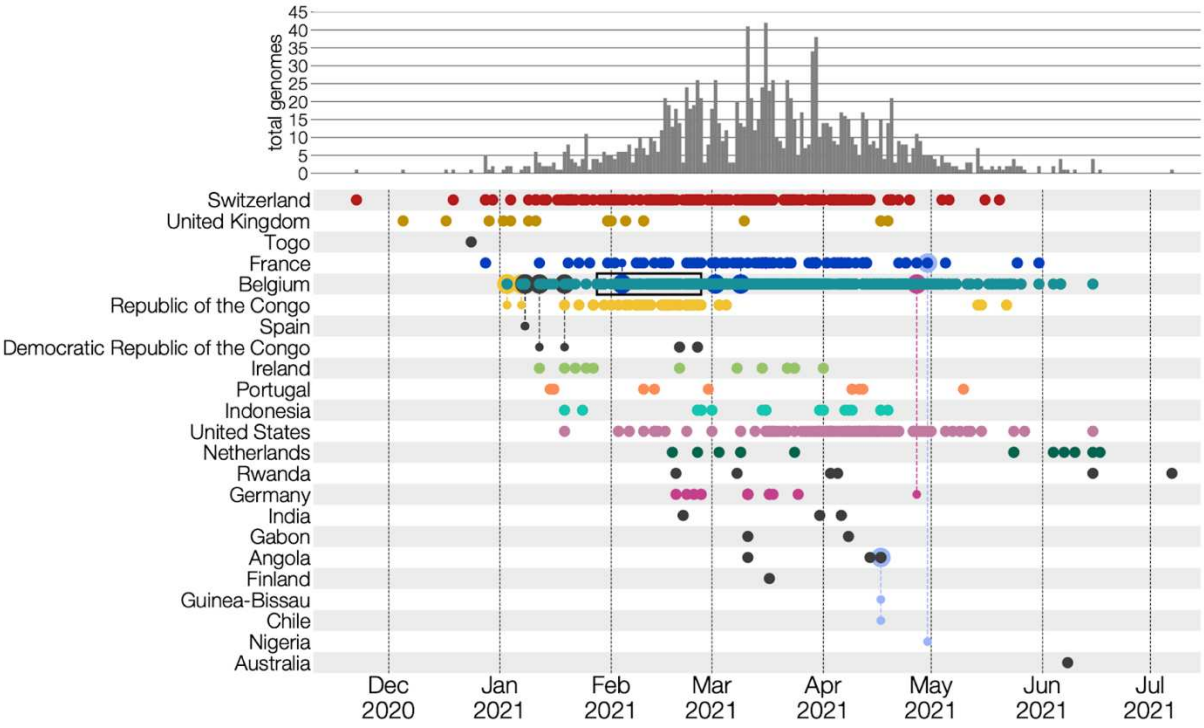


Figure 1. International Dispersal of SARS-CoV-2 Lineage, B.1.214.2. *Dots represent sequenced cases deposited to GISAID from November 2020 to July 2021. The majority of sequenced cases were collected between February and June 2021. Colored countries represent countries with 10 or more*

Andrew Holtz

PhD Evolutionary Biology / 2023

Frontiére de l'Innovation en Recherche et Education

Université de Paris Cité

sequences. Dotted lines between dots represent a travel-history included in the analysis. Countries that are included only by result of travel-history interviews are colored light blue. The overall accumulation of genomes from the time period is represented above by a bar chart showing total B.1.2142 genomes by week. The time range of the nursing home outbreak in Belgium can be identified by a grey box.

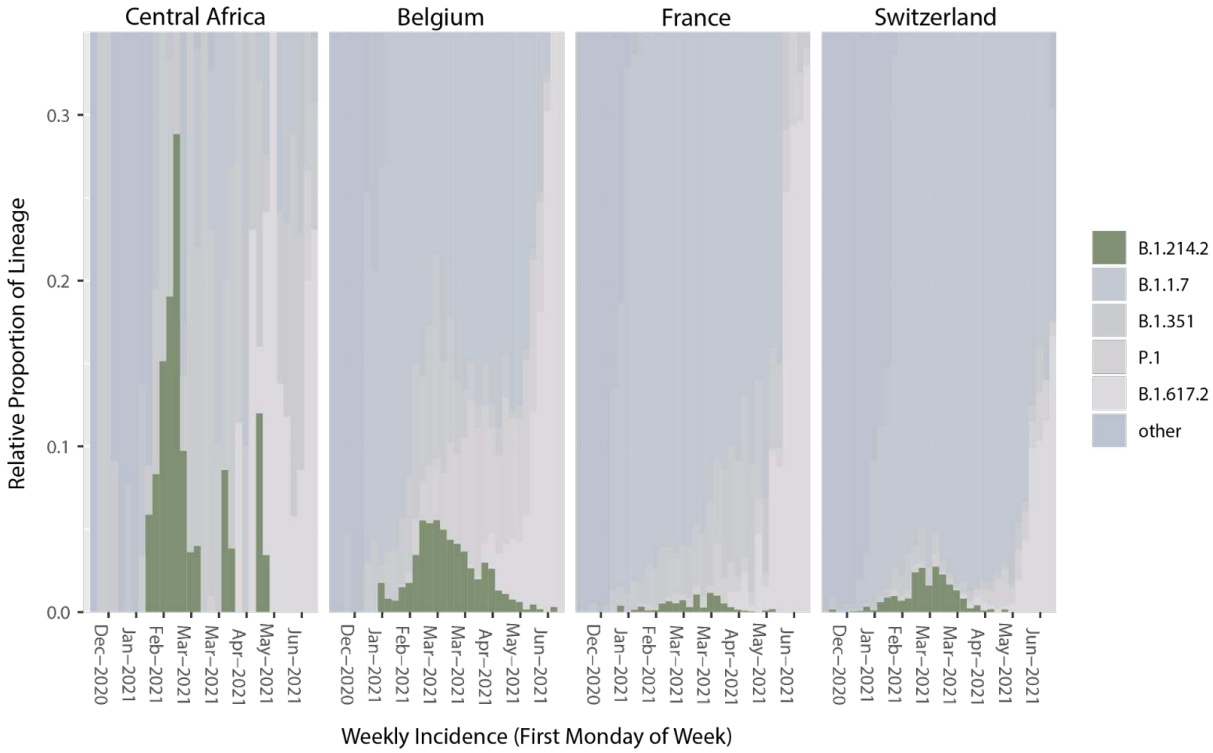


Figure 2. Relative proportions of SARS-CoV-2 lineages from GISAID in four regions. *Sequences deposited to GISAID from Belgium, France, Switzerland, or Central Africa. Due to a limited number of sequences, the Republic of the Congo, the Democratic Republic of the Congo, Angola, and Gabon are aggregated into one 'Central Africa' definition. Dark green color represents B.1.214.2 sequences. A high proportion of B.1.214.2 sequences is visible in Central Africa during January and February 2021. Peaks in Europe occurred in February and March 2021. The prevalence is clearer in Belgium and in Switzerland than in France due to the presence of other lineages.*

Andrew Holtz

PhD Evolutionary Biology / 2023

Frontiére de l'Innovation en Recherche et Education

Université de Paris Cité

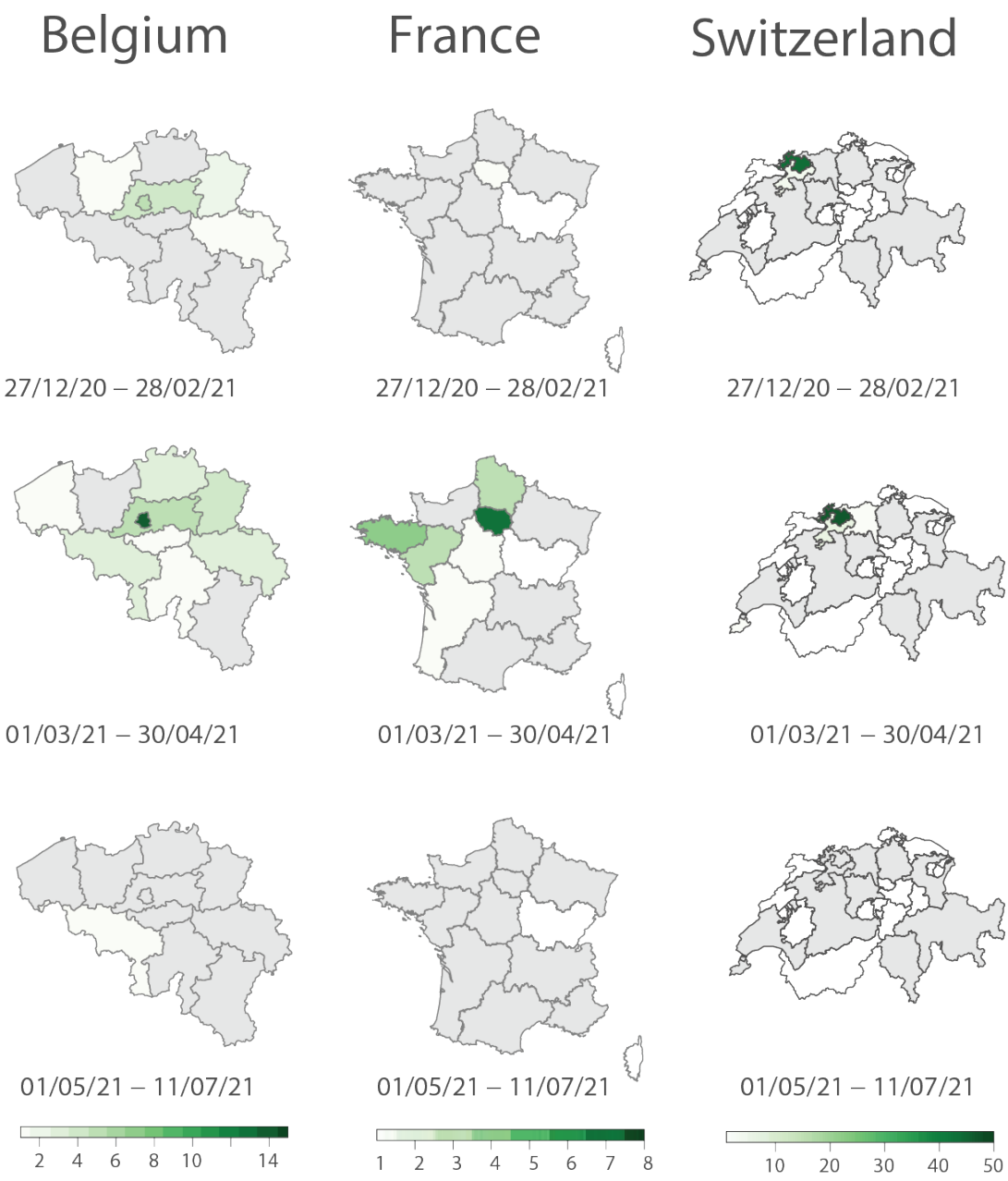


Figure 3: Incidence values (cases/10⁵ people) of B.1.214.2 in Belgian, French, and Swiss regions, grouped into three successive time periods: 27/12/2021 - 28/02/2021, 01/03/2021-30/04/2021, 01/05/2021-11/07/2021. *The B.1.214.2 has the highest density in the Île-de-France region in March and April. In Belgium, the highest density is found in the Brussels region between March and the end of April. B.1.214.2 has the highest density in the Basel-Stadt canton (here merged with Basel-Landschaft for presentation) in the first two time periods. Among all three countries, B.1.214.2 incidence quickly dissipates the last time period as B.1.617.2 begins expanding (see Supp Fig 2). -*

Andrew Holtz

PhD Evolutionary Biology / 2023

Frontiére de l'Innovation en Recherche et Education

Université de Paris Cité

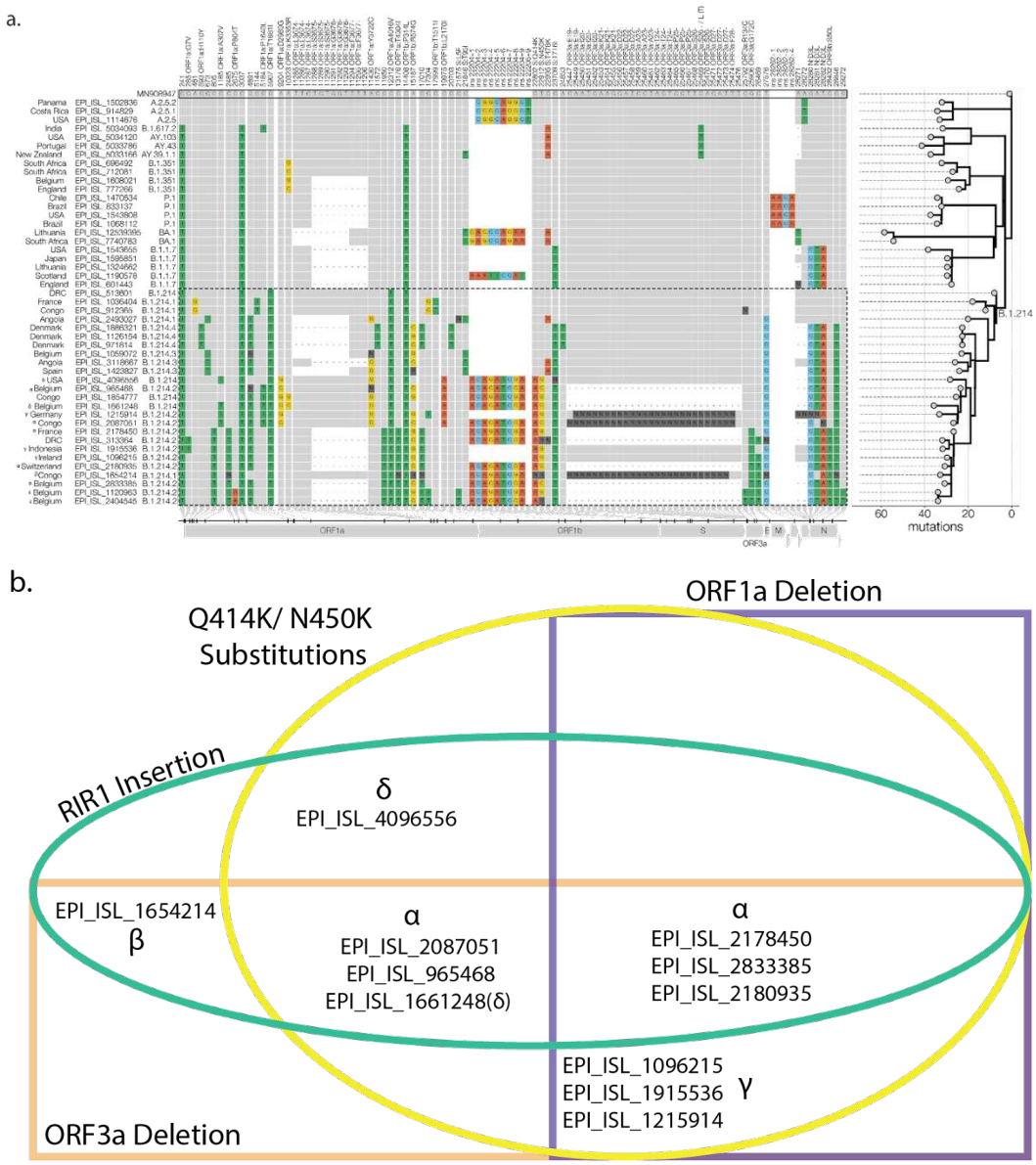


Fig. 4 Lineage-defining SNPs and insertion of lineage B.1.214.2. *a. Genomes belonging to the B.1.214 clade are defined inside the dashed lined box. SNPs that differentiate from the reference (GenBank accession MN908947) and are found in at least two B.1.214 sequences are shown in the condensed SNP alignment. Nucleotides that are shared with the reference strain are shown in grey, while changes from the reference are colored and ambiguities are shown in dark grey. The phylogeny (branch lengths in the number of mutations) on the right shows the relationships between depicted genomes and was rooted on the reference sequence. The long branch defining B.214 is labelled as 'B.1.214'. Greek letters represent groups of sequences with noteworthy mutational profiles. α =*

Andrew Holtz

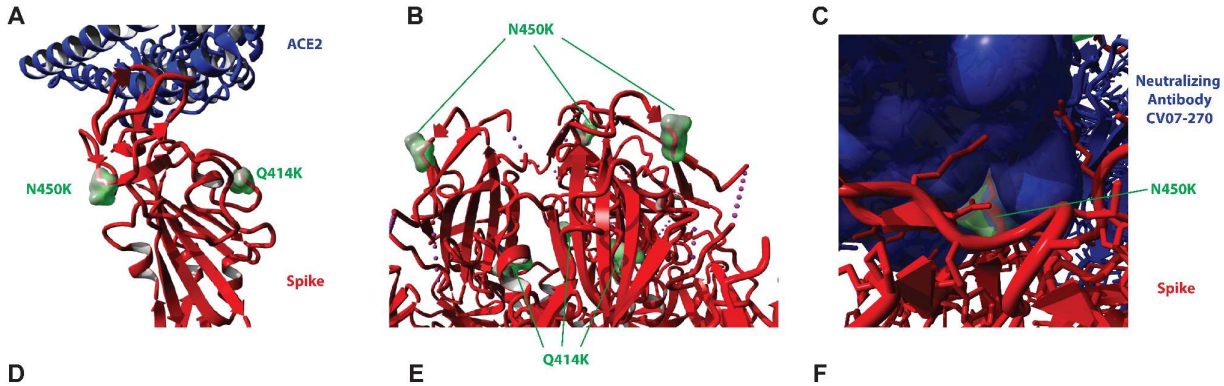
PhD Evolutionary Biology / 2023

Frontiére de l'Innovation en Recherche et Education

Université de Paris Cité

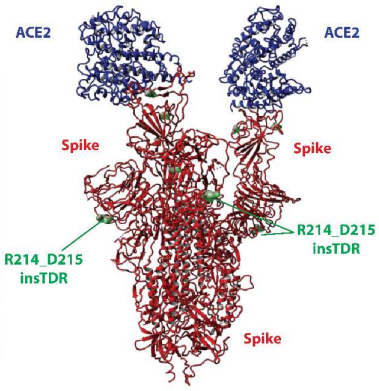
pangolin definition B.1.214.2 (Q414K, N450K, Insertion, and orf1a (50%) and orf3a deletions), β = B.1.214.1 with insertion sequence but missing critical Q414K and N450K, δ = B.1.214 with insertion, Q414K, N450K- possible misclassification by pangolin, ε = Belgian nursing home outbreak; γ = B.1.214.2 that lack insertion sequence but contain Q414K, N450K. b. Venn diagram separated by different lineage defining mutations. Greek letters represent the same sequences as described above. α represents B.1.214.2 both with and without the orf1a deletion.

Andrew Holtz
PhD Evolutionary Biology / 2023
Frontiére de l'Innovation en Recherche et Education
Université de Paris Cité



D

Pdb	Description	Mutations	FoldX ΔΔG (kcal/mol)	FoldX ΔΔG Interaction (kcal/mol)
6M0J	Spike with ACE2 receptor	N450K	1.5	0.0
		Q414K	-0.5	0.0
		N450K,Q414K	1.1	0.0
7A97	Spike trimer with two ACE2 receptors	N450K	1.1	0.0
		Q414K	-1.5	-0.1
		N450K,Q414K	-0.3	-0.1
6VXX	Spike trimer - closed conformation	N450K	1.8	0.0
		Q414K	-1.2	-0.5
		N450K,Q414K	0.7	-0.6
6XKP	Spike protein with neutralizing antibody	NA450K	7.7	6.4
		QA414K	-0.4	0.0
		NA450K,QA414K	8.4	6.9



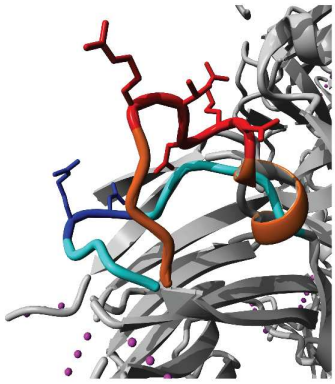


Figure 5. Structural changes of B.1.214.2 in the Spike protein and their functional effects on ACE2 receptor and neutralizing antibodies*. (A). Spike protein in open conformation bound to ACE2. PDB ID: 6VXX (https://doi.org/10.1016/j.cell.2020.02.058). Red: Spike protein, Blue: ACE2 receptor, Green: Mutations site. (B). Spike protein trimer in closed conformation (not bound to ACE2). PDB ID: 6M0J (https://doi.org/10.1038/s41586-020-2180-5). Red: Spike protein, Blue: ACE2 receptor, Green: Mutations site. (C) The N450K mutation would reduce or obliterate the binding affinity of some neutralizing antibodies. Crystal structure 6XKP (https://doi.org/10.1016/j.cell.2020.09.049) is shown as an example. Blue: antibody, Red: Spike protein, Green: N450, which when mutated to Lysine (N450K), would likely decrease the binding affinity of the antibody to the Spike protein and would therefore decrease their neutralizing effect. (D) Effect of mutations on thermodynamic stability and interaction energy predicted by FoldX for different types of complexes. Values are in kcal/mol, above 0.5 is deemed destabilizing and below -0.5 is deemed stabilizing, in between -0.5 and 0.5 is considered to have no effect. (E) Structure of the complex between the Spike protein and the ACE2 receptor (PDBid: 7A97 (https://doi.org/10.1038/s41586-020-2772-0)). Red: Spike protein, Blue: ACE2 receptor, Green: Insertions site. Modeling of effect of Belgian Spike insertion compared to WT Spike protein. (F) The structure of the wildtype loop is shown in cyan, the insertion site is highlighted in dark blue. The structure of the modeled Belgian loop is shown in orange, insertion is highlighted in red.*

Andrew Holtz

PhD Evolutionary Biology / 2023

Frontiére de l'Innovation en Recherche et Education

Université de Paris Cité

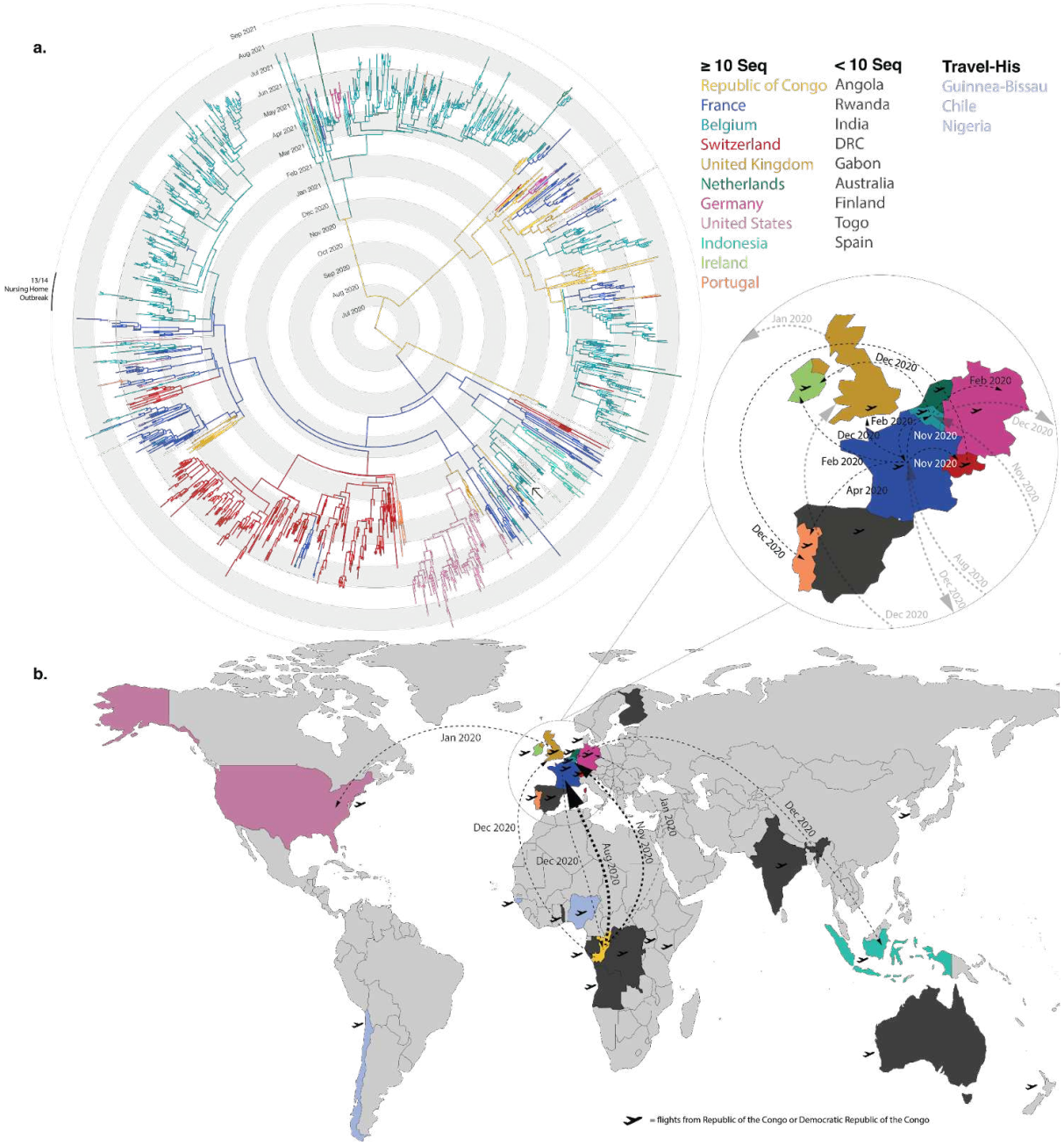


Figure 6. Geographic spread of SARS-CoV-2 lineage B.1.214.2. *a) MCC tree from travel history-aware phylogeographic analysis presenting ancestral country estimations of the B.1.214.2 variant. Colored branches and tips indicate the country of branch origin. Timescale is displayed radially as month and year starting from June 2020. Countries with less than 10 sequences are shown in dark grey. Countries added to analysis from travel-histories are presented in light blue. Tips in the tree that represent sequences from the nursing-home outbreak in Belgium are indicated. The solo arrow indicates one nursing-home sequence which is separate from the rest b) World Map displaying the*

Andrew Holtz

PhD Evolutionary Biology / 2023

Frontiére de l'Innovation en Recherche et Education

Université de Paris Cité

countries with B.1.214.2 cases. Colors follow the same as above, while light grey countries are absent of B.1.214.2 cases. The plane icon next to the country indicates inbound flights from the Republic of the Congo (RC) and the Democratic Republic of the Congo (DRC). Arrows show the first two introductions to each country or transmissions that result in 15 or more tips from the MCC tree. The date of introduction is shown along the arrow. Specifics are shown in Supp. Table 2. Liechtenstein was compiled with Switzerland and Luxembourg compiled with Belgium, as sequencing was performed across borders.

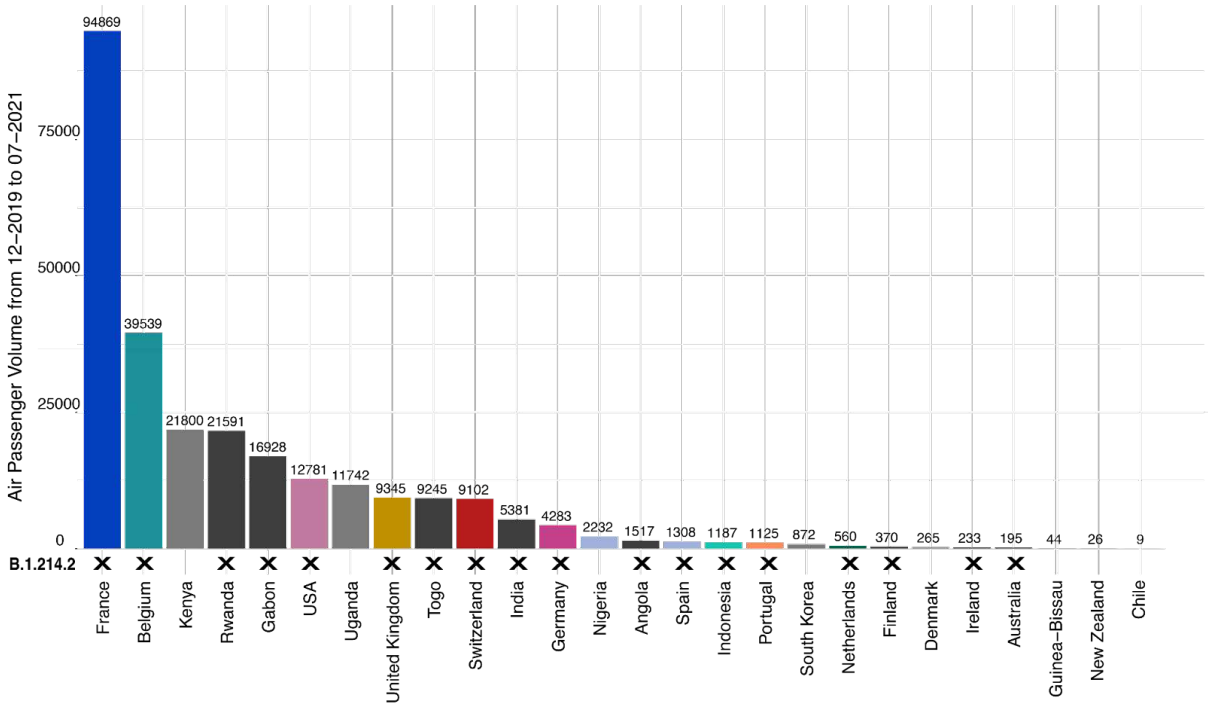


Figure 7. Countries with inbound flights from Republic of the Congo (RC) or the Democratic Republic of the Congo (DRC) are sorted by total passenger volume between December 2019 and July 2021. *RC and DRC are combined since the two airports in their capital cities are used interchangeably by the citizens of both countries. Countries that have submitted at least one B.1.214.2 sequence are marked by 'X'. Only nine of 27 countries with inbound flights from RC and DRC do not report B.1.214.2 cases.*

Andrew Holtz

PhD Evolutionary Biology / 2023

Frontiére de l'Innovation en Recherche et Education

Université de Paris Cité

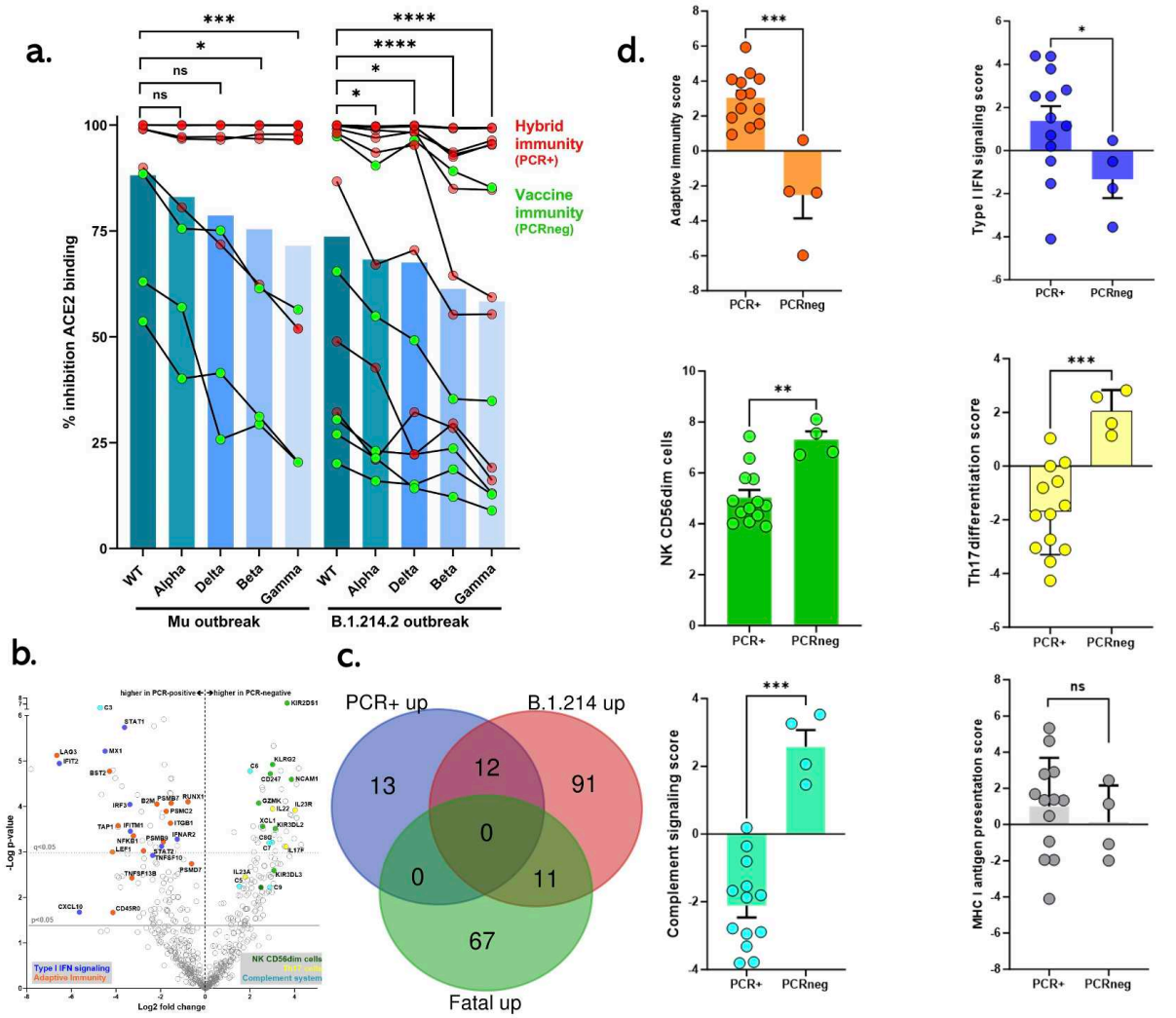


Figure 8: Cross-comparison of large nursing home outbreaks reveals similar systemic neutralizing antibody levels but divergent upper airway immune signature of B.1.214.2 in high-risk elderly. *a. Vaccine-induced (PCR-negative cases) and hybrid (PCR-positive cases) humoral immunity was highly similar in age- and sex-matched residents from B.1.214.2 (n=15) and Mu (n=9) nursing home outbreaks. In both outbreaks, significant decreases in cross-neutralizing antibodies follow the same order of antigenic distance across VOC (WT>Alpha>Delta>Gamma>Beta). Each line represents a single individual, bars represent the median;. Kruskal-Wallis test with FDR correction for multiple testing, *p<0.05, **p<0.01, ***p<0.001, ****p<0.0001. b. Volcano plot of differentially expressed genes in upper airway of age- and sex-matched B.1.214.2-infected (PCR-positive, n=13) and uninfected (PCR-negative, highly exposed, n=4) nursing home residents. Grey lines show raw p-value <0.05 and FDR-corrected q-value <0.05 (dotted line). Individual genes belonging to significantly enriched cell types or signaling pathways are highlighted. c. Venn diagram shows a significant overlap (enrichment p<0.01) in upregulated immune genes shared between B.1.214.2-infected PCR+ ("B.1.214 up") upper airway samples and matched samples of mild/moderate ("PCR+ up")*

Andrew Holtz

PhD Evolutionary Biology / 2023

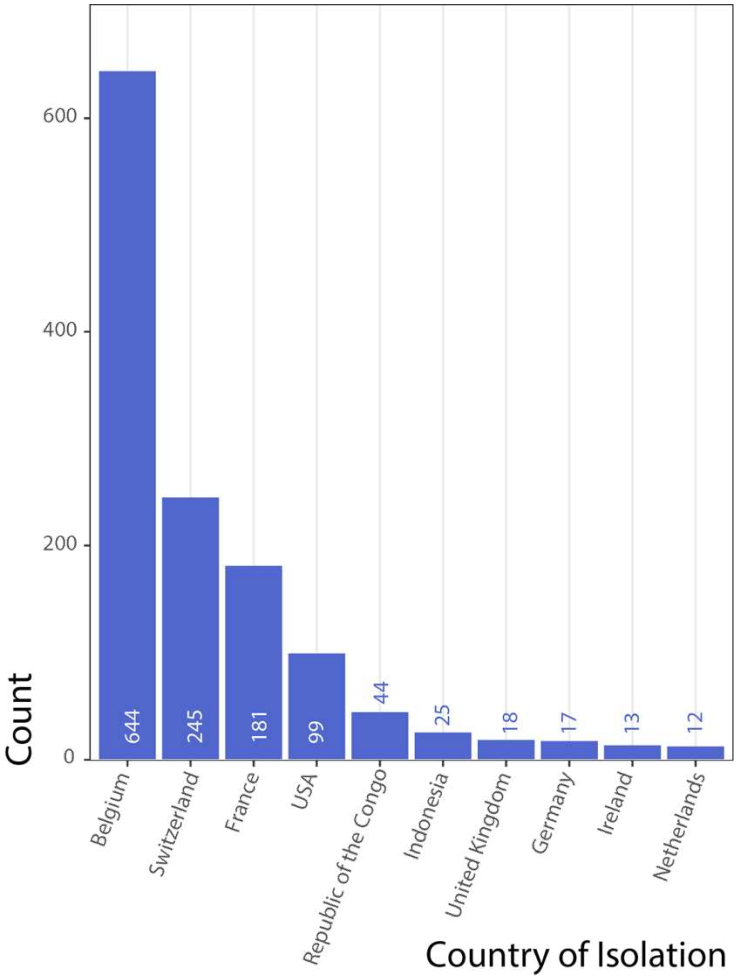
Frontiére de l'Innovation en Recherche et Education

Université de Paris Cité

*Gamma/Delta/Mu nursing home outbreaks, but not with immune genes upregulated in matched fatal cases ("fatal up"). d. Increased Adaptive immunity and type I IFN signaling but decreased NK CD56"dim" cells, Th17 differentiation and complement system in B.1.214.2 PCR+ (n=13) vs. PCR-negative highly exposed (n=4) residents. Mann-Whitney test *p<0.05, **p<0.01, ***p<0.001.*

Andrew Holtz
PhD Evolutionary Biology / 2023
Frontiére de l'Innovation en Recherche et Education
Université de Paris Cité

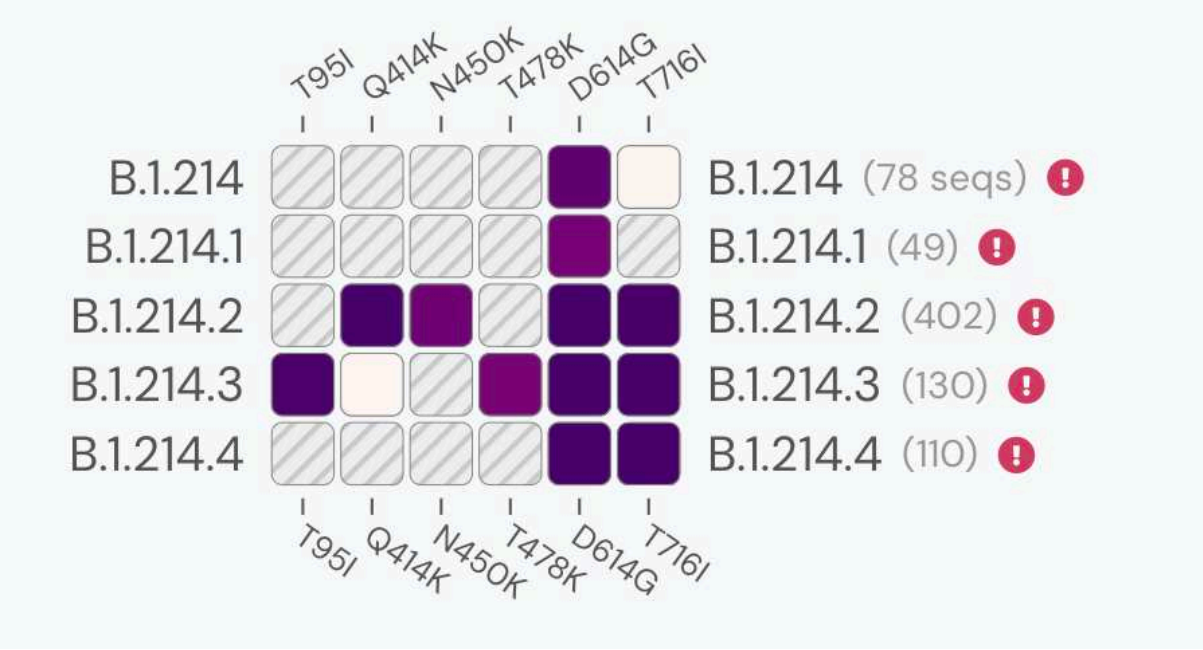
3.8 Supplemental Figures



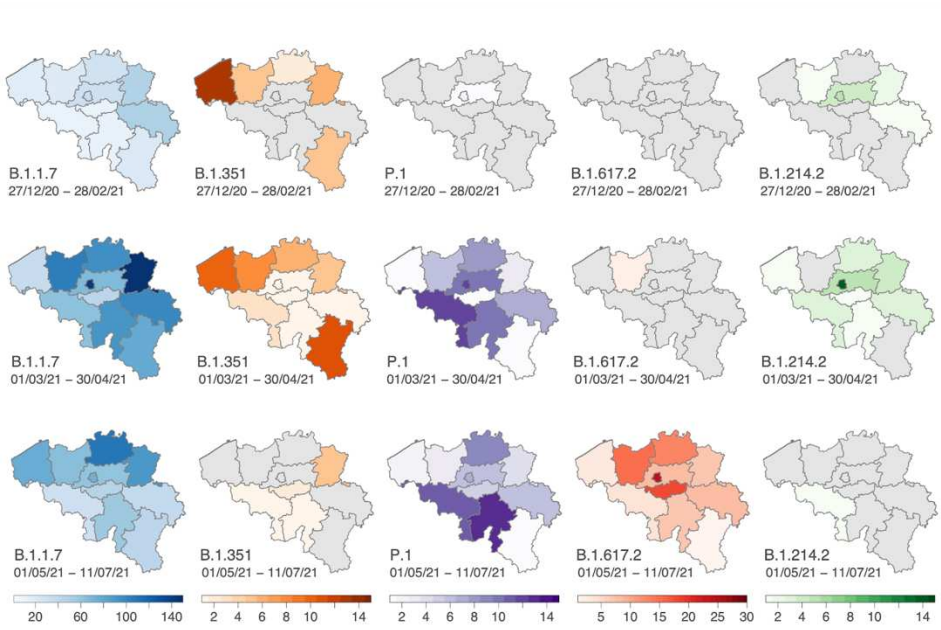
Country	# Sequences
Belgium	644
Switzerland	245
France	181
USA	99
Republic of the Congo	44
Indonesia	25
United Kingdom	18
Germany	17
Ireland	13
Netherlands	12
Portugal	11
Liechtenstein	10
Angola	6
Rwanda	6
India	3
Democratic Republic of the Congo	2
Gabon	2
Australia	1
Finland	1
Togo	1

Supp. Figure 1. Sequenced cases of B.1.214.2 per country. Number of sequences are shown by country. The ten countries with the most B.1.214.2 sequences are shown in the bar plot on the left. The table on the right shows the total number of countries and number of sequences of the variant submitted to GISAID. This does not include travel-history additions.

Andrew Holtz
PhD Evolutionary Biology / 2023
Frontiére de l'Innovation en Recherche et Education
Université de Paris Cité



Supp. Figure 2. Spike protein mutational prevalence across B.1.214 descending lineages. *This figure was taken from outbreak.info*

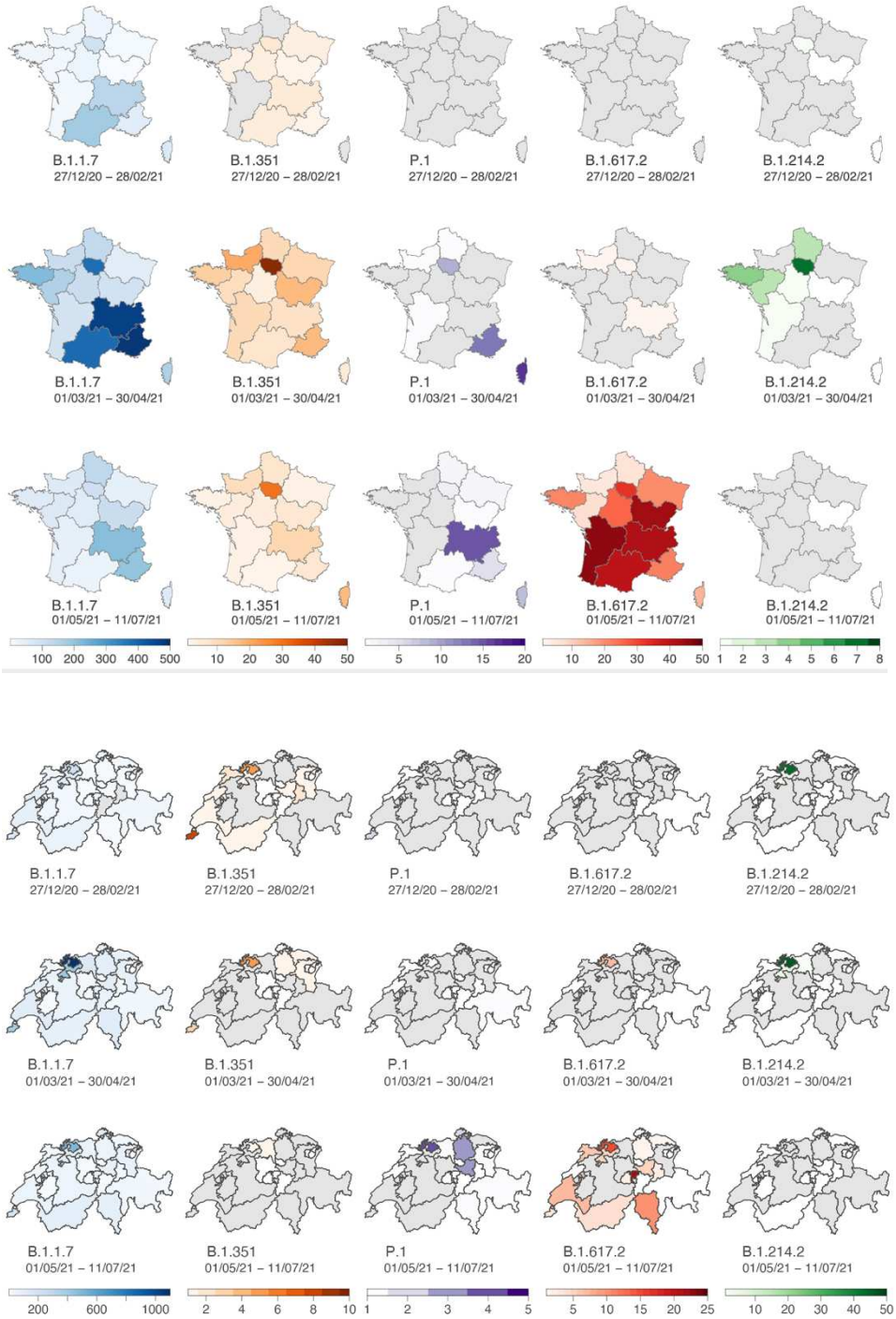


Andrew Holtz

PhD Evolutionary Biology / 2023

Frontiére de l'Innovation en Recherche et Education

Université de Paris Cité



Supp. Fig. 3: Incidence values of each VOC in Belgian Provinces (cases/10⁵ people), French regions (cases/10⁵ people), and Swiss Canton (cases/10⁵ people) grouped into three successive time periods: 27/12/2021 - 28/02/2021, 01/03/2021- 30/04/2021, 01/05/2021-11/07/2021. *Incidence values are calculated by the number of sequences deposited on GISAID for that region divided by the population of that region/province/canton (2021). Brussels in Belgium, Ile-de-France*

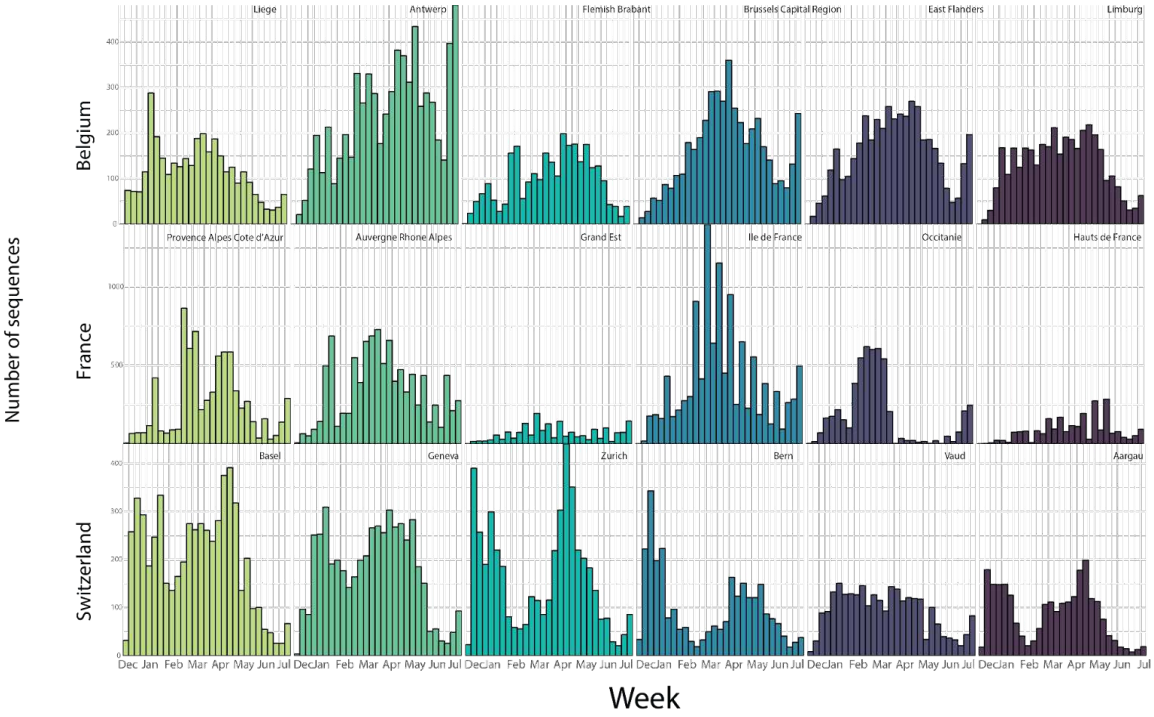
Andrew Holtz

PhD Evolutionary Biology / 2023

Frontiére de l'Innovation en Recherche et Education

Université de Paris Cité

in France and Basel canton (here merged with Basel-Lanschaft and Basel-Stadt for presentation) in Switzerland show the highest incidence of B.1.214.2. The variant quickly dies off in the last time period as B.1.617.2 begins expanding. B.1.214.2 have relatively low incidence values outside of the Basel, Brussels, and Ile-de-France regions.



Supp Figure 4. Number of Sequenced Cases of SARS-CoV-2 in Belgium, France, and Swizerland. *Each country is representing by the sequence totals per week by the six regions with the most number of sequences. Sequences submitted to GISAID were grouped by French region, Belgian Province, and Swiss Canton. In Switzerland, Basel is the combination of Basel-Stadt and Basel-Landschaft.*

Spike protein variations.

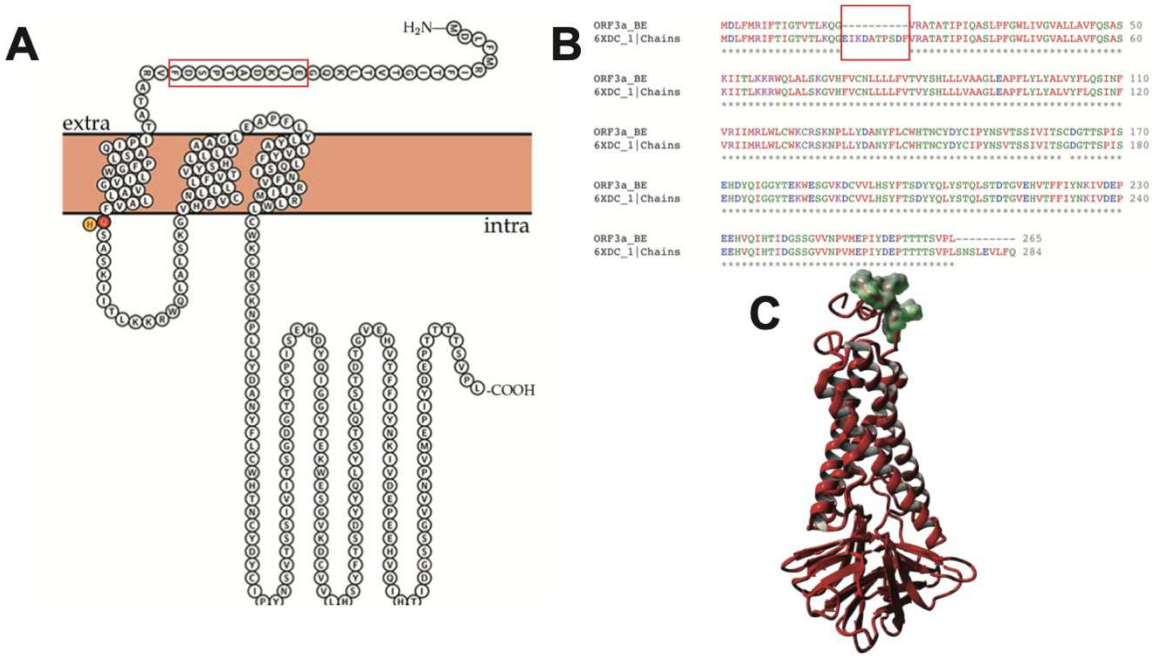
Supp. Figure 5. Multiple sequence alignment of wildtype, Belgian variants and a Costa Rica variant. *The area of insertion is shown in the red box.*

Andrew Holtz

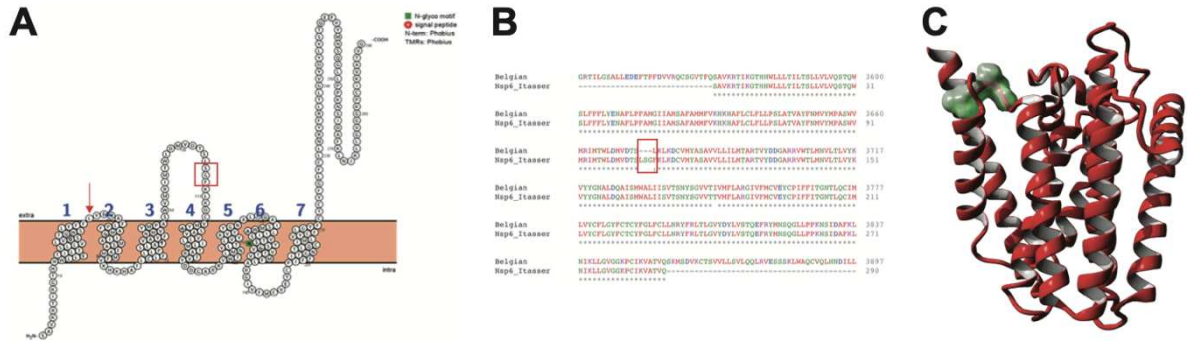
PhD Evolutionary Biology / 2023

Frontiére de l'Innovation en Recherche et Education

Université de Paris Cité



Supp. Figure 6. Deletion of 10 amino acids (30bp) in ORF3a. *A. Schematic of ORF3a protein, red box indicated deleted amino acids. B. Sequence alignment between ORF3a_BE and the "wildtype" ORF3a. C. Structure visualisation of ORF3a, in green the modelled deleted region is highlighted.*



Supp. Figure 7. Deletion of 3 amino acids (9bp) in Orf1a*. A. Schematic of Nsp6 protein (ion channel), red box indicated deleted amino acids.* ***B****. Sequence alignment between Nsp6_BE and the "wildtype" Nsp6.* ***C.*** *Structure visualisation of Nsp6, in green the modelled deleted region is highlighted.*

Andrew Holtz

PhD Evolutionary Biology / 2023

Frontiére de l'Innovation en Recherche et Education

Université de Paris Cité

Supp. Table 1. Travel-history associate with B.1.214.2 sequences. These travel histories were collected from GISAID metadata and from contacting country sequencing laboratories. Travel days are suspected number of days individual was in the country of travel before the sample date.

Name	Location	Sampling Date	Travel History	Travel Days	Prior Mean	Prior Stdev
hCoV-19/Belgium/rega-1632/2021|EPI_ISL_890291|2021-01-03	Belgium	03/01/2021	Republic of the Congo	NA	10	3
hCoV-19/Belgium/rega-1638/2021|EPI_ISL_890294|2021-01-03	Belgium	03/01/2021	Republic of the Congo	NA	10	3
hCoV-19/Belgium/ULG-11260/2021|EPI_ISL_833185|2021-01-07	Belgium	07/01/2021	Republic of the Congo	3	NA	NA
hCoV-19/Belgium/rega-1778/2021|EPI_ISL_894200|2021-01-08	Belgium	08/01/2021	Spain	NA	10	3
hCoV-19/Belgium/rega-1784/2021|EPI_ISL_894201|2021-01-08	Belgium	08/01/2021	Spain	NA	10	3
hCoV-19/Belgium/rega-1865/2021|EPI_ISL_912424|2021-01-12	Belgium	12/01/2021	Democratic Republic of	NA	10	3
hCoV-19/Belgium/ULG-12537/2021|EPI_ISL_1123370|2021-01-19	Belgium	19/01/2021	Democratic Republic of	0	NA	NA
hCoV-19/Belgium/Jessa_55-2105-001118/2021|EPI_ISL_1128129|2021-02-04	Belgium	04/02/2021	France	12	NA	NA
hCoV-19/Belgium/rega-5194/2021|EPI_ISL_1382699|2021-03-02	Belgium	02/03/2021	France	NA	10	3
hCoV-19/Belgium/WHT-UMONS-CV2100704647/2021|EPI_ISL_1524721|2021-03-09	Belgium	09/03/2021	France	38	10	3
hCoV-19/Angola/CERI-KRISP-K014923/2021|EPI_ISL_2493007|2021-04-17	Angola	17/04/2021	Guinea-Bissau	NA	10	3
hCoV-19/Angola/CERI-KRISP-K014924/2021|EPI_ISL_2492994|2021-04-17	Angola	18/04/2021	Chile	NA	10	3
hCoV-19/Belgium/rega-8462/2021|EPI_ISL_2833326|2021-04-27	Belgium	27/04/2021	Germany	NA	10	3
hCoV-19/France/NOR-IPP11606/2021|EPI_ISL_2259092|2021-04-30	France	30/04/2021	Nigeria	NA	10	3

Andrew Holtz
PhD Evolutionary Biology / 2023
Frontiére de l'Innovation en Recherche et Education
Université de Paris Cité

Supp. Table 2. First country introductions and the country transitions up to the root. *First country introductions from the MCC tree are shown here with subtrees larger than or greater to 15 tips. If country introductions are smaller than 15 tips, the two introductions with the largest subtree per country are shown. Country jumps are not nodes, but rather they indicate country change.*

	Country	Country	Country	Final	Tips	Date	Node	95%	95%
Root				RC	1360	10/06/2020	2020.443	2020.164	2020.681
Branch			RC	France	827	22/08/2020	2020.646	2020.490	2020.777
Branch	RC	RC	France	Switzerland	247	03/11/2020	2020.845	2020.791	2020.888
Branch	RC	RC	France	Belgium	59	13/11/2020	2020.868	2020.793	2020.933
Branch	RC	RC	France	Belgium	225	28/11/2020	2020.911	2020.853	2020.976
Branch			RC	Belgium	18	29/11/2020	2020.913	2020.994	2020.831
Branch			RC	Belgium	219	03/12/2020	2020.922	2020.987	2020.854
Branch	RC	France	Belgium	Indonesia	25	03/12/2020	2020.929	2020.838	2021.005
Branch		RC	France	UK	86	09/12/2020	2020.940	2020.885	2020.984
Branch		RC	France	RC	7	16/12/2020	2020.958	2020.880	2021.024
Branch			RC	UK	23	16/12/2020	2020.959	2020.915	2020.992
Branch			RC	France	20	26/12/2020	2020.986	2020.917	2021.045
Branch	RC	France	Switzerland	Portugal	6	28/12/2020	2020.992	2020.944	2021.030
Branch	RC	France	Belgium	Ireland	8	29/12/2020	2020.994	2020.943	2021.028
Branch			RC	Germany	8	03/01/2021	2021.005	2020.914	2021.076
Branch			RC	Belgium	118	05/01/2021	2021.010	2021.010	2021.008
Branch	RC	France	UK	USA	80	06/01/2021	2021.014	2020.961	2021.058
Branch			RC	Belgium	41	12/01/2021	2021.031	2021.035	2020.997
Branch		RC	Belgium	France	20	02/02/2021	2021.088	2021.042	2021.133
Branch		RC	Belgium	Netherlands	4	15/02/2021	2021.124	2021.094	2021.145
Branch		RC	Belgium	Germany	10	17/02/2021	2021.130	2021.099	2021.161
Branch		RC	France	Ireland	3	25/02/2021	2021.149	2021.108	2021.178
Branch		RC	France	Portugal	4	01/04/2021	2021.248	2021.228	2021.265
Branch		RC	Belgium	Netherlands	3	29/05/2021	2021.406	2021.371	2021.428

Andrew Holtz

PhD Evolutionary Biology / 2023

Frontiére de l'Innovation en Recherche et Education

Université de Paris Cité

Supp. Table 3. *Overlap between upregulated genes in B.1.214.2 infected nursing home residents and Gamma/Delta/Mu-infected nursing home residents with mild/moderate vs. fatal outcome*

Gene Set	Num. of genes	Gene Symbol
B.1.214 up PCR+ up	12	CXCL10, SERPING1, CCL5, FYN, GZMB, TAP1, BST2, LAG3, CIITA, LILRB1, CCL2, IFIT2
B.1.214 up Fatal up	11	IFI35, TP53, NOS2, UBE2L3, CDKN1A, ILF3, IL6ST, ATG5, CCL7, CTNNB1, PSMC2
PCR+ up	13	PRF1, ZAP70, CTLA4_all, C1S, C1R, CARD9, EBI3, KLRC3, PDCD1, CXCL11, CD2, IRF8, LAIR1
B.1.214 up	91	FKBP5, LGALS3, IFI16, CLEC7A, TOLLIP, LAMP3, CD45R0, TNFSF10, TNFSF13B, CEACAM1, MUC1, ARG2, LEF1, GBP1, CUL9, STAT2, IDO1, IFNAR2, CD46, IRF5, MALT1, ITGAE, TMEM173, CTSS, TNF, MX1, TRAF4, PSMB8, C1QBP, RAF1, CX3CL1, IKZF2, MYD88, IFITM1, HLA-DRB1, CASP2, CXCL9, ITGB1, IRF3, CD40, CD59, CXCL13, PPBP, CFH, HLA-DMB, CLEC5A, MRC1, KIT, IRAK2, CFB, PSMB5, CXCL1, JAK2, CHUK, IL10, TBK1, CASP1, STAT3, ATG7, MAPK1, CCL20, MCL1, RELA, PSMB7, SOCS1, IL13RA1, CCND3, STAT1, IFIH1, CFD, SMAD3, IFNAR1, RUNX1, IKBKAP, TRAF5, TCF4, B2M, PSMD7, IRAK3, TAP2, NFKB1, CD58, CD164, AHR, DUSP4, CASP3, PSMB9, NOTCH2, FAS, C3, ITGA4
Fatal up	67	TRAF2, XCL1, KCNJ2, ITGAM, IFNG, CD79B, IL17F, C14orf166, TNFAIP6, C8G, IL26, BID, C4A/B, IFNA1/13, LILRA4, CR2, IFNB1, TNFSF15, TLR2, CCL16, TIGIT, TGFB1, RAG2, EOMES, CCL3, BCAP31, CD19, ABL1, B3GAT1, PDGFB, CCR1, ITGA2B, NCAM1, THY1, PTPN22, IL6R, CDH5, CCR5, CD99, TLR1, CTSG, IL28A, LILRA5, IL17B, CD80, IL22, RORC, ICAM5, CD1A, KIR3DL2, ARG1, STAT5A, CCL19, C6, PTGER4, KIR3DL3, SRC, FCAR, NT5E, C9, PLA2G2A, CD79A, CSF2, CCRL2, AIRE, CCL13, KLRC1

Andrew Holtz
PhD Evolutionary Biology / 2023
Frontiére de l'Innovation en Recherche et Education
Université de Paris Cité

4 General Discussion, Outcomes and Perspectives

4.1 Project Development

The primary objective of this thesis was to identify predictors influencing the spread of RABV (and other viruses). However, during this exploration, we recognized a gap in our comprehensive understanding and in the literature of RABV's historical geographic distribution. This realization propelled me into an exhaustive study of rabies, employing various analytical methods and parameters, including maximum likelihood and BEAST, and evaluating diverse sample sizes. I focused on specific regions, such as Africa and Europe, and utilized the generalized linear model in BEAST to discern potential factors contributing to the virus's emergence. Ultimately, I developed a method via maximum likelihood to analyze all RABV including all sequences and countries in one analysis. Using this amount of data allowed us to investigate human migration impact on the spread of RABV.

Upon establishing our methodology for deciphering rabies phylogeography and highlighting the potential impact of human migration over long distance in the spread of canine RABV, I concentrated on the mechanics of BEAST phylogeography, particularly the use of the generalized linear model for predictor testing. This curiosity, coupled with the emergence of the SARS-CoV-2 pandemic, shifted my attention to B.1.214.2.

Combined, these two projects underscore the significance of using phylogenetic, geographic, and human migration data in tandem for large-scale phylogeographic analyses. By integrating insights from these two projects, I was able to conceptualize a novel method that uses the generalized linear model within the maximum likelihood framework to study predictors and their influence on large-scale phylogeographic analyses.

In this discussion, I will begin by exploring the proposed GLM method within the ML framework. Following that, I will delve into the overarching themes present in the two chapters. I will address the challenges and advantages inherent in large-scale phylogeographic studies, emphasizing the transformative power of thorough data integration in our models. Additionally, I'll touch upon the

limitations of the two chapters and the profound role of human migration in disease propagation. In closing, I'll discuss the broader epidemiological implications of the research, highlight challenges in applying findings to real-world scenarios, and underscore our works potential to shape the future landscape of disease research.

4.2 Future work: Generalize Linear Model in a ML Framework

As discussed in section 1.2.4.1.3, the Generalized Linear Model is a statistical model that can be used to model relations between different data types based on the assumption that the response variable is linearly related to the predictor variables through a link function. The GLM can be estimated using a variety of methods, BIC (Bayesian information criterion) or using maximum likelihood. Once the model has been estimated, the regression coefficients can be used to interpret the relationship between the predictor variables and the response variable.

In phylogeography, the GLM can be used to statistically quantify the relationship between various predictors and rates of viral movement or dispersion across geographic space. This is done by estimating the transition rates between two locations, or countries, by building a model between the predictor variables between two countries and an optimized weight on that predictor variable.

	A	B	C	…
A	-	d_{AB}	d_{AC}	…
B	d_{BA}	-	d_{BC}	…
C	d_{CA}	d_{CB}	-	…
…	…	…	…	-

--

$$\text{Log}(\Lambda_{AB}) = \boldsymbol{\delta}_1 \boldsymbol{\beta}_1 \mathbf{P}_{1AB} + \boldsymbol{\delta}_2 \boldsymbol{\beta}_2 \mathbf{P}_{2AB} + \boldsymbol{\delta}_3 \boldsymbol{\beta}_3 \mathbf{P}_{3AB} + \boldsymbol{\delta}_n \boldsymbol{\beta}_n \mathbf{P}_{nAB}$$

Transition rate from A to B = $\boldsymbol{\Sigma}$(*n* Predictors for AB* coefficient weight *binary support)

Andrew Holtz
PhD Evolutionary Biology / 2023
Frontiére de l'Innovation en Recherche et Education
Université de Paris Cité

Here, the response variable is the rate of viral movement between locations (AB…n). The predictors (P_n) are various factors that are hypothesized to influence the spread of disease such as geographic distance, air transit data or ecological data like temperature or humidity. Each rate of movement between locations is parameterized as a log-linear function of the predictors, meaning that the natural log of the rate is linearly dependent on the predictors. For each predictor (P), there is a coefficient **β**, which quantifies the effect size of the predictor on the rate of movement. Lastly, in BEAST, a Bayesian stochastic search variable selection (BSSVS) is used to assess the relevance of each predictor of the data, providing a probability that a predictor should be included in the model. This weight influences the predictor variables, represented by matrices. The summation of the product of each predictor and its weight yields a new rate matrix between two countries, which is then employed in BEAST to estimate phylogeographic probabilities.

In 2014, Lemey et al. first reported using the generalized linear model (GLM) to estimate the influence of air transportation on the spatial spread of H3N2. Subsequent to their study, researchers have applied the GLM model in various disease contexts such as rabies [34,38], Zika [232], West Nile virus [125], and SARS-CoV-2 [26]. In Dellicour et al. the authors used temperature, precipitation and greenness index as predictor for West Nile virus genetic diversity through time, highlighting the impact of a multivariate analysis by testing all four of their covariates at a time [125]. In Dudas et al. researchers studied the factors that helped push and sustain the Ebola epidemic of 2013-2016. They used 1610 whole genome Ebola virus sequences and reconstructed the phylogenetic history of the virus between Sierra Leone, Guinea, and Liberia. They then tested 25 geography-related factors for their impact on viral spread ranging from geographic distance to language sharing, to population size, precipitation, and temperature. Of the 25 tested, their GLM estimated categorical support for five indicating that geographically close regions are likely to share Ebola diversity. They also found that geographic borders tended to limit the spread of EBOV [90].

However, in BEAST, this approach can be computationally demanding, given a large number of sequences, geographic locations, and test predictors. Dudas et al. showcases the GLM in BEAST and also highlights the upper limit for the number of sequences used, as 1610 WGS, has been largest sequence dataset used in a BEAST GLM model. For a sustained epidemic, this has been so far sufficient. The Ebola epidemic underscored the value of whole genome sequencing in

Andrew Holtz
PhD Evolutionary Biology / 2023
Frontiére de l'Innovation en Recherche et Education
Université de Paris Cité

understanding epidemic emergence and maintenance, as well as the importance of international laboratory support and sequencing efforts. Yet, the SARS-CoV-2 global pandemic has further cemented the role of sequencing as a paramount tool for disease investigation, signaling its continued growth in importance. However, current analytical tools, designed for smaller epidemics or dependent on extensive subsampling, face limitations when confronted with such vast data sets. As of September 2023, with 16 million sequences of SARS-CoV-2 submitted in GISAID, the challenge becomes clear. To effectively utilize these data sets in phylogeographic reconstructions using a GLM model, we must develop tools adept at managing larger numbers of sequencing, thus increasing evolutionary, temporal, and geographic diversity.

Recently, a team developed a tool that highlights potential relationships between predictors (covariates) and dispersal rates of viral lineages. Named PhyCovA (Phylogeographic Covariate Analysis) [233], this tool takes a dated tree annotated previously with discrete state annotations and determines the relationship between covariates and the phylogeography via univariate and multivariate linear regression analysis [233]. Although this highlights potential relationships, the tool does not modify the ancestral character reconstruction, but rather determines a relation between the phylogeographic reconstruction and test predictors after a previous ACR step. This tool, similar to TempEst for temporal signal investigation, works as a preliminary step prior to a formal and time-consuming generalized linear model with probabilistic inferencing.

The GLM model for phylogeographic reconstruction is computationally intensive. It not only conducts mathematical analysis but also optimizes and estimates parameters through MCMC, making it more demanding than other Bayesian phylogeographic inferences. As outlined in section 1.2.3, phylogeographic reconstruction using maximum likelihood operates in stages. First, an evolutionary tree is estimated using ML software. This tree is then dated with methods such as Least-Squares Dating or TreeTime. Only after this step can ancestral characters be estimated within the tree. This step-wise approach reduces the computational workload in a maximum likelihood model for ACR. Moreover, the maximum likelihood calculation is less computationally intensive than its Bayesian counterpart, as it doesn't rely on prior information in its search space. However, for analyses involving vast numbers of sequences, a maximum likelihood method might be adequate since larger datasets naturally provide more inference power.

Andrew Holtz
PhD Evolutionary Biology / 2023
Frontiére de l'Innovation en Recherche et Education
Université de Paris Cité

With this in mind, I present the generalize linear model in the maximum likelihood framework. Stemming from the generalized linear model in maximum likelihood in BEAST, I propose the same formulaic calculation of transition rates between locations but within maximum likelihood, namely within the ancestral character reconstruction software, PastML (see section 1.2.3.1.4).

GLM in Maximum likelihood

PastML is a software that uses decision theorist concepts to infer a set of likely states (locations) for each tree node with the highest marginal posterior probability. PastML requires a dated-tree reconstruction and state location information, and from this it optimizes parameters to best estimate ancestral nodes. Transition rates between locations can be inferred in PastML, or they can also be given by the user. I explored the PastML code both computationally and mathematically to determine how we can repackage the user defined rate matrix parameter to take on the generalized linear model.

In the user defined rate matrix parameter, a user provides a set of predictors given as symmetric *nxn* matrices, where *n* represents the number of countries represented in the dataset, like shown below:

Share Economic Area

	Germany	France	UK
Germany	1	1	0
France	1	1	0
UK	0	0	1

Direct Train to Capital City

	Germany	France	UK
Germany	1	0	0
France	0	1	1
UK	0	1	1

The goal is to determine a new rate matrix based on the generalized linear model which we present below:

$$\mathrm{Log}(\Lambda_{AB}) = \boldsymbol{\delta}_1 \boldsymbol{\beta}_1 \mathbf{P}_{1AB} + \boldsymbol{\delta}_2 \boldsymbol{\beta}_2 \mathbf{P}_{2AB} + \boldsymbol{\delta}_3 \boldsymbol{\beta}_3 \mathbf{P}_{3AB} + \boldsymbol{\delta}_n \boldsymbol{\beta}_n \mathbf{P}_{nAB}$$

Andrew Holtz
PhD Evolutionary Biology / 2023
Frontiére de l'Innovation en Recherche et Education
Université de Paris Cité

Log(Λ_{AB}) is the new rate between locations A and B given the predictors above, so in the example above, we have the following equation for the rate between France and Germany.

$$\begin{aligned} \text{Log}(\Lambda_{\text{France->Germany}}) &= \boldsymbol{\delta}_{\text{train}}\,\boldsymbol{\beta}_{\text{train}}\,\mathbf{P}_{\text{train(F-G)}} + \boldsymbol{\delta}_{\text{EU}}\,\boldsymbol{\beta}_{\text{EU}}\,\mathbf{P}_{\text{EU(F-G)}} \\ &= \boldsymbol{\delta}_{\text{train}}\,\boldsymbol{\beta}_{\text{train}}\,[\mathbf{O}]_{\text{train(F-G)}} + \boldsymbol{\delta}_{\text{EU}}\,\boldsymbol{\beta}_{\text{EU}}\,[\mathbf{1}]_{\text{EU(F-G)}} \end{aligned}$$

The result is therefore the new transition rate between the two countries and will be used for the ancestral reconstruction. However, this is very simplified since the probabilistic modeling for determining the probabilities of the weights (**δ**) of the predictors, coefficients, and the inclusion and exclusion coefficients (**β**) has not yet been determined. PastML uses an optimizer algorthim to optimize many parameters such as frequencies and scaling factor. The idea is to use the same optimizer for the weights. The weights would then be reported, and the values of the weights would determine how PastML estimates the potential impact of that predictor on the ancestral reconstruction. Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) are selection models which evaluate the goodness of fit penalizing for additional parameters. This can be used to determine which combination of covariates should be selected, since we only want to keep the covariates that have a probability of influencing viral spread in the phylogeny. Future builds could include machine learning probability models such as L1 or L2 regularization. The result is a set of coefficients that are statistically estimated to influence viral dispersal, given the starting dated phylogeny.

4.3 **Outcomes and Perspectives**

Breaking the bottleneck in phylogenetics

In this thesis, I explored methods that harness the recent surge in available sequencing data. With advancements in data processing and increased international collaboration toward data sharing, the scientific community faces a unique opportunity for novel analyses. However, this abundance of data introduces a major bottleneck in phylogenetic analysis. The challenge in running analyses is not in data availability, but the need for adequate computational power and methodologies to derive meaningful insights from it.

I demonstrated how the growing amount of data available for phylogenetic studies can be harnessed for phylogeographic and predictor inferencing, especially when it relates to how human migration can influence viral dispersion. The first chapter focuses on the issues of sequence and geographic bias in RABV phylogeographic projects. As reported in the literature, case data of canine RABV does not proportionally represent the number of sequences in international databases. Instead, countries with high case counts of RABV are underrepresented by sequencing data[96,110,111]. To combat this issue, I developed and published a novel method to concatenate partial sequences and aggregate these along with whole-genome sequences. From this I achieved an increased phylogeographic signal and validated our confidence in this method by comparing our results to those from various other methods, conducting subsampling, and reviewing relevant previous studies. Given the amount of statistically validated ancestral estimations and the geographic range, I can estimate how human migration over extensive distances contributes to our phylogeographic results.

This method has significance beyond rabies epidemiology. It can be used for many other viral sequence databases that are fortunate to have a large history of partial sequence submissions, including viruses such as dengue and WNV, and Zika (Introduction, Figure 2). By using all the sequences available, we can reduce sequence and geographic bias. I found that in RABV, there is regional and continental bias in the regions sequenced. Therefore, during phylogenetic and phylogeographic reconstruction, that bias permeates into the results. Especially for epidemic origin investigations, ignoring sequences from neighboring countries can result in biased ancestral estimations.

In Chapter 3, I emphasize that geographic bias in sequencing isn't limited to historical studies with partial sequence databases, as evidenced by our investigation into the SARS-CoV-2 variant B.1.214.2. I used a previously developed method by Lemey et al. to incorporate travel history information and related countries from patients to incorporate countries of travel into the estimation of ancestral origin [26]. This has been used especially for SARS-CoV-2. At the beginning of the pandemic, it was clear that there was a sampling bias in sequencing. Phylogeographic analyses at that time did not take travel information into account, and countries without sequencing data were ignored. Similarly, countries with strong sequencing capacity were overrepresented [91]. This method to incorporate travel history goes beyond just SARS-CoV-2 and

Andrew Holtz
PhD Evolutionary Biology / 2023
Frontiére de l'Innovation en Recherche et Education
Université de Paris Cité

can be used across many viruses, especially those that transit easily across international borders and through air travel. In my study, patient travel history allowed for an understanding of the early spread of variant B.1.214.2 and helped explain why there were countries with direct flights to the Republic of Congo but without B.1.214.2 sequencing.

This thesis offers an additional method to address the bottleneck in phylogenetics. As described earlier, I discuss the potential for a generalized linear model in a maximum likelihood framework. This presents a significant methodological advancement, enhancing our capacity to infer how predictors viral dispersion. With a PastML implementation, users could analyze a larger number of sequences, predictors, and geographic locations facilitating large-scale phylogeographic analysis. This advancement has the potential to restructure how predictors are assessed in viral spread dynamics.

In the context of canine rabies, I was able to demonstrate a relationship between a country's colonial history and viral spread. This approach is not rooted in intricate probabilistic computations, but derives from ancestral character estimations. With a GLM model in PastML a large-scale analysis using more than 14,000 RABV sequences could be used to test defined covariates, similar to the smaller-scale Ebola, West Nile virus, and RABV studies[34,90,125].

Previous studies have used the GLM in BEAST to highlight the effect of human migration and human population density on RABV spread based on finite geographic regions (either within a country or region) [26,34,90]. By using a much larger set of sequences, similar to our study, it is conceivable to provide a global understanding of the factors that drive canine RABV dispersion, including past and present human migration patterns, human density data, global trade networks, and ecological and topographical information such as riverways, deserts, land use, deforestation, etc.

Similarly, a GLM model in PastML could be used on large data sets of SARS-CoV-2 sequences across a large group of countries. The sequence bottleneck has been especially evident for analyses during this pandemic because of the plethora of sequences available. This method offers the potential to use the immense sequence data available along with geographic locations to understand factors that have previously been studied such as air passenger data, land use,

contact with wildlife, and more. Additionally, we can potentially quantify the impact of early SARS-CoV-2 containment measures on viral spread, shedding light on both national and international successes and failures in controlling the virus.

4.4 Limitations

We have discussed the limitations of each study individually, but it is important to highlight some of the major limitations to consider more broadly. Phylogeography in both studies examined countries with fixed borders; however, in both historical geography and in the modern day, borders are fluid. In historical studies involving viruses with dated phylogenies dating back hundreds or thousands of years, country definitions have less meaning, since these countries may not have existed or the borders changed (eg. Russian Empire, European colonial expansion, recent African border definitions). In addition, national efforts in eliminating infectious disease further complicates the story, since sequenced samples might not exist in countries that could have previously been vital to spread (eg. the UK in the canine RABV investigation) but that no longer have cases. In terms of modern fluidity in the country borders, I highlighted how borders have a range in their potential to curb disease spread. For example, it was determined that the international borders during the Ebola epidemic of 2013-2016 were effective in curbing disease spread. This was potentially due to additional language and tribal differences at borders [90]. In contrast, European borders early in the SARS-CoV-2 pandemic fluctuated on their penetrability. As shown in this thesis, viral transmission was still very much possible due to human movements across borders, particularly the French-Belgian and French-Swiss borders. Therefore, in phylogeographic origin studies, it is important to present the assumptions made. Although I estimate the origin of B.1.214.2 is likely the Republic of the Congo, in reality this could be a general symbol for a greater region including the Democratic Republic of the Congo (who shared a capital city border with the Republic of the Congo), and other Central African countries, perhaps those with less sequencing capacity.

Relying on travel history for phylogeographic inference might also introduce inaccuracies in the tree, potentially misattributing travel history as the geographic origin of a strain. Even though travel history provides additional data, in the reconstruction it appears as a new geographic location rather than a hypothetical. A notable example is a transmission sequence in the

B.1.214.2 phylogeographic reconstruction that appears to go from Angola to Chile and then back to Angola. The data suggests that a journey to Chile resulted in a subsequent minor outbreak in Angola. However, considering the original ancestral strain was from Angola, it is more logical to infer that the strain in Chile did not originate there, but was introduced from Angola.

Understanding and acting on drivers of infectious disease spread

Both studies analyzed the impact of human migration on the spread of viral disease. I reviewed the results of each study in detail. However, it is important to highlight how human migration and an increase in global travel has increased the spread of infectious diseases in general. In the introduction, I discussed the danger of zoonoses and how their emergent risk will increase over time. The two studies of this thesis attempt to better understand how viruses can spread, with the hope that this can impact epidemiological decision-making. Better collaboration between science and politics is required; even if methodologies exist that can explain epidemic emergence, the ability to transfer that information to political decision makers is vital. In infectious disease epidemiology, the solutions are often known, but the difficulty relies in implementation—for example, RABV is a vaccine-preventable disease, but the logistics to provide education and vaccines to vulnerable areas is lacking. Even if we uncover novel predictors for viral spread, economic, social, and medical factors affect the way we can translate these predictors into action.

As an example of this, the first SARS epidemic of 2002-2003 highlighted the dangers of live animal markets in the spread of viruses to humans. Live animal markets bring together several species that in the wild remain separate, increasing the potential for viral host-species jumping. Because of this, live animal markets in Guangdong province were banned, but due to social and economic reasons, this ban was only temporary [234]. Similar trends can be seen throughout Asia and Africa, where containment measures are often understood but social and economic pressures prevent epidemiological control.

Importantly, the SARS-CoV-2 pandemic also underscored the capacity for international collaboration and highlighted our limited understanding of disease emergence and dynamics. This investigation, along with other research, equips researchers with tools to grasp epidemic emergence more comprehensively. In addition, collaboration in disease control goals such as the WHO's goal to eliminate rabies by 2030 helps set guidelines, expectations and accountability for

control methods and collaborative efforts. These goals also help establish better collaborations across fields, which highlights the One Health approach for zoonoses control. Our goal is not only to prevent future infectious disease outbreaks but also to understand and control existing ones. Ideally, insights from novel methodologies will enhance our comprehension and potentially mitigate the spread of diseases projected to expand their geographic reach in upcoming decades. This encompasses mosquito-borne viruses like Zika, Dengue, Chikungunya, and West Nile virus; tick-borne diseases including Crimean-Congo Hemorrhagic Fever virus, Lyme disease, and Rocky Mountain Spotted Fever; and other insect vector-driven diseases, such as Leishmaniasis transmitted by sandflies [19].

4.5 **Perspectives**

The utilization of diverse data sets and mathematical approaches in this thesis has underscored the significant benefits of adaptability and diversity in phylogeographic methodologies. I took two contrasting approaches. In one project, I analyzed an endemic RABV that transmits mostly by dogs and wildlife species using maximum likelihood methods to understand historical transmission. In the other project, I focused on the contemporary SARS-CoV-2 using a Bayesian framework and using novel methods, integrating patient travel data with phylogeographic data. Importantly, I demonstrated that using varied approaches to understand phylogeography and factor testing yield nuanced insights, enhancing the depth and breadth of our understanding of disease spread. This adaptability facilitates a more robust and comprehensive analysis, capable of accommodating the evolving nature of infectious diseases and the multifaceted influences of human migration. Consequently, it fosters the development of more effective, data-informed disease control and prevention strategies, tailored to the specific characteristics and challenges posed by different infectious diseases.

Although, not achievable during this thesis, the two projects along with the proposed GLM method in maximum likelihood presents a novel approach to phylogeographic studies with large databases. While I discuss how using a GLM method in maximum likelihood could allow for large covariate testing in SARS-CoV-2 dispersion, a combination of the methods presented in this thesis offers a potential approach in covariate testing for large databases of endemic viruses (Figure 2).

Neglected tropical endemic viruses still receive less financial and programmatic support. These include, Dengue, chikungunya, Japanese encephalitis, rabies, avian influenzas, zika, and West Nile viruses. This is, however, ripe for change due to the introduction of many of these tropical diseases into regions of the world with more financial capacity. Due to the discussed factors in our growing world, many of these endemic viruses will be play a greater role in upper-middle economic regions.

I have already discussed extensively the different factors that can impact the emergence of these zoonotic viruses. To launch a phylogeographic analysis understanding these factors for one virus, a diverse and well sampled sequence sample is required, and for this, the concatenation method (presented in Chapter 1) can be utilized to cover the entire genome and every location that has sequenced a portion of the genome. Yielding, in the cases of RABV and dengue, over 20,000 sequences, a GLM approach in phylogeographic predictor (presented in Chapter 2) testing can be done in a maximum likelihood framework (presented in the first section of the Discussion). Harnessing these methods into this approach could bring new approaches to the control of these viruses. The potential factors subject to testing are vast and diverse, encompassing a wide array of variables that can influence disease transmission and spread. Historical and political factors, including the dynamics of government relationships, colonization history, visa policies, and levels of international cooperation, play a pivotal role in shaping the patterns of disease dissemination across borders and regions. Cultural elements, such as shared languages and religious practices, also contribute to the complexity of disease spread, influencing interpersonal interactions and communal behaviors that can either mitigate or exacerbate transmission risks. Environmental variables, including fluctuations in temperature and humidity, as well as the increasing incidence of tropical storms and flooding can lead to the emergence of infectious agents. Refugee migration patterns, the emergence and overlap of new insect vectors, healthcare capacity, urbanization trends, city growth, and deforestation contribute to the multifaceted landscape of epidemiological dynamics. Each of these elements, individually and collectively, shapes the trajectory of disease spread, underscoring the necessity for comprehensive testing of these diverse factors.

In addition, epidemiological control methods could also be evaluated by testing the method as a factor. It is known that public health intervention evaluation is a difficult undertaking, and many interventions either forgo evaluations or they undergo unsupported and bias ones. Although

there are strong studies on the effect of dog vaccination in the curbing of RABV emergence [96,235], there are very few phylogeographic quantitative methods in the evaluation of public health intervention that effect the emergence and spread of disease. While GLM methods have been in use since 2014 [26], there use in evaluating public health programming has not been used. I suggest using a large database of sequences with the GLM in maximum likelihood to allow for the evaluation of large-scale public health initiatives such as vaccination, mass drug administration, or vector control.

4.6 Conclusions

In this thesis, I explore how shifting migration patterns contribute to the emergence and spread of zoonoses. Armed with advancements in genomic sequencing and an increased emphasis on global collaboration, we can adopt new methodologies for large-scale phylogeographic projects to dissect factors driving viral dispersion. I have highlighted the role of human migration in the dissemination of viruses such as rabies virus and SARS-CoV-2, emphasizing innovative phylogeographic methods and pioneering techniques for large-scale data inclusion. This thesis underscores the importance of integrating genomic data with socio-epidemiological insights to impact strategies for disease control, particularly in the contexts of canine RABV and SARS-CoV-2. History has a tendency to repeat itself, but equipped with novel phylogeographic tools, we can understand our historical errors and better prepare for current and future zoonotic challenges.

5 References

1. Suchard, M. A. *et al.* Bayesian phylogenetic and phylodynamic data integration using BEAST 1.10. *Virus Evol.* **4**, vey016 (2018).
2. Price, M. N., Dehal, P. S. & Arkin, A. P. FastTree 2 - Approximately maximum-likelihood trees for large alignments. *PLoS ONE* **5**, e9490 (2010).
3. Pond, S. L. K., Frost, S. D. W. & Muse, S. V. HyPhy: hypothesis testing using phylogenies. *Bioinforma. Oxf. Engl.* **21**, 676–679 (2005).
4. Minh, B. Q. *et al.* IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. *Mol. Biol. Evol.* **37**, 1530–1534 (2020).
5. Letunic, I. & Bork, P. Interactive Tree Of Life (iTOL) v5: an online tool for phylogenetic tree display and annotation. *Nucleic Acids Res.* **49**, W293–W296 (2021).
6. To, T. H., Jung, M., Lycett, S. & Gascuel, O. Fast Dating Using Least-Squares Criteria and Algorithms. *Syst. Biol.* **65**, 82–97 (2016).
7. Katoh, K., Misawa, K., Kuma, K. & Miyata, T. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res.* **30**, 3059–3066 (2002).
8. Aksamentov, I., Roemer, C., Hodcroft, E. B. & Neher, R. A. Nextclade: clade assignment, mutation calling and quality control for viral genomes. *J. Open Source Softw.* **6**, 3773 (2021).
9. O'Toole, Á. *et al.* Assignment of epidemiological lineages in an emerging pandemic using the pangolin tool. *Virus Evol.* **7**, veab064 (2021).
10. Ishikawa, S. A., Zhukova, A., Iwasaki, W. & Gascuel, O. A Fast Likelihood Method to Reconstruct and Visualize Ancestral Scenarios. *Mol. Biol. Evol.* **36**, 2069–2085 (2019).
11. Campbell, K. *et al.* Making genomic surveillance deliver: A lineage classification and nomenclature system to inform rabies elimination. *PLOS Pathog.* **18**, e1010023 (2022).
12. Rambaut, A., Lam, T. T., Max Carvalho, L. & Pybus, O. G. Exploring the temporal structure of heterochronous sequences using TempEst (formerly Path-O-Gen). *Virus Evol.* **2**, vew007 (2016).
13. Sagulenko, P., Puller, V. & Neher, R. A. TreeTime: Maximum-likelihood phylodynamic analysis. *Virus Evol.* **4**, vex042 (2018).
14. Khan, P. Y. *et al.* Transmission of drug-resistant tuberculosis in HIV-endemic settings. *Lancet Infect. Dis.* **19**, e77–e88 (2019).
15. Bell, L. C. K. & Noursadeghi, M. Pathogenesis of HIV-1 and Mycobacterium tuberculosis co-infection. *Nat. Rev. Microbiol.* **16**, 80–90 (2018).
16. Manna, S. *et al.* Synergism and Antagonism of Bacterial-Viral Coinfection in the Upper Respiratory Tract. *mSphere* **7**, e0098421 (2022).
17. Celentano, D. D. & Szklo, M. *Gordis Epidemiology*. (Elsevier, 2019).
18. World Health Organization, Food and Agriculture Organization of the United Nations, World Organisation for Animal Health, & United Nations Environment Programme. *One health joint plan of action (2022–2026): working together for the health of humans, animals, plants and the environment*. (World Health Organization, 2022).
19. Jánová, E. Emerging and threatening vector-borne zoonoses in the world and in Europe: a brief update. *Pathog. Glob. Health* **113**, 49–57 (2019).
20. Karesh, W. B. *et al.* Ecology of zoonoses: natural and unnatural histories. *Lancet Lond. Engl.* **380**, 1936–1945 (2012).
21. Allen, T. *et al.* Global hotspots and correlates of emerging zoonotic diseases. *Nat. Commun.* **8**, 1–10 (2017).

Andrew Holtz

PhD Evolutionary Biology / 2023

Frontiére de l'Innovation en Recherche et Education

Université de Paris Cité

22. Ba, J. *et al.* Zoonosis emergence linked to agricultural intensification and environmental change. *Proc. Natl. Acad. Sci. U. S. A.* **110**, (2013).
23. Mode of Communication of Cholera(John Snow, 1855). https://www.ph.ucla.edu/epi/snow/snowbook.html.
24. Gozdzielewska, L. *et al.* The effectiveness of hand hygiene interventions for preventing community transmission or acquisition of novel coronavirus or influenza infections: a systematic review. *BMC Public Health* **22**, 1283 (2022).
25. Badri, S. *et al.* Disparities and Temporal Trends in COVID-19 Exposures and Mitigating Behaviors Among Black and Hispanic Adults in an Urban Setting. *JAMA Netw. Open* **4**, e2125187 (2021).
26. Lemey, P. *et al.* Unifying Viral Genetics and Human Transportation Data to Predict the Global Transmission Dynamics of Human Influenza H3N2. *PLoS Pathog.* **10**, e1003932 (2014).
27. Cleaveland, S., Haydon, D. T. & Taylor, L. Overviews of pathogen emergence: which pathogens emerge, when and why? *Curr. Top. Microbiol. Immunol.* **315**, 85–111 (2007).
28. Cottontail, V. M., Wellinghausen, N. & Kalko, E. K. V. Habitat fragmentation and haemoparasites in the common fruit bat, Artibeus jamaicensis (Phyllostomidae) in a tropical lowland forest in Panamá. *Parasitology* **136**, 1133–1145 (2009).
29. Klitting, R. *et al.* Predicting the evolution of the Lassa virus endemic area and population at risk over the next decades. *Nat. Commun.* **13**, 5596 (2022).
30. Deal, A. *et al.* Migration and outbreaks of vaccine-preventable disease in Europe: a systematic review. *Lancet Infect. Dis.* **21**, e387–e398 (2021).
31. Amabo, F. C., Seukap, E. C., Mathieu, E. & Etoundi, G. A. Evaluation of diarrheal disease surveillance in the Minawao refugee camp, Cameroon, 2016. *Int. J. Infect. Dis. IJID Off. Publ. Int. Soc. Infect. Dis.* **82**, 9–14 (2019).
32. Rabaa, M. A. *et al.* Phylogeography of Recently Emerged DENV-2 in Southern Viet Nam. *PLoS Negl. Trop. Dis.* **4**, e766 (2010).
33. Parker, J., Rambaut, A. & Pybus, O. G. Correlating viral phenotypes with phylogeny: accounting for phylogenetic uncertainty. *Infect. Genet. Evol. J. Mol. Epidemiol. Evol. Genet. Infect. Dis.* **8**, 239–246 (2008).
34. Dellicour, S. *et al.* Using phylogeographic approaches to analyse the dispersal history, velocity and direction of viral lineages — Application to rabies virus spread in Iran. *Mol. Ecol.* **28**, 4335–4350 (2019).
35. Campbell, L. P. *et al.* Climate change influences on global distributions of dengue and chikungunya virus vectors. *Philos. Trans. R. Soc. Lond. B. Biol. Sci.* **370**, 20140135 (2015).
36. Licoppe, A. *et al.* Management of a Focal Introduction of ASF Virus in Wild Boar: The Belgian Experience. *Pathog. Basel Switz.* **12**, 152 (2023).
37. Fusaro, A. *et al.* The introduction of fox rabies into Italy (2008-2011) was due to two viral genetic groups with distinct phylogeographic patterns. *Infect. Genet. Evol. J. Mol. Epidemiol. Evol. Genet. Infect. Dis.* **17**, 202–209 (2013).
38. Dellicour, S. *et al.* Using Viral Gene Sequences to Compare and Explain the Heterogeneous Spatial Dynamics of Virus Epidemics. *Mol. Biol. Evol.* **34**, 2563–2571 (2017).
39. Andersen, L. K. & Davis, M. D. P. Climate change and the epidemiology of selected tick-borne and mosquito-borne diseases: update from the International Society of Dermatology Climate Change Task Force. *Int. J. Dermatol.* **56**, 252–259 (2017).
40. Estrada-Peña, A. & Venzal, J. M. Climate niches of tick species in the Mediterranean region: modeling of occurrence data, distributional constraints, and impact of climate change. *J. Med. Entomol.* **44**, 1130–1138 (2007).

Andrew Holtz

PhD Evolutionary Biology / 2023

Frontiére de l'Innovation en Recherche et Education

Université de Paris Cité

41. CDC. Tickborne disease surveillance data summary | CDC. *Centers for Disease Control and Prevention* https://www.cdc.gov/ticks/data-summary/index.html (2022).
42. Olafsdottir, B. & Askling, H. H. Increasing spread of borreliosis in Europe. *New Microbes New Infect.* **48**, 101022 (2022).
43. European Centre for Disease Prevention and Control. Tick-borne diseases. https://www.ecdc.europa.eu/en/tick-borne-diseases (2017).
44. Gossner, C. M. *et al.* Dengue virus infections among European travellers, 2015 to 2019. *Eurosurveillance* **27**, 2001937 (2022).
45. Rossati, A. Global Warming and Its Health Impact. *Int. J. Occup. Environ. Med.* **8**, 7–20 (2017).
46. Amuasi, J. H., Lucas, T., Horton, R. & Winkler, A. S. Reconnecting for our future: The Lancet One Health Commission. *The Lancet* **395**, 1469–1471 (2020).
47. Swedberg, C., Mazeri, S., Mellanby, R. J., Hampson, K. & Chng, N. R. Implementing a One Health Approach to Rabies Surveillance: Lessons From Integrated Bite Case Management. *Front. Trop. Dis.* **3**, 829132 (2022).
48. Global report on neglected tropical diseases 2023. https://www.who.int/publications-detail-redirect/9789240067295.
49. Lv, M.-M. *et al.* Dynamic analysis of rabies transmission and elimination in mainland China. *One Health* **17**, 100615 (2023).
50. Subedi, D., Chandran, D., Subedi, S. & Acharya, K. P. Ecological and Socioeconomic Factors in the Occurrence of Rabies: A Forgotten Scenario. *Infect. Dis. Rep.* **14**, 979–986 (2022).
51. Jou, W. M., Haegeman, G., Ysebaert, M. & Fiers, W. Nucleotide Sequence of the Gene Coding for the Bacteriophage MS2 Coat Protein. *Nature* **237**, 82–88 (1972).
52. Fiers, W. *et al.* Complete nucleotide sequence of bacteriophage MS2 RNA: primary and secondary structure of the replicase gene. *Nature* **260**, 500–507 (1976).
53. Jain, M., Olsen, H. E., Paten, B. & Akeson, M. The Oxford Nanopore MinION: delivery of nanopore sequencing to the genomics community. *Genome Biol.* **17**, 239 (2016).
54. Brunker, K. *et al.* Rapid in-country sequencing of whole virus genomes to inform rabies elimination programmes. *Wellcome Open Res.* **5**, 3 (2020).
55. Shabalina, S. A. & Spiridonov, N. A. The mammalian transcriptome and the function of non-coding DNA sequences. *Genome Biol.* **5**, 105 (2004).
56. Rhodes, T. D., Nikolaitchik, O., Chen, J., Powell, D. & Hu, W.-S. Genetic Recombination of Human Immunodeficiency Virus Type 1 in One Round of Viral Replication: Effects of Genetic Distance, Target Cells, Accessory Genes, and Lack of High Negative Interference in Crossover Events. *J. Virol.* **79**, 1666–1677 (2005).
57. Sanjuán, R. & Domingo-Calap, P. Mechanisms of viral mutation. *Cell. Mol. Life Sci.* **73**, 4433–4448 (2016).
58. Perelson, A. S. Modelling viral and immune system dynamics. *Nat. Rev. Immunol.* **2**, 28–36 (2002).
59. Dudas, G. & Bedford, T. The ability of single genes vs full genomes to resolve time and space in outbreak analysis. *BMC Evol. Biol.* **19**, 232 (2019).
60. Ou, C. Y. *et al.* Molecular epidemiology of HIV transmission in a dental practice. *Science* **256**, 1165–1171 (1992).
61. Blanchard, A., Ferris, S., Chamaret, S., Guétard, D. & Montagnier, L. Molecular evidence for nosocomial transmission of human immunodeficiency virus from a surgeon to one of his patients. *J. Virol.* **72**, 4537–4540 (1998).
62. Troupin, C. *et al.* Large-Scale Phylogenomic Analysis Reveals the Complex Evolutionary History of Rabies Virus in Multiple Carnivore Hosts. *PLoS Pathog.* **12**, (2016).

Andrew Holtz

PhD Evolutionary Biology / 2023

Frontiére de l'Innovation en Recherche et Education

Université de Paris Cité

63. Information on Swine/Variant Influenza | CDC. https://www.cdc.gov/flu/swineflu/index.htm (2023).
64. CDCespanol. Information on Avian Influenza. *Centers for Disease Control and Prevention* https://t.cdc.gov/VAY (2023).
65. Markov, P. V. *et al.* The evolution of SARS-CoV-2. *Nat. Rev. Microbiol.* **21**, 361–379 (2023).
66. Holtz, A., Baele, G., Bourhy, H. & Zhukova, A. Integrating full and partial genome sequences to decipher the global spread of canine rabies virus. *Nat. Commun.* **14**, 4247 (2023).
67. Notredame, C., Higgins, D. G. & Heringa, J. T-Coffee: A novel method for fast and accurate multiple sequence alignment. *J. Mol. Biol.* **302**, 205–217 (2000).
68. Edgar, R. C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* **32**, 1792–1797 (2004).
69. Hennig, W. *Grundzüge einer Theorie der phylogenetischen Systematik*. (Deutscher zentralverlag, 1950).
70. Zimmerman, W. Arbeitsweise der botanischen Phylogenetik und anderer Cruppierungswissenschaften. *E. Abderhalden (ed.), Handbuch der biologischen Arbeitsmethoden. Urban and Schwarzenberg, Germany.* 941–1053 (1931).
71. Zuckerkandl, E. & Pauling, L. Molecules as documents of evolutionary history. *J. Theor. Biol.* **8**, 357–366 (1965).
72. *The Phylogenetic Handbook: A Practical Approach to Phylogenetic Analysis and Hypothesis Testing*. (Cambridge University Press, 2009). doi:10.1017/CBO9780511819049.
73. Yang, Z. *Computational Molecular Evolution*. (Oxford University PressOxford, 2006). doi:10.1093/acprof:oso/9780198567028.001.0001.
74. Minin, V. N. & Suchard, M. A. Counting labeled transitions in continuous-time Markov models of evolution. *J. Math. Biol.* **56**, 391–412 (2008).
75. Tavare, S. Some probabilistic and statistical problems in the analysis of DNA sequences. *Some Math. Quest. Biol. DNA Seq. Anal. Ed. Robert M Miura* (1986).
76. Jukes, T. H. & Cantor, C. R. Evolution of Protein Molecules. in *Mammalian Protein Metabolism* 21–132 (Elsevier, 1969). doi:10.1016/B978-1-4832-3211-9.50009-7.
77. Yang, Z. Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: Approximate methods. *J. Mol. Evol.* **39**, 306–314 (1994).
78. Soubrier, J. *et al.* The Influence of Rate Heterogeneity among Sites on the Time Dependence of Molecular Rates. *Mol. Biol. Evol.* **29**, 3345–3358 (2012).
79. Yang, Z. A space-time process model for the evolution of DNA sequences. *Genetics* **139**, 993–1005 (1995).
80. Churchill, G., von Haeseler, A. & Navidi, W. Sample Size for Phylogenetic Inference. *Mol. Biol. Evol.* **9**, 753–69 (1992).
81. Felsenstein, J. *Inferring Phylogenies*. (Sinauer, 2003).
82. Zhukova, A. *et al.* Efficiently Analysing Large Viral Data Sets in Computational Phylogenomics To cite this version : HAL Id : hal-02536435 Chapter 5 . 3 Efficiently Analysing Large Viral Data Sets in Computational Phylogenomics. 0–43 (2020).
83. Nguyen, L.-T., Schmidt, H. A., von Haeseler, A. & Minh, B. Q. IQ-TREE: A Fast and Effective Stochastic Algorithm for Estimating Maximum-Likelihood Phylogenies. *Mol. Biol. Evol.* **32**, 268–274 (2015).
84. Price, M. N., Dehal, P. S. & Arkin, A. P. FastTree: Computing Large Minimum Evolution Trees with Profiles instead of a Distance Matrix. *Mol. Biol. Evol.* **26**, 1641–1650 (2009).
85. Ho, S. Y. W. & Duchêne, S. Molecular-clock methods for estimating evolutionary rates and timescales. *Mol. Ecol.* **23**, 5947–5965 (2014).

Andrew Holtz

PhD Evolutionary Biology / 2023

Frontiére de l'Innovation en Recherche et Education

Université de Paris Cité

86. Gilks, W. R., Richardson, S. & Spiegelhalter, D. *Markov Chain Monte Carlo in Practice*. (CRC Press, 1995).
87. Hastings, W. K. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika* **57**, 97–109 (1970).
88. Bouckaert, R. *et al.* BEAST 2: A Software Platform for Bayesian Evolutionary Analysis. *PLOS Comput. Biol.* **10**, e1003537 (2014).
89. Dellicour, S., Vrancken, B., Trovão, N. S., Fargette, D. & Lemey, P. On the importance of negative controls in viral landscape phylogeography. *Virus Evol.* **4**, vey023 (2018).
90. Dudas, G. *et al.* Virus genomes reveal factors that spread and sustained the Ebola epidemic. *Nature* **544**, 309–315 (2017).
91. Lemey, P. *et al.* Accommodating individual travel history and unsampled diversity in Bayesian phylogeographic inference of SARS-CoV-2. *Nat. Commun.* **11**, 5110 (2020).
92. King, A. A. *Historical Perspective of Rabies in Europe and the Mediterranean Basin: A Testament to Rabies by Dr. Arthur A. King*. (World Organisation for Animal Health, 2004).
93. Steele, J. H. & Fernandez, P. J. History of Rabies and Global Aspects. in *The Natural History of Rabies* (Routledge, 1991).
94. Baer, G. M. *The Natural History of Rabies*. (Routledge, 2017). doi:10.1201/9780203736371.
95. Hampson, K. *et al.* Modelling to inform prophylaxis regimens to prevent human rabies. *Vaccine* **37**, A166–A173 (2019).
96. Hampson, K. *et al.* Estimating the Global Burden of Endemic Canine Rabies. *PLoS Negl. Trop. Dis.* **9**, (2015).
97. Fooks, A. R. *et al.* Rabies. *Nat. Rev. Dis. Primer* **3**, 1–19 (2017).
98. Anothaisintawee, T. *et al.* Cost-effectiveness modelling studies of all preventive measures against rabies: A systematic review. *Vaccine* **37 Suppl 1**, A146–A153 (2019).
99. Kunkel, A. *et al.* The urgency of resuming disrupted dog rabies vaccination campaigns: a modeling and cost-effectiveness analysis. *Sci. Rep.* **11**, 12476 (2021).
100. Tarantola, A., Tejiokem, M. C. & Briggs, D. J. Evaluating new rabies post-exposure prophylaxis (PEP) regimens or vaccines. *Vaccine* **37**, A88–A93 (2019).
101. Mbilo, C. *et al.* Rabies in dogs, livestock and wildlife: a veterinary perspective. *Rev. Sci. Tech. Int. Off. Epizoot.* **37**, 331–340 (2018).
102. Bourhy, H., Dautry-Varsat, A., Hotez, P. J. & Salomon, J. Rabies, Still Neglected after 125 Years of Vaccination. *PLoS Negl. Trop. Dis.* **4**, e839 (2010).
103. High-Risk Countries for Dog Rabies | Bringing an Animal into U.S. | Importation | CDC. https://www.cdc.gov/importation/bringing-an-animal-into-the-united-states/high-risk.html (2022).
104. Badrane, H. & Tordo, N. Host Switching in Lyssavirus History from the Chiroptera to the Carnivora Orders. *J. Virol.* **75**, 8096–8104 (2001).
105. Bourhy, H. *et al.* The origin and phylogeography of dog rabies virus. *J. Gen. Virol.* **89**, 2673–2681 (2008).
106. Yakovchits, N. V. *et al.* Fox rabies outbreaks in the republic of Buryatia: Connections with neighbouring areas of Russia, Mongolia and China. *Transbound. Emerg. Dis.* **68**, 427–434 (2021).
107. Tao, X.-Y., Li, M.-L., Guo, Z.-Y., Yan, J.-H. & Zhu, W.-Y. Inner Mongolia: A Potential Portal for the Spread of Rabies to Western China. *Vector-Borne Zoonotic Dis.* **19**, 51–58 (2019).
108. Muleya, W. *et al.* Genetic diversity of rabies virus in different host species and geographic regions of Zambia and Zimbabwe. *Virus Genes* **55**, 713–719 (2019).
109. Bourhy, H. *et al.* Ecology and evolution of rabies virus in Europe. *J. Gen. Virol.* **80 (Pt 10**, 2545–2557 (1999).

Andrew Holtz

PhD Evolutionary Biology / 2023

Frontiére de l'Innovation en Recherche et Education

Université de Paris Cité

110. Layan, M., Dellicour, S., Baele, G., Cauchemez, S. & Bourhy, H. Mathematical modelling and phylodynamics for the study of dog rabies dynamics and control: A scoping review. *PLoS Negl. Trop. Dis.* **15**, e0009449 (2021).
111. Nahata, K. D. *et al.* On the Use of Phylogeographic Inference to Infer the Dispersal History of Rabies Virus: A Review Study. *Viruses* **13**, (2021).
112. Wu, F. *et al.* A new coronavirus associated with human respiratory disease in China. *Nature* **579**, 265–269 (2020).
113. Harvey, W. T. *et al.* SARS-CoV-2 variants, spike mutations and immune escape. *Nat. Rev. Microbiol.* **19**, 409–424 (2021).
114. Jonathan E. Pekar *et al.* The recency and geographical origins of the bat viruses ancestral to SARS-CoV and SARS-CoV-2. *bioRxiv* 2023.07.12.548617 (2023) doi:10.1101/2023.07.12.548617.
115. Carabelli, A. M. *et al.* SARS-CoV-2 variant biology: immune escape, transmission and fitness. *Nat. Rev. Microbiol.* **21**, 162–177 (2023).
116. Tracking SARS-CoV-2 variants. https://www.who.int/activities/tracking-SARS-CoV-2-variants.
117. Tian, D., Sun, Y., Zhou, J. & Ye, Q. The Global Epidemic of the SARS-CoV-2 Delta Variant, Key Spike Mutations and Immune Escape. *Front. Immunol.* **12**, (2021).
118. ECDC. SARS-CoV-2 variants of concern as of 21 September 2023. https://www.ecdc.europa.eu/en/covid-19/variants-concern (2023).
119. Cui, Z. *et al.* Structural and functional characterizations of infectivity and immune evasion of SARS-CoV-2 Omicron. *Cell* **185**, 860-871.e13 (2022).
120. Kaleta, T. *et al.* Antibody escape and global spread of SARS-CoV-2 lineage A.27. *Nat. Commun.* **13**, 1152 (2022).
121. Dudas, G. *et al.* Emergence and spread of SARS-CoV-2 lineage B.1.620 with variant of concern-like mutations and deletions. *Nat. Commun.* **12**, 5769 (2021).
122. Brito, A. F. *et al.* Global disparities in SARS-CoV-2 genomic surveillance. *Nat. Commun.* **13**, 7003 (2022).
123. Hepp, C. M. Towards Translational Epidemiology: Next-Generation Sequencing and Phylogeography as Epidemiological Mainstays. *mSystems* **4**, e00119-19 (2019).
124. Nahata, K. D. *et al.* On the Use of Phylogeographic Inference to Infer the Dispersal History of Rabies Virus: A Review Study. *Viruses* **13**, 1628 (2021).
125. Dellicour, S. *et al.* Epidemiological hypothesis testing using a phylogeographic and phylodynamic framework. *Nat. Commun.* **11**, 5620 (2020).
126. Oude Munnink, B. B. *et al.* Rapid SARS-CoV-2 whole-genome sequencing and analysis for informed public health decision-making in the Netherlands. *Nat. Med.* **26**, 1405–1410 (2020).
127. Hatcher, E. L. *et al.* Virus Variation Resource - improved response to emergent viral outbreaks. *Nucleic Acids Res.* **45**, D482–D490 (2017).
128. Smith, G. J. D. *et al.* Origins and evolutionary genomics of the 2009 swine-origin H1N1 influenza A epidemic. *Nature* **459**, 1122–1125 (2009).
129. Lemey, P., Rambaut, A., Drummond, A. J. & Suchard, M. A. Bayesian Phylogeography Finds Its Roots. *PLOS Comput. Biol.* **5**, e1000520 (2009).
130. Faria, N. R. *et al.* Establishment and cryptic transmission of Zika virus in Brazil and the Americas. *Nature* **546**, 406–410 (2017).
131. Bedford, T. *et al.* Cryptic transmission of SARS-CoV-2 in Washington State. *MedRxiv Prepr. Serv. Health Sci.* (2020) doi:10.1101/2020.04.02.20051417.
132. World Health Organization. *WHO expert consultation on rabies: third report*. (World Health Organization, 2018).

Andrew Holtz

PhD Evolutionary Biology / 2023

Frontiére de l'Innovation en Recherche et Education

Université de Paris Cité

133. Vega, S., Lorenzo-Rebenaque, L., Marin, C., Domingo, R. & Fariñas, F. Tackling the Threat of Rabies Reintroduction in Europe. *Front. Vet. Sci.* **7**, (2021).
134. Gibson, A. D. *et al.* Elimination of human rabies in Goa, India through an integrated One Health approach. *Nat. Commun.* **13**, 2788 (2022).
135. Zero human deaths from dog-mediated rabies by 2030: perspectives from quantitative and mathematical modelling. *Gates Open Res.* **3**, 1564 (2020).
136. Carnieli, P., Ruthner Batista, H. B. C., de Novaes Oliveira, R., Castilho, J. G. & Vieira, L. F. P. Phylogeographic dispersion and diversification of rabies virus lineages associated with dogs and crab-eating foxes (Cerdocyon thous) in Brazil. *Arch. Virol.* **158**, 2307–2313 (2013).
137. Tian, H. *et al.* Transmission dynamics of re-emerging rabies in domestic dogs of rural China. *PLOS Pathog.* **14**, e1007392 (2018).
138. Kobayashi, Y. *et al.* Evolutionary history of dog rabies in Brazil. *J. Gen. Virol.* **92**, 85–90 (2011).
139. Brunker, K. *et al.* Elucidating the phylodynamics of endemic rabies virus in eastern Africa using whole-genome sequencing. *Virus Evol.* **1**, 11 (2015).
140. Velasco-Villa, A. *et al.* The history of rabies in the Western Hemisphere. *Antiviral Res.* **146**, 221–232 (2017).
141. http://rabv-glue.cvr.gla.ac.uk. (2022).
142. Holtz, A. Deciphering the global spread of canine rabies virus in the modern era. *Zenodo* https://doi.org/10.5281/zenodo.8047854 (2023) doi:10.21203/rs.3.rs-2648592/v1.
143. Arel-Bundock, V., Enevoldsen, N. & Yetman, C. countrycode: An R package to convert country names and country codes. *J. Open Source Softw.* **3**, 848 (2018).
144. WDI - Home. https://datatopics.worldbank.org/world-development-indicators/.
145. Hayman, D. T. S. *et al.* Evolutionary History of Rabies in Ghana. *PLoS Negl. Trop. Dis.* **5**, e1001 (2011).
146. Maio, N. D., Wu, C.-H., O'Reilly, K. M. & Wilson, D. New Routes to Phylogeography: A Bayesian Structured Coalescent Approximation. *PLOS Genet.* **11**, e1005421 (2015).
147. Mollentze, N., Biek, R. & Streicker, D. G. The role of viral evolution in rabies host shifts and emergence. *Curr. Opin. Virol.* **8**, 68–72 (2014).
148. Holmes, E. C., Woelk, C. H., Kassis, R. & Bourhy, H. Genetic constraints and the adaptive evolution of rabies virus in nature. *Virology* **292**, 247–257 (2002).
149. Deviatkin, A. A. & Lukashev, A. N. Recombination in the rabies virus and other lyssaviruses. *Infect. Genet. Evol. J. Mol. Epidemiol. Evol. Genet. Infect. Dis.* **60**, 97–102 (2018).
150. Chaiklin, M. & Gooding, P. T. M. Animal Trade Histories in the Indian Ocean World. *Eds Martha Chaiklin Philip Gooding Gwyn Campbell Cham CH Palgrave* (2020).
151. Finlay, R. The Voyages of Zheng He: Ideology, State Power, and Maritime Trade in Ming China. *J. Hist. Soc.* **8**, 327–347 (2008).
152. Colombi, D., Poletto, C., Nakouné, E., Bourhy, H. & Colizza, V. Long-range movements coupled with heterogeneous incubation period sustain dog rabies at the national scale in Africa. *PLoS Negl. Trop. Dis.* **14**, e0008317 (2020).
153. Malerczyk, C., DeTora, L. & Gniel, D. Imported Human Rabies Cases in Europe, the United States, and Japan, 1990 to 2010. *J. Travel Med.* **18**, 402–407 (2011).
154. Lemoine, F. & Gascuel, O. Gotree/Goalign: toolkit and Go API to facilitate the development of phylogenetic workflows. *NAR Genomics Bioinforma.* **3**, lqab075 (2021).
155. Shimodaira, H. & Hasegawa, M. Multiple Comparisons of Log-Likelihoods with Applications to Phylogenetic Inference. *Mol. Biol. Evol.* **16**, 1114 (1999).
156. Faith, D. P. Conservation evaluation and phylogenetic diversity. *Biol. Conserv.* **61**, 1–10 (1992).

Andrew Holtz

PhD Evolutionary Biology / 2023

Frontiére de l'Innovation en Recherche et Education

Université de Paris Cité

157. 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* **14**, 587–589 (2017).
158. Chernomor, O., von Haeseler, A. & Minh, B. Q. Terrace Aware Data Structure for Phylogenomic Inference from Supermatrices. *Syst. Biol.* **65**, 997–1008 (2016).
159. Hoang, D. T., Chernomor, O., von Haeseler, A., Minh, B. Q. & Vinh, L. S. UFBoot2: Improving the Ultrafast Bootstrap Approximation. *Mol. Biol. Evol.* **35**, 518–522 (2018).
160. Sand, A. *et al.* tqDist: a library for computing the quartet and triplet distances between binary or general trees. *Bioinformatics* **30**, 2079–2080 (2014).
161. Smith, M. R. Quartet: Comparison of Phylogenetic Trees Using Quartet and Bipartition Measures. (2020) doi:10.5281/zenodo.4117460.
162. Murrell, B. *et al.* Detecting Individual Sites Subject to Episodic Diversifying Selection. *PLOS Genet.* **8**, e1002764 (2012).
163. Kosakovsky Pond, S. L. & Frost, S. D. W. Not So Different After All: A Comparison of Methods for Detecting Amino Acid Sites Under Selection. *Mol. Biol. Evol.* **22**, 1208–1222 (2005).
164. Smith, M. D. *et al.* Less Is More: An Adaptive Branch-Site Random Effects Model for Efficient Detection of Episodic Diversifying Selection. *Mol. Biol. Evol.* **32**, 1342–1353 (2015).
165. Mayer, T. & Zignago, S. Notes on CEPII's distances measures: the GeoDist database. https://mpra.ub.uni-muenchen.de/36347/ (2011).
166. van Dorp, L. *et al.* Emergence of genomic diversity and recurrent mutations in SARS-CoV-2. *Infect. Genet. Evol.* **83**, 104351 (2020).
167. Korber, B. *et al.* Tracking Changes in SARS-CoV-2 Spike: Evidence that D614G Increases Infectivity of the COVID-19 Virus. *Cell* (2020) doi:10.1016/j.cell.2020.06.043.
168. Zhang, L. *et al.* SARS-CoV-2 spike-protein D614G mutation increases virion spike density and infectivity. *Nat. Commun.* **11**, 6013 (2020).
169. Shah, M. & Woo, H. G. Omicron: A Heavily Mutated SARS-CoV-2 Variant Exhibits Stronger Binding to ACE2 and Potently Escapes Approved COVID-19 Therapeutic Antibodies. *Front. Immunol.* **12**, (2022).
170. Yang, W.-T. *et al.* SARS-CoV-2 E484K Mutation Narrative Review: Epidemiology, Immune Escape, Clinical Implications, and Future Considerations. *Infect. Drug Resist.* **15**, 373–385 (2022).
171. Tan, T. S. *et al.* Dissecting Naturally Arising Amino Acid Substitutions at Position L452 of SARS-CoV-2 Spike. *J. Virol.* **96**, e01162-22.
172. Chakraborty, C., Bhattacharya, M., Sharma, A. R. & Mallik, B. Omicron (B.1.1.529) - A new heavily mutated variant: Mapped location and probable properties of its mutations with an emphasis on S-glycoprotein. *Int. J. Biol. Macromol.* **219**, 980–997 (2022).
173. Kuzmina, A. *et al.* SARS-CoV-2 spike variants exhibit differential infectivity and neutralization resistance to convalescent or post-vaccination sera. *Cell Host Microbe* **29**, 522-528.e2 (2021).
174. Blanco, J. D., Hernandez-Alias, X., Cianferoni, D. & Serrano, L. In silico mutagenesis of human ACE2 with S protein and translational efficiency explain SARS-CoV-2 infectivity in different species. *PLOS Comput. Biol.* **16**, e1008450 (2020).
175. Gerdol, M., Dishnica, K. & Giorgetti, A. Emergence of a recurrent insertion in the N-terminal domain of the SARS-CoV-2 spike glycoprotein. *Virus Res.* **310**, 198674 (2022).
176. Gangavarapu, K. *et al.* Outbreak.info genomic reports: scalable and dynamic surveillance of SARS-CoV-2 variants and mutations. *Nat. Methods* **20**, 512–522 (2023).
177. Mfoutou Mapanguy, C. C. *et al.* SARS-CoV-2 B.1.214.1, B.1.214.2 and B.1.620 are predominant lineages between December 2020 and July 2021 in the Republic of Congo. *IJID Reg.* **3**, 106–113 (2022).

Andrew Holtz

PhD Evolutionary Biology / 2023

Frontiére de l'Innovation en Recherche et Education

Université de Paris Cité

178. Tsueng, G. *et al.* Outbreak.info Research Library: a standardized, searchable platform to discover and explore COVID-19 resources. *Nat. Methods* **20**, 536–540 (2023).
179. Cuypers, L. *et al.* Immunovirological and environmental screening reveals actionable risk factors for fatal COVID-19 during post-vaccination nursing home outbreaks. *Nat. Aging* **3**, 722–733 (2023).
180. Haseltine, W. A. New Belgian Variant Illustrates The Versatility Of SARS-CoV-2 In Escaping Immune Suppression. *Forbes* https://www.forbes.com/sites/williamhaseltine/2021/03/31/new-belgian-variant-illustrates-the-versatility-of-sars-cov-2-in-escaping-immune-suppression/.
181. Redactie & |Redactie|. Van Ranst: "'Nieuwe' coronavariant B.1.214 dook eind december al op in België". *hln.be* https://www.hln.be/binnenland/van-ranst-nieuwe-coronavariant-b-1-214-dook-eind-december-al-op-in-belgie~ab92da5b/ (2021).
182. Welle (www.dw.com), D. Belgian researchers identify new coronavirus variant | DW | 29.03.2021. *DW.COM* https://www.dw.com/en/belgian-researchers-identify-new-coronavirus-variant/a-57042412.
183. Novel lineage with key amino acid SNPs and a 9bp spike insertion · Issue #20 · cov-lineages/pango-designation. *GitHub* https://github.com/cov-lineages/pango-designation/issues/20.
184. Liu, Y. & Rocklöv, J. The reproductive number of the Delta variant of SARS-CoV-2 is far higher compared to the ancestral SARS-CoV-2 virus. *J. Travel Med.* **28**, taab124 (2021).
185. Yi, C. *et al.* Comprehensive mapping of binding hot spots of SARS-CoV-2 RBD-specific neutralizing antibodies for tracking immune escape variants. *Genome Med.* **13**, 164 (2021).
186. Alouane, T. *et al.* Genomic Diversity and Hotspot Mutations in 30,983 SARS-CoV-2 Genomes: Moving Toward a Universal Vaccine for the "Confined Virus"? *Pathogens* **9**, 829 (2020).
187. Liu, C. *et al.* The antibody response to SARS-CoV-2 Beta underscores the antigenic distance to other variants. *Cell Host Microbe* **30**, 53-68.e12 (2022).
188. Hu, Y.-F. *et al.* Computation of Antigenicity Predicts SARS-CoV-2 Vaccine Breakthrough Variants. *Front. Immunol.* **13**, (2022).
189. Gruell, H. *et al.* SARS-CoV-2 Omicron sublineages exhibit distinct antibody escape patterns. *Cell Host Microbe* **30**, 1231-1241.e6 (2022).
190. Hu, Y.-F. *et al.* Rational design of a booster vaccine against COVID-19 based on antigenic distance. *Cell Host Microbe* **31**, 1301-1316.e8 (2023).
191. Pairo-Castineira, E. *et al.* GWAS and meta-analysis identifies 49 genetic variants underlying critical COVID-19. *Nature* **617**, 764–768 (2023).
192. Lee, G. C. *et al.* Immunologic resilience and COVID-19 survival advantage. *J. Allergy Clin. Immunol.* **148**, 1176–1191 (2021).
193. Chua, R. L. *et al.* COVID-19 severity correlates with airway epithelium–immune cell interactions identified by single-cell analysis. *Nat. Biotechnol.* **38**, 970–979 (2020).
194. Unterman, A. *et al.* Single-cell multi-omics reveals dyssynchrony of the innate and adaptive immune system in progressive COVID-19. *Nat. Commun.* **13**, 440 (2022).
195. Yoo, J.-S. *et al.* SARS-CoV-2 inhibits induction of the MHC class I pathway by targeting the STAT1-IRF1-NLRC5 axis. *Nat. Commun.* **12**, 6602 (2021).
196. Vigón, L. *et al.* Impaired Cytotoxic Response in PBMCs From Patients With COVID-19 Admitted to the ICU: Biomarkers to Predict Disease Severity. *Front. Immunol.* **12**, 665329 (2021).
197. Castro de Moura, M. *et al.* Epigenome-wide association study of COVID-19 severity with respiratory failure. *EBioMedicine* **66**, 103339 (2021).

Andrew Holtz

PhD Evolutionary Biology / 2023

Frontiére de l'Innovation en Recherche et Education

Université de Paris Cité

198. Tegally, H. *et al.* The evolving SARS-CoV-2 epidemic in Africa: Insights from rapidly expanding genomic surveillance. *Science* **378**, eabq5358 (2022).
199. Wilkinson, E. *et al.* A year of genomic surveillance reveals how the SARS-CoV-2 pandemic unfolded in Africa. *Science* **374**, 423–431 (2021).
200. Hamisi, N. M., Dai, B. & Ibrahim, M. Global Health Security amid COVID-19: Tanzanian government's response to the COVID-19 Pandemic. *BMC Public Health* **23**, 205 (2023).
201. Viana, R. *et al.* Rapid epidemic expansion of the SARS-CoV-2 Omicron variant in southern Africa. *Nature* **603**, 679–686 (2022).
202. Suñer, C. *et al.* A retrospective cohort study of risk factors for mortality among nursing homes exposed to COVID-19 in Spain. *Nat. Aging* **1**, 579–584 (2021).
203. Zhang, Q. *et al.* Inborn errors of type I IFN immunity in patients with life-threatening COVID-19. *Science* **370**, eabd4570 (2020).
204. Zhang, Q., Bastard, P., COVID Human Genetic Effort, Cobat, A. & Casanova, J.-L. Human genetic and immunological determinants of critical COVID-19 pneumonia. *Nature* **603**, 587–598 (2022).
205. Goncalves, D. *et al.* Antibodies against type I interferon: detection and association with severe clinical outcome in COVID-19 patients. *Clin. Transl. Immunol.* **10**, e1327 (2021).
206. Bastard, P. *et al.* Preexisting autoantibodies to type I IFNs underlie critical COVID-19 pneumonia in patients with APS-1. *J. Exp. Med.* **218**, e20210554 (2021).
207. Bastard, P. *et al.* Autoantibodies against type I IFNs in patients with life-threatening COVID-19. *Science* **370**, eabd4585 (2020).
208. van der Wijst, M. G. P. *et al.* Type I interferon autoantibodies are associated with systemic immune alterations in patients with COVID-19. *Sci. Transl. Med.* **13**, eabh2624 (2021).
209. Elbe, S. & Buckland-Merrett, G. Data, disease and diplomacy: GISAID's innovative contribution to global health. *Glob. Chall. Hoboken NJ* **1**, 33–46 (2017).
210. Blanco, J. D., Radusky, L., Climente-González, H. & Serrano, L. FoldX accurate structural protein–DNA binding prediction using PADA1 (Protein Assisted DNA Assembly 1). *Nucleic Acids Res.* **46**, 3852–3863 (2018).
211. Delgado, J., Radusky, L. G., Cianferoni, D. & Serrano, L. FoldX 5.0: working with RNA, small molecules and a new graphical interface. *Bioinformatics* **35**, 4168–4169 (2019).
212. Yang, J. *et al.* The I-TASSER Suite: protein structure and function prediction. *Nat. Methods* **12**, 7–8 (2015).
213. Krieger, E. & Vriend, G. YASARA View—molecular graphics for all devices—from smartphones to workstations. *Bioinformatics* **30**, 2981–2982 (2014).
214. Omasits, U., Ahrens, C. H., Müller, S. & Wollscheid, B. Protter: interactive protein feature visualization and integration with experimental proteomic data. *Bioinformatics* **30**, 884–886 (2014).
215. FoldX web server: an online force field | Nucleic Acids Research | Oxford Academic. https://academic.oup.com/nar/article/33/suppl_2/W382/2505499.
216. Hadfield, J. *et al.* NextStrain: Real-time tracking of pathogen evolution. *Bioinformatics* **34**, 4121–4123 (2018).
217. Duchene, S. *et al.* Bayesian Evaluation of Temporal Signal in Measurably Evolving Populations. *Mol. Biol. Evol.* **37**, 3363–3379 (2020).
218. Baele, G., Lemey, P. & Suchard, M. A. Genealogical Working Distributions for Bayesian Model Testing with Phylogenetic Uncertainty. *Syst. Biol.* **65**, 250–264 (2016).
219. Hong, S. L., Lemey, P., Suchard, M. A. & Baele, G. Bayesian Phylogeographic Analysis Incorporating Predictors and Individual Travel Histories in BEAST. *Curr. Protoc.* **1**, e98 (2021).

Andrew Holtz

PhD Evolutionary Biology / 2023

Frontiére de l'Innovation en Recherche et Education

Université de Paris Cité

220. Lemey, P. *et al.* Accommodating individual travel history, global mobility, and unsampled diversity in phylogeography: a SARS-CoV-2 case study. *BioRxiv Prepr. Serv. Biol.* (2020) doi:10.1101/2020.06.22.165464.
221. Beard, R., Magee, D., Suchard, M. A., Lemey, P. & Scotch, M. Generalized linear models for identifying predictors of the evolutionary diffusion of viruses. *AMIA Jt. Summits Transl. Sci. Proc. AMIA Jt. Summits Transl. Sci.* **2014**, 23–8 (2014).
222. BlueDot: Outbreak Intelligence Platform. *BlueDot* https://bluedot.global/.
223. Baele, G., Gill, M. S., Lemey, P. & Suchard, M. A. Hamiltonian Monte Carlo sampling to estimate past population dynamics using the skygrid coalescent model in a Bayesian phylogenetics framework. *Wellcome Open Res.* **5**, 53 (2020).
224. Rambaut, A., Drummond, A. J., Xie, D., Baele, G. & Suchard, M. A. Posterior Summarization in Bayesian Phylogenetics Using Tracer 1.7. *Syst. Biol.* **67**, 901–904 (2018).
225. Heled, J. & Bouckaert, R. Looking for trees in the forest: Summary tree from posterior samples. *BMC Evol. Biol.* **13**, 221 (2013).
226. Decru, B. *et al.* IgG Anti-Spike Antibodies and Surrogate Neutralizing Antibody Levels Decline Faster 3 to 10 Months After BNT162b2 Vaccination Than After SARS-CoV-2 Infection in Healthcare Workers. *Front. Immunol.* **13**, 909910 (2022).
227. Menezes, S. M., Braz, M., Llorens-Rico, V., Wauters, J. & Van Weyenbergh, J. Endogenous IFNβ expression predicts outcome in critical patients with COVID-19. *Lancet Microbe* **2**, e235–e236 (2021).
228. Lloréns-Rico, V. *et al.* Clinical practices underlie COVID-19 patient respiratory microbiome composition and its interactions with the host. *Nat. Commun.* **12**, 6243 (2021).
229. Fukutani, K. F. *et al.* In situ Immune Signatures and Microbial Load at the Nasopharyngeal Interface in Children With Acute Respiratory Infection. *Front. Microbiol.* **9**, 2475 (2018).
230. Quick, J. *nCoV-2019 sequencing protocol v1*. https://www.protocols.click/view/ncov-2019-sequencing-protocol-bbmuik6w (2020) doi:10.17504/protocols.io.bbmuik6w.
231. Freed, N. E., Vlková, M., Faisal, M. B. & Silander, O. K. Rapid and inexpensive whole-genome sequencing of SARS-CoV-2 using 1200 bp tiled amplicons and Oxford Nanopore Rapid Barcoding. *Biol. Methods Protoc.* **5**, bpaa014 (2020).
232. Grubaugh, N. D., Saraf, S., Isern, S., Michael, S. F. & Andersen Correspondence, K. G. Travel Surveillance and Genomics Uncover a Hidden Zika Outbreak during the Waning Epidemic In Brief. *Cell* **178**, 1057-1071.e11 (2019).
233. Blokker, T., Baele, G., Lemey, P. & Dellicour, S. Phycova — a tool for exploring covariates of pathogen spread. *Virus Evol.* **8**, veac015 (2022).
234. Cheng, M. H. SARS source back on the menu. *Lancet Infect. Dis.* **7**, 14 (2007).
235. Sambo, M. *et al.* Scaling-up the delivery of dog vaccination campaigns against rabies in Tanzania. *PLoS Negl. Trop. Dis.* **16**, e0010124 (2022).

Andrew Holtz

PhD Evolutionary Biology / 2023

Frontiére de l'Innovation en Recherche et Education

Université de Paris Cité