UNIVERSITÉ DU QUÉBEC À MONTRÉAL

UNE ÉTUDE DE L'ÉVOLUTION DU SARS-COV-2 À L'AIDE DES MÉTHODES BIOINFORMATIQUES

MÉMOIRE

PRÉSENTÉ

COMME EXIGENCE PARTIELLE

DE LA MAITRISE EN INFORMATIQUE

PAR

YASMINE KHELIL

DECEMBRE 2023

REMERCIEMENTS

Je souhaite exprimer ma profonde gratitude à mon directeur de recherche Vladimir Makarenkov pour son accompagnement précieux tout au long de mon parcours en bioinformatique à l'UQAM. Sa contribution inestimable, son expertise et son dévouement ont joué un rôle essentiel dans la réalisation de mon mémoire. Je tiens à souligner combien sa fonction de superviseur a été déterminant dans ma réussite académique. Son soutien constant, ses conseils éclairés et sa disponibilité ont été d'une importance capitale pour mener à bien mon projet de recherche. Sa capacité à partager ses vastes connaissances et son expérience avec générosité m'a permis d'acquérir une compréhension approfondie de la discipline. Ses commentaires constructifs, sa rigueur scientifique et son attention méticuleuse aux détails ont grandement amélioré la qualité de mon travail. Grâce à son encadrement avisé, j'ai pu surmonter les obstacles et relever les défis qui se sont présentés.

Je souhaite également adresser mes remerciements chaleureux à tous mes professeurs de la maîtrise en informatique ainsi que le diplôme d'études supérieures spécialisées DESS en bioinformatique. Leurs échanges et leurs idées ont été une source inestimable d'apprentissage et d'inspiration.

Je tiens à exprimer ma gratitude envers le département d'informatique de l'UQAM. L'environnement stimulant qu'il offre, ainsi que les ressources et les collaborations enrichissantes qu'il propose, ont grandement contribué à ma formation et à mon épanouissement en tant qu'étudiante.

Enfin, je voudrais aussi remercier ma famille. Leur soutien indéfectible, leurs encouragements constants et leur compréhension ont été essentiels pour surmonter les moments difficiles de cette aventure académique. Leur présence bienveillante a été une source de force et de motivation.

DÉDICACE

A ma famille

TABLE DES MATIÈRES

LISTE DES FIGURES

LISTE DES TABLEAUX

LISTE DES ABRÉVIATIONS, DES SIGLES ET DES ACRONYMES

ADN	Acide Désoxyribonucléique
ADNdB	ADN double brin
ADNsb	ADN simple brin
AIC	Akaike Information Criterion
ARN	Acide Ribonucléique
ATG	Agents de Transfert de Gènes
BEAST	Bayesian Evolutionary Analysis Sampling Trees
BIC	Bayesian Information Criterion
BLAST	Basic Local Alignment Search Tool
COVID	Coronavirus disease
COVID-19	Coronavirus disease 2019
ICTV	International Committee on Taxonomy of Viruses
MCMC	Markov Chain Monte Carlo
MEGA	Molecular Evolutionary Genetics Analysis using Maximum Likelihood
MERS-CoV	Middle East Respiratory Syndrome Coronavirus
MRCA	Most Recent Common Ancestor
NJ	Neighbor Joining
Nsp14	Nonstructural protein
OMS	Organisation Mondiale de la Santé
ORF	Open Reading Frame
PP	Probabilité Postérieure
RdRp	RNA-dépendant RNA-polymérase
SARS-CoV-2	Syndrome Respiratoire Aigu Sévère Coronavirus 2
SimPlot	Similarity Plot
THG	Transfert Horizontal de Gènes
T-Rex	Tree Explorer

LISTE DES SYMBOLES ET DES UNITÉS

E	Envelope protein (protéine d'enveloppe)
M	Membrane protein (protéine de membrane)
N	Nucleocapsid protein (protéine de nucléocapside)
NC	Nucléocapside
S	Spike protein (protéine de spicule)
SRT	Séquences Régulatrices de la Transcription

RÉSUMÉ

Notre étude présente les résultats clés d'une investigation visant à examiner l'origine et l'évolution du SARS-Cov-2, en utilisant des analyses phylogénétiques et des comparaisons de séquences génomiques. L'objectif principal de cette recherche était de retracer l'histoire évolutive du SARS-Cov-2 en analysant les modifications génétiques observées dans différentes séquences génomiques. Une recherche bibliographique approfondie a été réalisée pour examiner les connaissances existantes sur l'origine et l'évolution du SARS-Cov-2, l'ADN et l'alignement des séquences, la phylogénie, le transfert horizontal de gènes, la recombinaison génétique, ainsi que l'impact du temps sur l'évolution de SARS-Cov-2. Les étapes expérimentales ont inclus une analyse des différentes espèces en utilisant des outils bioinformatique tels que MEGA, T-REX et SimPlot++. Les résultats ont révélé des similitudes élevées entre le génome complet du SARS-CoV-2 et les gènes de coronavirus humain trouvés chez les chauves-souris et les pangolins, soutenant l'hypothèse d'une origine chez les chauves-souris et d'une transmission aux humains par un hôte intermédiaire. Des événements de recombinaisons avec d'autres souches virales ont également été identifiés, suggérant leur rôle dans l'émergence du SARS-CoV-2. En général, nos résultats sont en accord avec ceux de Makarenkov et al. 2021, sauf les transferts horizontaux complets et partiels, trouvés pour le gène E, renforçant ainsi la crédibilité de cette étude. En conclusion, cette étude offre de nouvelles perspectives sur l'origine du SARS-CoV-2 et met en évidence l'importance des analyses phylogénétiques et des comparaisons de séquences génomiques dans la compréhension de son évolution.

Mots clés : SARS-CoV-2, origine, évolution, analyses phylogénétiques, comparaisons de séquences.

INTRODUCTION

Le concept pandémique n'a rien de nouveau pour l'homme. L'histoire humaine a d'ailleurs été jalonnée de maladies épidémiques. Mais si autrefois celles-ci étaient ralenties par la fréquence des déplacements, ce même facteur semble aujourd'hui constituer la cause primaire de la propagation des nouvelles épidémies. De longue date, les affluences se limitaient aux échanges commerciaux et aux guerres, toutefois le développement des voyages et leurs moyens procurent aujourd'hui aux situations épidémiques une nouvelle dimension, celle du temps. (Sardon, 2020)

Parallèlement, l'homme a beaucoup progressé depuis en matière de détection précoce et de réponse rapide à ces épidémies, faisant qu'il réussisse à instaurer les bons gestes et traitements pour limiter, à chaque fois, l'impact de ces épidémies sur l'humanité. Pourtant, le Covid19 a été l'un des plus grands défis sanitaires que l'homme moderne a rencontrés.

En matière de pandémies, les ancêtres du coronavirus humain ont beaucoup appris sur les maladies épidémiques. La peste et le choléra figurent parmi les pandémies les plus foudroyantes qui ont frappé l'humanité, changeant depuis le cours du monde. D'une épidémie à l'autre, l'homme a appris qu'outre les taux de mortalité de ces maladies, elles sont responsables de bouleversements majeurs. Tel était le cas pour les famines succédant à chaque pandémie autrefois. C'est ainsi depuis l'air des temps que l'humanité est amenée à faire face à des pandémies, la défiant en continuant d'exister tantôt en s'adaptant aux conditions et tantôt en essayant de les améliorer par l'innovation et la recherche. Mais l'homme n'est pas le seul être vivant à apprendre de ses erreurs et à progresser. Le monde microscopique réussit parallèlement à faire de même et s'adapter aux différentes résistances acquises de l'homme, c'est le cas de la mutation virale. Le coronavirus humain en est la preuve la plus statiquement significative. (Battin, 2020)

C'est en décembre 2019 que des cas de pneumonie d'étiologie inconnue ont été recensés à Wuhan en Chine. Il s'agit du troisième Coronavirus humain (CoV) se manifestant durant les 20 dernières années. Très rapidement reconnue, une alerte mondiale est lancée par l'Organisation mondiale de la santé (OMS) au titre d'une nouvelle souche de coronavirus : Le SARS-COV-2. Signant le début de la pandémie du siècle. Ce nouveau coronavirus humain a depuis suscité l'intérêt de toute la communauté scientifique et médicale.

C’est ainsi que très vite après le signalement des premiers cas de coronavirus humain, plusieurs travaux ont été publiés où, paradoxalement, l’origine du SARS-COV-2 était peu discutée. (Domingo, 2021)

À ce jour, beaucoup de travaux vont dans la direction d’une origine zoonotique, cependant la problématique de l’origine du SARS-COV-2 reste imparfaitement élucidée et ne permet pas de conclure en une définition propre à ce virus. Néanmoins, les recherches continuent à ce jour, où l’OMS admet que toutes les hypothèses sont à considérer. Notamment, en vue des variations génomiques du SARS-COV-2 consécutives à ses mutations. Là encore, une porte pour de nouvelles hypothèses est permise et qui est en rapport avec la fréquence accrut des mutations de ce virus. En effet, de manière générale, les CoV présentent un génome exceptionnellement long d’environ 30 000 nucléotides (le virus du Sida comporte 10 000, celui d’Ebola en comprend 19 000). Un génome d’une telle longueur est parfaitement possible pour un virus à ARN en raison du système de correction d’erreurs et de réplication uniques dont disposent les CoV qui semblent pour le moindre aider et orienter les recherches visant à enfin conclure à son origine (Sallard et al., 2020).

C’est là que le rôle du séquençage génomique prend toute son ampleur dans la riposte au SARS-COV-2. En admettant la théorie de la transmission interespèces qui est d’ailleurs à ce jour un sujet à controverses, on se retrouve face à un questionnement tout aussi important concernant le dernier hôte animal ayant été infecté avant que le virus ne soit transmis à l’homme. Mais si de loin, répondre à cette question semble pouvoir résoudre le mystère de l’origine de la pandémie du siècle, les mutations perpétuelles du SARS-COV-2 sont quant à elles une preuve réelle du contraire. Effectivement, les CoV se caractérisent par une coévolution avec leurs hôtes, cette adaptation est au final le résultat de mutations ponctuelles et des recombinaisons caractérielles. C’est ainsi que depuis son émergence, toutes les inférences phylogénétiques se basant sur le génome complet se sont avérées biaisées par un amalgame de fragments génomiques donnant suite à des trajectoires évolutives du coronavirus humain. Ces études n’ont cependant pas été vaines et ont permis d’identifier certaines zones génomiques non affectées par les recombinaisons. Il est aujourd’hui indispensable de mener une étude phylogénétique sur chacune des régions recombinées du génome des CoV de chiroptères, une souche récurrente dans les différentes études ciblant à élucider l’origine sue SARS-COV responsable de SRAS en 2002 où le génome est dit en mosaïque comprenant différents morceaux d’au moins deux CoV différents de chauves-souris. Cette architecture génomique est ce qui semble être à l’origine de la diversification exponentielle du nombre de

séquences disponibles pour un même virus. Là encore, il faudra faire face au facteur temps et intercepter les mutations du coronavirus à la même vitesse que celles-ci se produisent.

Cette étude vise à clarifier la relation entre les diverses séquences du SARS-COV-2 provenant de différentes espèces en analysant leurs fragments génomiques qui ont subi des modifications. À ce jour, aucune preuve de l'origine du coronavirus humain n'a été établie et notre objectif est de souligner l'importance des analyses bioinformatiques dans la résolution du questionnement central sur l'origine du virus qui a marqué cette décennie. L'étude phylogénétique que nous avons effectuée a révélé de nouvelles informations concernant l'impact de l'évolution réticulé (Makarenkov & Legendre, 2000; Makarenkov et al., 2004) sur SARS-COV-2, notamment en ce qui concerne la recombinaison génétique et les transferts horizontaux de gènes.

CHAPITRE 1
ORIGINE ET ÉVOLUTION DES CORONAVIRUS

Malgré les freins imposés par le manque de méthodes et matériel d'études autrefois, les premières infections à coronavirus ont été identifiées dans les débuts du siècle précédent. Cependant, jusqu'aux deux dernières décennies, on associait les CoV à des infections respiratoires bénignes. Les humains ont assisté à une transformation marquée par l'émergence de trois nouveaux coronavirus provoquant désormais des syndromes respiratoires graves. Le SARS-COV-1 est apparu pour la première fois en 2003 touchant environ 8000 individus et occasionnant plus de 800 décès, le virus a été maitrisé en quelques mois mettant fin à cette première pandémie. 9 ans plus tard, en 2012, le MERS-CoV a enregistré plus de 2500 cas avec un taux de mortalité de 35 % (Segondy, 2020). Mais c'est en 2019 que la plus grande pandémie à coronavirus a été enregistrée, avec un bilan en évolution de 757 millions de cas depuis son apparition et un important taux de létalité particulièrement élevé chez les personnes âgées et/ou présentant des comorbidités (World Health Organization [WHO], 2023).

Depuis son apparition, le SARS-COV-2 a suscité une activité de recherche scientifique sans précédent, avec près de 2300 articles publiés par semaine (Corvol, 2021). Pourtant, peu de publications semblent aborder la question clé sur son origine.

1.1 Généralités sur les virus

Le mot « virus » est dérivé du latin signifiant poison. Il s'agit en réalité d'un microorganisme qui est beaucoup plus petit que la bactérie, à peine visible au microscope optique. Les virus sont communément connus sous le nom de parasite cellulaire obligatoire, car il s'agit d'une entité incapable de se reproduire de façon autonome. Ainsi, pour se reproduire, un virus nécessite une cellule qu'on appellera une « cellule hôte » afin d'en utiliser les constituants. Les virus sont infectieux, c'est-à-dire qu'ils occasionnent des infections chez leur hôte. La taille moyenne d'une particule virale varie entre 100 et 150 nm, ils sont principalement constitués d'un acide nucléique (ADN ou ARN) et sont responsables de leur propre reproduction et la synthèse de composantes virales. Les particules virales appelées « virions » se forment à l'issue de l'agrégation des composantes virales nouvellement synthétisées dans la cellule hôte et sont responsables de la transmission du génome à la prochaine cellule hôte ou au prochain organisme. Le virion est ensuite désassemblé au niveau de la cellule hôte enclenchant le début du prochain cycle infectieux. En plus de son patrimoine génétique responsable de sa réplication, un virus est typiquement constitué d'une

coque protéique « capside » dont le rôle est de protéger le génome viral (Pellett et al., 2014) (Segondy, 2020).

1.1.1 Les coronavirus

Les coronavirus sont des descendants de la famille des Coronaviridae, ils sont des virus dont l'appellation renvoie à leur aspect en microscopie électronique où leurs spicules forment une couronne autour de la particule virale. Ce sont des virus à ARN, ils possèdent le plus grand génome jamais retrouvé chez des virus à ARN. D'une taille de 30 kilos de bases lui conférant un fort potentiel évolutif. Les CoV se caractérisent par une adaptabilité sans précédent par mutations, délétions et recombinaisons leur permettant de facilement franchir la barrière de l'espèce (Segondy, 2020).

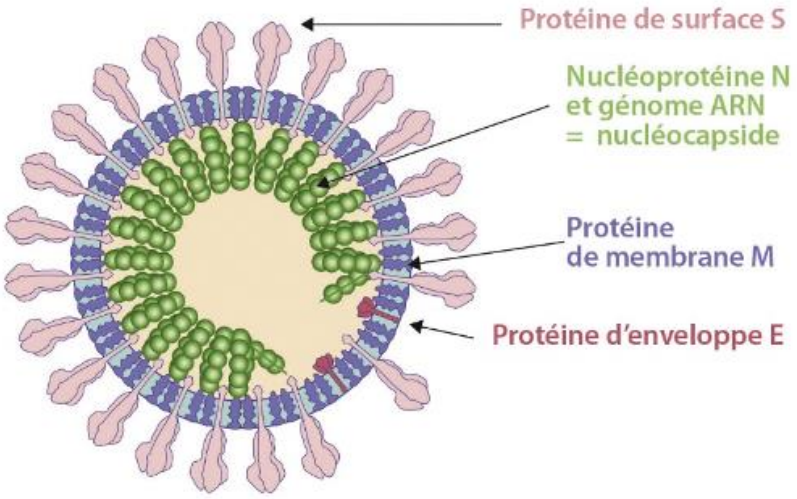

Figure 1.1 Structure du Sars-CoV-2

Source : Lefeuvre et al. (2020)

La particule virale du coronavirus est constituée d'une nucléocapside (NC) formée par l'ARN génomique lié à la protéine N, l'enveloppe formée par les protéines membranaires E, M et la protéine S conférant aux CoV leur structure caractéristique en forme de couronne (Tratner, 2003) :

La protéine N : forme avec l'ARN génomique la nucléocapside, elle présente également des interactions avec les protéines de l'enveloppe.

La protéine E : c'est la plus petite des protéines de structure, elle intervient dans la production et la maturation des particules virales

La protéine M : c'est la protéine la plus abondante des protéines de structures, son rôle se manifeste au moment de l'assemblage des particules virales et dans la forme de l'enveloppe. La protéine M interagit avec toutes les protéines de structure, notamment la protéine N permettant de stabiliser les nucléocapsides.

La protéine S : c'est la protéine clé dans l'expression du pouvoir infectieux du virus. Cette glycoprotéine se lie au récepteur cellulaire de la cellule hôte afin de permettre la fusion entre l'enveloppe virale et la membrane cellulaire amenant éventuellement la pénétration du virus dans la cellule. (Tratner, 2003)

L'enveloppe caractéristique des coronavirus les rend en réalité vulnérables en milieu extérieur. En effet, l'enveloppe virale codant le pouvoir infectieux des CoV est très vite dégradée par la chaleur, la dessiccation, les détergents et les solvants. Ainsi, la contamination par les CoV se fait principalement d'un individu à individu par voie respiratoire. Pour le SARS-CoV-2, le temps de demi-vie est estimé à 8 heures sur des surfaces comme l'inox ou le plastique avec un pouvoir infectieux se prolongeant jusqu'à 72 h. (Segondy, 2020)

1.1.1.1 Origine des coronavirus

En décembre 2019 une émergence de cas de pneumonie sévère a été notée à Wuhan, province du Hubei, en Chine, dont l'origine a été dès le départ lié à un marché ouvert vendant des animaux vivants. Rapidement, les autorités sanitaires du pays ont déclaré un état d'alerte. C'est en janvier 2020 que le pathogène responsable de cette affection pulmonaire sévère a été identifié, c'est le coronavirus subséquemment appelé le SARS-CoV-2 par le ICTV (International Committee on Taxonomy of Viruses). C'est en mars 2020 que la pandémie du coronavirus a été déclarée. Depuis, les théories se sont succédé pour expliquer l'origine du virus responsable de la plus grande pandémie à COV jamais enregistrée. (Boni et al., 2020)

Aussitôt la source des pneumonies identifiée, l'urgence était de contenir la situation. L'origine de cette nouvelle espèce devait être alors au plus vite élucidée. En s'appuyant sur les données disponibles sur les autres espèces de coronavirus, l'origine animale était la première théorie émise.

Depuis le temps, le coronavirus est connu pour des infections chez l'espèce animale. La famille des coronaviridae compte ainsi quatre genres : Alphacoronavirus, Betacoronavirus, Gammacoronavirus et Deltacoronavirus. Chaque genre compte plusieurs sous-genres qui sont à leur tour constitués d'espèces. On recense à ce jour 7 différents coronavirus pouvant provoquer des infections humaines faisant principalement partie des deux genres : Alphacoronavirus et Betacoronavirus.

L'origine des coronavirus a toujours été admise comme étant de nature animale, en effet, c'est à partir des chauves-souris et/ou de certains rongeurs que les CoV ont pu se développer et muter acquérant la capacité de toucher d'autres espèces, notamment l'homme. (Segondy, 2020)

Tableau 1.1 Hôtes naturels et intermédiaires des coronavirus infectant l'homme

Coronavirus humains	Hôtes naturels	Hôtes intermédiaires
HCoV-NL63	Chauves-souris	?
HCoV-229E	Chauves-souris	Dromadaire
HCoV-OC43	Rongeurs	Bovins
HCoV-HKU1	Rongeurs	?
Sars-CoV-1	Chauves-souris	Civette palmiste masquée
Sars-CoV-2	Chauves-souris	Pangolin ?
MERS-CoV	Chauves-souris	Dromadaire

Source : Segondy. (2020)

Cependant, si l'origine primaire des coronavirus a depuis toujours été reconnue comme étant zoonotique, la pandémie du COVID19 semble avoir remis en question cette théorie pour l'espèce SARS-COV-2. Depuis, plusieurs théories ont émergé. Pourtant, la théorie de l'origine chiroptérique (famille de mammifères comptant plus de 1200 espèces de chauves-souris) reste la plus prépondérante même si elle n'a jamais été concrètement confirmée. Le marché ouvert de Wuhan qui a abrité les premiers cas enregistrés de SARS-COV-2 a très rapidement été fermé après le recensement des premières affections pulmonaires ne laissant aucune piste à explorer qui pourrait justement confirmer ou réfuter la théorie zoonotique.

Pourtant, de toutes les théories, l'origine animale du SARS-Cov-2 a été la plus redondante. En effet, c'est en janvier 2020 que Torres-López a suggéré une origine zoonotique à ce nouveau virus pandémique. Comme pour toute pandémie à origine animale, l'identification de la source d'infection est primordiale,

car elle permettrait de séparer l'humain de la population animale responsable de la transmission de l'infection et ainsi y mettre fin. Toutefois, en vue de la rapidité de réplication du SARS-CoV-2, les efforts engagés à cet effet étaient vains, surtout qu'aucune preuve scientifiquement significative n'appuyait cette première théorie. Dès lors, la priorité a été d'identifier l'origine de ce virus dans l'espoir de maitriser cette épidémie (Domingo, 2021).

1.1.1.2 SARS-COV-2: histoire des théories

Des controverses continuent cependant à jaillir au sujet de l'origine chiroptérique du Covid 19. Comme pour tous les CoV humains, la barrière d'espèce ne peut être franchie qu'à travers le passage du virus par une espèce dite « intermédiaire » qui permettrait entre autres au virus de se développer de manière à s'adapter aux récepteurs humains. C'est ce qui a été démontré lors de l'étude des relations phylogénétiques comparant les nouveaux virus et ceux isolés à partir des espèces animales se trouvant au niveau des régions originaires du virus. Il est vrai qu'à ce jour on ne note aucune épidémie causée par une transmission directe de la chauve-souris à l'homme, cependant, plus de 60 CoV ont été identifiés en 2017 comme capables d'infecter les cellules de l'homme lors de cultures in vitro, certains étaient d'ailleurs très similaires au SARS-COV et c'est justement ce qui a fait pencher les recherches scientifiques vers la théorie de la transmission directe. C'est ce qui a récemment été prouvé par l'identification du Sarbecovirus chez la chauve-souris qui serait capable de se multiplier dans la partie supérieure de l'appareil respiratoire de l'homme comme des pangolins (Boni et al., 2020).

Tableau 1.2 Histoire des théories sur l'origine du SARS-CoV-2

Auteur (s)	Date de publication	Théorie
Torres-López	2020	Origine zoonotique
Lyons-Weiler	2020	Recombinaison génétique (fuite de laboratoire)
Hao et al.	2020	Réfutation de la théorie de fuite de laboratoire
Ye et al.	2020	Origine zoonotique
Guo et al.	2020	Identification de la chauve-souris comme hôte naturel du SARS-CoV-2
Zhang et al.	2020	Identification du Pangolin comme réservoir naturel au SARS-CoV-2
Dallavilla et al.	2020	Origine zoonotique probablement chiroptérique
Qiu et al.	2020	Élimination de la théorie de la transmission directe chauve-souris — homme
Lauxmann et al.	2020	La chauve-souris, hôte réservoir du SARS-CoV-2
Lau et al.	2020	Origine zoonotique : chauve-souris fer à cheval +pangolins
Lundstrom et al.	2020	Théorie d'adaptation du virus lors de la transmission humain-humain
Wong et al.	2020	Origine zoonotique (chauve-souris)
Huang et al.	2020	Pangolin, hôte intermédiaire

Hassanin et al.	2021	Origine liée au commerce illégal de mammifères sauvages
Makarenkov et al.	2021	Origine zoonotique, pangolins hôtes intermédiaires du SARS-CoV-2
Pei et Yau.	2021	Théorie sur l'origine géographique du Covid 19 : France, Inde, Pays-Bas, Royaume-Uni et É.-U. et non pas la Chine

CHAPITRE 2
CADRE THÉORIQUE DE L'ÉTUDE

2.1 ADN et alignement des séquences

2.1.1 Notions de base

Typiquement, un virus est constitué d'un noyau de matériel génétique et d'une enveloppe extérieure pour le protéger. Le support biochimique de l'information génétique chez les virus peut être sous forme d'acide désoxyribonucléique (ADN), ou sous forme d'acide ribonucléique (ARN). Les virus à ADN ont un génome à ADN double ou à simple brin, tandis que les virus à ARN sont toujours monocaténaires, c'est-à-dire à simple brin. Chaque brin est constitué de quatre éléments de base appelés nucléotides, à savoir l'adénine (A), la cytosine (C), la guanine (G) et la thymine (T), cette dernière étant remplacée par l'uracile (U) dans l'ARN.

L'ARN de certains virus peut être à sens positif, ce qui signifie qu'il peut servir de modèle direct pour la synthèse des protéines, ou à sens négatif, ce qui signifie qu'il doit être converti en ARN à sens positif avant de pouvoir être utilisé comme modèle. Le type de matériel génétique présent dans un virus peut influencer sa biologie, notamment sa stratégie de réplication et sa gamme d'hôtes.

2.1.1.1 Le génome du SARS-COV-2

Tout comme les autres CoV, le génome du SARS-CoV-2 est un ARN monocaténaire à un sens positif contenant un peu moins de 30 kb. Il comprend deux régions non codantes en 5′ et en 3′. La partie codante comprend 14 cadres de lectures ouverts (Open Reading Frames ORFs) encodant :

Des protéines non structurales (NSP) intervenant dans la réplication du virus et dans les processus d'assemblage.

Des protéines structurales, dont la protéine Spike S, l'enveloppe E, la protéine de membrane M, la nucléocapside N. à côté d'autres protéines accessoires.

Les deux premières Orfs contiennent approximativement 65 % du génome viral et se traduisent soit en polyprotéine pp1a (ORF1a) ou en pp1ab (ORF1b) qui codent le complexe de réplication-transcription dont le gène RNA-dépendant RNA Polymerse (RdRp) qui chiffre l'ARN polymérase ARN-dépendante jouant un

rôle crucial dans la réplication virale. La dernière ORF codifie les protéines de structures (S, E, M, N) ainsi que des protéines accessoires (Lefeuvre et al., 2020 ; Mohamadian et al., 2021).

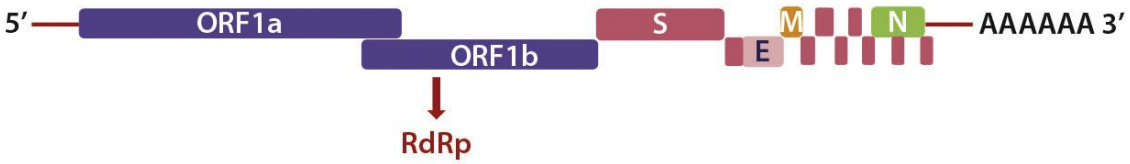

Figure 2.1 Organisation génomique du Sars-CoV-2. ORF : open reading frame; RdRp : gène codant l'ARN polymérase ARN-dépendante ; S, E, M, N : gènes codant les protéines de structure (S [surface], E [enveloppe], M [membrane], N [nucléoprotéine]))

Source Lefeuvre et al. (2020)

2.1.1.1.1 Séquençage du matériel génétique du SARS-COV-2

Après son apparition pour la première fois en décembre 2019, moins de deux semaines ont été nécessaires pour que la première séquence de référence du SARS-CoV-2 soit publiée le 12 janvier 2020. Aussi évidente et anodine que cette information puisse sembler, il est important de souligner qu'autre fois, il aurait fallu des mois pour obtenir une information de cette envergure. Les outils en matière de bioinformatique dont nous disposons aujourd'hui sont le résultat de travaux initiés durant les années 1980-1990 et qui continuent à ce jour à être développés.

2.1.1.1.2 Isolement des lectures d'intérêt

Avant de parler du processus d'identification des gènes et de l'alignement des séquences, nous allons décrire brièvement l'ensemble des étapes qui permettent d'obtenir à partir d'un simple écouvillon comportant un amas de microorganismes extrêmement variés, le génome du virus qui nous intéresse, en l'occurrence le SARS-CoV-2. En effet, il s'agit du premier obstacle rencontré lors de l'acquisition des données de séquençage ; l'échantillon clinique est en réalité un mélange de micro-organismes et de cellules humaines environnantes. La première étape consistera donc à effectuer un tri. Cette étape est possible grâce à des moteurs de recherche génomique, tels que Blast, qui permettent de traiter et de comparer de grandes quantités de séquences en s'appuyant sur des bases de données génomiques qui

comptent l'ensemble des microorganismes connus (virus, bactéries, champignons et parasites). C'est à partir de ce processus que l'on peut enfin identifier ce qu'on appelle les « lectures d'intérêt ».

2.1.1.1.3 Assemblage du génome

Après l'identification des lectures d'intérêt, l'étape suivante consiste à reconstruire la séquence du génome à partir de l'ensemble des lectures préalablement obtenues. Seulement, la technologie utilisée ne permet pas d'obtenir des séquences d'ADN ou d'ARN de bout en bout, elles se limitent à produire de courts fragments d'environ 200 nucléotides. Ces lectures sont alors en réalité un ensemble de courtes séquences nucléiques qui codent de manière aléatoire le génome initial. Grâce aux progrès de la bioinformatique, il existe aujourd'hui des logiciels qui permettent de facilement réaliser l'assemblage des petits génomes, comme ceux des particules virales par exemple. Les méthodes appliquées actuellement se basent sur les graphes de Bruijn. À la fin, l'assemblage du génome permet de reconstruire en totalité la séquence génomique.

L'étape de la détermination de la séquence génomique, bien qu'elle soit obligatoire et inévitable, reste peu informative et devra être complétée par une analyse du génome qui permettra à son tour d'ouvrir la voie aux hypothèses concernant son origine, mais aussi d'identifier les protéines codées régissant le fonctionnement du virus (Lemaitre et al., 2022 ; Touzet et al., 2022).

2.1.1.1.4 Alignement des séquences

L'analyse de séquence est un terme qui représente de manière exhaustive l'analyse computationnelle d'une séquence d'ADN, d'ARN ou de peptide, afin d'extraire des connaissances sur ses propriétés, sa fonction biologique et son évolution. Le séquençage de l'ADN est une technique qui consiste à déterminer l'ordre des bases nucléotidiques des molécules d'ADN. Si le séquençage s'intéresse à l'ADN du génome de l'organisme en question, on parle alors de séquençage génomique.

La détermination du génome d'un virus est une étape fondamentale dans l'étude de la maladie qui lui est reliée. Ordinairement, un séquençage génétique est nécessaire pour identifier l'agent causal d'une infection. Il s'agit alors de définir la séquence des nucléotides composant son génome. Une fois la séquence génomique établie, le séquençage à haut débit jumelé à la bioinformatique permet de comparer la séquence obtenue aux bases de données virales disponibles, c'est le processus d'alignement des séquences.

Par définition, l'alignement de séquences est le processus de comparaison et de détection des similitudes entre des séquences biologiques (Prjibelski et al., 2019). Les « similitudes » détectées dépendent des objectifs du processus d'alignement particulier, mais aussi de la base de données utilisée pour effectuer les comparaisons. La façon la plus simple de comparer deux séquences de même longueur est de calculer le nombre de symboles correspondants. La valeur qui mesure le degré de similarité des séquences est appelée le score d'alignement de deux séquences. La valeur opposée, correspondant au degré de dissemblance entre les séquences, est généralement appelée la distance entre les séquences.

Il convient toutefois de noter que la comparaison des caractères d'une séquence, position par position, telle que décrite ci-dessus, peut difficilement être qualifiée de processus d'alignement, puisqu'elle ne permet pas d'obtenir des résultats précis d'alignement. Cette méthode ne tient pas compte d'événements biologiques typiques tels que les délétions et les insertions. La notion classique de l'alignement de séquences inclut le calcul de la distance dite d'édition, qui correspond généralement au nombre minimal de substitutions, d'insertions et de suppressions nécessaires pour transformer une séquence en une autre (Prjibelski et al., 2019).

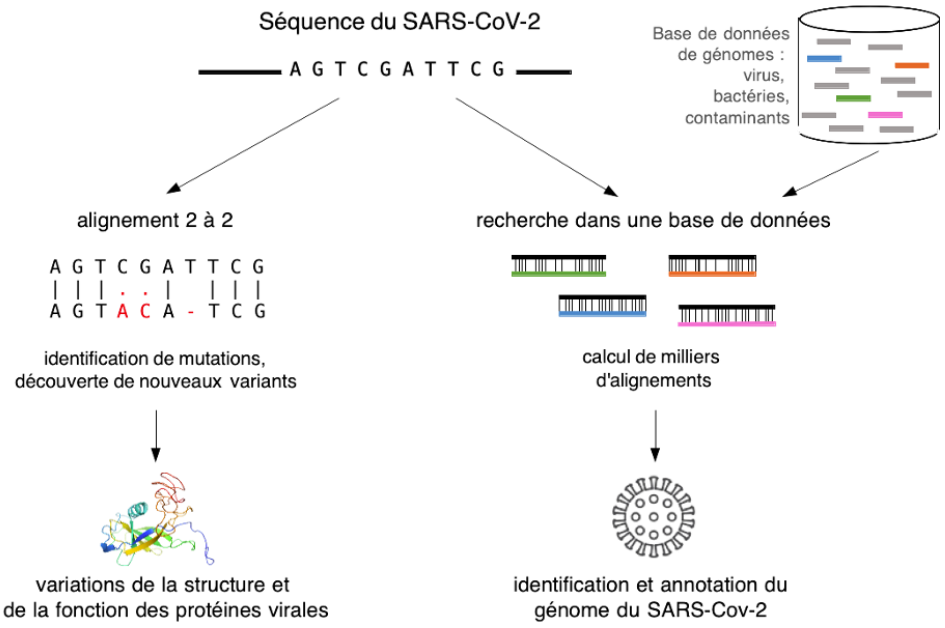

Figure 2.2 Alignement de séquences biologiques

Source : Lemaitre et el » (2020)

C'est ainsi que peu après l'identification de son génome, l'alignement des séquences du SARS-CoV-2 a rapidement pu révéler qu'il s'agit d'un bêta-coronavirus identique à 85 % à plusieurs CoV chiroptériques,

bien qu'il s'éloigne des autres bêta-coronavirus humains déjà découverts. L'analyse des séquences a montré que la SARS-CoV-2 présente 79 % de similitude avec le SARS-CoV, à l'origine de l'épidémie de SRAS apparue en Asie en 2003, et 50 % de similitude avec le MERS-CoV responsable du syndrome respiratoire au Moyen-Orient. Ces données sont fondamentales dans l'étude de l'origine du virus et ses caractéristiques et ont permis de comprendre le fonctionnement du virus peu après son apparition. (Fung et al., 2020)

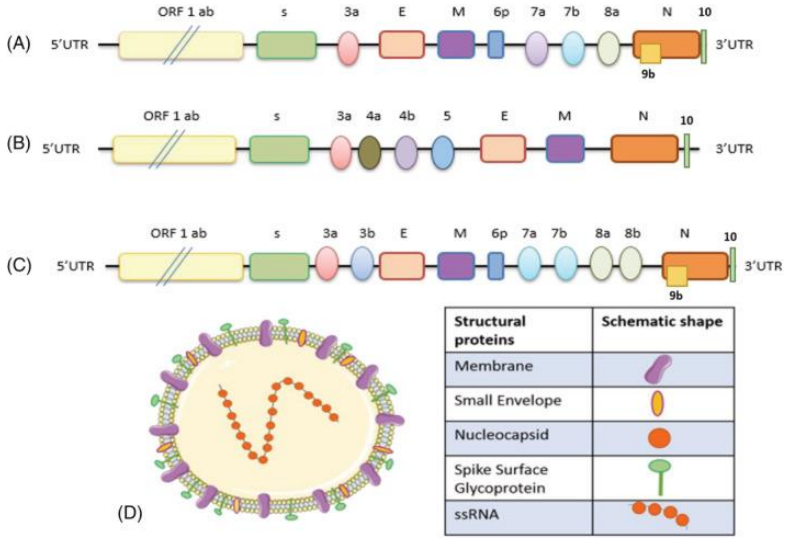

Figure 2.3 Schématisation de la structure génomique du coronavirus. (A) Coronavirus humain. (B) MERS-CoV. (C) SARS-CoV. (D) Coronavirus codant pour des protéines structurelles quatre gènes structurels, y compris les gènes de la pointe Spike, de l'enveloppe, de la membrane et de la nucléocapside, ainsi que des protéines accessoires (3 a, 3 b, 6, 7a, 7 b, 8 b, 9 b et ORF)

Source : Mohamadian et al. (2021)

2.2 Phylogénie et arbres

Les efforts engagés à découvrir le réel réservoir du coronavirus humain ont jusqu'ici été compromis. Aucun consensus n'est disponible sur l'origine et le mode d'évolution. Toutefois, nous disposons aujourd'hui d'assez de recul sur les précédentes pandémies causées par la même famille de virus, sur la base desquelles les études phylogéniques peuvent s'appuyer pour enfin élucider la réelle origine du virus responsable de l'une des plus grandes pandémies de ce millénaire.

2.2.1 Introduction à la phylogénie

Les phylogénies ont été historiquement utilisées pour classer les organismes dans des groupes évolutifs naturels basés sur les relations ancêtre-descendant. L'un des arbres phylogéniques les plus célèbres est « l'arbre de vie » dessiné par Charles Darwin comme seule illustration de son livre « L'origine des espèces » (Charleston, 2013) ; ouvrage duquel s'inspire Ernst Haeckel pour publier en 1866 un arbre unique rassemblant l'ensemble des organismes vivants connus de l'époque : l'arbre de Haeckel, premier prototype de l'arbre phylogénétique tel que nous le connaissons aujourd'hui (Lecointre et al., 2016).

Par définition, la phylogénie est un modèle de relation évolutive historique entre les espèces et les taxons de niveau supérieur qui est souvent représenté par un diagramme d'arbre, appelé arbre phylogénétique. Cette étude permet de restituer des relations de parenté entre les espèces et l'évolution de leurs traits dans le temps (Grandcolas et al., 2013). Le but ultime de la phylogénétique est en effet d'établir la parenté évolutive et les relations entre les gènes, les traits, ou les organismes. Servant une nécessité théorique qui découle des fondements de la biologie, qui voudrait que les similarités observées entre les êtres vivants remontent à des liens de parenté communs entre les espèces. La biologie évolutive utilise alors ces similarités ou caractéristiques communes entre différents organismes afin de reconstruire des arbres phylogénétiques dans le but d'élucider le lien de parenté entre les caractéristiques étudiées.

À ses débuts, on parlait de phylogénie morphologique étant donné que les caractères exploités pour les comparaisons étaient à caractère physique, c'est-à-dire visible comme les ossatures d'animaux par exemple. Aujourd'hui, on parle de phylogénie moléculaire qui exploite les séquences d'ADN et de protéines comme caractères de comparaison révolutionnant ainsi l'approche méthodologique de l'analyse phylogénétique. En se basant sur les molécules d'ADN dans son approche, la phylogénie moléculaire se base en fait sur une constante universelle : les bases nucléotidiques composant les molécules d'ADN.

Vraisemblablement, pour tout être vivant, la séquence ADN est constituée des mêmes bases nucléotidiques, le code génétique constituant les protéines est donc le même pour la quasi-totalité des organismes. Ce qui fournit une base de données tout à fait unique pour les analyses phylogéniques (Bonnin et lombard, 2019).

2.2.2 L'analyse phylogénétique

2.2.2.1 Sélection des homologues

La première étape de l'analyse phylogénique vise à déterminer un ancêtre commun dont la séquence génétique peut être exploitée pour la comparaison avec l'organisme en question ; c'est la sélection des homologues. Par définition, deux caractères sont dits homologues s'ils ont hérité d'un dernier ancêtre commun. En ce qui concerne les caractères moléculaires, l'homologie est déterminée de manière statistique étant donné la taille des séquences, le nombre des variantes possibles, et l'ampleur des bases de données des séquences ADN et des protéines. Faisant que la recherche de séquence dans le cadre de l'homologie ne puisse se faire manuellement. Des méthodes plus heuristiques sont aujourd'hui disponibles et permettent d'explorer les « meilleures possibilités » dans les meilleurs délais. De nos jours, les outils les plus utilisés à cet effet sont le Basic local Alignement Search Tool (BLAST) et ses dérivés. Cet outil permet de calculer un score de similarité entre deux séquences supposées homologues. En plus de calculer le nombre de sites identiques entre les deux séquences, le score de similarité du BLAST permet également d'évaluer la présence d'éventuelles mutations. (Bonnin & Lombard, 2019)

2.2.2.2 Alignement des séquences

La seconde étape de l'analyse phylogénétique consiste à aligner les séquences homologues préalablement identifiées. La première phase de l'alignement consistera principalement à comparer les séquences homologues entre elles en déterminant l'homologie entre chacun des nucléotides. La seconde phase dite de « découpage » vise à éliminer des analyses les sites des séquences contenant trop d'espaces ou trop de variabilités, pour à la fin ne garder que les sites susceptibles de comporter des informations phylogénétiques qui pourraient être exploitées. Le découpage des séquences est souvent obtenu à l'aide de logiciels spécialisés qui utilisent des critères plus ou moins complexes afin d'exclure de manière automatique les sites dont l'homologie est incertaine. À l'issue de cette étape, on obtient une base comparative sur laquelle se construira l'arbre phylogénétique (Bonnin & Lombard, 2019 ; Prjibelski et al., 2019).

2.2.2.3 Construction des arbres

Le but de cette étape est de représenter les relations évolutives entre les séquences sous forme d'un arbre phylogénétique. Pour ce faire, il existe deux types de méthodes :

Les méthodes basées sur les caractères des séquences : qui sont au nombre de trois

La méthode de maximum de parcimonie : calcule le nombre minimal de changements de nucléotides ou d'acides aminés nécessaires pour expliquer les données en utilisant chaque topologie d'arbre possible. La topologie d'arbre présentant le plus petit nombre d'évènements évolutifs est appelée l'arbre le plus parcimonieux et constitue l'estimation de la phylogénie de l'espèce. C'est une méthode rarement utilisée en phylogénie moléculaire en vue de la difficulté de distinction entre les différents arbres également parcimonieux.

La méthode de maximum de vraisemblance : L'application des méthodes de vraisemblance à la reconstruction phylogénétique est devenue de plus en plus populaire au cours de la dernière décennie, en grande partie en raison de leur plus grande précision et cohérence dans la récupération d'une phylogénie correcte, et à l'augmentation significative de la capacité et de la vitesse de calcul. Le maximum de vraisemblance tente d'identifier la topologie de l'arbre avec la plus grande « vraisemblance » compte tenu des données de séquence fournies (Stavrinides & Ochman, 2009).

La méthode bayésienne : L'approche bayésienne de la reconstruction phylogénétique est considérée par beaucoup comme une alternative idéale à l'approche par la méthode de vraisemblance. Cette méthode s'appuie sur le théorème de Bayes, qui prend en compte les données de séquences fournies (vraisemblances) et les hypothèses que l'on peut formuler sur les données pour calculer une probabilité postérieure (PP) pour une topologie donnée. Les analyses bayésiennes utilisent un algorithme connu sous le nom de Markov Chain Monte-Carlo (MCMC), qui explore et échantillonne « l'espace des arbres » (l'ensemble complet des arbres possibles pour un ensemble de données défini) et estime le PP de chacun d'entre eux. Contrairement aux approches basées sur la distance, l'approche bayésienne ne produit pas un arbre unique, mais échantillonne plutôt une série de topologies d'arbres probables en fonction de l'ensemble de données et des séquences fournies (Bonnin & Lombard, 2019 ; Stavrinides & Ochman, 2009).

Les méthodes basées sur la distance évolutive entre les séquences : ici il s'agit de calculer la distance génétique entre chaque paire d'espèces sur la base de l'alignement des séquences. L'arbre phylogénétique est ensuite construit de manière itérative à l'aide de la matrice de distances. Ce type de méthodes n'est cependant pas l'idéal lorsqu'il s'agit d'espèces éloignées, car plus la distance est importante, plus il devient difficile de l'estimer (Makarenkov & Leclerc, 1996; Leclerc & Makarenkov, 1998; Bonnin & Lombard, 2019 ; Kapli et al., 2020).

L'arbre phylogénétique obtenu est constitué de nœuds et de branches. Les nœuds représentent les ancêtres communs qui, dans ce cas, sont hypothétiques, reconstruits à partir des informations disponibles sur les espèces en amont et en aval du nœud. Les branches représentent des lignées évolutives et retracent l'histoire des unités taxonomiques depuis leurs ancêtres communs jusqu'aux espèces actuelles. La quantité de changement que l'on estime s'être produit entre l'ancêtre commun et l'espèce actuelle est codifiée par la longueur de la branche (Bonnin & Lombard, 2019).

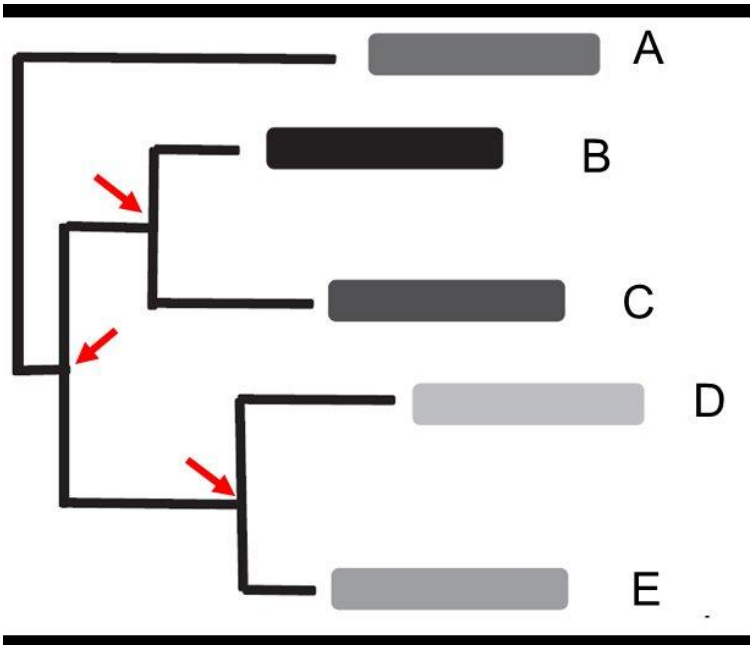

Figure 2.4 Exemple schématique d'arbre phylogénétique.

Les flèches indiquent les nœuds représentant d'hypothétiques ancêtres communs entre les séquences comparées
Source : Bonnin & Lonbard. (2019)

2.2.2.4 L'analyse de l'arbre phylogénétique

L'analyse phylogénétique emploie une combinaison d'approches moléculaires et de statistiques pour déduire ou estimer les relations entre les espèces. Elle fournit une méthode crédible pour explorer la relation entre la similarité des séquences et la fonction des protéines appartenant à la même famille (Song et al., 2018). Le résultat final de l'analyse phylogénétique est validé en comparant les résultats obtenus avec d'autres analyses où d'autres méthodes d'alignement, de découpage et de reconstruction d'arbres sont appliquées sur les mêmes données de base ou encore en employant d'autres données (Bonnin & Lombard, 2019 ; Song et al., 2018).

2.3 Transfert horizontal de gènes

Le transfert horizontal de gènes (THG) a été décrit pour la première fois en 1928 (Keese, 2008), et ce n'est que deux décennies plus tard que son rôle dans les processus d'adaptation des eucaryotes a été élucidé. Cette découverte a depuis ouvert de nouveaux horizons à l'étude de la variabilité génétique des virus, des procaryotes et des eucaryotes où on a commencé à accorder de plus en plus d'intérêt au développement des méthodes de détection du HGT. Depuis, il s'est avéré que bon nombre de duplications génétiques apparentes étaient en réalité le résultat d'un THG et non d'une duplication génique autochtone. Ces découvertes ont fait qu'on ne parle plus d'arbre de vie en constante bifurcation, mais d'une toile de la vie où il y a tout autant de transfert génétique horizontal que vertical, si ce n'est plus (Soucy et al., 2015).

Contrairement au transfert vertical des gènes, le THG est une source importante de variation génétique où les gènes sont transmis entre les espèces et non pas du parent à la progéniture. C'est un processus qui permet à une espèce d'acquérir le matériel génétique d'une autre espèce afin de pouvoir mieux s'adapter à son environnement, comme c'est souvent le cas chez les procaryotes (Acar Kirit et al., 2022 ; Doolittle, 1999 ; Ochman et al., 2000). La faculté du partage de matériel génétique qu'offrent les THG aux espèces a de nombreuses applications potentielles dans l'industrie de l'agriculture et la production chimique (Keese, 2008), mais le THG n'est certainement pas sans danger pour l'humain. En effet, ce genre d'échanges peut être à l'origine de nouveaux gènes résistants aux antibiotiques, de nouveaux virus et bactéries résultant de recombinaison de deux ou plusieurs gènes pathologiques et même d'insertion d'ADN transgénique dans les cellules humaines se traduisant par des cancers (Ho, 2002). C'est dire qu'un ADN transgénique résultant d'un THG est plus différent, plus dangereux et plus imprévisible qu'un ADN ordinaire.

2.3.1 Les mécanismes de transfert horizontal de gènes

Il existe principalement trois mécanismes de THG chez les procaryotes

2.3.1.1 La conjugaison

Il s'agit d'un mécanisme de transmission d'ADN à sens unique, où le matériel génétique est transmis par contact physique entre le donneur et le récepteur via le pilus de conjugaison. La conjugaison est largement retrouvée chez les procaryotes, c'est en effet le mécanisme le plus important en matière de transfert d'ADN chez les bactéries. Dans la nature, la conjugaison se limite aux souches d'une même espèce, mais il n'est pas rare qu'il se produise entre deux espèces différentes comme les archées et les bactéries. (Daubin & Szöllősi, 2016 ; Keese, 2008 ; Soucy et al., 2015)

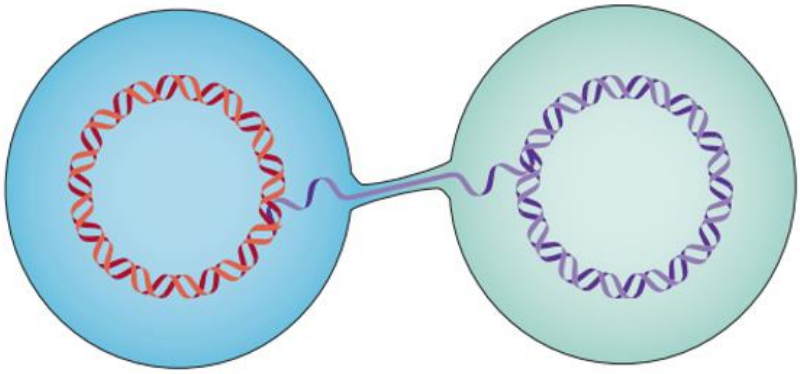

Figure 2.5 Schématisation du mécanisme de conjugaison

Source : Soucy et al. (2015)

2.3.1.2 La transformation

La transformation est un mécanisme actif par lequel l'ADN libre présent dans le milieu, généralement dérivé d'organismes morts, est absorbé dans le cytoplasme de la cellule. Ce mécanisme se produit principalement à des fins nutritionnelles, mais certaines bactéries sont très sélectives quant au type d'ADN qu'elles laissent entrer dans la cellule, ce qui suggère que cela sert également à favoriser la recombinaison génétique. (Daubin & Szöllősi, 2016) C'est un mécanisme souvent rencontré chez les bactéries et les archées. (Soucy et al., 2015).

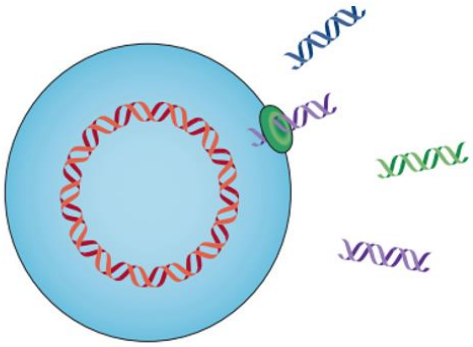

Figure 2.6 Schématisation du mécanisme de transformation

Source : Soucy et al. (2015)

2.3.1.3 La transduction

La transduction est un type de transfert de matériel génétique qui se produit via un bactériophage (virus attaquant les bactéries) qui transmet l'ADN d'une cellule à une autre. À la fin de son cycle de réplication, la cellule hôte subit une lyse et l'ADN fragmenté du génome de l'hôte est parfois emballé à l'intérieur de

particules infectieuses. Cet ADN peut alors être injecté à un autre individu, à la place de l'ADN du virus. Certaines espèces de bactéries ont détourné ce mécanisme à leur avantage et ont recruté des gènes de bactériophages pour faciliter les échanges génétiques. Ces capsides de phages défectueuses, présentes notamment chez de nombreuses α-protéobactéries, sont appelées « agents de transfert de gènes (ATG) » (Daubin & Szöllősi, 2016 ; Soucy et al., 2015).

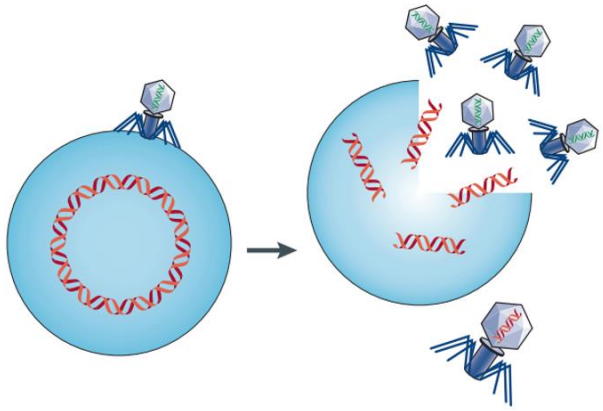

Figure 2.7 Schématisation du mécanisme de transduction

Source : Soucy et al. (2015)

Une fois le transfert du matériel génétique accompli, l'ADN se trouvant alors au niveau du cytoplasme du récepteur est soit :

– Détruit par les systèmes de dégradation d'ADN présents dans le cytoplasme de la cellule hôte (enzymes de restriction, ADNases… etc.)

– Préservé en tant qu'entité de réplication autonome, comme les plasmides

– Intégré en partie ou en totalité dans le chromosome hôte. Cette intégration va dépendre de plusieurs paramètres comme le degré de similitude entre les deux ADN dans la recombinaison homologue, ou l'association physique avec d'autres séquences comme les gènes de bactériophages. Lorsque la recombinaison homologue se produit, la nouvelle séquence d'ADN remplace les séquences homologues existantes dans le génome de l'hôte — c'est ce qui se passe dans le cas du pneumocoque. A l'opposé, lorsque l'ADN est intégré au génome par d'autres moyens, il est souvent simplement inséré comme un gène entièrement nouveau (Daubin & Szöllősi, 2016).

2.3.2 Détection du transfert horizontal de gènes

À l'heure actuelle, deux méthodes sont admises pour la détection du THG :
La méthode phylogénétique : elle prend un ensemble relativement important de séquences codantes homologues (provenant d'un ancêtre commun), construit leur phylogénie correspondante et la compare à la phylogénie de leur espèce d'origine. Cette comparaison se fait généralement à l'aide de la distance de Robinson et Foulds (Makarenkov & Leclerc, 2000). Lorsque des incongruences sont retrouvées entre les deux arbres, elles sont expliquées par l'introduction des HGT. Si cette approche présente l'avantage d'identifier des événements relativement anciens, elle repose sur une hypothèse très stricte quant à l'endroit où rechercher ces événements. Enfin, elle nécessite également un alignement multiple des séquences et l'inférence d'un arbre des espèces fiable, deux problèmes majeurs en soi (Sevillya et al., 2020).

La méthode basée sur la composition nucléotidique : Le THG déduit de la composition anormale des nucléotides est basé sur le principe que les génomes évoluent vers des valeurs spécifiques aux espèces en raison des biais de réplication, de transcription et de traduction, des variations dans les groupes de nucléotides et d'acides aminés et des préférences de réparation de l'ADN. Ainsi, si l'utilisation des codons, le contenu en G-C ou les signatures oligonucléotidiques diffèrent significativement de la moyenne pour un génome donné, alors le THG peut être évoqué. Selon ces critères, près de 20 % de génomes de procaryotes ont récemment été classés comme résultants de THG. L'avantage de la détection des THG par la composition anormale des nucléotides réside dans le fait qu'elle ne nécessite qu'un seul génome à examiner et qu'elle est très rapide à évaluer. Cependant, elle détecte généralement des changements plus récents et exige que le génome de l'organisme donneur soit distinct de celui du récepteur (Keese, 2008). Cette approche souffre du fait que les espèces concernées peuvent partager des modèles de composition similaires. De plus, la longueur d'un segment transféré peut être trop courte pour révéler de manière fiable ces différences. Comme le concluent Lawrence & Ochman (2002), « le contenu G+C atypique et le modèle d'utilisation des codons ne sont pas des indicateurs fiables des événements de transfert horizontal de gènes » (Sevillya et al., 2020).

De manière générale, toutes les méthodes de détection de THG mettent en évidence des ensembles différents de gènes et peuvent passer à côté de véritables événements de THG (faux négatifs) et identifier d'autres comme étant de vrais THG (faux positifs). Par conséquent, une combinaison d'approches peut s'avérer nécessaire pour identifier et confirmer un THG (Keese, 2008).

2.3.3 Impact du transfert horizontal de gènes sur une phylogénie

On peut observer de nombreuses incongruences entre des phylogénies de gènes et d'espèces. La figure 2.8 schématise une phylogénie d'espèce et deux autres phylogénies pour les gènes 1 et 2 étudiés séparément. Bien qu'étant déduite du même arbre d'espèces, la présence de deux transferts (entre C et A et entre B et F) a fait que les arbres de gènes soient complètement incongruents. Ce THG a fait que les deux gènes ne partagent plus de nœud interne commun et par conséquent aucune histoire d'évolution commune. C'est ainsi que le THG peut altérer la topologie d'un arbre phylogénétique (Philippe & Douady, 2003).

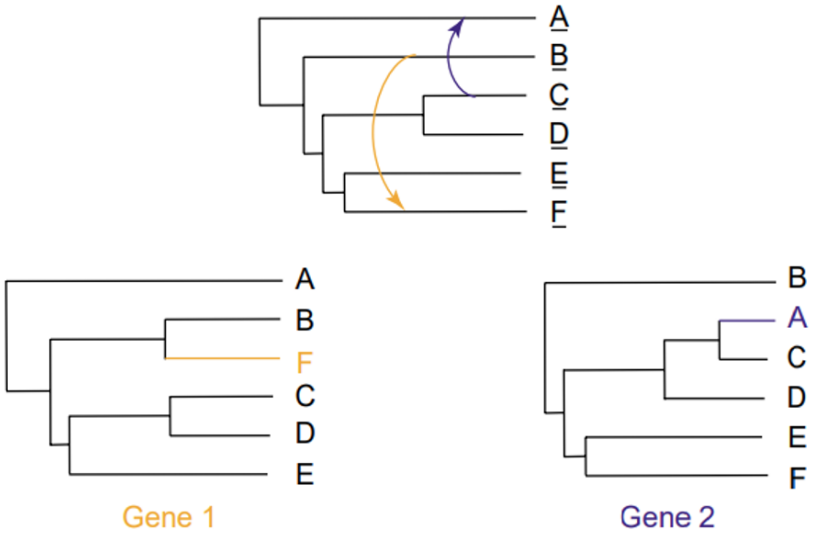

Figure 2.8 Incongruence phylogénétique générée par le THG

Source : Philippe & Douady. (2003)

2.4 La recombinaison Génétique

Si la question autour de l'origine du SARS-CoV-2 a fait l'objet d'une polémique et de nombreuses controverses au sein de la population scientifique, la question sur son évolution n'en a certainement pas fait moins. Étant un virus à ARN, on savait d'emblée qu'il s'agissait d'une entité qui peut être sujette à de nombreuses mutations et recombinaisons génétiques. Très peu après sa découverte, la communauté scientifique appréhendait déjà l'évolution future du SARS-CoV-2. La théorie de recombinaison génétique suscitait d'autant plus d'intérêt qu'on savait que cela pouvait bousculer tout ce que nous savions déjà sur ce virus, étiquetant ainsi tous les vaccins développés d'une date de péremption certaine (Amoutzias et al., 2022).

En biologie moléculaire, la recombinaison est un échange d'information génétique entre deux brins d'acide nucléique, permettant de créer de nouvelles combinaisons génétiques (Stapley et al., 2017). Il s'agit d'un phénomène d'une importance fondamentale dans la diversité moléculaire. Les données actuelles permettent d'associer la recombinaison à l'émergence de nouveaux organismes, l'augmentation de l'adaptabilité des virus, l'intensification de la virulence microbienne ou encore la suppression de mutations dangereuses (Del Amparo et al., 2023).

2.4.1 La recombinaison génétique chez les virus

La recombinaison génétique chez les virus leur permet de rapidement s'adapter aux nouvelles situations environnementales. Mais si le rôle de la recombinaison dans la diversité génétique est connu depuis près de 100 ans, la capacité des virus à ARN à échanger du matériel génétique par recombinaison n'a été découverte, quant à elle, que bien plus tard. La recombinaison est depuis reconnue comme un mécanisme évolutif qui permet de promouvoir la diversité virale par la formation de nouveaux génomes chimériques. Ce processus et ses conséquences, par exemple la génération de virus présentant de nouveaux phénotypes, ont historiquement été étudiés par l'analyse des produits finaux. Plus récemment, avec la prise de conscience de l'existence de mécanismes réplicatifs et non-réplicatifs, et avec de nouvelles approches et techniques d'analyse des produits intermédiaires, les facteurs viraux et cellulaires qui influencent le processus commencent à être compris.

Cependant, il est clair que la recombinaison est un processus propre aux virus à ARN positif et qu'elle est très rarement observée chez les virus à ARN négatif. Cela s'explique par le fait que l'ARN viral négatif est

présent sous forme de ribonucléoprotéines, c'est-à-dire que le génome n'est jamais à l'état nu dans la cellule, créant ainsi des conditions défavorables à sa recombinaison (Bouloy, 2002).

La première preuve expérimentale de la recombinaison des virus à ARN positif a été apportée par des études sur le poliovirus en 1962, mais il a depuis été démontré qu'elle se produit dans un large éventail de virus à ARN positif, affectant tous les types d'organismes ; des humains et des animaux aux plantes et aux bactéries. Il semble probable que la recombinaison soit omniprésente dans les virus à ARN à brin positif (Bentley & Evans, 2018 ; White et al., 2011).

2.4.2 Types de recombinaison virale

La recombinaison se produit lorsque deux ou plusieurs virus co-infectent une même cellule et échangent des segments, donnant souvent naissance à une descendance hybride viable. Il existe de nombreux mécanismes de recombinaisons, dont certains sont uniques aux virus à ARN. Selon la structure de la molécule de l'ARN viral et les sites au niveau desquels la recombinaison a lieu, la recombinaison des virus à ARN peut être classifiée en trois types :

1. Recombinaison homologue : Elle se produit entre deux molécules d'ARN parentales dont les séquences sont similaires. Le site de la recombinaison est le même pour les deux séquences. L'ARN recombiné qui en résulte à la même séquence et la même structure que la molécule d'ARN parent.
2. Recombinaison illégitime : La recombinaison illégitime nécessite également deux molécules d'ARN parentales ayant des séquences similaires. Cependant, dans ce cas, l'événement de recombinaison a lieu dans une région non apparentée proche de la séquence homologue. La recombinaison illégitime peut donner lieu à des molécules d'ARN comportant des insertions, des délétions ou des duplications et, de ce fait, elle produit souvent des génomes viraux défectueux.
3. Recombinaison non homologue (site spécifique) : La recombinaison non homologue ne dépend pas de l'homologie de séquence entre les molécules d'ARN parentales recombinées, et la sélection des sites de recombinaison reste à ce jour mal élucidée (Lai, 1992 ; Wang et al., 2022).

2.4.3 La recombinaison génétique chez les coronavirus

Les CoV représentent l'une des plus importantes sous-familles de virus à ARN positif. Et tous comme beaucoup de virus à ARN, ils ont une grande prédisposition aux recombinaisons. En effet, bien que les CoV aient un taux de mutation relativement faible par rapport aux autres virus à ARN, ils présentent tout de même un taux de recombinaison relativement élevé. L'exonucléase NSP s14 est responsable de la

relecture et, par conséquent, de la fidélité de la réplication du génome. Cependant, cette enzyme est également le médiateur de la recombinaison, tandis que son inactivation diminue la fréquence et modifie les modèles de recombinaison.

De plus, la transcription des CoVs implique des ARNm sous génomique qui sont formés par un changement de gabarit entre les séquences régulatrices de la transcription (SRT) des différentes Orfs (désignée SRT-B : B pour body.) et la SRT au 5′ UTR (désignée SRT-L : L pour leader). Par conséquent, les CoVs sont intrinsèquement enclins à la recombinaison du fait de ce mécanisme de transcription caractéristique. De nombreuses études ont d'ailleurs mis en évidence le rôle crucial de la recombinaison homologue dans l'évolution des CoV, et elles ont été largement documentées avant même la pandémie de COVID-19 (Amoutzias et al., 2022).

Des événements de recombinaison entre les génomes du SARS-CoV-2 actuellement en circulation continuent à se produire. Cependant, de tels événements ont été difficiles à détecter au début de la pandémie en raison du haut niveau de similitude des séquences entre les génomes en circulation. Mais à mesure que la pandémie progressait et que des souches plus divergentes de SARS-CoV-2 circulaient, il était devenu plus facile de détecter les événements de recombinaison intra-SARS-CoV-2.

Une analyse de 1,6 million de génomes du SARS-CoV-2 a montré que 2,7 % des séquences circulantes appartiennent à une lignée recombinante et que les points de rupture de la recombinaison se situent de manière disproportionnée dans la région Spike (Turkahia et al., 2021). Au fur et à mesure que des mutations ponctuelles se produisent dans les divers variants du SARS-CoV-2, il est possible que la diversité de cette espèce augmente encore davantage par le biais d'un brassage recombinatoire intratypique. Il est donc possible que le SARS-CoV-2 se trouve encore dans la première phase d'une évolution plutôt lente, principalement dirigée par des mutations ponctuelles et des insertions/délétions. Cette phase pourrait être suivie d'une autre phase de divergence rapide, ce qui aurait des répercussions importantes sur l'efficacité des vaccins (Amoutzias et al., 2022).

2.5 Le temps et les virus

L'une des questions les plus importantes en biologie virale évolutive est de déterminer la vitesse à laquelle un virus peut évoluer et de comprendre la variation de cette évolution d'un virus à l'autre et son développement dans le temps. Bien que l'évolution soit un processus dépendant de plusieurs facteurs, la théorie voudrait que la vitesse de l'évolution moléculaire soit directement déterminée par le taux d'apparition des mutations spontanées (Sanjuán, 2012). Parallèlement, il a été établi, il y a déjà plus d'un demi-siècle, qu'il existerait une corrélation inverse entre le taux d'évolution morphologique et l'intervalle de temps sur lequel ce taux était mesuré. Autrement dit, les taux de mutation spontanée, mesurés sur un petit nombre de générations, semblaient plus rapides lorsqu'ils étaient mesurés sur des échelles de temps plus courtes (Ho et al., 2011). Les virus ont été dans ce cadre un excellent modèle d'étude, car leur taux de mutations varie par ordre de magnitude et leur évolution est compatible avec un cadre temporel humainement décelable (Sanjuán, 2012).

2.5.1 La phylodynamique virale

L'abondance des données en virologie a donné naissance à un nouveau champ disciplinaire appelé « la phylodynamique » qui associe la phylogénie et l'épidémiologie. Elle vise à exploiter les progrès en matière de séquençage génomique pour étudier et estimer des paramètres tels que le taux de croissance de la population virale, le nombre d'infections ou même leur durée moyenne (Alizon & Saulnier, 2018).

La phylodynamique virale est définie comme l'étude de la manière dont les processus épidémiologiques, immunologiques et évolutifs agissent et interagissent potentiellement pour aboutir à des phylogénies virales. Depuis l'invention de ce terme en 2004, la recherche sur la phylodynamique virale s'est concentrée sur la dynamique de transmission dans le but d'éclaircir l'impact de cette dynamique sur la variation génétique virale. La dynamique de transmission peut être considérée au niveau des cellules d'un hôte infecté, d'hôtes individuels au sein d'une population ou de populations entières d'hôtes (Volz et al., 2013).

Une connaissance du taux de l'évolution virale est capitale quant au processus de la reconstitution de l'histoire naturelle du virus. C'est un passage obligatoire pour calculer les différents paramètres de l'évolution en allant de l'âge viral à l'estimation de la taille de la population. De manière générale, les virus sont typiquement classés en deux sous-groupes :

– Les virus à évolution lente : C'est les virus à ADN, notamment ceux à doubles brins. Il a été démontré que de nombreux virus à ADNdb ont des liens extrêmement stables avec leurs hôtes et que, par conséquent, leurs dates de divergence peuvent être directement déduites de celles de leurs hôtes.

– Les virus à évolution rapide : C'est les virus à ARN qui sont souvent caractérisés par des transmissions interespèces. Ce qui rend souvent difficile de calibrer leur taux d'évolution à travers leurs hôtes.

Il ne s'agit cependant pas d'une règle générale. En effet, l'analyse d'ensemble de données moléculaires hétérochrones de virus à ADNdb et à ADN simple brin (ADNsb) a révélé que les virus à ADN sont en fait capables d'évoluer très rapidement sur de courtes échelles de temps, de façon comparable aux virus à ARN (Aiewsakun & Katzourakis, 2016).

2.5.2 Apport de l'horloge moléculaire à la phylogénie d'infections

C'est en 1960 que la dimension temporelle a été introduite pour la première fois dans la biologie évolutive, où la toute première hypothèse suggérait que le taux d'évolution d'une protéine donnée était constant dans le temps. La distance génétique, le nombre de substitutions entre séquences provenant d'espèces différentes, pourrait ainsi être convertie en temps de divergence entre les espèces (Makarenkov, 1997). Suggérant l'existence d'une « horloge moléculaire » cette hypothèse a fini par alimenter une nouvelle technique de datation moléculaire. Cette méthode permettrait d'estimer le taux de substitution et la date de chaque ancêtre commun le plus récent (MRCA pour Most Recent Common Ancestor) d'une phylogénie (Zuckerkandl & Pauling, 1965).

L'introduction du concept de l'horloge moléculaire en phylogénie a surtout impacté la phylogénie d'infections. En effet, contrairement aux phylogéniques d'espèces classiquement utilisées en biologie où l'âge des différents nœuds internes est exprimé en millions d'années. Les échelles de temps épidémiologique des phylogénies d'infections se comptent en années voire en mois notamment dans le cas de génomes microbiens tels que les virus à ARN. Cela crée une variation des taux d'évolution entre les lignées d'un arbre phylogénétique. Ignorées, ces variations peuvent conduire à des inférences incorrectes des taux et des dates d'évolution. Afin de pallier cela, un modèle d'horloge moléculaire dite « relâchée » basé sur les approches bayésiennes a été proposé, qui permettrait une variabilité des taux d'évolution d'une branche à l'autre dans le même arbre phylogénétique (Drummond et al., 2006).

Les modèles phylodynamiques peuvent aider à retracer les origines des épidémies et des pandémies dans le temps. Le taux d'évolution rapide des virus permet d'estimer des modèles d'horloge moléculaire à partir de séquences génétiques, fournissant ainsi un taux d'évolution du virus par année. Le taux d'évolution étant mesuré en unités de temps réels, il est possible de déduire la date du MRCA pour un ensemble de séquences virales. L'âge de l'MRCA de ces isolats est une limite inférieure ; l'ancêtre commun de l'ensemble de la population virale doit avoir existé avant l'MRCA de l'échantillon de virus. En avril 2009, l'analyse génétique de 11 séquences de grippe H1N1 d'origine porcine a suggéré que l'ancêtre commun existait au plus tard le 12 janvier 2009. Cette découverte a permis de faire une estimation précoce du nombre de reproductions de base R0 de la pandémie. De même, l'analyse génétique de séquences isolées chez un individu peut être utilisée pour déterminer le moment de l'infection de cet individu (Volz et al., 2013).

2.5.3 La phylodynamique du SARS-COV-2

L'épidémie actuelle du nouveau coronavirus (SRAS-CoV-2) marque une première en termes d'abondance de données génomiques dans un cadre pandémique. En effet, plus d'une cinquantaine de séquences étaient générées et publiquement partagées pas plus tard que la fin janvier 2020 (Duchene et al., 2020). Cela a permis de révéler que le SARS-CoV-2 présentait une très faible diversité chez les humains suggérant ainsi qu'il s'agisse d'un virus d'origine récente dans la population humaine, avec une introduction unique.

Semblablement aux autres virus à ARN, les coronavirus ont une capacité de mutation relativement rapide. Une première approche de datation a donc été d'utiliser l'accumulation régulière des mutations durant les infections comme une horloge moléculaire qui permettrait de dater les évènements épidémiologiques. Bien qu'elle soit classique, cette approche aurait permis, à l'aide des modèles phylogénétiques, d'estimer les caractéristiques du génome des ancêtres communs au virus. C'est ainsi que l'origine de l'épidémie a rapidement pu être cernée entre août et décembre 2019. Une approche plus moderne s'est basée sur l'étude phylodynamique a permis de quantifier l'évolution de l'épidémie (Elie & Alizon, 2020).

La phylodynamique permet ainsi de détailler certains aspects de la démographie virale, comme les taux d'évolution et de déclin des populations virales, la distribution des événements de ramification des arbres phylogénétiques, la structure des populations et la sélection immunitaire. Elle permet également d'estimer le nombre de reproductions de base (R0) directement à partir des données de séquences

génétiques, ce qui permet de faire le lien entre l'analyse évolutive des séquences génomiques et l'épidémiologie d'une maladie infectieuse (Cardona-Ospina et al., 2021).

Les données obtenues à partir des analyses phylodynamiques ont été cruciales dans l'étude de l'origine et l'évolution du SARS-CoV-2, mais elles restent malheureusement insuffisantes. La réalité derrière tous les chiffres est que les gouvernements et les secteurs de santé sont en perpétuel combat contre la montre. Car bien que les taux de mortalité du COVID19 sont relativement bas comparés à ces homologues de la famille des bêtacoronavirus, son taux de contagion est quant à lui sans précédent, ayant réussi à se propager à 185 pays en un temps record (Aboubakr et al., 2021).

CHAPITRE 3
Méthodologie

3.1 Description des données

Nous avons mené une étude de l'évolution de 25 coronavirus, en utilisant les mêmes données que celles utilisées par Makarenkov et al., 2021. Ces ensembles de données comprennent quatre génomes du SRAS-CoV-2 provenant de différentes régions géographiques, dont Wuhan (Chine), l'Italie, l'Australie et les États-Unis. Ces génomes ont été sélectionnés à partir de grappes dans l'arbre phylogénétique des coronavirus de Gisaid. De plus, deux génomes GD Pangolin CoV de pangolins malais qui sont morts lors d'opérations anti-contrebande dans la province chinoise du Guangdong, et cinq génomes Guangxi Pangolin CoV de l'Institut de microbiologie et d'épidémiologie de Pékin ont été inclus. Cette étude a également inclus le génome de chauve-souris CoV RaTG13 de la province du Yunnan en Chine, et le groupe Bat CoV Z de deux virus, dont les virus chauve-souris CoV ZC45 et ZXC21, collectés dans la province chinoise du Zhejiang en 2018 en 2010, ainsi que cinq chauves-souris CoV génomes échantillonnés entre 2006 et 2010 dans différentes provinces de Chine. Les génomes liés au SRAS-CoV des premières épidémies de SRAS, y compris le SRAS humain, Tor2, le SRAS-CoV BJ182-4 CoV et le CoV Rs3367 de chauve-souris trouvée dans Rhinolophus sinicus, ainsi que deux CoV différents de requins trouvés au Kenya et Bulgarie (BtKY72 et BM48 31 BGR 2008) ont été également examiné. La plupart de ces génomes de CoV ont été initialement étudiés par Lam et al., 2020. De plus, pour explorer plus avant l'analyse des événements putatifs de transfert de gènes-recombinaison affectant le domaine RB de la protéine de pointe et les événements de recombinaison intergénique (transferts de gènes complets) affectant tous les gènes étudiés, nous avons examiné les statistiques de 21 autres coronavirus, y compris les virus classés comme coronavirus de la grippe CoV Gisaid arbre phylogénétique et autres virus CoV étudiés par Prabakaran et al. Ces séquences sont disponibles dans la base de données GenBank. (Makarenkov et al., 2021).

Des informations plus détaillées sur la description des données utilisées dans ce travail, ainsi qu'un fichier téléchargeable des alignements de séquences multiples pour toutes les séquences de gènes et de génomes Cov en format Fasta est accessible dans l'article de Makarenkov et al., 2021.

Le tableau 3.1 présente les noms complets des organismes, les espèces hôtes et les numéros d'accès GenBank ou Gisaid pour tous les génomes Cov analysés dans cette étude.

Tableau 3.1 les noms complets des organismes, les espèces hôtes et les numéros d'accès GenBank ou Gisaid pour tous les génomes CoV analysés dans cette étude

Organism name	Abbreviation	Host	Accession number (GenBank, GISAID)
hCoV-19/Australia/VIC231/2020	Hu Australia VIC231 2020	Human	EPI_ISL_419926
hCoV-19/USA/UT-00346/2020	Hu USA UT 00346 20202	Human	EPI_ISL_420819
BetaCoV/Wuhan-Hu-1	Hu Wuhan 2020	Human	NC_045512.2
hCoV-19/Italy/TE4836/2020	Hu Italy TE4836 2020	Human	EPI_ISL_418260
Bat coronavirus RaTG13	RaTG13	Bat	MN996532.1
hCoV-19/pangolin/Guangdong/1/2019	Guangdong Pangolin 1 2019	Pangolin	EPI_ISL_410721
hCoV-19/pangolin/Guangdong/P2S/2019	Guangdong Pangolin P2S 2019	Pangolin	EPI_ISL_410544
PCoV_GX-P5E	Guangxi PangolinP 5E	Pangolin	MT040336
PCoV_GX-P2V	Guangxi PangolinP 2V	Pangolin	MT072864
PCoV_GX-P5L	Guangxi PangolinP 5L	Pangolin	MT040335
PCoV_GX-P1E	Guangxi PangolinP 1E	Pangolin	MT040334.1
PCoV_GX-P4L	Guangxi PangolinP 4L	Pangolin	MT040333
bat-SL-CoVZC45	Bat CoVZC45	Bat	MG772933.1
bat-SL-CoVZXC21	Bat CoVZXC21	Bat	MG772934.1
BtCoV/273/2005	BtCoV 273 2005	Bat	DQ648856.1
Bat SARS coronavirus Rf1	Rf1	Bat	DQ412042.1
Bat SARS coronavirus HKU3-12	HKU3-12	Bat	GQ153547.1
Bat SARS coronavirus HKU3-6	HKU3-6	Bat	GQ153541.1

BtCoV/279/2005	BtCoV 279 2005	Bat	DQ648857.1
SARS coronavirus BJ01	SARS	Human	AY278488.2
SARS coronavirus	Tor2	Human	NC_004718.3
SARS coronavirus BJ182-4	SARS-CoV BJ182-4	Human	EU371562
Bat SARS-like coronavirus Rs3367	Rs3367	Bat	KC881006.1
SARS-related coronavirus BtKY72	BtKY72	Bat	KY352407.1
Bat coronavirus BM48-31/BGR/2008	BM48 31 BGR 2008	Bat	GU190215.1
Human betacoronavirus 2c EMC/2012	MERS-CoV S	Human	JX869059
Bat coronavirus HKU5-1	Bat HKU5-1	Bat	NC_009020
Bat coronavirus HKU4-1	Bat HKU4-1	Bat	NC_009019
Feline infectious peritonitis virus	Feline per	Cat	NC_002306
Human coronavirus HKU1	HKU1	Human	NC_006577
Human coronavirus HKU1 (isolate N2)	HKU1 N2	Human	Q14EB0
Human coronavirus HKU1 (isolate N1)	HKU1 N1	Human	5GNB_A
Murine coronavirus RA59/R13	Murine RA59/R13	Mouse	ACN89689
Murine hepatitis virus strain 4	Murine hep 4	Mouse	P22432
Murine hepatitis virus strain JHM	Murine JHM	Mouse	NC_006852
Mouse hepatitis virus strain MHV-A59 C12 mutant	Murine A59	Mouse	NC_001846
Murine hepatitis virus	Murine virus	Mouse	ABS87264
Rat coronavirus Parker	Rat Parker	Rat	NC_012936
Rabbit coronavirus HKU14	Rabbit HKU14	Rabbit	NC_017083
Equine coronavirus	Equine NC99	Horse	NC_010327
Porcine hemagglutinating encephalomyelitis virus	Porcine v	Pig	NC_007732
Human coronavirus OC43	Human OC43	Human	NC_005147
Human enteric coronavirus strain 4408	Human ent 4408	Human	NC_012950
Bovine coronavirus	Bovine CoV	Calf	NC_003045
Bovine respiratory coronavirus AH187	Bovine AH187	Calf	NC_012948

Bovine respiratory coronavirus bovine/US/OH-440-TC/1996	Bovine OH440	Calf	NC_012949

Source : (Makarenkov et al., 2021)

3.2 Description des gènes/régions génomiques

Dans le cadre de nos recherches, nous nous sommes concentrés sur des régions génomiques spécifiques, jouant un rôle important dans la caractérisation évolutive et fonctionnelle des organismes étudiés. Ces gènes sélectionnés ont été choisis en fonction de leur pertinence et de leur capacité à fournir des informations significatives sur les relations évolutives.

Gène E

L'un des nombreux gènes du Sars-cov-2 est le gène E, également appelé gène de l'enveloppe. Il code une protéine qui participe à la formation de l'enveloppe virale et joue un rôle important dans la structure et la fonction du coronavirus humain. Il contribue à la libération de particules virales et peut avoir un impact sur la virulence et la force de l'infection. Sa conception en fait une cible potentielle pour les vaccins, et ses recherches sont utilisées pour la recherche phylogénétique et épidémiologique, fournissant des informations précieuses sur l'évolution virale et la transmission des maladies pour comprendre le contrôle des infections.

Gène M

Le gène M, ou gène matrice, fait partie intégrante du génome des virus à ARN comme le coronavirus humain. Il code une protéine de la matrice virale qui joue un rôle important dans la structure et la stabilité des particules. La protéine matricielle lie la membrane lipidique et la capside, contribuant ainsi à la réplication et à l'assemblage du virus. Le gène M présente un intérêt particulier dans la recherche en biologie et le développement de vaccins. Ses recherches pourraient fournir des informations importantes sur la structure bactérienne et les mécanismes de réplication, aidant à comprendre à la fois les agents pathogènes et la résistance aux maladies infectieuses.

Gène N

Le génome des virus à ARN tels que les coronavirus comprend le gène N également appelé gène de la nucléocapside. Il contrôle la production de la protéine nucléocapside, qui est essentielle à la liaison et à la protection des gènes viraux. La nucléocapside virale est formée par cette protéine qui encapsule le génome viral. Le gène N est crucial pour les études virologiques, la compréhension des mécanismes de

l'infection, le criblage des médicaments et la protection du développement de vaccins contre le virus en raison de son rôle central dans l'écologie virale. Ses recherches offrent des données.

Gène N2

Le gène N2, également connu sous le nom de gène de la nucléocapside 2, est un composant important du génome du coronavirus. Ce gène est responsable de la synthèse d'une protéine de nucléocapside unique, qui joue un rôle important dans la structure et la fonction virales. L'analyse des séquences génomiques des gènes N2 à l'aide du logiciel MEGA11 offre la possibilité de rechercher des mutations génétiques spécifiques liées à ce gène dans différentes souches de coronavirus humain.

Gène ORF1ab

Le cadre de lecture ouverte 1AB (orf1AB) fait partie intégrante du génome du coronavirus, une classe de virus à ARN. Il comprend les régions ORF1a et ORF1b qui codent pour des polyprotéines précoces impliquées dans la réplication virale et la synthèse des enzymes nécessaires à cette réplication.

L'analyse de Orf1AB est essentielle pour comprendre la biologie des coronavirus. Il fournit des informations précieuses sur la façon dont le virus se réplique, se propage et évolue. En étudiant orf1AB, les chercheurs peuvent identifier des cibles potentielles pour les médicaments antiviraux et les vaccins. Les différences dans cette partie du génome peuvent également éclairer l'origine et les mécanismes de l'infection. Dans ce contexte, l'exploration des implications et des découvertes d'orf1AB est essentielle pour élargir notre compréhension des coronavirus et des mécanismes régissant leur comportement.

Gène ORF3a

Le génome du coronavirus comprend la région ORF3a, également appelée cadre de lecture ouverte 3a. Le code ORF3a est la protéine accessoire virale 3a. Des études suggèrent que la protéine 3a pourrait jouer plusieurs rôles importants dans le cycle de vie viral et la communication virale, bien que la fonction spécifique de la protéine ne soit pas bien comprise.

La protéine 3a pourrait être impliquée dans la modulation de la réponse immunitaire virale, principalement en inhibant la réponse antivirale des cellules infectées. Il peut également jouer un rôle dans la formation de membranes dans les cellules infectées, ce qui peut être important pour l'élimination des virus nouvellement répliqués. De plus, la protéine 3a a été étudiée pour ses effets sur la perméabilité membranaire, la régulation des ions et d'autres processus cellulaires.

L'analyse de la région ORF3a peut fournir des indices importants sur la virulence, la pathogenèse et les interactions avec l'hôte. Comprendre les fonctions de la protéine 3a peut avoir des implications pour le développement de médicaments antiviraux et de vaccins. L'exploration de cette région permet d'éclairer les mécanismes moléculaires de l'infection virale et de mieux comprendre les interactions complexes entre le virus et son hôte.

Gène S

Le gène S, également connu sous le nom de gène de pointe ou gène de la glycoprotéine S, est un composant important du génome du coronavirus. Il code pour la protéine de pointe, une glycoprotéine de type trimère qui réside à la surface bactérienne. La protéine de pointe lie le virus à des récepteurs spécifiques de la cellule hôte, permettant au virus de pénétrer dans la cellule et d'initier l'infection. La protéine de pointe joue un rôle important dans l'évolution des coronavirus, car elle est le déterminant clé de la spécificité virale. Les différences dans les séquences de protéines de pointe peuvent influencer la capacité du virus à infecter différents virus et peuvent également contribuer à la virulence. Par conséquent, le gène S est fréquemment ciblé pour les études épidémiologiques, phylogénétiques et vaccinales.

L'analyse génomique S peut suivre les changements dans les souches au fil du temps et identifier des souches nouvelles ou potentiellement différentes. Ces informations sont importantes pour surveiller la propagation du virus et modifier les stratégies de contrôle. De plus, le gène S est la cible principale des vaccins, car il peut conférer des réponses immunitaires spécifiques à la maladie qui peuvent protéger contre l'infection.

En résumé, le gène S est un élément important de la biologie des coronavirus, avec une virulence et des effets spécifiques à l'agent pathogène. Ses recherches fournissent des informations importantes pour comprendre la virulence, l'évolution et le développement de vaccins.

Génome complet

Le génome complet d'un organisme, qu'il soit bactérie, virus, végétal ou animal, représente l'ensemble de ses gènes. C'est une séquence de molécules d'ADN (ou d'ARN dans le cas d'une bactérie) qui code à la fois la nature et la fonction de cet organisme. Un génome complet contient les informations nécessaires à toutes les protéines et molécules nécessaires à la croissance, au développement et au fonctionnement de l'organisme. L'analyse pangénomique est une approche importante de la biologie moléculaire et de la

génétique. Il peut fournir des informations importantes sur la régulation des gènes, les régions non codantes, les facteurs de régulation, les mutations, les transgènes et les génotypes. L'analyse à l'échelle du génome permet de comprendre les relations évolutives entre les espèces, d'identifier les gènes responsables de traits spécifiques, de diagnostiquer les maladies génétiques et d'étudier les mécanismes moléculaires qui contrôlent la diversité biologique des races.

Dans le contexte des virus, l'analyse pangénomique est particulièrement importante pour comprendre leur structure, leur fonction, leur évolution et leur pathogenèse. L'analyse comparative des génomes bactériens permet de rechercher des origines, des mutations, des recombinaisons et des mutations qui définissent différentes espèces bactériennes. En outre, l'analyse pangénomique est importante pour la conception de vaccins, le développement de traitements et l'épidémiologie. En termes simples, le génome complet représente le code génétique de base d'un organisme et son analyse fournit des informations approfondies sur la biologie, l'évolution et la santé, avec des implications importantes dans les domaines de la recherche scientifique et médicale.

3.3 Descriptions des méthodes bioinformatiques

Dans cette étude, nous avons utilisé des techniques bioinformatiques sophistiquées pour élucider l'évolution des séquences génomiques. Tout d'abord, le logiciel MEGA (Kumar et al., 2018) a été utilisé pour déterminer les modèles évolutifs les plus appropriés pour chaque région génomique étudiée, fournissant des informations importantes sur les taux de mutation et les relations phylogénétiques. Ensuite, le concept T-REX (transfert, évolution et expansion des gènes) (Boc et al., 2012) a été appliqué pour étudier les niveaux élevés de transfert de gènes et les mécanismes évolutifs entre les lignées. Nous avons également utilisé l'outil SimPlot++ (Samson et al., 2022) pour visualiser et interpréter les similitudes et les différences entre les séquences génomiques, fournissant un aperçu visuel riche de leurs propriétés évolutives. Pour une analyse plus approfondie, le package BEAST (Drummont et Rambeau, 2007) a été adopté et des analyses phylogénétiques bayésiennes avancées ont été effectuées à l'aide de BEAGLE, ainsi que TreeAnnotator et FigTree pour résumer et visualiser les arbres phylogénétiques résultants. La combinaison de ces techniques bioinformatiques a permis une approche globale pour retracer les processus évolutifs et les interactions génétiques entre les séquences étudiées, apportant ainsi un éclairage nouveau sur l'histoire évolutive et l'écologie d'un être considéré sous diverses formes. (Drummont et Rambeau, 2007)

3.3.1 MEGA (Molecular Evolutionary Genetics Analysis)

MEGA (Molecular Evolutionary Genetics Analysis) est un logiciel de pointe qui joue un rôle important dans la compréhension des processus évolutifs des séquences de gènes. Conçu pour répondre aux besoins des chercheurs en biologie évolutive et en génomique, MEGA propose un ensemble complet d'outils d'analyse, de manipulation et de visualisation de données moléculaires. En utilisant des techniques de pointe, MEGA permet d'estimer des modèles évolutifs complexes, d'estimer des arbres phylogénétiques précis et de mesurer la différenciation génétique entre les séquences. Son interface conviviale le rend facile à prendre en main, même pour les novices, tout en offrant des fonctionnalités avancées pour les chercheurs.

3.3.2 T-REX

L'analyse du transfert horizontal de gènes (THG) avec T-REX (disponible sur le site : http://www.trex.uqam.ca/) permet de détecter les gènes qui ont été transférés de manière horizontale entre différentes espèces ou souches de virus (Smith et al., 2018). Cela peut être utilisé pour étudier les origines des gènes du coronavirus humain et pour comprendre comment ils ont évolué au cours du temps. T-REX utilise des méthodes bioinformatiques pour détecter le THG en comparant les séquences génomiques de différentes espèces ou souches de virus. Cela peut inclure des analyses de l'homologie des séquences, de l'expression génique et de la distribution des gènes dans les génomes. Dans le contexte des coronavirus, l'analyse du THG avec T-Rex peut être utilisée pour identifier les gènes qui ont été transférés de manière horizontale entre différentes souches du coronavirus ou entre le coronavirus et d'autres espèces de virus. Cela peut aider à comprendre comment les gènes du coronavirus ont évolué, notamment les gènes qui sont importants pour la virulence ou la transmissibilité, ainsi que les gènes qui ont été acquis récemment par les souches du coronavirus. Pour détecter la recombinaison génétique dans les génomes du SARS-cov-2, il est important de disposer de séquences génomiques provenant de différentes souches du coronavirus humain. Cela permet de comparer les séquences génomiques et de détecter les régions qui ont subi des changements importants, comme la recombinaison.

3.3.3 SimPlot++

Dans notre étude on a utilisé le logiciel SimPlot++ pour analyser les graphes et réseaux des résultats obtenus à partir des 43 séquences génétiques de différentes espèces du Coronavirus. L'objectif principal est de comprendre les schémas de recombinaison génétique et les événements de transfert horizontal de gènes entre les différentes souches du SARS-cov-2. Les graphes et les réseaux générés par le logiciel SimPlot++ fournissent des visualisations claires des relations génétiques et des échanges de matériel

génétique, offrant ainsi des informations essentielles sur l'évolution et la diversité du Coronavirus humain.Avec le modèle de la distance de Hamming dans le logiciel SimPlot++, nous avons utilisé les outils et les techniques de visualisation disponibles suivantes :

Graphe de SimPlot ++

SimPlot++ peut également afficher les résultats de l'analyse de la divergence génétique sous forme de graphe, qui consiste en une série de points disposés le long d'une ligne, chaque point représente une région du génome. La hauteur de chaque point est proportionnelle à la divergence génétique mesurée entre les séquences génomiques analysées.

Réseau de SimPlot++

SimPlot++ peut aussi afficher les résultats de l'analyse de la divergence génétique sous forme de réseau, qui consiste en une série de points reliés par des lignes, chaque point représente une séquence génomique. La longueur des lignes entre les points est proportionnelle à la divergence génétique mesurée entre les séquences, il peut également être utilisé pour visualiser les relations de parenté entre les séquences et pour identifier les régions du génome où la divergence génétique est la plus élevée.

3.3.4 BEAST (Bayesian Evolutionary Analysis by Sampling Trees)

BEAST (Bayesian Evolutionary Analysis by Sampling Trees) est un logiciel couramment utilisé pour estimer les temps de divergence dans les analyses phylogénétiques bayésiennes. L'estimation du temps de divergence est un aspect essentiel de l'étude des relations évolutives entre les espèces. BEAST utilise une approche bayésienne pour estimer les temps de divergence en combinant des informations provenant des séquences d'ADN et d'autres données évolutives. Il utilise une méthode de simulation par chaîne de Markov Monte-Carlo (MCMC) pour échantillonner l'espace des arbres et des paramètres du modèle. Cette approche permet d'obtenir des estimations probabilistes des temps de divergence, accompagnées d'intervalles de crédibilité. BEAST est très flexible et permet de modéliser diverses sources d'incertitude, telles que les taux d'évolution moléculaire variables et les décalages d'horloge moléculaire. Il fournit ainsi une estimation plus réaliste et précise des temps de divergence entre les espèces. (Suchard et al., 2018).

3.3.4.1 Beagle (Broadly Applicable Graphical Likelihood Engine)

Beagle (Broadly Applicable Graphical Likelihood Engine) est un moteur statistique performant utilisé pour fournir des estimations de probabilités rapides dans les analyses phylogénétiques bayésiennes. Il est conçu

pour accélérer le calcul de modèles complexes et à forte intensité de main-d'œuvre, permettant une analyse rapide et analytique sur de grands ensembles de données. BEAGLE est souvent utilisé en conjonction avec un logiciel d'analyse phylogénétique tel que BEAST, pour améliorer l'efficacité des calculs.

BEAST est un progiciel complet pour les analyses phylogénétiques bayésiennes, tandis que BEAGLE est un moteur de calcul qui accélère les calculs de probabilité dans ces analyses, améliorant leurs performances et leur évolutivité. Les deux outils sont souvent utilisés ensemble pour une analyse évolutive approfondie et précise.

3.3.4.2 Tree Annotator

Tree Annotator est un outil logiciel couramment utilisé dans le cadre des analyses phylogénétiques pour annoter les arbres de sortie générés par des programmes tels que BEAST. Son rôle principal est de prendre un ensemble d'arbres échantillonnés à partir d'une chaîne de Markov Monte-Carlo (MCMC) et de produire un arbre de consensus représentatif, souvent appelé « arbre d'échantillonnage postérieur ». L'arbre d'échantillonnage postérieur est généralement utilisé pour résumer les relations évolutives et les temps de divergence entre les espèces étudiées. Tree Annotator permet également d'ajouter des estimations de paramètres et de caractéristiques spécifiques aux nœuds de l'arbre, telles que les dates de divergence ou les valeurs de probabilité de soutien. Cela facilite l'interprétation des résultats phylogénétiques et la présentation visuelle des arbres. Tree Annotator est un outil pratique pour les chercheurs utilisant BEAST ou d'autres programmes similaires pour analyser les données phylogénétiques. (Bouckaert et al., 2014)

3.3.4.3 FigTree

FigTree est un logiciel largement utilisé pour visualiser, annoter et manipuler des arbres phylogénétiques. Il est conçu pour fournir une interface intuitive et conviviale pour explorer les résultats des analyses phylogénétiques et représenter graphiquement les relations évolutives entre les espèces. FigTree permet aux chercheurs de personnaliser la visualisation des arbres en ajustant les couleurs, les formes, les étiquettes et les styles de branche. Il offre également des fonctionnalités avancées telles que l'ajout d'échelles de temps, la rotation de l'arbre, la sélection et le regroupement des nœuds, ainsi que la création de figures de haute qualité pour la publication. FigTree est compatible avec différents formats de fichiers d'arbres, ce qui en fait un outil polyvalent pour la communauté scientifique travaillant dans le domaine de la phylogénétique. (Rambaut, A., 2016).

CHAPITRE 4
RÉSULTATS ET DISCUSSIONS

Les résultats de notre étude ont été façonnés par la combinaison minutieuse de plusieurs techniques bioinformatiques puissantes. À l'aide de MEGA (Molecular Evolutionary Genetics Analysis), nous avons effectué une analyse phylogénétique complète, examinant les relations évolutives entre les séquences de gènes étudiées. Cette approche a permis la construction d'arbres phylogénétiques précis, apportant ainsi un nouvel éclairage sur les relations entre parents et les trajectoires évolutives des organismes considérés. En parallèle, le transfert de gènes a été étudié à l'aide du nouveau concept T-REX (Transfer, Evolution, and Expansion of Genes), qui a permis d'accéder aux mécanismes par lesquels les gènes sont échangés entre les lignées en profondeur. L'outil SimPlot++ a ensuite été utilisé pour rechercher des scénarios de recombinaison, mettant en évidence d'éventuelles localisations de réarrangements génétiques dans les séquences étudiées. Enfin, pour mieux comprendre le temps de la divergence entre les espèces, l'approche d'analyse évolutive bayésienne par échantillonnage d'arbres (BEAST) a été appliquée, tirant parti de la capacité de BEAGLE à fournir des estimations de probabilité rapides. Les résultats ont ensuite été résumés et visualisés à l'aide de TreeAnnotator et FigTree. Ces approches ingénieuses ont permis de mieux comprendre les mécanismes évolutifs.

4.1 Analyse phylogénétique

Nous avons commencé notre analyse phylogénétique en cherchant le meilleur modèle ADN/Protéine, en évaluant les critères AIC (Akaike Information Criterion) ou le BIC (Bayesian Information Criterion) pour par la suite sélectionner celui qui présente la valeur la plus basse selon la méthode utilisée par Yves en 2016. Ensuite, on a inséré les séquences nucléotidiques en format FASTA. Nous avons choisi la méthode de construction d'arbres du Neighbour Joining (NJ) (Saitou et al., 1987), une méthode de distance rapide reconnue, qui est particulièrement adaptée à la construction d'arbres pour de grandes séquences, comme confirmée par Yoann en 2012. Pour ce faire, des calculs ont été effectués avec un estimateur du maximum de vraisemblance, avec un Bootstrap de 100 itérations pour évaluer la robustesse de l'arbre résultant.

Cette étude se concentre sur l'analyse des arbres génétiques de SARS-Cov-2 à partir de 43 séquences d'espèces différentes, à l'aide du logiciel MEGA11. L'objectif principal est d'élucider les relations évolutives des espèces grâce à une analyse approfondie des séquences génétiques.

Les modèles évolutifs les plus appropriés trouvés par le programme Mega11 dans l'analyse de 43 séquences de différentes espèces du SARS-Cov-2 pour les régions des gènes et le domaine RB qui ont

été sélectionnées dans notre étude, avec les paramètres optimaux correspondants Gamma (G) et Intensité (I), ainsi que les valeurs des BIC (Bayesian Information Criterion) ont été analysés et représentés dans le tableau 4.1.

Tableau 4.1 Statistiques des meilleurs modèles évolutifs avec leurs paramètres optimaux (G) (I) et valeurs (BIC) trouvées par MEGA11 pour l'analyse de 43 séquences de différentes espèces du Sars-Cov-2.

Région	Modèle évolutif	G	I	BIC
Gène ORF1ab	LG+F	--	--	787597.699
Gène S	LG+F	--	--	168763.113
Gène ORF3a	LG+F	--	--	23116.255
Gène E	LG+F	--	--	6562.227
Gène M	LG+F	--	--	22152.210
Gène N	LG+F	--	--	36447.916
Gène N 2 (18 espèces)	cpREV+F	--	--	4722.328
Domaine RB	WAG	--	--	12243.461

LG : Le et Gascuel, F : Fréquence, Cprev : Prévisions de caractéristiques, WAG : Whelan and Goldman

4.1.1 Résultats de l'analyse phylogénique

Dans cette section, on va présenter les résultats de l'analyse phylogénique obtenus avec MEGA 11 des gènes E, M, N, N2, ORF1AB, ORF3A et S, qui fournissent des informations précieuses sur les relations évolutives des séquences étudiées. Grâce à des modèles évolutifs appropriés et à des méthodes de construction d'arbres phylogénétiques, nous avons pu cartographier des modèles de mutations et de divergences génétiques uniques à chaque gène et au génome dans son ensemble.

Les figures ci-dessous présentent un arbre phylogénétique du coronavirus pour plusieurs gènes. Chaque arbre a été construit à partir de 43 séquences selon la méthode Neighbor Joining (NJ), en tenant compte des différentes distances génétiques entre les séquences (Willems et al., 2014). Les branches de l'arbre représentent les différentes variantes virales, tandis que les extrémités symbolisent les différentes espèces.

Gène E

La figure 4.1 présente Arbre phylogénétique de coronavirus pour le gène E, construit à partir de 43 séquences extraites de différentes espèces

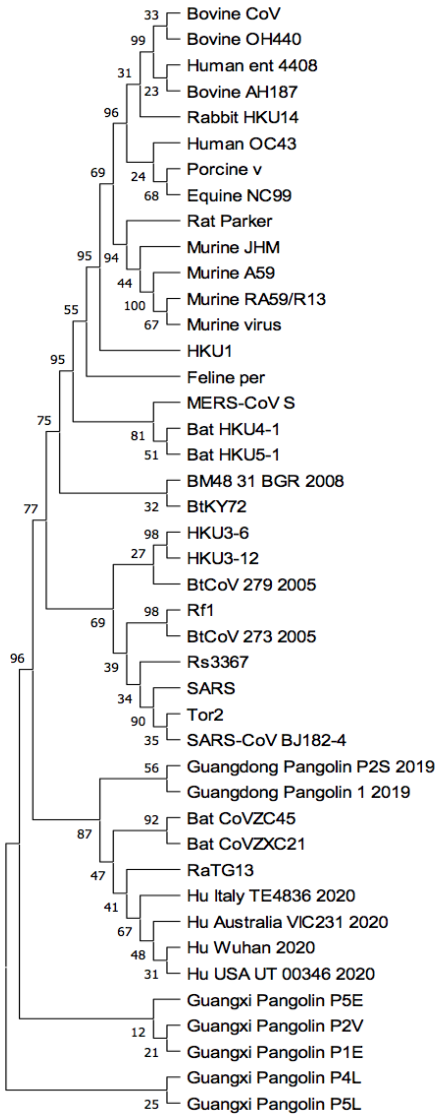

Figure 4.1 Arbre phylogénétique de coronavirus pour le gène E

Gène M

La figure 4.2 présente Arbre phylogénétique de coronavirus pour le gène M, construit à partir de 43 séquences extraites de différentes espèces.

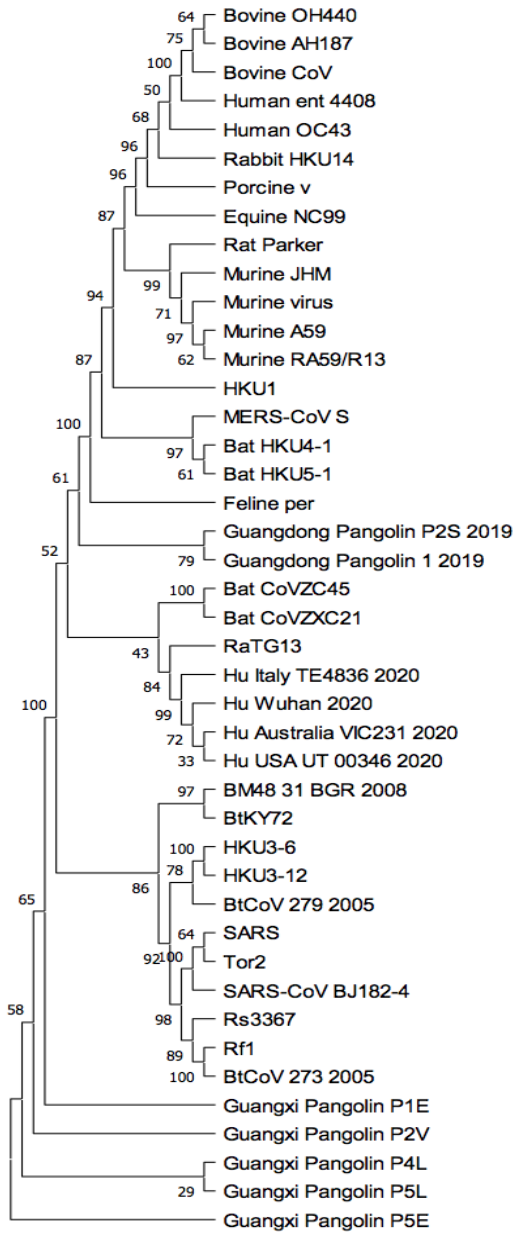

Figure 4.2 Arbre phylogénétique de coronavirus pour le gène M

Gène N

La figure 4.3 présente Arbre phylogénétique de coronavirus pour le gène N, construit à partir de 43 séquences extraites de différentes espèces.

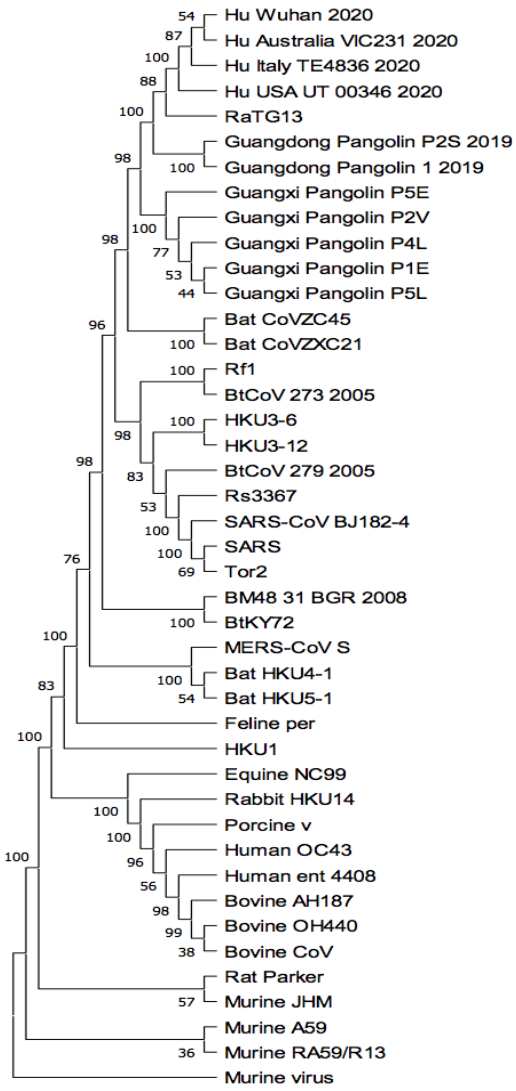

Figure 4.3 Arbre phylogénétique de coronavirus pour le gène N

Gène N2

La figure 4.4 présente Arbre phylogénétique de coronavirus pour le gène N2, construit à partir de 43 séquences extraites de différentes espèces.

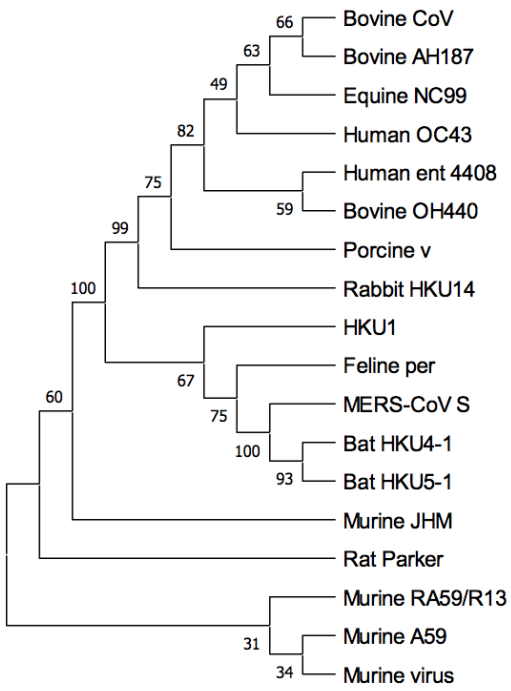

Figure 4.4 Arbre phylogénétique de coronavirus pour le gène N (N2)

Gène ORF1ab

La figure 4.5 présente Arbre phylogénétique de coronavirus pour le gène ORF1ab, construit à partir de 43 séquences extraites de différentes espèces.

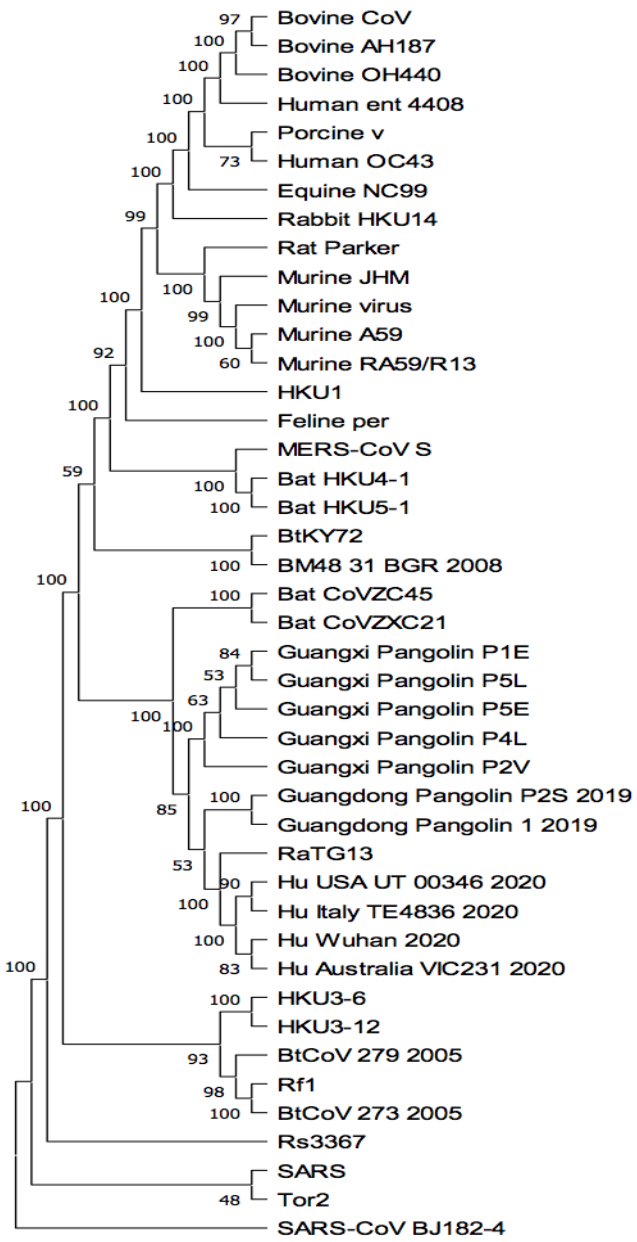

Figure 4.5 Arbre phylogénétique de coronavirus pour le gène ORF1ab

Gène ORF3a

La figure 4.6 présente Arbre phylogénétique de coronavirus pour le gène ORF3a, construit à partir de 43 séquences extraites de différentes espèces.

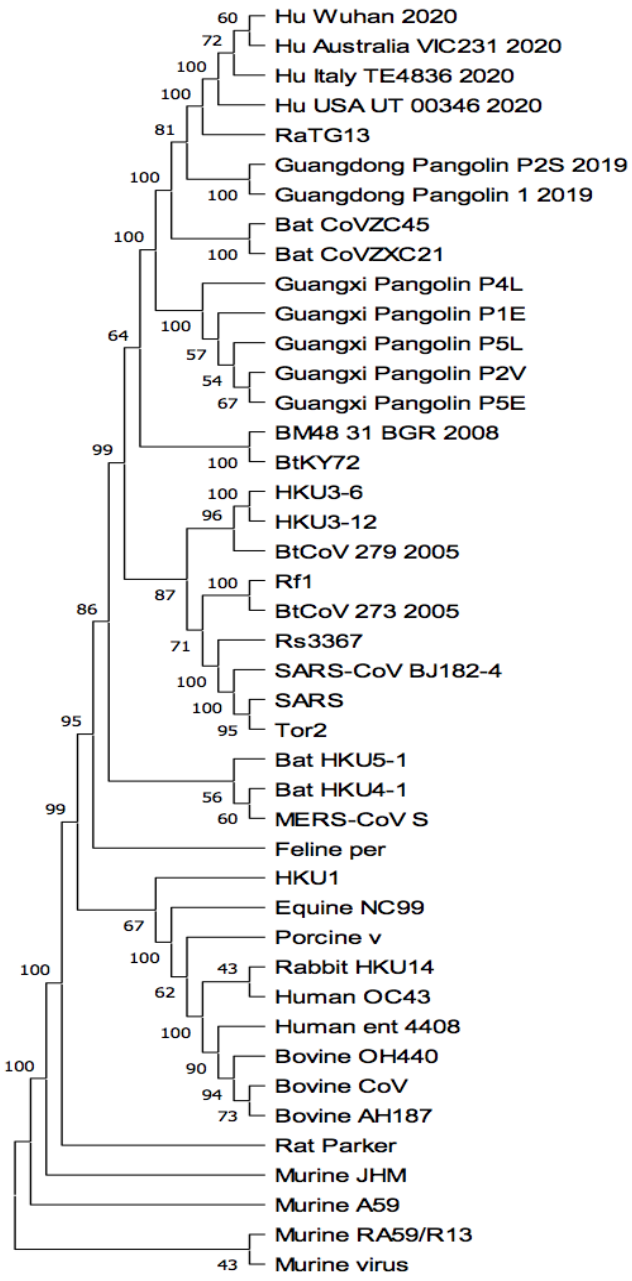

Figure 4.6 Arbre phylogénétique de coronavirus pour le gène ORF3a

Gène S

La figure 4.7 présente Arbre phylogénétique de coronavirus pour le gène S, construit à partir de 43 séquences extraites de différentes espèces.

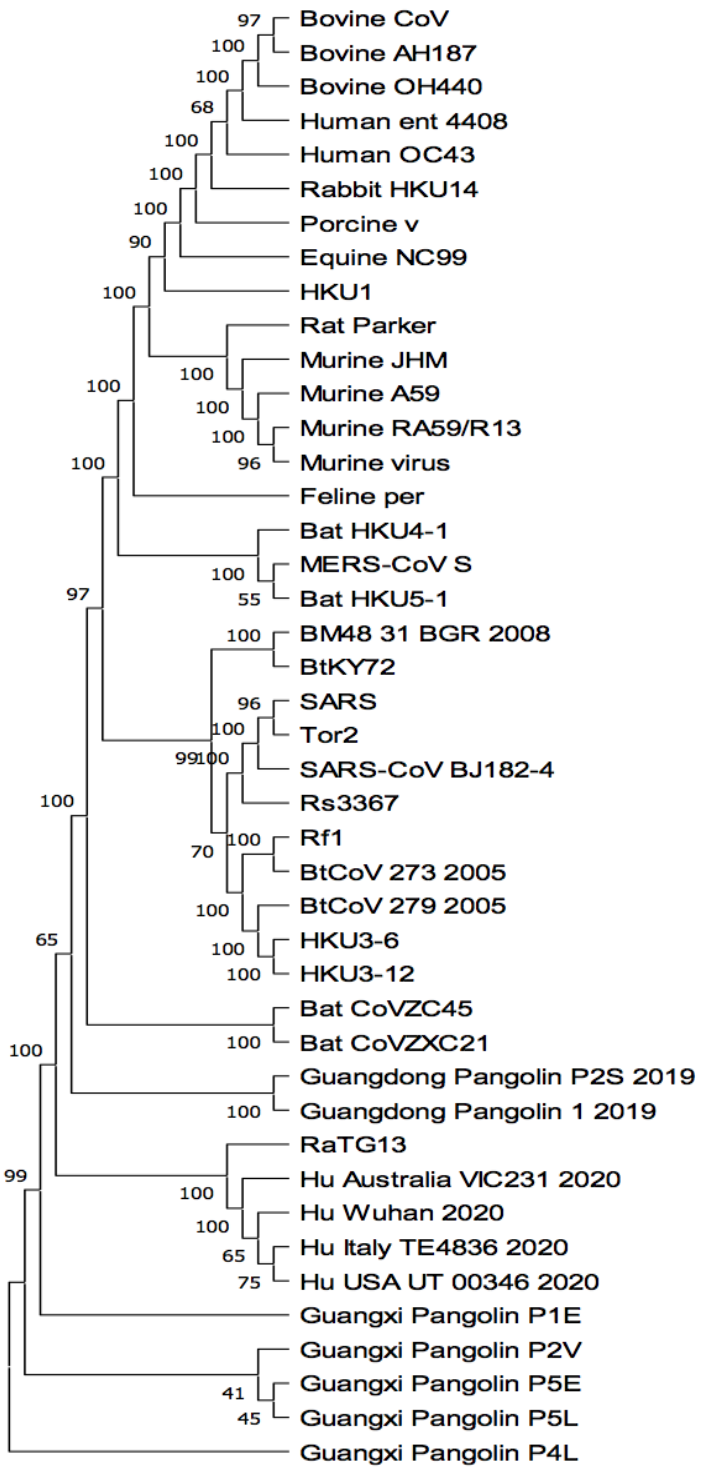

Figure 4.7 Arbre phylogénétique de coronavirus pour le gène S

Domaine RB

La figure 4.8 présente Arbre phylogénétique de coronavirus pour le domaine RB, construit à partir de 43 séquences extraites de différentes espèces.

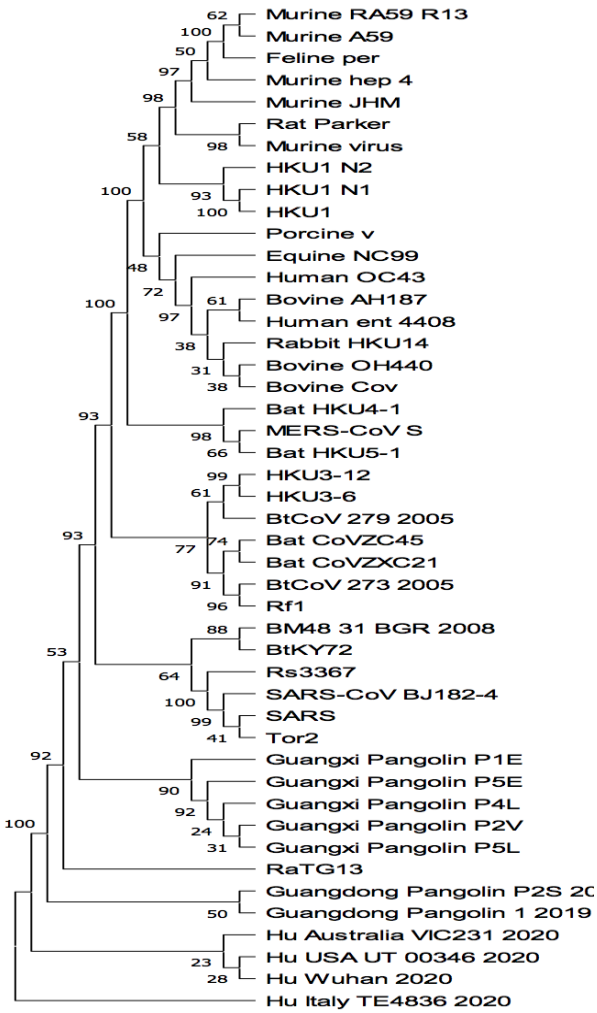

Figure 4.8 Arbre phylogénétique de coronavirus pour le domaine RB

4.1.2 Interprétation des résultats du Logiciel Méga

Dans cette section, on va interpréter et analyser les résultats des arbres phylogéniques obtenus à travers le logiciel Méga des différents gènes de coronavirus.

Gène E

Dans le cadre de notre analyse phylogénétique, l'arbre construit pour le gène E du coronavirus révèle des informations importantes sur les relations évolutives entre les différentes espèces. Chacune de ces branches représente une opportunité d'enquêter sur les origines et les relations génétiques.

L'arbre phylogénétique pour le gène E met en évidence 14 taxons distincts. Notamment, nous observons que le gène E du Bovine CoV et Bovine OH440, associés à des veaux, partagent un nœud externe avec une distance génétique de 33, soulignant une parenté proche. De même, le gène E du Human ent 4408 et celui du Bovine AH187, tous les deux liés à des veaux, présentent une similarité marquée avec une distance génétique également de 23. Des similitudes génétiques sont également apparentes entre d'autres souches, comme le gène E du Porcine v du porc et celui de l'Equine NC99 du cheval, avec une distance génétique de 68. Parmi les découvertes notables, nous notons que le gène E du Murine RA59/R13 partage une relation étroite avec le gène E du Murine virus de la souris, avec une distance génétique de 67. De même, les gènes E du Bat HKU4-1 et du Bat HKU5-1 de la chauve-souris, bien que légèrement distincts, présentent une proximité génétique considérable avec une distance de 51. Une observation similaire est observée entre le gène E du BM48 31 BGR 2008 et celui du BTKY72 de la chauve-souris, avec une distance de 22. L'analyse révèle également des similitudes intrigantes. Les gènes E du HKU3-6 et du HKU3-12 de la chauve-souris, bien que distincts, partagent un nœud externe avec une distance génétique de 98, indiquant une proximité significative. De même, le gène E du Rf1 et celui du BtCoV2732005 de la chauve-souris, ainsi que le gène E du SARS-CoV BJ182-4 et celui du Tor2 de l'humain, démontrent des similarités génétiques marquées avec une distance de 98 et 35 respectivement. En outre, l'arbre révèle des affinités entre diverses souches de pangolins et de chauve-souris. Par exemple, le gène E du Guangdong Pangolin P2S 2019 et du Guangdong Pangolin 1 2019 partagent un nœud externe avec une distance génétique de 56. De même, le gène E du Bat CoVZC45 et du Bat CoVZXC21 de la chauve-souris, bien que presque similaire, montre une légère distinction génétique avec une distance de 92. Le gène E du Hu Wuhan 2020 est plus proche du gène E du Hu USA UT 00346 2020 de l'humain et forment chacun un nœud externe avec une distance génétique égale à 31 et le gène E du Guangxi Pangolin P2V est plus proche du gène E de Guangxi Pangolin P1E du pangolin et forment chacun un nœud externe avec une distance génétique égale à 21. Le gène E du Guangxi Pangolin P4L est plus proche du gène E du Guangxi Pangolin P5L du pangolin et forment chacun un nœud externe avec une distance génétique égale à 25.

En analysant les ancêtres les plus proches dans notre analyse du gène E, nous constatons que le gène E du Rabbit HKU14 (lapin) partage des similitudes avec divers clades et présente une affinité avec un clade englobant le gène E du Bovine CoV et bovine OH440, ainsi que le gène du Human ent 4408 et bovine AH187. De même, le gène E du Human OC43 (humain) est étroitement lié au clade comprenant les gènes E du porc porcine v et du cheval Equine NC99. Le gène E du Rat Parker (rate) avait toutes les homologies des gènes E du Murine JHM, A59, RAS 59/R13. Le gène E du MERS-CoV S (humain) est plus proche de deux

clades : du veau qui comprend le gène E de la chauve-souris, Bat HKU4-1 et Bat HKU5-1. Le gène E de la chauve-souris RaTG13 avait toutes les homologies des gènes E de l'humain Hu (Wuhan 2020, USA UT 00346 2020, Italy TE4836 2020, Australia VIC231 2020).

Gène M

Notre analyse phylogénétique continue de se concentrer sur le gène M. En construisant un arbre phylogénétique, nous avons l'opportunité d'explorer le cœur des relations évolutives entre les espèces.

L'arbre génétique de coronavirus humain pour le gène M révèle la présence de 11 taxons distincts. Dans ce cas, une relation génétique surprenante est évidente. Par exemple, le gène M de Bovine AH187 et celui de Bovine OH440, qui est associé aux veaux, se chevauchent sur un nœud externe avec une distance génétique de 64. De même, le gène M du Murine RA59/R13 est plus proche du gène M de la Murine A59 de la souris et forment chacun un nœud externe avec une distance génétique égale à 62. La similitude génétique est également significative entre les gènes M de Bat HKU4-1 et Bat HKU5-1 de la chauve-souris, à savoir une proximité étroite avec une distance distincte de 61. De même, les gènes M de Guangdong Pangolin P2S 2019 et Guangdong Pangolin 1 2019 pour le pangolin partagent une relation génétique étroite, avec une distance de 79. Le gène M du Bat CoVZC45 est plus proche du gène M du Bat CoVZXC2 1 de la chauve-souris et chacun forme un nœud externe avec une distance génétique égale à 100, ils sont donc similaires. Une association intéressante apparaît également entre les gènes M humains Hu Australia VIC231 2020 et Hu USA UT 00346 2020, qui ont une proximité notable et une distance génétique de 33. De même, les gènes M de la chauve-souris BM48 31 BGR 2008 et BTKY72 présentent des similitudes génétiques frappantes avec une distance de 97. La relation génétique entre le gène M de la chauve-souris HKU3-6 et HKU3-12 est également intéressante, et la proximité est grande et la distance génétique est de 100. De même, le gène M du SRAS et le gène humain Tor2 présentent 64 relations génétiques proches et distantes. Le gène M du Rf1 est plus proche du gène M du BtCoV2732005 de la chauve-souris et forment chacun un nœud externe avec une distance génétique égale à 100, ils sont donc similaires et le gène M du Guangdong Pangolin P4L est plus proche du gène M de Guangdong Pangolin P5L du pangolin et forment chacun un nœud externe avec une distance génétique égale à 29.

En analysant les ancêtres les plus proches dans notre analyse du gène M, nous identifions des homologies génétiques d'intérêt. Le gène M de l'Equin NC99 (cheval) présente des homologies avec le clade du veau, qui comprend les gènes M du Bovin CoV et du Bovin OH440, et des feuilles représentant l'humain, l'Humain OC43 et l'Humain ent 4408, le porc, porcine v, le lapin et Rabbit HKU14. De même, le gène M de

Rat Parker (la rate) présente des similitudes avec les gènes M de Murine JHM, A59, RAS 59/R13. De plus, le génome MERS-CoV S (humain) présente une proximité avec deux clades distincts. Le premier clade est associé au veau, composé des gènes bat M, Bat HKU4-1 et Bat HKU5-1. Enfin, le gène M de la chauve-souris RaTG13 présente une similitude marquée avec les gènes M de l'humain Hu (Wuhan 2020, USA UT 00346 2020, Italie TE4836 2020, Australie VIC231 2020).

Gène N

En analysant l'arbre génétique de coronavirus humain pour le gène N, des relations génétiques complexes sont révélées. Parmi les 11 taxons présents, les similarités et similitudes génétiques sont clairement démontrées. Le gène N du Hu Wuhan 2020 établit une proximité remarquable avec les gènes N de Hu Australia VIC231 2020, fournissant un nœud externe avec une distance génétique de 54. De même, les gènes N de Guangdong Pangolin P2S 2019 et Guangdong Pangolin 1 2019 utilisent le pangolin comme un nœud externe avec une distance génétique de 100, soulignant leur similitude génétique. Une relation spécifique est également évidente. Le gène N du Guangxi Pangolin P1E partage une liaison avec les gènes Guangxi Pangolin P5L, les reliant à un nœud externe avec une distance génétique de 44. De même, le gène N de Bat CoVZC45 et les Bats CoVZXC2 1 de la chauve-souris s'unissent à un nœud externe avec une distance génétique de 100. D'autres relations se distinguent par la similarité génétique. Le gène N du Rf1 se regroupe avec celui de la chauve-souris BtCoV2732005 dans un nœud externe avec une distance de gène de 100, tout comme les gènes de chauve-souris HKU3-6 et HKU3-12 N avec une distance de gène de 100.

Cependant, des différences génétiques sont également évidentes. Le gène N du SRAS et le gène humain Tor2 partagent une distance génétique de 69. Le gène N du BM48 31 BGR 2008 est plus proche du gène N du BTKY72 de la chauve-souris et forment chacun un nœud externe avec une distance génétique égale à 100 ils sont donc similaires. En parallèle, d'autres interactions intéressantes émergent. Le gène N de Bat HKU4-1 s'hybride au gène N de Bat HKU5-1 de la chauve-souris, produisant un nœud externe avec une distance génétique de 54. De plus, une relation entre le gène N de la moelle épinière de Rat Paker et la souris Murine JHM a cartographié un écart génétique de 57, le gène N le gène N de la murine A59 de la rate est plus proche du gène N de la murine RA59/R13 de la souris et forment chacun un nœud externe avec une distance génétique égale à 36.

En analysant les ancêtres les plus proches dans notre analyse du gène N, des découvertes intéressantes émergent, tissant des relations génétiques complexes. Ils présentent une ascendance proche, révélant des relations complexes entre les espèces. Il a été démontré que le gène N, RaTG13 (chauve-souris) avait toutes les homologies des gènes N de l'humain (Hu: Australia VIC231 2020, USA UT 00346 20202, Wuhan 2020, Italy TE4836 2020). Le gène N, BtCoV 273 2005 (chauve-souris) avait toutes les homologies des gènes N de l'humain (SRAS, Tor2, SARS-Cov BJ182-4) et de la chauve-souris (RS3367). Le gène N, Guangxi Pangolin P5E (pangolin) avait toutes les homologies des gènes N du pangolin (Guangxi Pangolin : P2V, P4L, P1E, P5L). Le gène N du cheval, équine NC99 est plus proche du gène N du Rabbit HKU14 du lapin, porcine v du porc, de l'humain (Human OC43, Humane nt 4408) et du veau (Bovine CoV, Bovine AH187, Bovine OH440). Le gène N du MERS-CoV S (humain) est plus proche des deux clades : du veau qui comprend le gène N de la chauve-souris, bat HKU4-1 et bat HKU5-1.

Gène N2

Dans l'arbre génétique de coronavirus humain pour le gène N2, nous pouvons identifier quatre groupes distincts, le gène N du Bovine CoV est plus proche du gène N de la bovine AH187 du veau et forment chacun un nœud externe avec une distance génétique égale à 66 et le gène N du Human ent 4408 de l'humain est plus proche du gène N de la bovine OH440 du veau et forment chacun un nœud externe avec une distance génétique égale à 59. Le gène N du Bat HKU5-1 est plus proche du gène N du Bat HKU4-1 de la chauve-souris et chacun forme un nœud externe avec une distance génétique égale à 93, ils sont donc presque similaires. Le gène N de la murine A59 est plus proche du gène N de la murine virus de la souris et forment chacun un nœud externe avec une distance génétique égale à 34.

En analysant les ancêtres les plus proches dans notre analyse du gène N2, nous identifions des homologies génétiques d'intérêt. Le gène N du Human OC43 de l'humain avait toutes les homologies des gènes N de l'équine NC99 du cheval et du clade du veau qui comprend les gènes N : Bovine CoV, bovine AH187.Le gène N du HKU1 de l'humain avait toutes les homologies des gènes N du MERS-CoV S (humain), féline per (chat), bat HKU5-1 et bat HKU4-1 de la chauve-souris. Le gène N de la murine RA59/R 13 est plus proche du clade qui comprend les gènes N du murin A59, Murine virus de la souris.

Gène ORF1ab

Au total, il y a 11 taxons dans l'arbre génétique de coronavirus humain pour le gène ORF1ab. Le gène ORF1ab du Bovine CoV et le gène ORF1ab du Bovine AH187 chez le veau partagent un nœud externe avec une distance génétique de 97, indiquant ainsi une similarité remarquable. De même, le gène ORF1ab du Porcine V (porc) est adjacent au gène ORF1AB de Human OC43, chacun formant un nœud externe avec une longueur de distance génétique de 73. Le gène ORF1ab de la murine RA59/R13 est plus proche du gène ORF1ab de la murine A59 de la souris et forment chacun un nœud externe avec une distance génétique égale à 60. Le gène ORF1ab du Bat HKU4-1est plus proche du gène ORF1ab du Bat HKU5-1 de la chauve-souris et forment chacun un nœud externe avec une distance génétique égale à 100, ils sont donc similaires, cet arrangement similaire se retrouve également dans le gène ORF1ab du BM48 31 BGR 2008 et le gène ORF1ab du BTKY72 de la chauve-souris qui forment chacun un nœud externe avec une distance génétique égale à 100 tandis que le gène ORF1ab du Guangxi Pangolin P1E et Guangxi Pangolin P5L du pangolin sont également proches l'un de l'autre, formant un nœud externe avec une distance génétique de 84. De même, le gène ORF1ab dans le Guangdong Pangolin P2S 2019 est adjacent au gène ORF1ab dans le Guangdong Pangolin 1 2019, créant un nœud externe avec une distance génétique de 100. Le gène ORF1ab du Hu USA UT 00346 2020 est plus proche du gène ORF1ab du Hu Italy TE4836 2020 de l'humain et forment chacun un nœud externe avec une distance génétique égale à 90, ils sont donc presque similaires. Le gène ORF1ab du Hu Wuhan 2020 est plus proche du gène ORF1ab du Hu Australia VIC231 2020 de l'humain et forment chacun un nœud externe avec une distance génétique égale à 83. De même, le gène ORF1ab chez la chauve-souris HKU3-6 et HKU3-12 présente une similitude marquée, produisant un nœud avec une distance génétique de 100. Enfin, le gène ORF1ab avec le SRAS et les mutants Tor2 humains présente une proximité, formant un nœud externe avec une distance génétique de 48. De même, le gène ORF1ab de Rf1 est adjacent au gène ORF1ab de chauve-souris BtCoV2732005, permettant un nœud externe du génome qui est de 100.

En analysant les ancêtres les plus proches du gène ORF1ab, nous identifions des homologies génétiques d'intérêt. Le gène ORF1ab du Rabbit HKU14 (lapin) avait toutes les homologies du clade : du veau qui comprend le gène ORF1ab du Bovine CoV et bovine OH440 et les feuilles : d'humain, Human OC43et Human ent 4408. De porc, porcine v. et de cheval, équine NC99.Le gène ORF1ab du Rat Parker (la rate) avait toutes les homologies des gènes ORF1ab du Murine JHM, A59, RAS 59/R13, virus. Le gène ORF1ab du MERS-CoV S (humain) est plus proche des deux clades : du veau qui comprend le gène ORF1ab de la chauve-souris, bat HKU4-1 et bat HKU5-1. Le gène ORF1ab de la chauve-souris RaTG13 avait toutes les

homologies des gènes ORF1ab de l'humain Hu (Wuhan 2020, USA UT 00346 2020, Italy TE4836 2020, Australia VIC231 2020).

Gène ORF3a

L'arbre génétique de coronavirus humain pour le gène ORF3a comprend au total 12 taxons. Le gène ORF3a du Bovine CoV est plus proche du gène ORF3a du Bovine AH187 du veau et forment chacun un nœud externe avec une distance génétique égale à 73. Le gène ORF3a de la murine RA59/R13 est plus proche du gène ORF3a de la murine Virus de la souris et forment chacun un nœud externe avec une distance génétique égale à 43. Le gène ORF3a du Bat HKU4-1 est plus proche du gène ORF3a du MERS-CoV S de l'humain et forme chacun un nœud externe avec une distance génétique égale à 60. Le gène ORF3a du BM48 31 BGR 2008 est plus proche du gène ORF3a du BTKY72 de la chauve-souris et forme chacun un nœud externe avec une distance génétique égale à 100, ils sont donc similaires. Le gène ORF3a du Bat CoVZC45 est plus proche du gène ORF3a du Bat CoVZXC21 de la chauve-souris et forme chacun un nœud externe avec une distance génétique égale à 100, ils sont donc similaires. Le gène ORF3a du Guangxi Pangolin P2V est plus proche du gène ORF3a du Guangxi Pangolin P5E du pangolin et forment chacun un nœud externe avec une distance génétique égale à 67. Le gène ORF3a du Guangdong Pangolin P2S 2019 est plus proche du gène ORF3a du Guangdong Pangolin 1 2019 du pangolin et forment chacun un nœud externe avec une distance génétique égale à 100, ils sont donc similaires. Le gène ORF3a du lapin, Rabbit HKU14 est plus proche du gène ORF3a du Human OC43 de l'humain et forment chacun un nœud externe avec une distance génétique égale à 43. Le gène ORF3a du Hu Wuhan 2020 est plus proche du gène ORF3a du Hu Australia VIC231 2020 de l'humain et chacun forme un nœud externe avec une distance génétique égale à 60. Le gène ORF3a du HKU3-6 est plus proche du gène ORF3a du HKU3-12 de la chauve-souris et chacun forme un nœud externe avec une distance génétique égale à 100, ils sont donc similaires. Le gène ORF3a du SRAS est plus proche du gène ORF3a du Tor2 de l'humain et chacun forme un nœud externe avec une distance génétique égale à 95, ils sont donc presque similaires. Le gène ORF3a du Rf1 est plus proche et similaire au gène ORF3a du BtCoV2732005 de la chauve-souris et chacun forme un nœud externe avec une distance génétique égale à 100.

En analysant les ancêtres les plus proches du gène ORF3a, nous identifions des homologies génétiques d'intérêt. Le gène ORF3a du HKU1 (humain) avait toutes les homologies du clade : du veau qui comprend le gène ORF3a du Bovine CoV et bovine OH440 et les feuilles : Du porc, porcine v. du cheval, équine NC99.clade du lapin et de l'humain (Human OC43 et Rabbit HKU14). Le gène ORF3a du Rat Parker (rate)

avait toutes les homologies des gènes ORF3a du Murine JHM, A59, RAS 59/R13, virus. Le gène ORF3a du Bat HKU5-1 de la chauve-souris est plus proche des deux clades : du veau qui comprend le gène ORF3a de la chauve-souris, bat HKU4-1 et MERS-CoV S (humain). Le gène ORF3a de la chauve-souris RaTG13 avait toutes les homologies des gènes ORF3a de l'humain Hu (Wuhan 2020, USA UT 00346 2020, Italy TE4836 2020, Australia VIC231 2020).

Gène S

L'arbre génétique de coronavirus humain pour le gène S comprend 11 taxons. Le gène S du Bovine CoV est plus proche du gène S du Bovine AH187 du veau et forment chacun un nœud externe avec une distance génétique égale à 97, ils sont donc presque similaires. Le gène S de la murine RA59/R13 est plus proche du gène S de la murine Virus de la souris et forment chacun un nœud externe avec une distance génétique égale à 96, ils sont donc presque similaires. Le gène S du Bat HKU5-1 est plus proche du gène S du MERS-CoV S de l'humain et chacun forme un nœud externe avec une distance génétique égale à 55. Le gène S du BM48 31 BGR 2008 est plus proche du gène S du BTKY72 de la chauve-souris et chacun forme un nœud externe avec une distance génétique égale à 100, ils sont donc similaires. Le gène S du Bat CoVZC45 est plus proche du gène S du Bat CoVZXC21 de la chauve-souris et chacun forme un nœud externe avec une distance génétique égale à 100, ils sont donc similaires. Le gène S du Guangxi Pangolin P5L est plus proche du gène S du Guangxi Pangolin P5E du pangolin et forment chacun un nœud externe avec une distance génétique égale à 45. Le gène S du Guangdong Pangolin P2S 2019 est plus proche du gène S du Guangdong Pangolin 1 2019 du pangolin et forment chacun un nœud externe avec une distance génétique égale à 100, ils sont donc similaires. Le gène S du Hu Italy TE4836 2020est plus proche du gène S du Hu USA UT 00346 20202 de l'humain et forment chacun un nœud externe avec une distance génétique égale à 75. Le gène S du HKU3-6 est plus proche du gène S du HKU3-12 de la chauve-souris et chacun forme chacun un nœud externe avec une distance génétique égale à 100, ils sont donc similaires. Le gène S du SRAS est plus proche du gène S du Tor2 de l'humain et chacun forme un nœud externe avec une distance génétique égale à 96. Le gène S du Rf1 est plus proche du gène S du BtCoV2732005 de la chauve-souris et chacun forme un nœud externe avec une distance génétique égale à 100, ils sont donc similaires.

En analysant les ancêtres les plus proches du gène S, nous identifions des homologies génétiques d'intérêt. Le gène S du HKU1 (humain) avait toutes les homologies du clade : du veau qui comprend le gène S du Bovine CoV et bovine OH440 et les feuilles : Du porc, porcine v. du cheval, équine NC99, du lapin

(Rabbit HKU14) et de l'humain (Human). Le gène S de la chauve-souris RaTG13avait toutes les homologies des gènes S de l'humain Hu (Wuhan 2020, USA UT 00346 2020, Italy TE4836 2020, Australia VIC231 2020).

Domaine RB

L'arbre phylogénétique du domaine RB comprend 12 taxons. Le gène du domaine RB du Bovine AH187 du veau est plus proche du domaine RB du Human ent 4408 et et chacun forme un nœud externe avec une distance génétique égale à 61. Le domaine RB de la bovine OH 440 est plus proche du gène ORF3a de Bovine Cov du veau et chacun forme un nœud externe avec une distance génétique égale à 38. Le domaine RB de la murine RA59/R13 est plus proche du domaine RB de la murine A59 de la souris et chacun forme un nœud externe avec une distance génétique égale à 62. Le domaine RB du Rat Parker de la rate est plus proche du domaine RB de la murine Virus de la souris et chacun forme un nœud externe avec une distance génétique égale à 98, ils sont donc presque similaires. Le gène du domaine RB du HKU1 N1 est plus proche du domaine RB du HKU1 de l'humain et chacun forme un nœud externe avec une distance génétique égale à 100. Le gène du domaine RB du MERS-CoV S de l'humain est plus proche du domaine RB du Bat HKU5 1 de la chauve-souris et chacun forme un nœud externe avec une distance génétique égale à 66. Le domaine RB du BM48 31 BGR 2008 est plus proche du domaine RB du BTKY72 de la chauve-souris et chacun forme un nœud externe avec une distance génétique égale à 88. Le domaine RB du Bat CoVZC45 est plus proche du domaine RB du Bat CoVZXC21 de la chauve-souris et chacun forme un nœud externe avec une distance génétique égale à 74. Le domaine RB du Guangxi Pangolin P2V est plus proche du domaine RB du Guangdong Pangolin P5L du pangolin et forment chacun un nœud externe avec une distance génétique égale à 31. Le domaine RB du Guangdong Pangolin P2S 2019 est plus proche du domaine RB du Guangdong Pangolin 1 2019 du pangolin et forment chacun un nœud externe avec une distance génétique égale à 50. Le gène ORF3a du Hu Wuhan 2020 est plus proche du gène ORF3a du Hu USA UT 00346 20202 de l'humain et forment chacun un nœud externe avec une distance génétique égale à 28. Le domaine RB du HKU3-6 est plus proche du domaine RB de HKU3-12 de la chauve-souris et chacun forme un nœud externe avec une distance génétique égale à 99, ils sont donc presque similaires. Le gène du domaine RB du SRAS est plus proche du domaine RB du Tor2 de l'humain et chacun forme un nœud externe avec une distance génétique égale à 41. Le domaine RB du Rf1 est plus proche du domaine RB de BtCoV2732005 de la chauve-souris et chacun forme un nœud externe avec une distance génétique égale à 96, ils sont donc presque similaires.

En analysant les ancêtres les plus proches du gène S, nous identifions des homologies génétiques d'intérêt. Le domaine RB du HKU1 N2 (humain) avait toutes les homologies des feuilles de l'humain du domaine RB du HKU1 N1, HKU1. Le domaine RB du Bat HKU4-1 (chauve-souris) est plus proche du clade qui comprend le domaine RB de l'humain, MERS Cov S, et de la chauve-souris, Bat HKU5-1.

4.2 Analyse de transfert horizontal de gènes HGT

Le transfert horizontal de gènes (HGT) représente un processus important dans l'évolution biologique, permettant le transfert génétique entre espèces, indépendamment de l'ascendance commune. Ce processus joue un rôle important dans la modification génétique en introduisant de nouveaux traits, fonctions et changements dans les gènes. Dans HGT, les gènes peuvent être transférés d'un organisme à un autre par divers mécanismes tels que la mutation, la conjugaison ou la transduction, qui sont généralement influencées par des agents mobiles tels que des plasmides. HGT peut produire des traits bénéfiques, mais il peut également contribuer à la propagation de gènes nocifs, y compris ceux liés à la résistance aux antibiotiques. Comprendre le transfert de gènes de haut niveau est essentiel pour comprendre la complexité des interactions génétiques entre les espèces et pour faire la lumière sur la diversité génétique et les changements de processus biologiques dans leur environnement changeant.

Le tableau 4.2 résume les résultats des gènes, génomes, et domaines RB du coronavirus humain avec 6, 25, 43 séquences de différentes espèces, nous avons désigné les gènes, les génomes, et les domaines RB du coronavirus humain avec 6 séquences avec leurs noms et le chiffre 1, de 25 séquences par leurs noms et le chiffre 2, et de 43 séquences par leurs noms et le chiffre 3.

Tableau 4.2 Résultats des analyses du THG avec T-REX Online des gènes, génomes, et domaine RB du SARS-Cov-2

Arbres des gènes		Nombre total de THG	Les THGs	
			De sous arbres	À sous arbres
E1	E2	2	Hu_Wuhan_2020, RaTG13	Bat_CoVZC45, Bat_CoVZXC21
			Bat_CoVZC45, Bat_CoVZXC21, Hu_Wuhan_2020, RaTG13	Guangdong_Pangolin_1_2019
E2	E3	6	BtCoV_273_2005, Rf1, Rs3367, SARS, SARS-CoV_BJ182-4, Tor2	BtCoV_279_2005
			Bat_CoVZC45, Bat_CoVZXC21	Hu_Australia_VIC231_2020, Hu_Italy_TE4836_2020, Hu_USA_UT_00346_2020, Hu_Wuhan_2020, RaTG13
			Guangdong_Pangolin_1_2019, Guangdong_Pangolin_P2S_2019	Bat_CoVZC45, Bat_CoVZXC21, Hu_Australia_VIC231_2020, Hu_Italy_TE4836_2020, Hu_USA_UT_00346_2020, Hu_Wuhan_2020, RaTG13
			Bat_CoVZC45, Bat_CoVZXC21, Guangdong_Pangolin_1_2019, Guangdong_Pangolin_P2S_2019, Hu_Australia_VIC231_2020, Hu_Italy_TE4836_2020, Hu_USA_UT_00346_2020, Hu_Wuhan_2020, RaTG13	Guangxi_Pangolin_P1E, Guangxi_Pangolin_P2V, Guangxi_Pangolin_P4L, Guangxi_Pangolin_P5E, Guangxi_Pangolin_P5L

			Bat_CoVZC45, Bat_CoVZXC21, Guangdong_Pangolin_1_2019, Guangdong_Pangolin_P2S_2019, Guangxi_Pangolin_P1E, Guangxi_Pangolin_P2V, Guangxi_Pangolin_P4L, Guangxi_Pangolin_P5E, Guangxi_Pangolin_P5L, Hu_Australia_VIC231_2020, Hu_Italy_TE4836_2020, Hu_USA_UT_00346_2020, Hu_Wuhan_2020, RaTG13	BM48_31_BGR_2008, BtKY72
			SARS, SARS-CoV_BJ182-4, Tor2	Rs3367
Génome complet 1	Génome complet 2	2	Hu_Wuhan_2020, RaTG13	Bat_CoVZC45, Bat_CoVZXC21
			Bat_CoVZC45, Bat_CoVZXC21, Hu_Wuhan_2020, RaTG13	Guangdong_Pangolin_1_2019
M2	M3	2	Hu_USA_UT_00346_2020	Hu_Australia_VIC231_2020
			BM48_31_BGR_2008, BtCoV_273_2005, BtCoV_279_2005, BtKY72, Guangxi_Pangolin_P1E, Guangxi_Pangolin_P2V, Guangxi_Pangolin_P4L, Guangxi_Pangolin_P5E, Guangxi_Pangolin_P5L, HKU3-12, HKU3-6, Rf1, Rs3367, SARS, SARS-CoV_BJ182-4, Tor2	Bat_CoVZC45, Bat_CoVZXC21, Hu_Australia_VIC231_2020, Hu_Italy_TE4836_2020, Hu_USA_UT_00346_2020, Hu_Wuhan_2020, RaTG1
N2	N3	4	Hu_Wuhan_2020	Hu_Australia_VIC231_2020
			Guangxi_Pangolin_P1E	Guangxi_Pangolin_P5L
			Hu_Australia_VIC231_2020, Hu_Wuhan_2020	Hu_Italy_TE4836_2020

			Guangxi_Pangolin_P4L	Guangxi_Pangolin_P1E, Guangxi_Pangolin_P5L
ORF1ab 1	ORF1ab 2	2	Hu_Wuhan_2020, RaTG13	Bat_CoVZC45, Bat_CoVZXC21
			Bat_CoVZC45, Bat_CoVZXC21, Hu_Wuhan_2020, RaTG13	Guangdong_Pangolin_1_2019
ORF1ab 2	ORF1AB 3	2	Guangxi_Pangolin_P4L	Guangxi_Pangolin_P5E
			Hu_Australia_VIC231_2020	Hu_Wuhan_2020
ORF3a 1	ORF3a 2	1	Bat_CoVZC45, Bat_CoVZXC21	Hu_Wuhan_2020, RaTG13
ORF3a 2	ORF3a 3	1	Hu_Wuhan_2020	Hu_Australia_VIC231_2020
S1	S2	2	Bat_CoVZC45, Bat_CoVZXC21, Hu_Wuhan_2020, RaTG13	Guangdong_Pangolin_1_2019
			Bat_CoVZC45, Bat_CoVZXC21	Hu_Wuhan_2020, RaTG13
S2	S3	1	Guangxi_Pangolin_P5L	Guangxi_Pangolin_P2V
Domaine RB 1	Domaine RB 2	1	Guangdong_Pangolin_1_2019, Guangdong_Pangolin_P2S_2019	Hu_Wuhan_2020
Domaine RB 2	Domaine RB 3	1	Guangxi_Pangolin_P2V	Guangxi_Pangolin_P5E
ORF6 1	ORF6 2	2	Hu_Wuhan_2020, RaTG13	Bat_CoVZC45, Bat_CoVZXC21
			Bat_CoVZC45, Bat_CoVZXC21, Hu_Wuhan_2020, RaTG13	Guangdong_Pangolin_1_2019
ORF7a 1	ORF7a 2	3	Bat_CoVZC45, Bat_CoVZXC21	RaTG13
			Bat_CoVZC45, Bat_CoVZXC21, RaTG13	Hu_Wuhan_2020
			Bat_CoVZC45, Bat_CoVZXC21, Hu_Wuhan_2020, RaTG13	Guangdong_Pangolin_1_2019

ORF8 1	ORF8 2	1	Bat_CoVZC45, Bat_CoVZXC21, Hu_Wuhan_2020, RaTG13	Guangdong_Pangolin_1_2019
ORF10 1	ORF10 2	2	Bat_CoVZXC21	Guangdong_Pangolin_1_2019, Guangdong_Pangolin_P2S_2019
			RaTG13	Hu_Wuhan_2020

4.2.1 Résultats de l'analyse de THG

Dans cette section, on va présenter les résultats obtenus à partir de l'analyse de 6, 25 et 43 séquences effectuée en utilisant le logiciel T-REX.

La figure 4.9 montre une détection de THG des 2 gènes : E1 qui contient 6 séquences et E2 qui contient 25 séquences de différentes espèces.

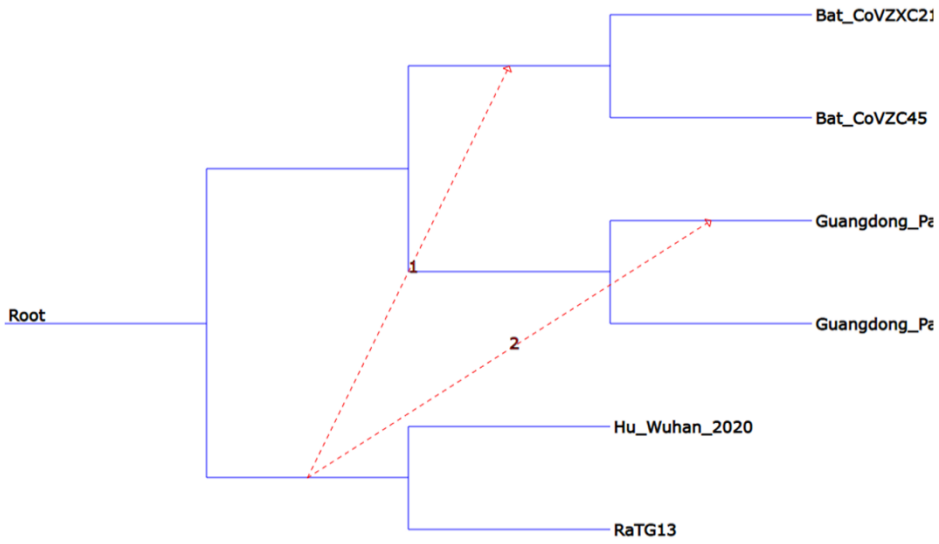

Figure 4.9 Événements de transfert horizontal de gènes de E1 et E2.

La figure 4.10 montre une détection de THG des 2 gènes : E2 qui contient 25 séquences et E3 qui contient 43 séquences de différentes espèces.

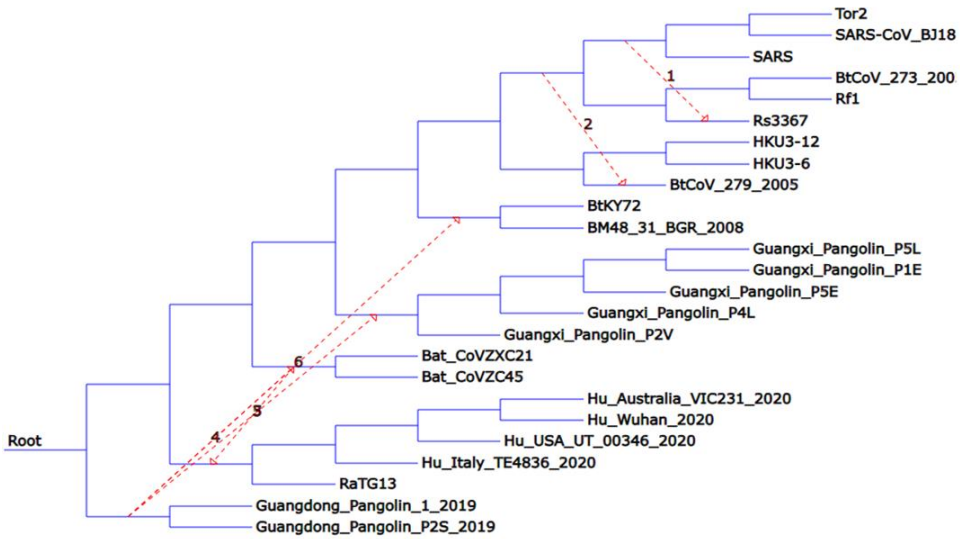

Figure 4.10 Événements de transfert horizontal de gènes de E2 et E3

La figure 4.11 montre une détection de THG des 2 génomes complets : génome complet 1 qui contient 6 séquences et génome complet 2 qui contient 25 séquences de différentes espèces.

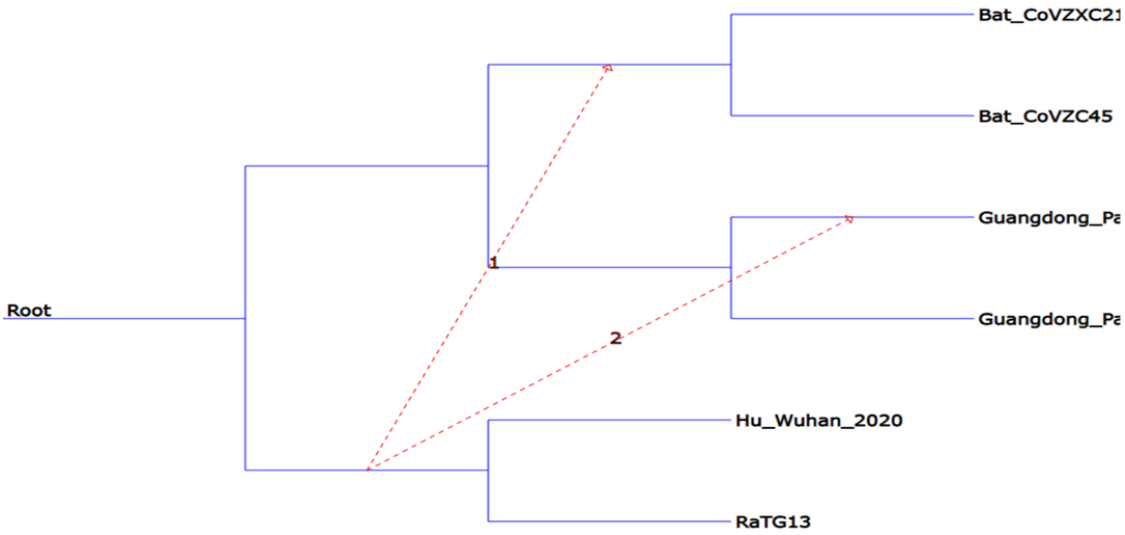

Figure 4.11 Événements de transfert horizontal de gènes des deux génomes complets

La figure 4.12 montre une détection de THG des 2 gènes : M 2 qui contient 25 séquences et M 3 qui contient 43 séquences de différentes espèces.

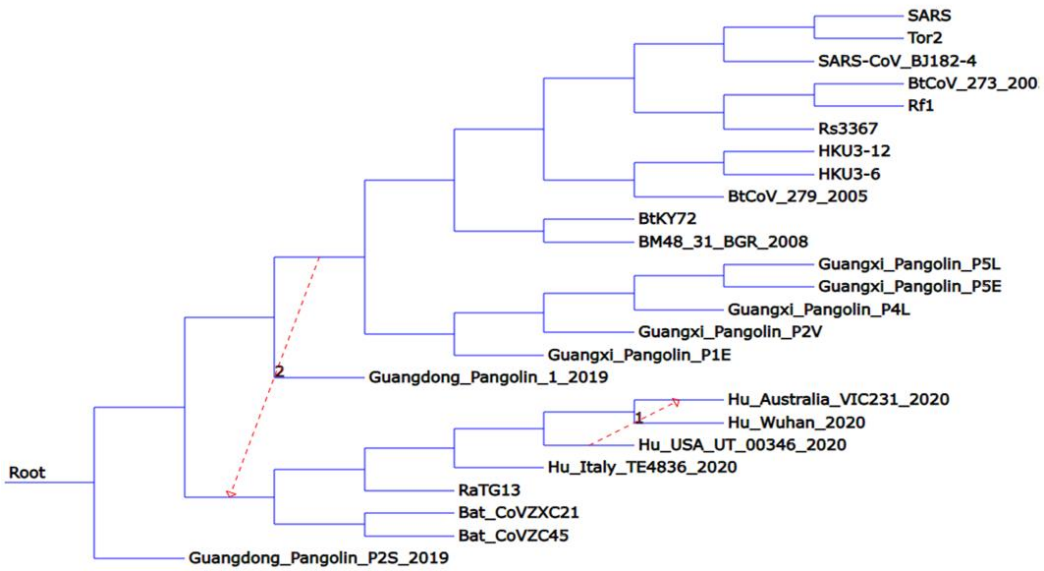

Figure 4.12 Événements de transfert horizontal de gènes de M2 et M3

La figure 4.13 montre une détection de THG des 2 gènes : N 2 qui contient 25 séquences et N 3 qui contient 43 séquences de différentes espèces.

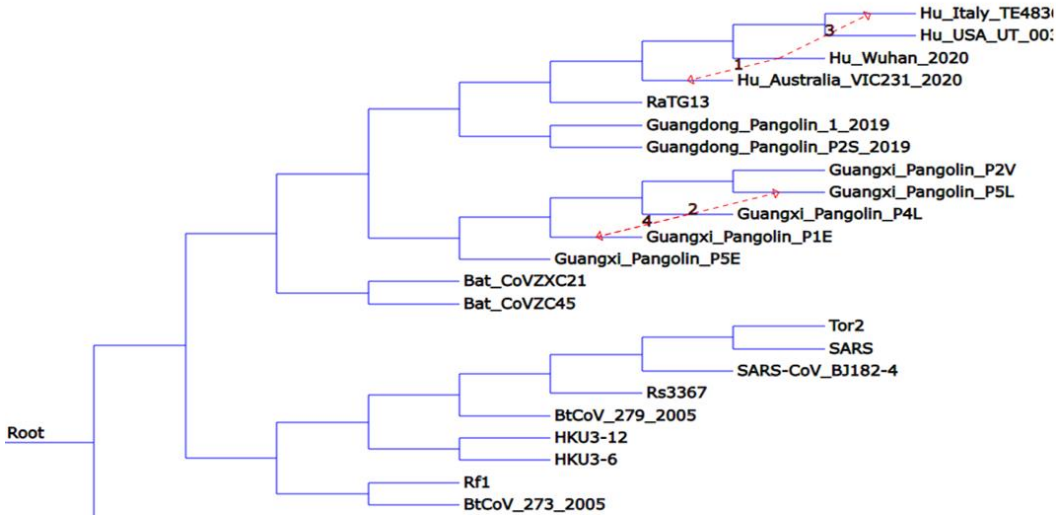

Figure 4.13 Événements de transfert horizontal de gènes de N2 et N3

La figure 4.14 montre une détection de THG des 2 gènes : ORF1ab 1 qui contient 6 séquences et ORF1ab 2 qui contient 25 séquences de différentes espèces.

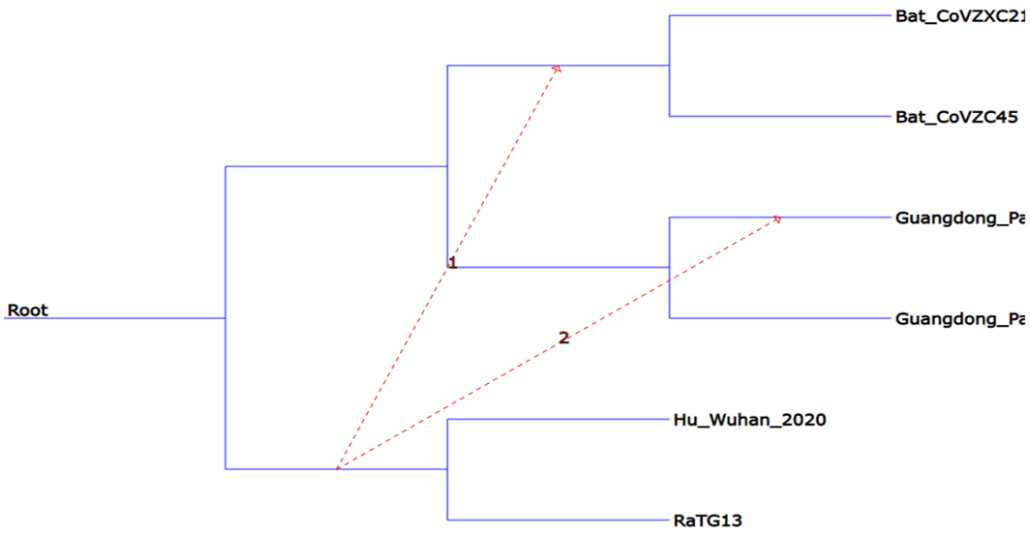

Figure 4.14 Événements de transfert horizontal de gènes de ORF1ab 1 et ORF1ab 2

La figure 4.15 montre une détection de THG des 2 gènes : ORF1ab 2 qui contient 25 séquences et ORF1ab 3 qui contient 43 séquences de différentes espèces.

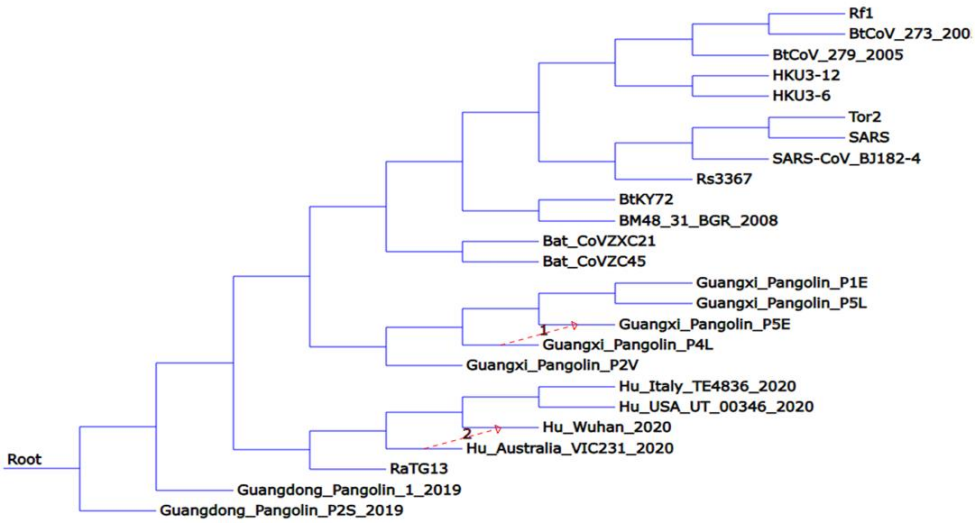

Figure 4.15 Événements de transfert horizontal de gènes de ORF1ab 2 et ORF1ab 3

La figure 4.16 montre une détection de THG des 2 gènes : ORF3a 1 qui contient 6 séquences et ORF3a 2 qui contient 25 séquences de différentes espèces.

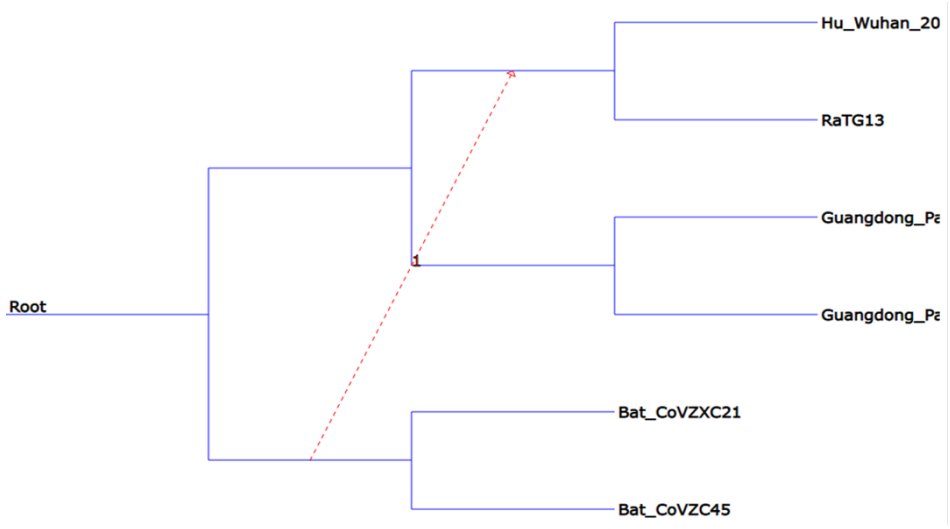

Figure 4.16 Événements de transfert horizontal de gènes de ORF3a et ORF3a 2

La figure 4.17 montre une détection de THG des 2 gènes : ORF3a 2 qui contient 25 séquences et OR3a 3 qui contient 43 séquences de différentes espèces.

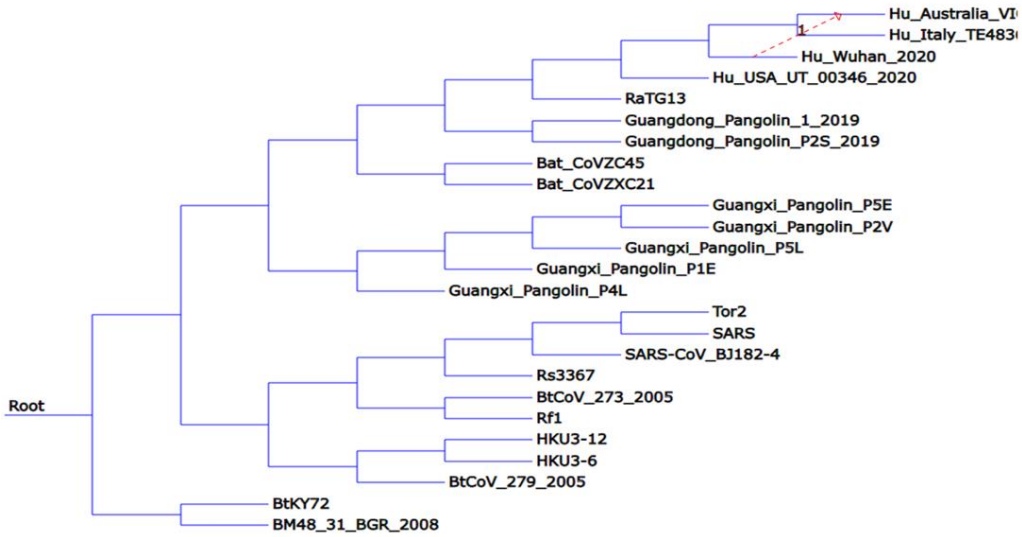

Figure 4.17 Événements de transfert horizontal de gènes de ORF3a 2 et ORF3a 3

La figure 4.18 montre une détection de THG des 2 gènes : S 1 qui contient 6 séquences et S 2 qui contient 25 séquences de différentes espèces.

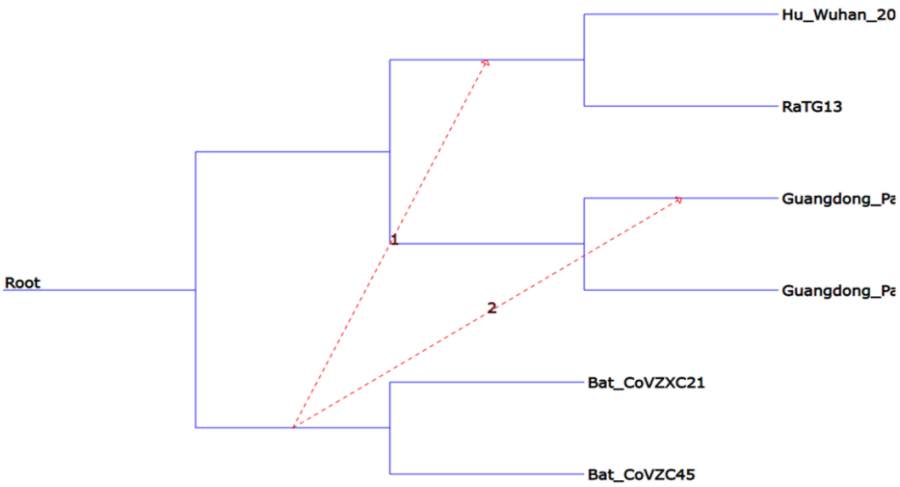

Figure 4.18 Événements de transfert horizontal de gènes de S 1 et S 2

La figure 4.19 montre une détection de THG des 2 gènes : S 2 qui contient 25 séquences et S 3 qui contient 43 séquences de différentes espèces.

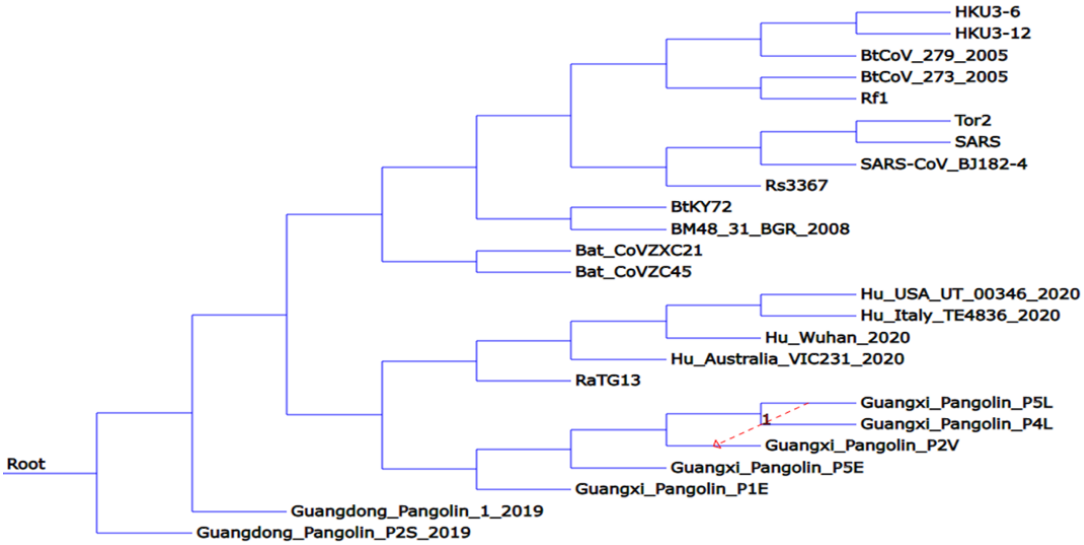

Figure 4.19 Événements de transfert horizontal de gènes de S 2 et S 3

La figure 4.20 montre une détection de THG des 2 domaines RB : domaine RB 1 qui contient 6 séquences et domaine RB 2 qui contient 25 séquences de différentes espèces.

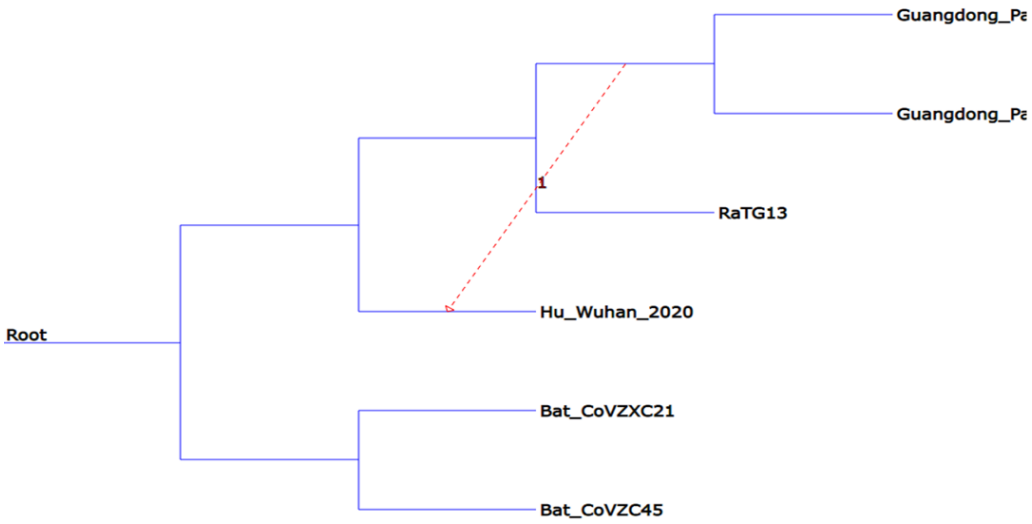

Figure 4.20 Événements de transfert horizontal de gènes de domaine RB 1 et domaine RB 2

La figure 4.21 montre une détection de THG des 2 domaines RB : domaine RB 2 qui contient 25 séquences et domaine RB 3 qui contient 43 séquences de différentes espèces.

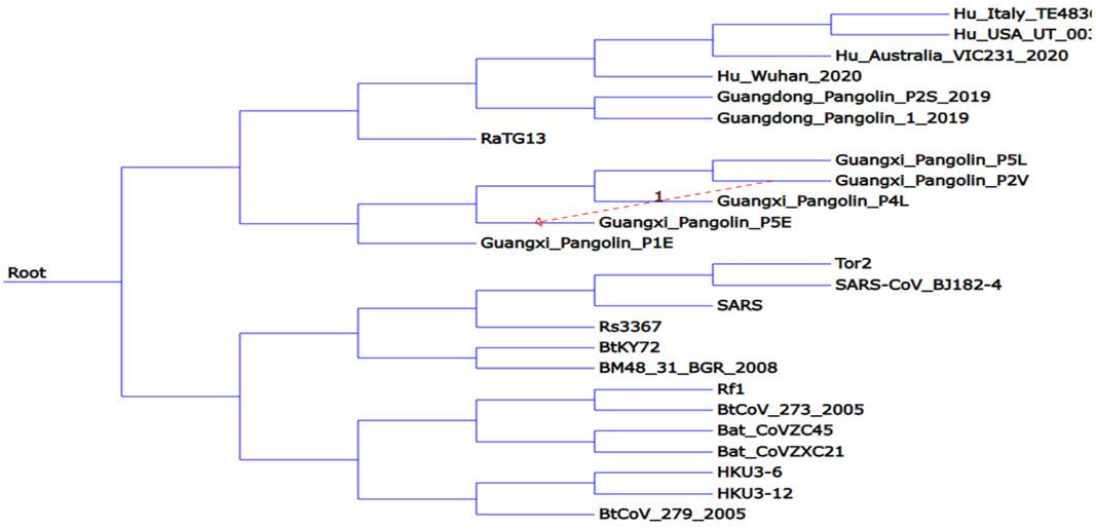

Figure 4.21 Événements de transfert horizontal de gènes des deux domaines RB (RB 3 ET RB 2)

La figure 4.22 montre une détection de THG des 2 gènes : ORF6 1 qui contient 6 séquences et ORF6 2 qui contient 25 séquences de différentes espèces.

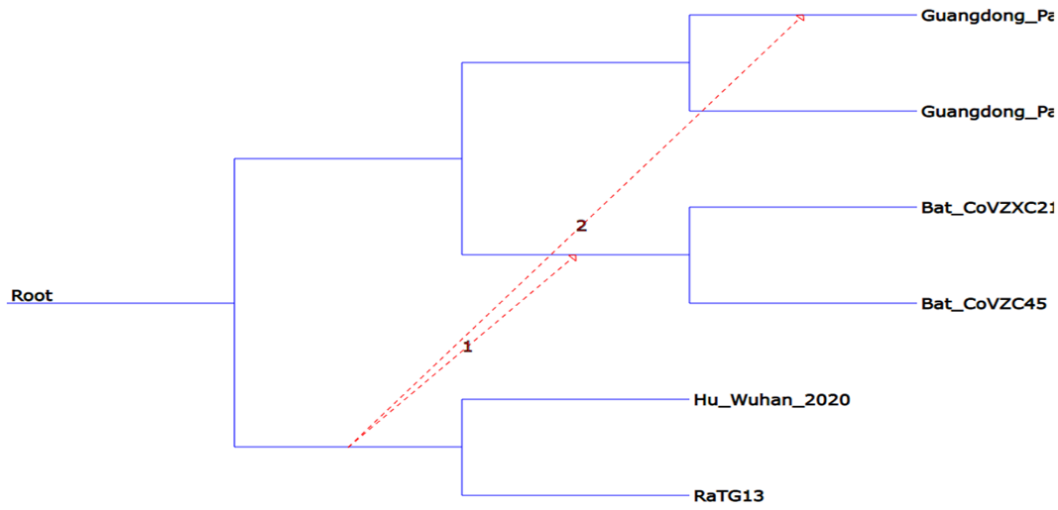

Figure 4.22 Événements de transfert horizontal de gènes de ORF6 1 et ORF6 2

La figure 4.23 montre une détection de THG des 2 gènes : ORF7a 1 qui contient 6 séquences et ORF7a 2 qui contient 25 séquences de différentes espèces.

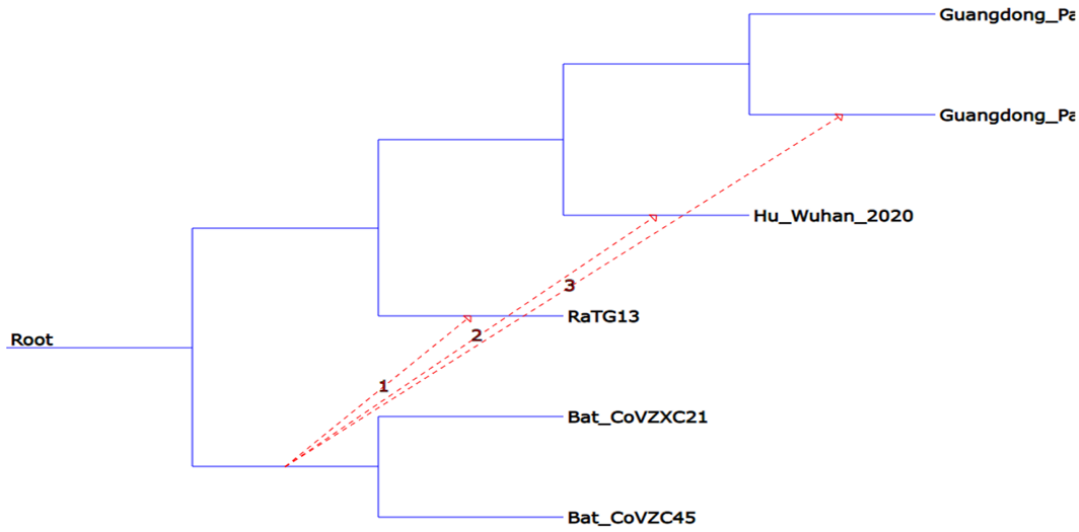

Figure 4.23 Événements de transfert horizontal de gènes DE ORF7a 1 et ORF7a 2

La figure 4.24 montre une détection de THG des 2 gènes : ORF8 1 qui contient 6 séquences et ORF8 2 qui contient 25 séquences de différentes espèces.

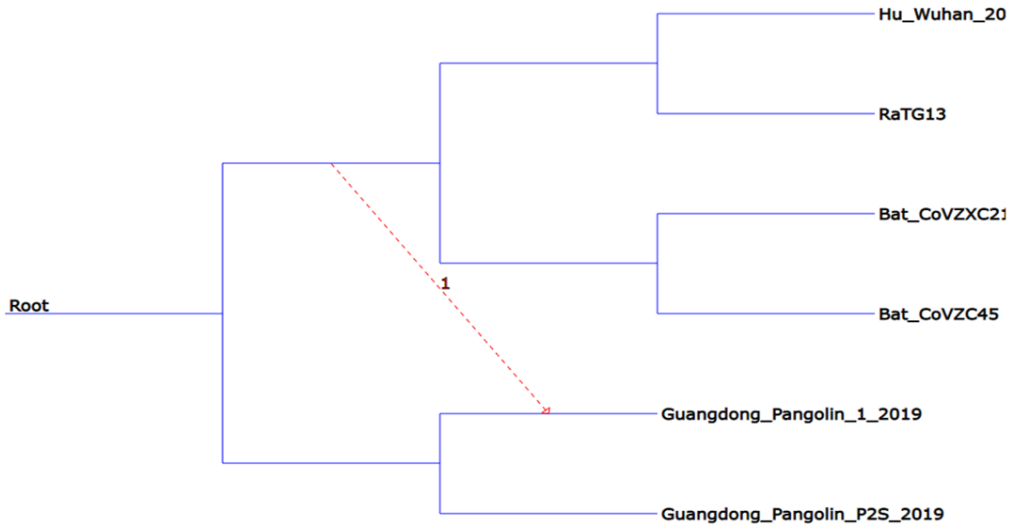

Figure 4.24 Événements de transfert horizontal de gènes ORF8 1 et ORF8 2

La figure 4.25 montre une détection de THG des 2 gènes : ORF10 1 qui contient 6 séquences et ORF10 2 qui contient 25 séquences de différentes espèces.

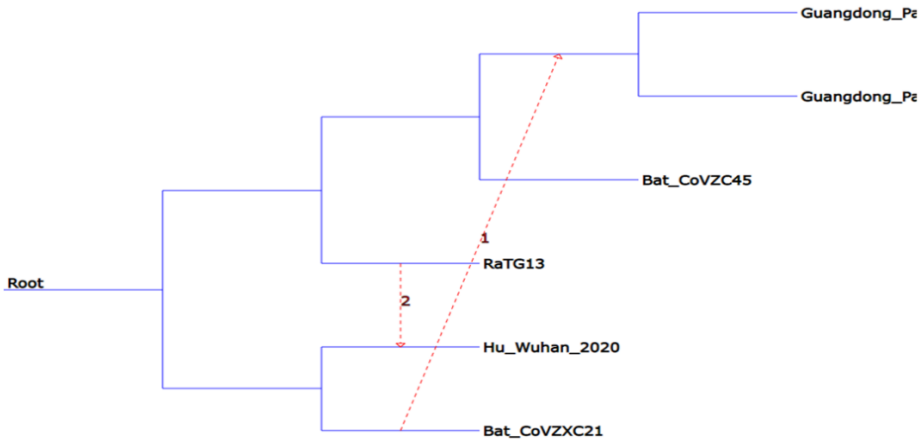

Figure 4.25 Événements de transfert horizontal de gènes ORF10 1 et ORF10 2

4.2.2 Interprétations des résultats obtenus par T-REX

Au regard des résultats obtenus précédemment par T-REX (le tableau et les arbres phylogénétiques qui présentent les THG des gènes, les génomes et les domaines RB du coronavirus humain), nous concluons que :

Le gène E

Des transferts horizontaux du gène E ont eu lieu entre différentes espèces. Un transfert de l'espèce humaine (Hu_Wuhan_2020) et de la chauve-souris (RaTG13) vers l'espèce de chauve-souris (Bat_CoVZC45, Bat_CoVZXC21). Également, il a été transféré de l'espèce de chauve-souris (Bat_CoVZC45, Bat_CoVZXC21, RaTG13) et l'humain (Hu_Wuhan_2020) vers l'espèce du pangolin (Guangdong_Pangolin_1_2019). De plus, des transferts horizontaux ont eu lieu entre l'espèce de chauve-souris (BtCoV_273_2005, Rf1, Rs3367) et l'humain (Tor2, SRAS, SARS-CoV_BJ182-4) vers l'espèce de chauve-souris (BtCoV_279_2005). En outre, il y a eu des transferts horizontaux entre l'espèce de chauve-souris (Bat_CoVZC45, Bat_CoVZXC21) vers l'espèce humaine (Hu_Australia_VIC231_2020, Hu_Italy_TE4836_2020, Hu_USA_UT_00346_2020, Hu_Wuhan_2020) et de chauve-souris (RaTG13). Enfin, des transferts horizontaux ont également été observés entre l'espèce du pangolin (Guangdong_Pangolin_1_2019, Guangdong_Pangolin_P2S_2019), l'espèce humaine (Hu_Australia_VIC231_2020, Hu_Italy_TE4836_2020, Hu_USA_UT_00346_2020, Hu_Wuhan_2020) et la chauve-souris (RaTG13, Bat_CoVZC45, Bat_CoVZXC21) vers l'espèce du pangolin (Guangxi_Pangolin_P1E, Guangxi_Pangolin_P2V, Guangxi_Pangolin_P4L, Guangxi_Pangolin_P5E, Guangxi_Pangolin_P5L). De plus, il y a eu des transferts horizontaux entre l'espèce humaine (SRAS, SARS-CoV_BJ182-4, Tor2) et l'espèce de chauve-souris (BM48_31_BGR_2008, BtKY72).

Le génome complet

Il y a eu un transfert horizontal du génome complet de l'espèce humaine (Hu_Wuhan_2020) et de l'espèce de chauve-souris (RaTG13) vers l'espèce de chauve-souris (Bat_CoVZC45, Bat_CoVZXC21). De plus, il a été transféré horizontalement de l'espèce humaine (Hu_Wuhan_2020) et de l'espèce de chauve-souris (RaTG13, Bat_CoVZC45, Bat_CoVZXC21) vers l'espèce du pangolin (Guangdong_Pangolin_1_2019).

Le gène M

Il y a eu un transfert horizontal de l'espèce humaine (Hu_USA_UT_00346_2020) vers l'espèce humaine (Hu_Australia_VIC231_2020) et un transfert horizontal de l'espèce du pangolin (Guangxi_Pangolin_P1E, Guangxi_Pangolin_P2V, Guangxi_Pangolin_P4L, Guangxi_Pangolin_P5E, Guangxi_Pangolin_P5L),

d'humain (SARS, SARS-CoV_BJ182-4, Tor2) et de chauve-souris (HKU3-12, HKU3-6, Rf1, Rs3367, BM48_31_BGR_2008, BtCoV_273_2005, BtCoV_279_2005, BtKY7) vers l'espèce de chauve-souris (Bat_CoVZC45, Bat_CoVZXC21, RaTG1) et d'humain (Hu_Australia_VIC231_2020, Hu_Italy_TE4836_2020, Hu_USA_UT_00346_2020, Hu_Wuhan_2020).

Le gène N

Le gène N a été sujet à des transferts horizontaux notables. Tout d'abord, il a été transféré de l'espèce humaine (Hu_USA_UT_00346_2020) à une autre souche humaine (Hu_Australia_VIC231_2020) et un transfert horizontal de l'espèce du pangolin (Guangxi_Pangolin_P1E) à une autre souche de pangolin (Guangxi_Pangolin_P5L). Également, on a constaté des transferts horizontaux de ce gène depuis différentes souches humaines (Hu_Australia_VIC231_2020, Hu_Wuhan_2020) vers une autre souche humaine (Hu_Italy_TE4836_2020). Enfin, des transferts horizontaux de gène N ont été observés de l'espèce du pangolin (Guangxi_Pangolin_P4L) vers diverses souches de pangolin (Guangxi_Pangolin_P1E, Guangxi_Pangolin_P5L).

Gène ORF1ab

Le transfert horizontal du gène ORF1ab s'est produit de diverses manières. Initialement, il s'est déplacé horizontalement de l'espèce humaine (Hu_Wuhan_2020) et de la chauve-souris (RaTG13) vers l'espèce de chauve-souris (Bat_CoVZC45, Bat_CoVZXC21). De plus, ce gène a également été transféré horizontalement de l'espèce humaine (Hu_Wuhan_2020) et de la chauve-souris (RaTG13, Bat_CoVZC45, Bat_CoVZXC21) vers l'espèce de pangolin (Guangdong_Pangolin_1_2019). De plus, nous avons constaté un transfert horizontal de l'espèce de pangolin (Guangxi_Pangolin_P4L) à l'espèce de pangolin (Guangxi_Pangolin_P5E). Enfin, il y a eu un transfert horizontal de l'espèce humaine (Hu_Australia_VIC231_2020) vers une autre souche humaine (Hu_Wuhan_2020).

Gène ORF3a

Le gène ORF3a a été transféré horizontalement de l'espèce de chauve-souris (Bat_CoVZC45, Bat_CoVZXC21) à l'espèce de chauve-souris (RaTG13) et de l'humain (Hu_Wuhan_2020). De plus, il a été transféré horizontalement de l'espèce humaine (Hu_Wuhan_2020) à une autre souche humaine (Hu_Australia_VIC231_2020).

Gène S

Le gène S a été transféré horizontalement de l'espèce humaine (Hu_Wuhan_2020) et de la chauve-souris (RaTG13, Bat_CoVZC45, Bat_CoVZXC21) à l'espèce du pangolin (Guangdong_Pangolin_1_2019). De plus, il a été transféré horizontalement de l'espèce de chauve-souris (Bat_CoVZC45, Bat_CoVZXC21) à l'espèce humaine (Hu_Wuhan_2020) et de la chauve-souris (RaTG13). Également, nous avons observé un transfert horizontal de l'espèce du pangolin (Guangxi_Pangolin_P5L) à l'espèce du pangolin (Guangxi_Pangolin_P2V).

Domaine RB

Le domaine RB a été transféré horizontalement de l'espèce du pangolin (Guangdong_Pangolin_1_2019, Guangdong_Pangolin_P2S_2019) à l'espèce humaine (Hu_Wuhan_2020). De plus, il a été transféré horizontalement de l'espèce du pangolin (Guangxi_Pangolin_P2V) à l'espèce du pangolin (Guangxi_Pangolin_P5E).

Gène ORF6

Le gène ORF6 a été transféré horizontalement de l'espèce humaine (Hu_Wuhan_2020) et de chauve-souris (RaTG13, Bat_CoVZC45, Bat_CoVZXC21) à l'espèce du pangolin (Guangdong_Pangolin_1_2019). De plus, il a été transféré horizontalement de l'espèce humaine (Hu_Wuhan_2020) et de chauve-souris (RaTG13) à l'espèce de chauve-souris (Bat_CoVZC45, Bat_CoVZXC21).

Gène ORF7a

Le gène ORF7a a été transféré horizontalement de l'espèce de chauve-souris (Bat_CoVZC45, Bat_CoVZXC21) à l'espèce de chauve-souris (RaTG13). De plus, il a été transféré horizontalement de l'espèce de chauve-souris (Bat_CoVZC45, Bat_CoVZXC21, RaTG13) à l'espèce humaine (Hu_Wuhan_2020). Enfin, il a été transféré horizontalement de l'espèce de chauve-souris (Bat_CoVZC45, Bat_CoVZXC21, RaTG13) et d'humain (Hu_Wuhan_2020) à l'espèce du pangolin (Guangdong_Pangolin_1_2019).

Gène ORF8

Le gène ORF8 a été transféré horizontalement de l'espèce de chauve-souris (Bat_CoVZC45, Bat_CoVZXC21, RaTG13) et d'humain (Hu_Wuhan_2020) à l'espèce du pangolin (Guangdong_Pangolin_1_2019).

Gène ORF10

Le gène ORF10 a été transféré horizontalement de l'espèce de chauve-souris Bat_CoVZXC21 à l'espèce du pangolin Guangdong_Pangolin_1_2019, Guangdong_Pangolin_P2S_2019. Il a également été transféré horizontalement de l'espèce de chauve-souris RaTG13 à l'espèce humaine Hu_Wuhan_2020.

4.3 Analyse de recombinaison

L'analyse de recombinaison permet de mesurer la différence entre les séquences génomiques de deux espèces ou de deux individus, et peut être utilisée pour détecter les régions du génome où la recombinaison a eu lieu.

Dans cette analyse, des séquences ont été regroupées en fonction de leurs espèces. Ensuite, le modèle de distance de Hamming a été appliqué, lequel est couramment utilisé pour construire des graphes et des réseaux phylogénétiques à travers le logiciel SimPlot++. Cette approche bioinformatique vise à évaluer la divergence génétique entre deux séquences génomiques, elle quantifie le nombre de positions ou les nucléotides diffèrent au sein des deux séquences. En d'autres termes, elle permet de dénombrer les substitutions de nucléotides requises pour transformer l'une des séquences en l'autre.

La figure 4.26 représente une interface SimPlot++ pour une classification des groupes du gène E du SARS-Cov-2 à partir de 6 séquences extraites de différentes espèces.

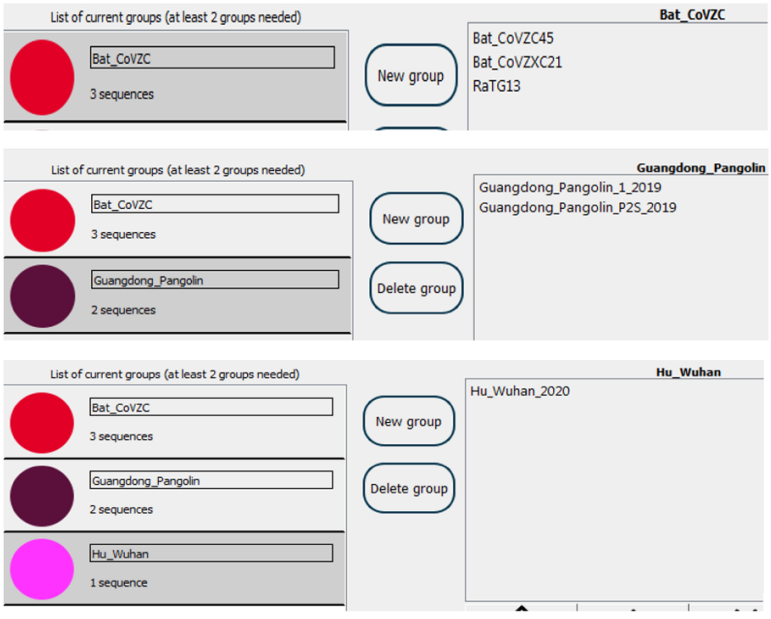

Figure 4.26 Interface SimPlot ++ pour une classification des groupes du gène E

La figure 4.27 représente une interface SimPlot++ qui montre les paramètres utilisés pour construire le graphe SimPlot++.

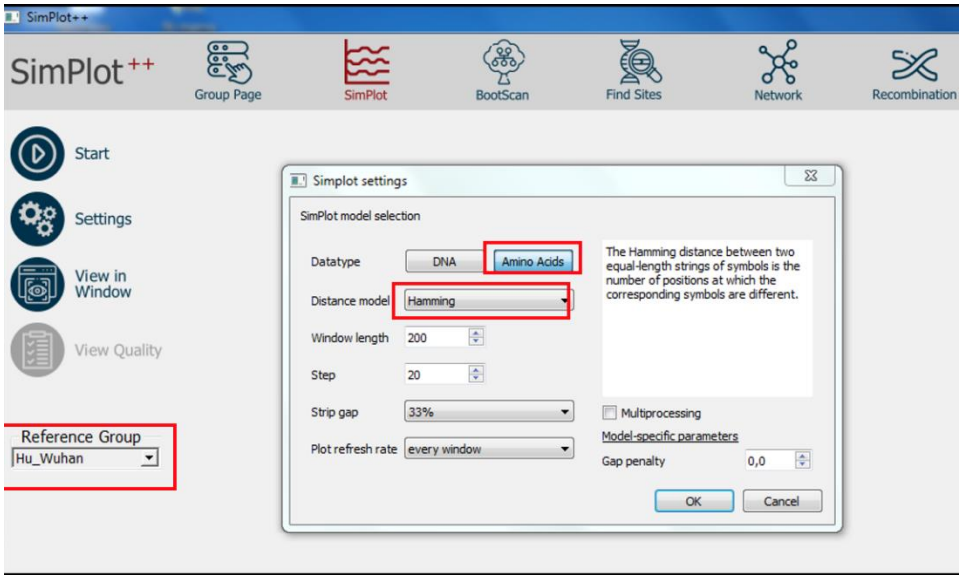

Figure 4.27 Interface du logiciel SimPlot++

4.3.1 Résultats de l'analyse de recombinaison

Dans cette section, on va présenter les graphes et les réseaux générés par le logiciel SimPlot++ des différents gènes du coronavirus humain. Ces modèles ont été construits à partir de 43 séquences mosaïques d'espèces différentes, tirant leur clarté des distances de Hamming. Cette visualisation offre une fenêtre incroyable sur les similitudes et les différences entre ces séquences génomiques, révélant des informations uniques sur l'évolution et la diversité bactériennes.

Gène E

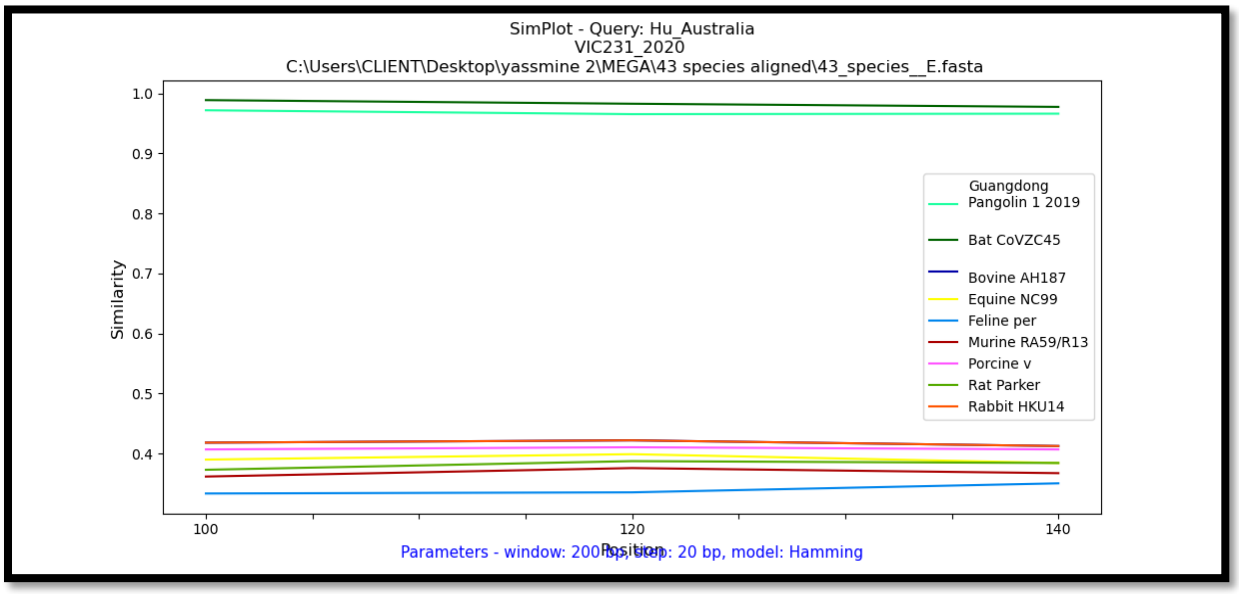

Figure 4.28 Graphe SimPlot++ du coronavirus pour le gène E

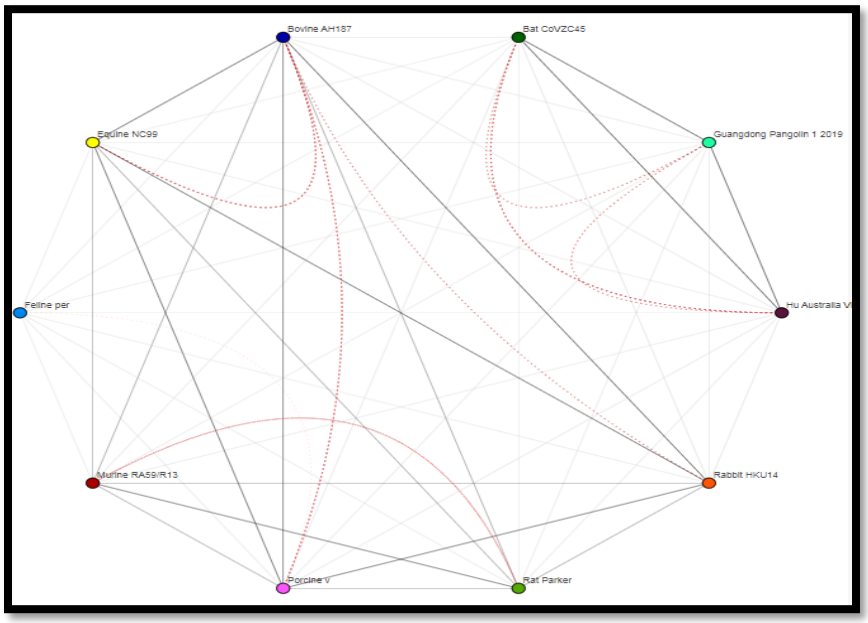

Figure 4.29 Réseau SimPlot++ du coronavirus pour le gène E

Gène M

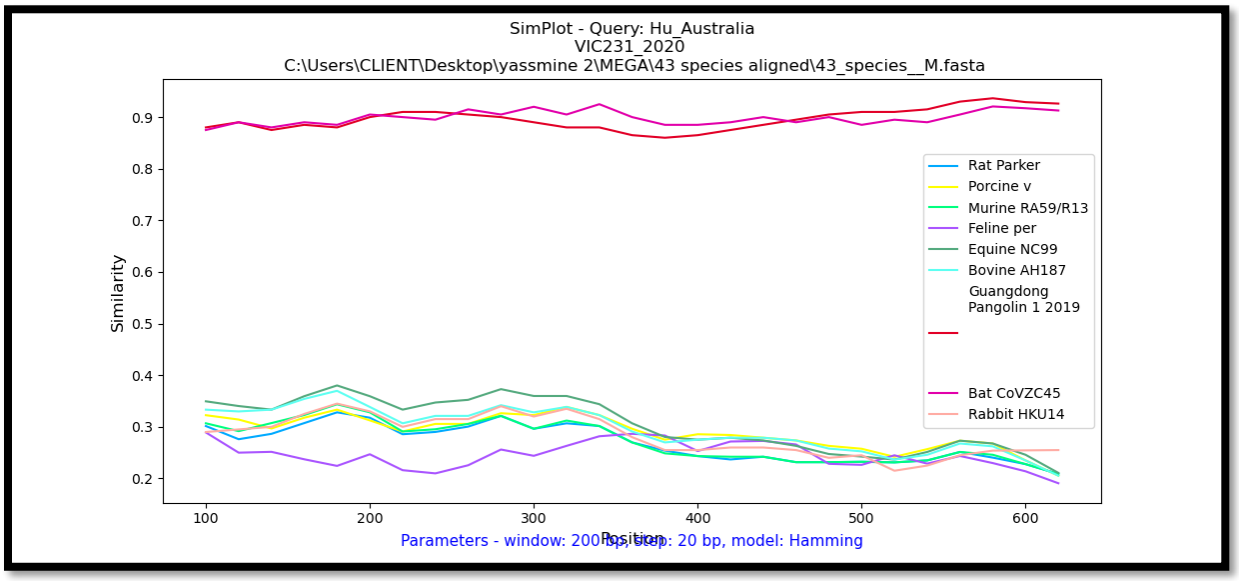

Figure 4.30 Graphe SimPlot++ du coronavirus pour le gène M

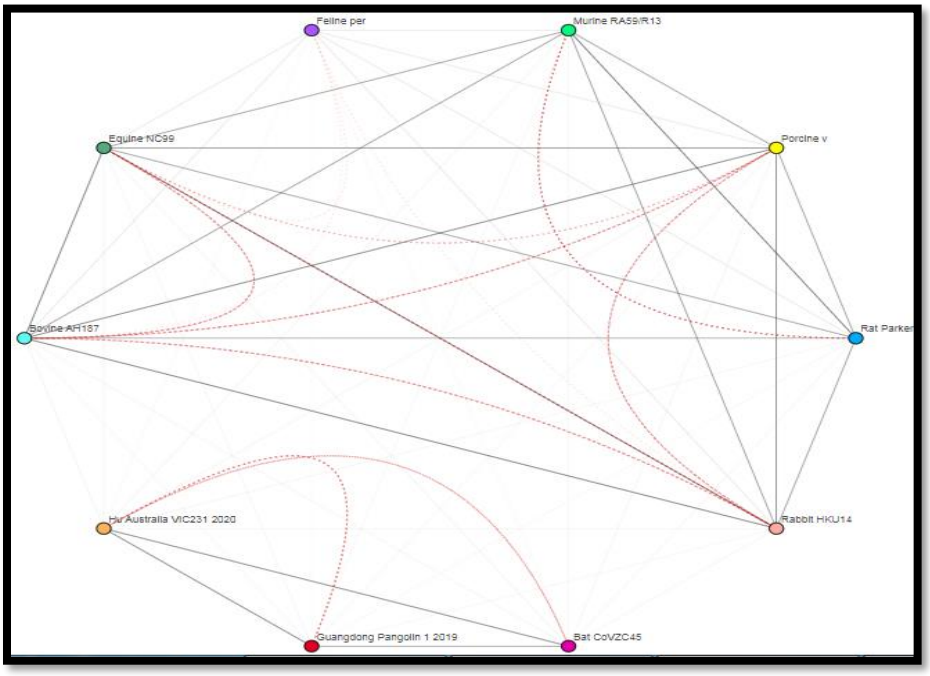

Figure 4.31 Réseau SimPlot++ du coronavirus pour le gène M

Gène N

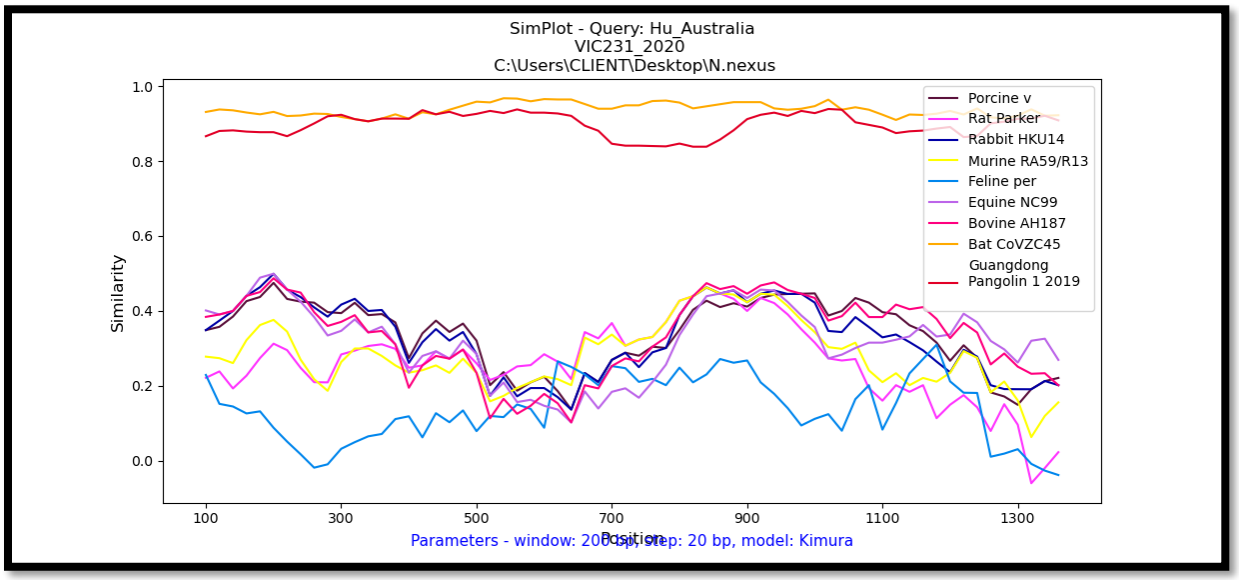

Figure 4.32 Graphe SimPlot++ du coronavirus pour le gène N

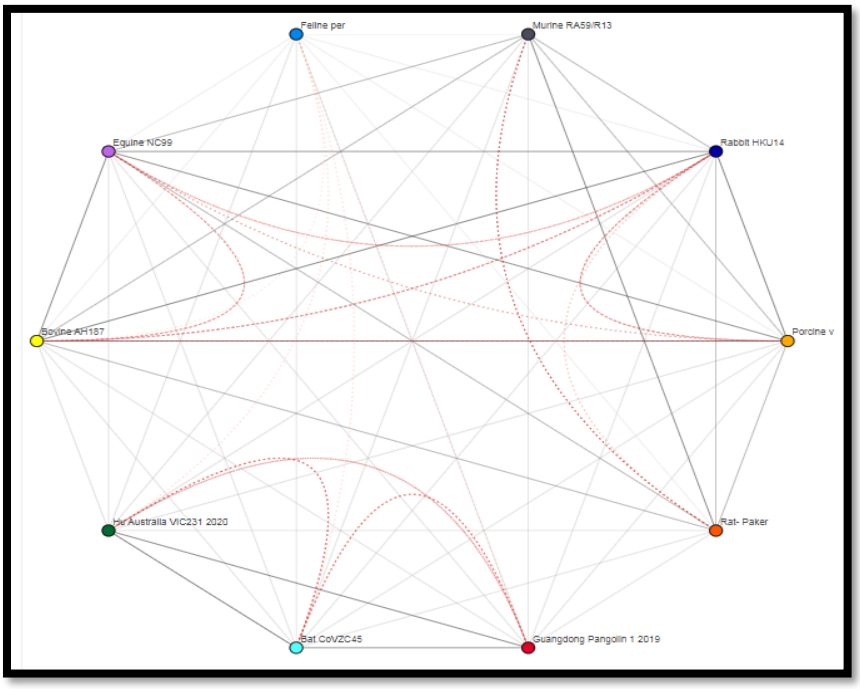

Figure 4.33 Réseau SimPlot++ du coronavirus pour le gène N

Gène N2

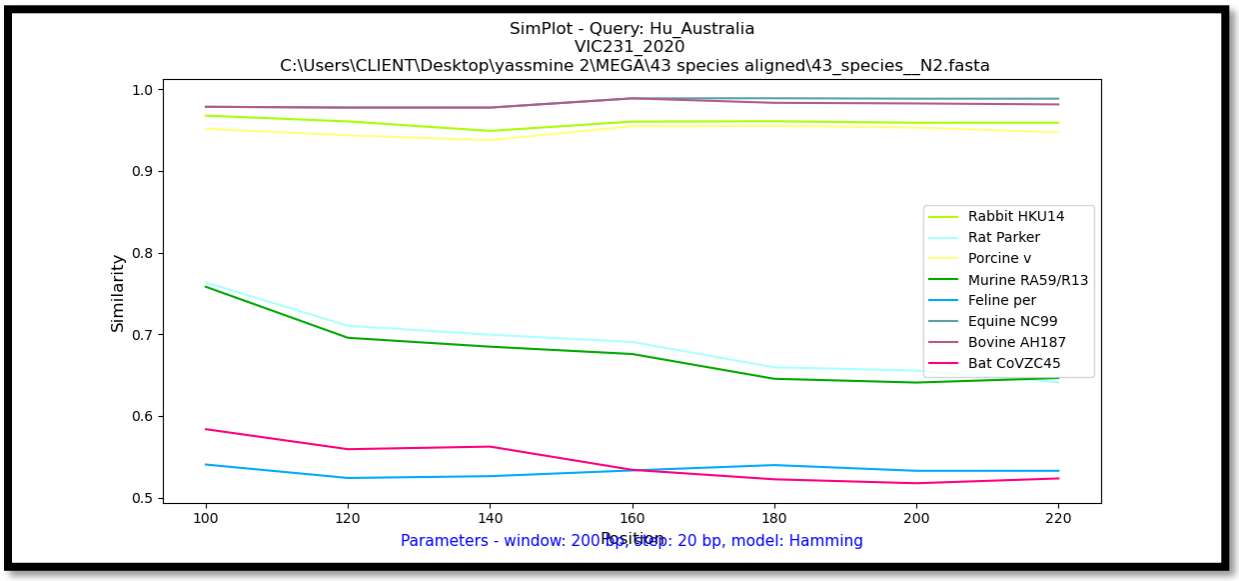

Figure 4.34 Graphe SimPlot++ du coronavirus pour le gène N2

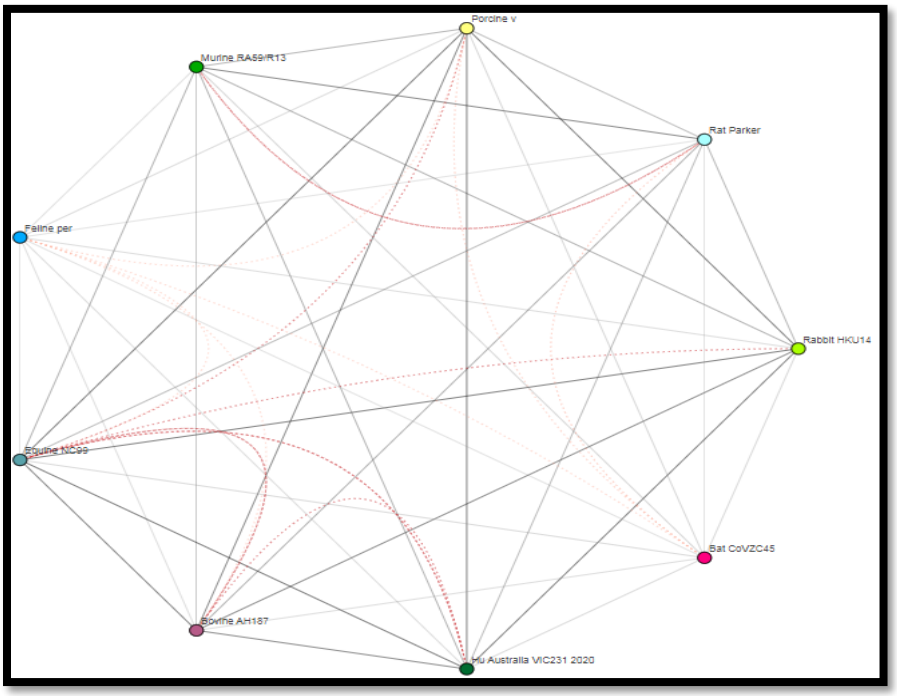

Figure 4.35 Réseau SimPlot++ du coronavirus pour le gène N2

Gène ORF1ab

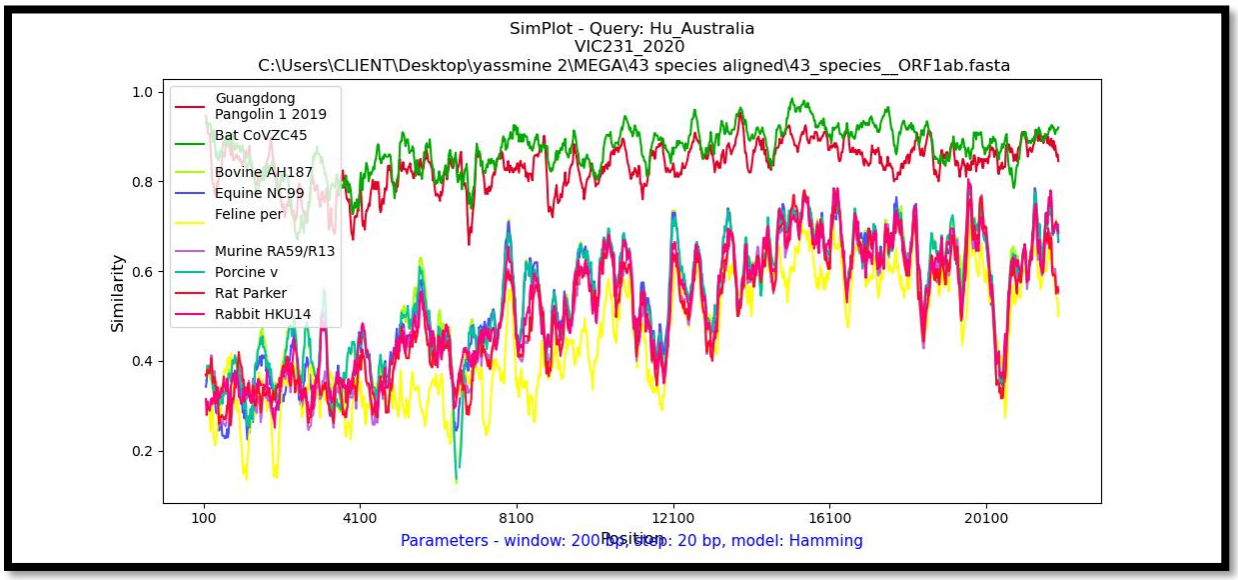

Figure 4.36 Graphe SimPlot++ du coronavirus pour le gène ORF1ab

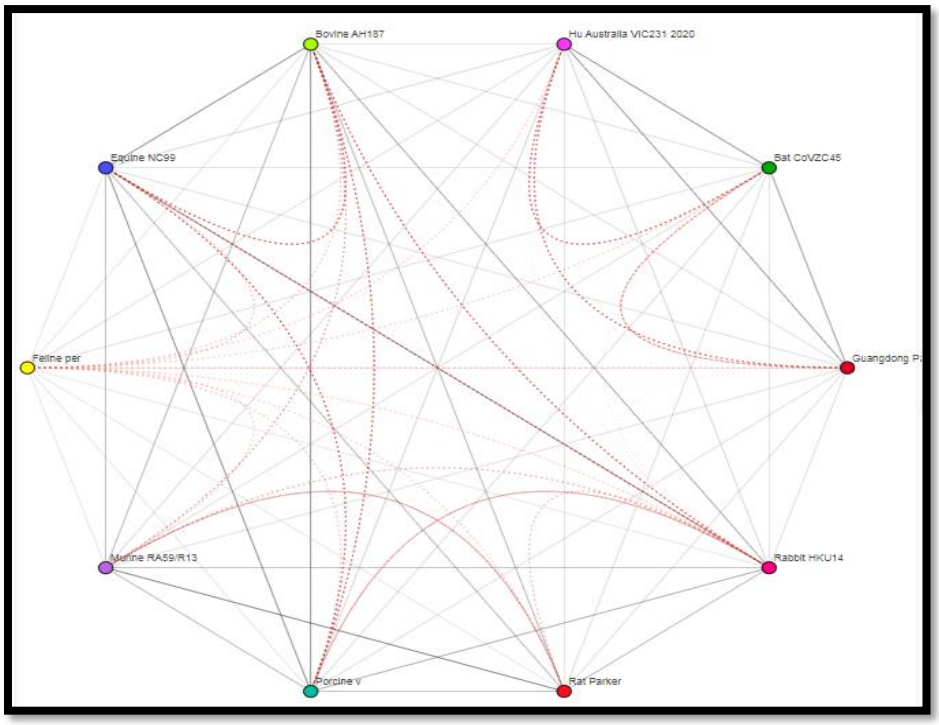

Figure 4.37 Réseau SimPlot++ du coronavirus pour le gène ORF1ab

Gène ORF3a

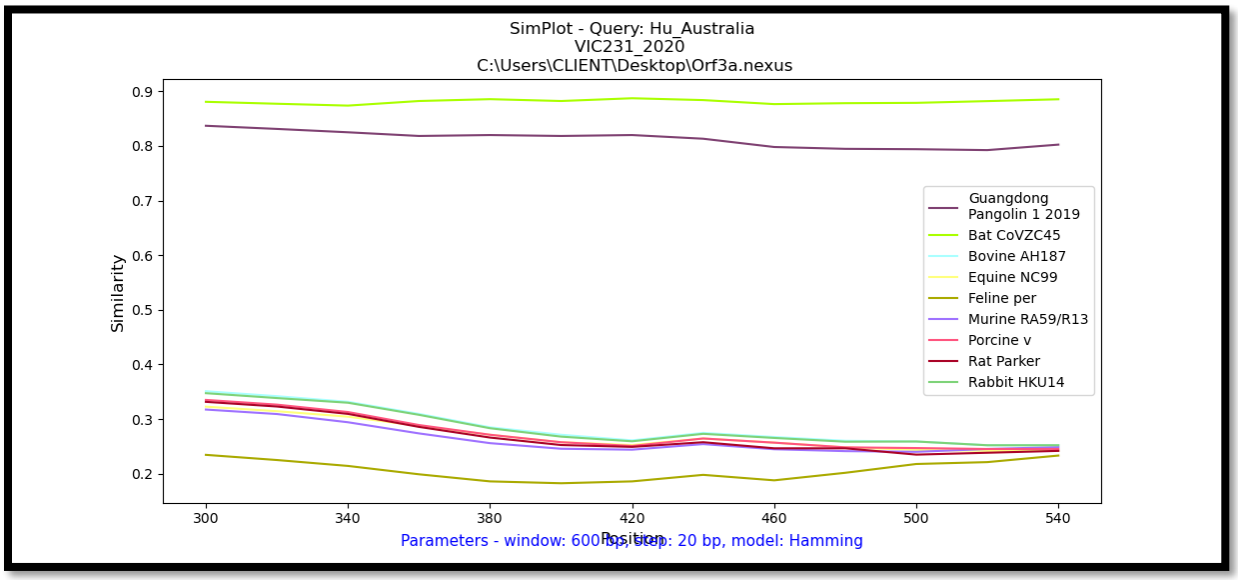

Figure 4.38 Graphe SimPlot++ du coronavirus pour le gène ORF3a

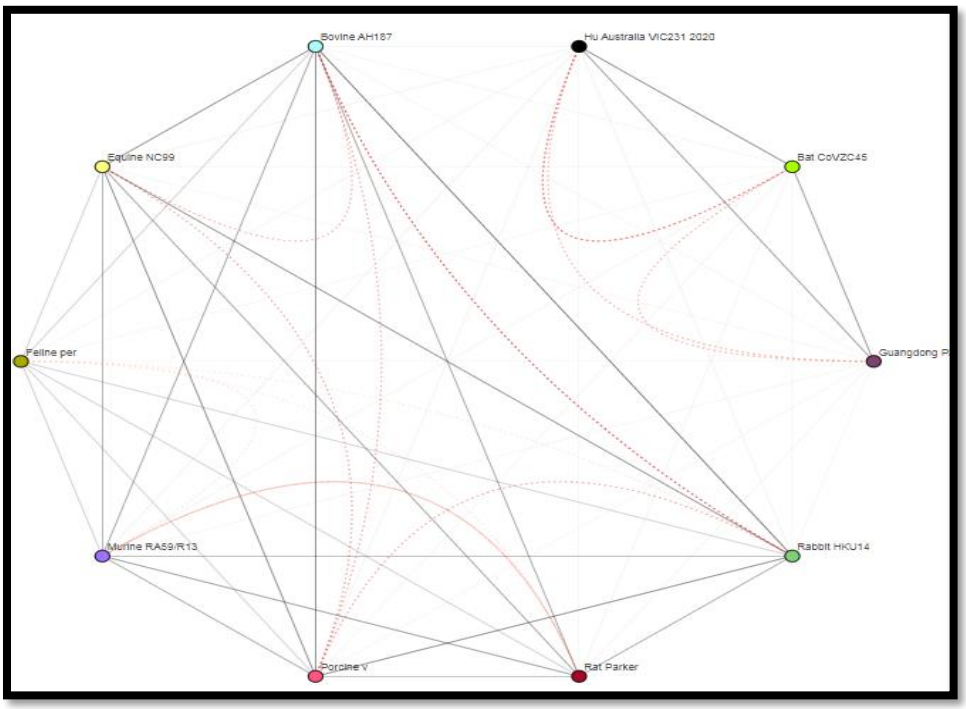

Figure 4.39 Réseau SimPlot++ du coronavirus pour le gène ORF3a

Gène S

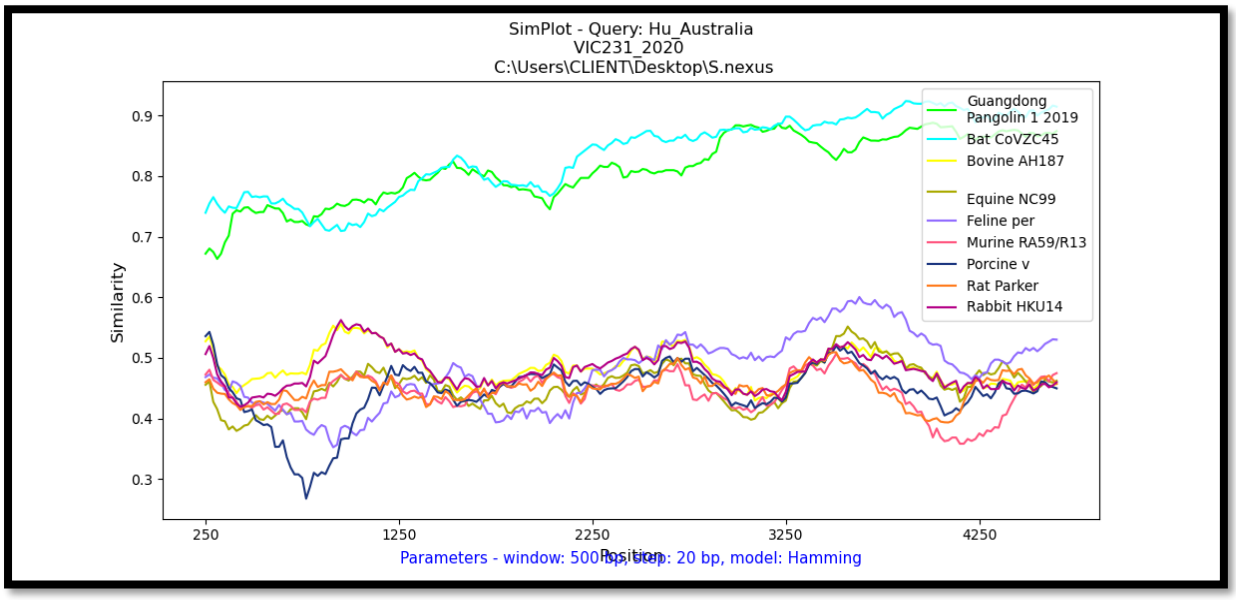

Figure 4.40 Graphe SimPlot++ du coronavirus pour le gène S

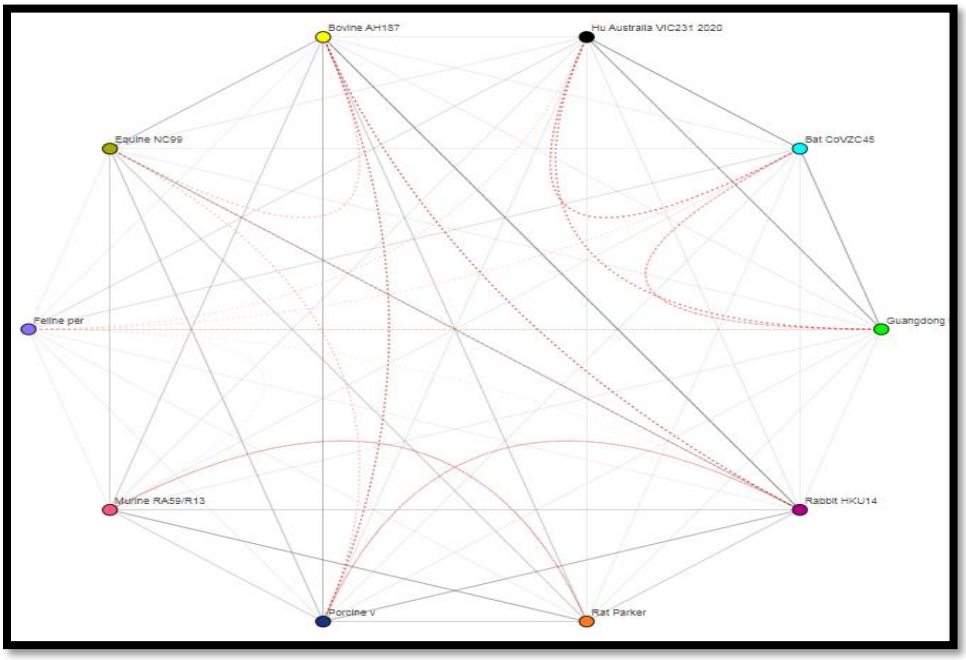

Figure 4.41 Réseau SimPlot++ du coronavirus pour le gène S

Génome complet

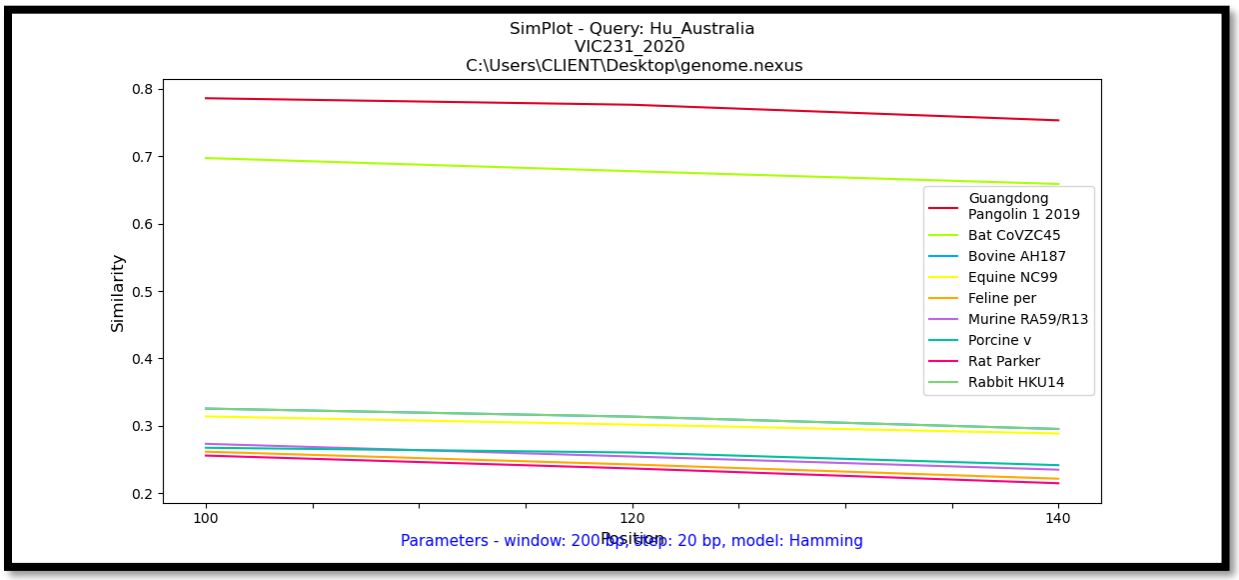

Figure 4.42 Graphe SimPlot++ du coronavirus pour le génome complet

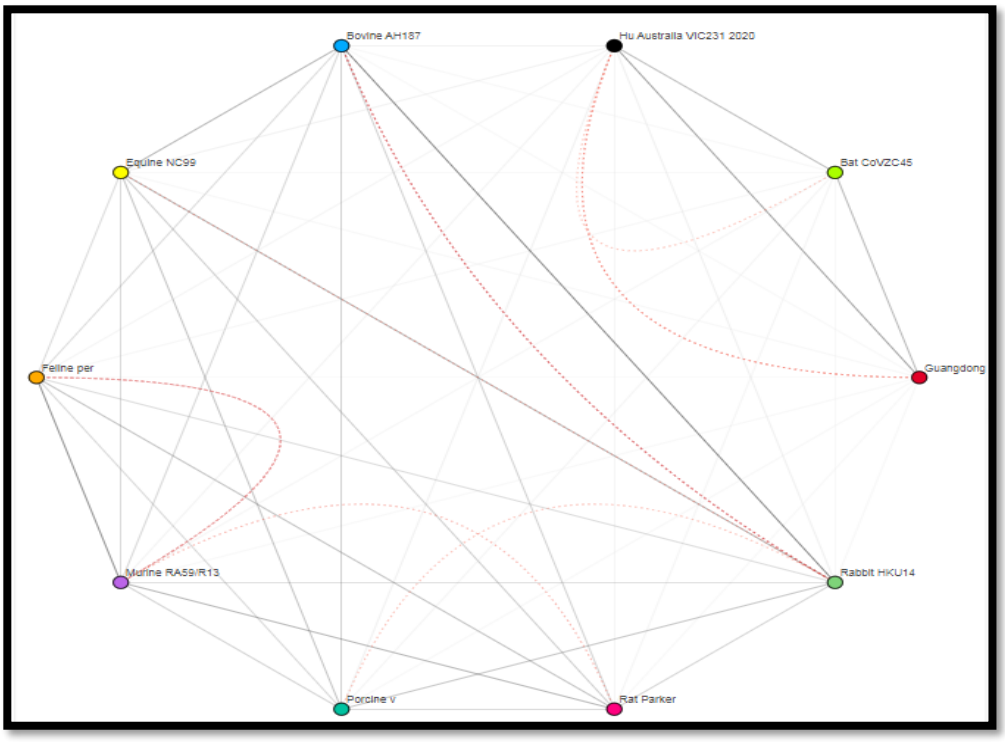

Figure 4.43 Réseau SimPlot++ du coronavirus pour le génome complet

4.3.2 Interprétation des résultats obtenus par SimPlot++

Dans cette section, on va analyser les graphiques et les réseaux résultants de l'application de la distance de Hamming par SimPlot++. Cette analyse se base sur l'interaction et la divergence génétique entre les séquences pour pouvoir interpréter et comprendre les variations génétiques présentes au sein de 43 séquences des différentes espèces.

Gène E

Le Graphe (Figure 4.28) montre que la divergence génétique entre les séquences de tous les groupes est différente, car les points sont de hauteurs différentes. La divergence génétique la plus élevée est celle des deux groupes du pangolin et de la chauve-souris. Elle est plus élevée dans le groupe de chauves-souris par rapport à celui du pangolin, mais reste presque similaire. Pour les 6 autres groupes, la divergence génétique est plus basse. Le taux le plus élevé de ces derniers est celui du groupe du lapin avec presque 0,42 de similarités et le plus bas est celui du groupe du chat.

Le Réseau (Figure 4.29) montre que les similarités globales sont égales à 75 % entre les groupes. La divergence génétique est élevée pour les groupes des chauves-souris, d'humains, du cheval, du veau et du lapin, car les lignes entre eux sont longues. La divergence génétique entre les séquences des groupes de chevaux, de veaux, humains et pangolins, chauve-souris est faible, car les lignes entre elles sont courtes. Les régions de recombinaison sont représentées par des lignes brisées.
Les lignes brisées entre deux points du réseau : groupe du pangolin et de chauve-souris, groupe de l'humain et du pangolin et groupe de l'humain et de chauve-souris, groupe du cheval et du veau, groupe du veau et du lapin et groupe du veau et du porc, signifient que les séquences génomiques reliées par ces lignes changent les régions de leurs génomes par recombinaison. C'est également le cas pour le groupe du chat et du porc où la ligne brisée est de couleur claire.

Gène M

Graphe (Figure 4.30) monte que la divergence génétique entre les séquences de tous les groupes est différente, car les points sont de hauteurs différentes. La divergence génétique est la plus élevée et presque similaire dans les deux groupes du pangolin et de la chauve-souris. Les 6 autres groupes ont une

divergence génétique plus basse et presque similaire. Le taux le plus élevé de ces derniers est celui du groupe du cheval et le plus bas est celui du groupe du chat.

Réseau (Figure 4.31) montre que les similarités globales sont égales à 75 % entre les groupes. La divergence génétique est élevée pour les groupes de chauves-souris, d'humains, du cheval, du veau, du lapin, de la souris et du porc, car les lignes entre eux sont longues. La divergence génétique entre les séquences des groupes du cheval, du veau et du pangolin, de l'humain et du pangolin, et de la chauve-souris est faible, car les lignes entre elles sont courtes. Les régions de recombinaison sont représentées par des lignes brisées. Les lignes brisées entre deux points du réseau : groupe du pangolin et de la chauve-souris, groupe de l'humain et du pangolin et groupe de l'humain et de la chauve-souris, groupe du cheval et du veau, groupe du veau et du lapin et groupe du veau et du porc, groupe du veau et du porc, groupe du porc et du lapin, groupe du porc et du cheval, rate, souris, signifient que les séquences génomiques reliées par ces lignes changent les régions de leurs génomes par recombinaison. C'est également le cas pour le groupe du chat, du veau et chat, du cheval et chat, et du lapin où la ligne brisée est de couleur claire.

Gène N

Graphe (Figure 4.32) montre que la divergence génétique entre les séquences de tous les groupes est différente, car les points sont de hauteurs différentes. La divergence génétique la plus élevée est celle des deux groupes du pangolin et de la chauve-souris ; elle est plus élevée dans le groupe de la chauve-souris que dans celui du pangolin, mais est presque similaire. Les 6 autres groupes ont la divergence génétique la plus basse où le taux le plus bas est celui du groupe du chat.

Réseau (Figure 4.33) montre que les similarités globales sont égales à 75 % entre les groupes. La divergence génétique est élevée pour les groupes du pangolin, de l'humain, du cheval, du veau, du lapin, de la souris, et du porc, car les lignes entre eux sont longues. La divergence génétique entre les séquences des groupes du cheval, veau et chauve-souris, humain et pangolin et lapin, porc et porc sont faibles, car les lignes entre elles sont courtes. Les régions de recombinaison sont représentées par des lignes brisées. Les lignes brisées entre deux points (groupe du pangolin et de la chauve-souris) (groupe de l'humain et du pangolin) et (groupe de l'humain et de la chauve-souris) (groupe du cheval et du veau) (groupe du veau et du lapin) et (groupe du veau et du porc) (groupe du porc, et du lapin) (groupe porc, cheval) (rate, souris)

du réseau, signifient que les séquences génomiques reliées par ces lignes changent les régions de leurs génomes par recombinaison.

Gène N2

Graphe (Figure 4.34) montre que la divergence génétique entre les séquences de tous les groupes est différente, car les points sont de hauteurs différentes. La divergence génétique la plus élevée est celle des deux groupes du cheval et du veau, suivie par celle des groupes du lapin et du porc qui sont tous presque similaires. Quant aux 4 autres groupes ; les souris et des rates ont une divergence génétique moyenne, alors que ceux du chat et de la chauve-souris ont la divergence génétique la plus basse.

Réseau (Figure 4.35) montre que les similarités globales sont égales à 75 % entre les groupes. La divergence génétique est élevée pour les groupes du cheval et de veau et du lapin et porcs, car les lignes entre eux sont longues. Les régions de recombinaison sont représentées par des lignes brisées. Les lignes brisées entre deux points du réseau : groupe de l'humain et veau, groupe de l'humain et cheval et groupe du veau et cheval, groupe du cheval et lapin, groupe de rate et souris et groupe du cheval et porc, groupe de la chauve-souris et porc, groupe de la chauve-souris et du chat, groupe du porc, cheval, et groupe rate, chauve-souris, signifient que les séquences génomiques reliées par ces lignes changent les régions de leurs génomes par recombinaison.

Gène ORF1ab

Graphe (Figure 4.36) montre que la divergence génétique entre les séquences de tous les groupes est différente, car les points sont de hauteurs différentes. La divergence génétique la plus élevée est celle des deux groupes du pangolin et de la chauve-souris. Elle est plus élevée dans le groupe de chauves-souris par rapport à celui du pangolin, mais presque similaire. Les 6 autres groupes ont la divergence génétique la plus basse.

Réseau (Figure 4.37) montre que les similarités globales sont égales à 75 % entre les groupes. La divergence génétique est élevée pour les groupes du pangolin, humains, chevaux, veaux, lapins, et porcs, car les lignes entre eux sont longues. La divergence génétique entre les séquences des groupes du cheval, veau et chauve-souris, humain et pangolin, chauve-souris est faible, car les lignes entre elles sont courtes.

Les régions de recombinaison sont représentées par des lignes brisées. Les lignes brisées entre deux points du réseau : groupe du pangolin et chauve-souris, groupe de l'humain et du pangolin, groupe de l'humain et de chauve-souris, groupe du cheval et veau, groupe du veau et lapin, groupe veau et porc, groupe veau et cheval, groupe porc et lapin, groupe porc et cheval, signifient que les séquences génomiques reliées par ces lignes changent les régions de leurs génomes par recombinaison.

Gène ORF3a

Graphe (Figure 4.38) montre que la divergence génétique entre les séquences de tous les groupes est différente, car les points sont de hauteurs différentes. La divergence génétique la plus élevée est celle des deux groupes du pangolin et de la chauve-souris. Elle est plus élevée dans le groupe de chauves-souris que dans celui du pangolin, mais est presque similaire. Les 6 autres groupes ont la divergence génétique la plus basse, où le taux élevé est celui du groupe du lapin et le plus bas est celui du groupe du chat.

Réseau (Figure 4.39) montre que les similarités globales sont égales à 75 % entre les groupes. La divergence génétique est élevée pour les groupes du pangolin, humain, cheval, veau, lapin, et porc, car les lignes entre eux sont longues. La divergence génétique entre les séquences des groupes du cheval, veau et chauve-souris, humain et pangolin, chauve-souris est faible, car les lignes entre elles sont courtes. Les régions de recombinaison sont représentées par des lignes brisées. Les lignes brisées entre deux points (groupe du pangolin et chauve-souris) (groupe de l'humain et pangolin) et (groupe de l'humain et chauve-souris) (groupe du cheval et veau) (groupe du veau et lapin), et (groupe veau et porc) (groupe porc, cheval), du réseau, signifient que les séquences génomiques reliées par ces lignes changent les régions de leurs génomes par recombinaison.

Gène S

Graphe (Figure 4.40) montre que la divergence génétique entre les séquences de tous les groupes est différente, car les points sont de hauteurs différentes. La divergence génétique est plus élevée et presque similaire dans les deux groupes du pangolin et de la chauve-souris. Les 6 autres groupes ont la divergence génétique la plus basse.

Réseau (Figure 4.41) montre que les similarités globales sont égales à 75 % entre les groupes. La divergence génétique est élevée pour les groupes du pangolin, humains, chevaux, veaux, lapins et porcs,

car les lignes entre eux sont longues. La divergence génétique entre les séquences des groupes du cheval, veau et chauve-souris, humain et pangolin, et chauve-souris sont faibles, car les lignes entre elles sont courtes. Les régions de recombinaison sont représentées par des lignes brisées. Les lignes brisées entre deux points (groupe du pangolin et chauve-souris) (groupe de l'humain et pangolin) et (groupe de l'humain et chauve-souris) (groupe du cheval et veau) (groupe du veau et lapin), et (groupe veau et porc) (groupe porc, cheval), du réseau, signifient que les séquences génomiques reliées par ces lignes changent les régions de leurs génomes par recombinaison.

Génome complet

Graphe (Figure 4.42) montre que la divergence génétique entre les séquences de tous les groupes est différente, car les points sont de hauteurs différentes. La divergence génétique la plus élevée est celle du groupe du pangolin suivi par celle de chauve-souris. Pour les groupes où elle est la moins élevée, celui du lapin et du cheval sont presque similaire, suivi par le groupe de veau et de souris. Enfin, la divergence génétique la plus basse est celle des groupes de chat et de rates.

Réseau (Figure 4.43) montre que les similarités globales sont égales à 75 % entre les groupes. La divergence génétique est élevée pour les groupes du pangolin, humain, cheval, veau, et du lapin, car les lignes entre eux sont longues. La divergence génétique entre les séquences des groupes du cheval, veau et chauve-souris, humain et pangolin, chauve-souris est faible, car les lignes entre elles sont courtes. Les régions de recombinaison sont représentées par des lignes brisées. Les lignes brisées entre deux points du réseau : groupe de l'humain et pangolin, groupe de l'humain et chauve-souris, groupe de veau et lapin, groupe souris et chat, signifient que les séquences génomiques reliées par ces lignes changent les régions de leurs génomes par recombinaison.

4.4 Estimation du temps de divergence bayésienne

L'estimation du temps de divergence bayésienne est une technique puissante en biologie évolutive pour reconstruire les moments clés de l'évolution, telle que la divergence entre les espèces ou les lignées génétiques. Cette approche est basée sur des principes statistiques bayésiens et utilise des modèles moléculaires pour analyser les séquences de gènes et estimer le temps écoulé depuis le dernier ancêtre commun.

4.4.1 Résultats de calcul de vraisemblance

Dans cette partie, les résultats obtenus par BEAUti (Bayesian Evolutionary Analysis Utility) seront présentés. BEAUti est un logiciel qui fait partie du package BEAST, largement utilisés en biologie évolutive pour effectuer des analyses phylogénétiques bayésiennes complètes. BEAUti joue un rôle clé dans la préparation et l'annotation des données critiques pour l'analyse, en particulier dans le développement de modèles évolutifs et d'horloges moléculaires appropriées.

Gène E

La figure 4.44 montre les résultats de Beagle pour le gène E qui contient 43 séquences extraites de différentes espèces, en utilisant les résultats de BEAUti par le logiciel BEAST.

Operator analysis

Operator	Tuning	Count	Time	Time/Op	Pr(accept)
scale(kappa)	0.53	93471	9536	0.1	0.2369
frequencies	0.076	93764	9855	0.11	0.2401
scale(nodeHeights(treeModel))	0.728	279672	39975	0.14	0.2335
subtreeSlide(treeModel)	0.021	2805035	88463	0.03	0.2319
Narrow Exchange(treeModel)		2800709	96289	0.03	0.2981
Wide Exchange(treeModel)		280555	7899	0.03	0.0115
wilsonBalding(treeModel)		280020	14822	0.05	0.0111
scale(treeModel.rootHeight)	0.143	280319	9939	0.04	0.2412
uniform(nodeHeights(treeModel))		2806605	131025	0.05	0.4961
scale(constant.popSize)	0.395	279850	1030	0.0	0.2381

8.4313 minutes

Figure 4.44 Résultat de beagle pour gène E

Gène M

La figure 4.45 montre les résultats de Beagle pour le gène M qui contient 43 séquences extraites de différentes espèces, en utilisant les résultats de BEAUti par le logiciel BEAST.

Operator analysis

Operator	Tuning	Count	Time	Time/Op	Pr(accept)
scale(kappa)	0.663	93584	22311	0.24	0.2352
frequencies	0.046	93638	22342	0.24	0.2387
scale(nodeHeights(treeModel))	0.847	279448	78288	0.28	0.2331
subtreeSlide(treeModel)	0.019	2804410	163329	0.06	0.2318
Narrow Exchange(treeModel)		2806710	172456	0.06	0.1275
Wide Exchange(treeModel)		279586	15173	0.05	0.0039
wilsonBalding(treeModel)		280115	26060	0.09	0.0034
scale(treeModel.rootHeight)	0.37	280685	13791	0.05	0.2387
uniform(nodeHeights(treeModel))		2802418	228477	0.08	0.306
scale(constant.popSize)	0.334	279406	1343	0.0	0.2374

14.251616666666667 minutes

Figure 4.45 Résultat de beagle pour gène M

Gène N

La figure 4.46 montre les résultats de Beagle pour le gène N qui contient 43 séquences extraites de différentes espèces, en utilisant les résultats de BEAUti par le logiciel BEAST.

Operator analysis

Operator	Tuning	Count	Time	Time/Op	Pr(accept)
scale(kappa)	0.78	93688	34872	0.37	0.233
frequencies	0.036	93871	37828	0.4	0.2373
scale(nodeHeights(treeModel))	0.871	280190	119435	0.43	0.2333
subtreeSlide(treeModel)	0.017	2803054	283873	0.1	0.2316
Narrow Exchange(treeModel)		2804904	279018	0.1	0.1433
Wide Exchange(treeModel)		279847	24813	0.09	0.0048
wilsonBalding(treeModel)		280810	43244	0.15	0.0034
scale(treeModel.rootHeight)	0.295	280669	21331	0.08	0.2388
uniform(nodeHeights(treeModel))		2802272	387300	0.14	0.2683
scale(constant.popSize)	0.386	280695	1435	0.01	0.2381

22.615766666666666 minutes

Figure 4.46 Résultat de beagle pour gène N

Gène N2

La figure 4.47 montre les résultats de Beagle pour le gène N2 qui contient 43 séquences extraites de différentes espèces, en utilisant les résultats de BEAUti par le logiciel BEAST.

Operator analysis

Operator	Tuning	Count	Time	Time/Op	Pr(accept)
scale(kappa)	0.478	92892	4906	0.05	0.2379
frequencies	0.087	93910	5411	0.06	0.2409
scale(nodeHeights(treeModel))	0.76	280258	18732	0.07	0.2351
subtreeSlide(treeModel)	0.037	2803418	68413	0.02	0.2317
Narrow Exchange(treeModel)		2805513	76349	0.03	0.1963
Wide Exchange(treeModel)		279851	5682	0.02	0.0168
wilsonBalding(treeModel)		280367	9924	0.04	0.0174
scale(treeModel.rootHeight)	0.148	280269	6952	0.02	0.2397
uniform(nodeHeights(treeModel))		2803094	97931	0.03	0.4466
scale(constant.popSize)	0.317	280428	972	0.0	0.2385

6.096283333333333 minutes

Figure 4.47 Résultat de beagle pour gène N2

Gène ORF1ab

La figure 4.48 montre les résultats de Beagle pour le gène ORF1ab qui contient 43 séquences extraites de différentes espèces, en utilisant les résultats de BEAUti par le logiciel BEAST.

Operator analysis

Operator	Tuning	Count	Time	Time/Op	Pr(accept)
scale(kappa)	0.945	93565	602814	6.44	0.2309
frequencies	0.007	92678	594491	6.41	0.2328
scale(nodeHeights(treeModel))	0.974	280447	1819569	6.49	0.2311
subtreeSlide(treeModel)	0.005	2803081	3544549	1.26	0.2317
Narrow Exchange(treeModel)		2805528	3297950	1.18	0.0355
Wide Exchange(treeModel)		280935	314777	1.12	0.0009
wilsonBalding(treeModel)		280568	533496	1.9	0.0005
scale(treeModel.rootHeight)	0.771	279728	135989	0.49	0.2336
uniform(nodeHeights(treeModel))		2803707	4470332	1.59	0.1146
scale(constant.popSize)	0.377	279763	3679	0.01	0.2388

4.3490238888888895 hours

Figure 4.48 Résultat de beagle pour gène ORF1ab

Gène ORF3a

La figure 4.49 montre les résultats de Beagle pour le gène ORF3a qui contient 43 séquences extraites de différentes espèces, en utilisant les résultats de BEAUti par le logiciel BEAST.

Operator analysis

Operator	Tuning	Count	Time	Time/Op	Pr(accept)
scale(kappa)	0.703	94098	24483	0.26	0.2344
frequencies	0.046	92763	23071	0.25	0.2384
scale(nodeHeights(treeModel))	0.859	280458	81722	0.29	0.2328
subtreeSlide(treeModel)	0.03	2804512	187395	0.07	0.2318
Narrow Exchange(treeModel)		2802297	195474	0.07	0.1144
Wide Exchange(treeModel)		281187	16455	0.06	0.0034
wilsonBalding(treeModel)		279784	28760	0.1	0.0038
scale(treeModel.rootHeight)	0.116	280494	14554	0.05	0.2405
uniform(nodeHeights(treeModel))		2804738	262695	0.09	0.3059
scale(constant.popSize)	0.414	279669	1260	0.0	0.2371

15.673883333333333 minutes

Figure 4.49 Résultat de beagle pour gène ORF3a

Domaine RB

La figure 4.50 montre les résultats de Beagle pour le domaine RB qui contient 43 séquences extraites de différentes espèces, en utilisant les résultats de BEAUti par le logiciel BEAST.

Operator analysis

Operator	Tuning	Count	Time	Time/Op	Pr(accept)
scale(nodeHeights(treeModel))	0.819	286294	1246602	4.35	0.2334
subtreeSlide(treeModel)	0.053	2857412	1525963	0.53	0.2317
Narrow Exchange(treeModel)		2854728	1608313	0.56	0.1488
Wide Exchange(treeModel)		286023	173565	0.61	0.0046
wilsonBalding(treeModel)		285560	302688	1.06	0.0038
scale(treeModel.rootHeight)	0.207	285473	59004	0.21	0.2407
uniform(nodeHeights(treeModel))		2859080	2166504	0.76	0.371
scale(constant.popSize)	0.438	285430	1126	0.0	0.2376

2.0021066666666667 hours

Figure 4.50 Résultat de beagle pour le domaine RB

Gène S

La figure 4.51 montre les résultats de Beagle pour le gène S qui contient 43 séquences extraites de différentes espèces, en utilisant les résultats de BEAUti par le logiciel BEAST.

Operator analysis

Operator	Tuning	Count	Time	Time/Op	Pr(accept)
scale(kappa)	0.892	93005	155730	1.67	0.2323
frequencies	0.011	93047	155942	1.68	0.2358
scale(nodeHeights(treeModel))	0.915	280738	487179	1.74	0.2321
subtreeSlide(treeModel)	0.014	2805756	864802	0.31	0.2316
Narrow Exchange(treeModel)		2804889	846163	0.3	0.0581
Wide Exchange(treeModel)		280259	82168	0.29	0.0019
wilsonBalding(treeModel)		279781	137096	0.49	0.0014
scale(treeModel.rootHeight)	0.552	280111	51727	0.18	0.2356
uniform(nodeHeights(treeModel))		2802536	1128086	0.4	0.1648
scale(constant.popSize)	0.459	279878	1873	0.01	0.2376

1.1336283333333335 hours

Figure 4.51 Résultat de beagle pour gène S

4.4.2 Résultats de TreeAnnotator

Cette section présente les résultats de l'analyse des gènes du SARS-Cov-2, en se concentrant sur les 43 séquences extraites de différentes espèces. L'analyse a été réalisée en utilisant le logiciel BEAST, et les résultats ont été traités à l'aide de TreeAnnotator.

Gène E

La figure 4.52 montre les résultats de TreeAnnotator pour le gène E du Coronavirus humain qui contient 43 séquences extraites de différentes espèces, en utilisant les précédents résultats de BEAST.

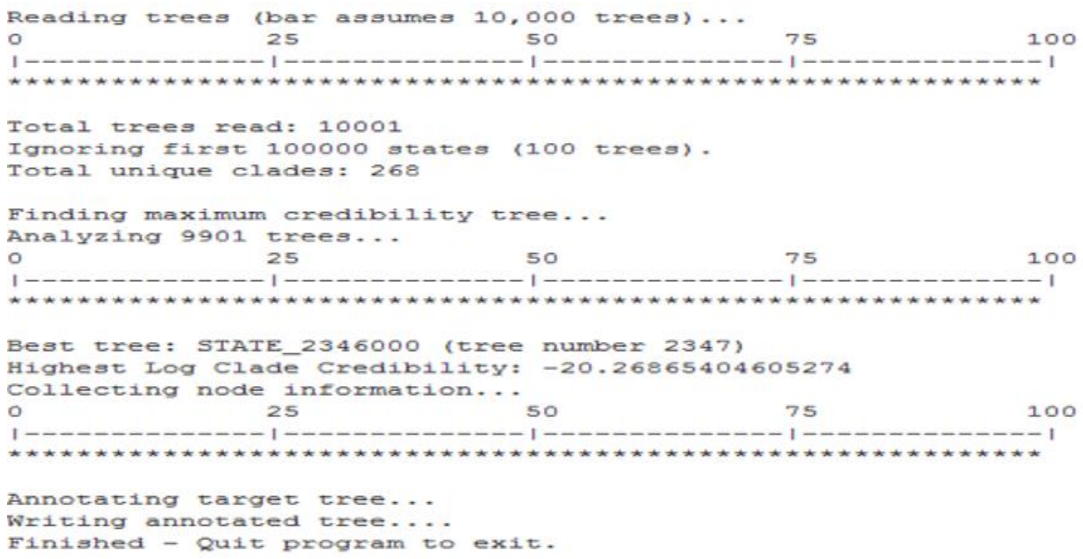

Figure 4.52 Résultat de TreeAnnotator du gène E du Coronavirus

Gène M

La figure 4.53 montre les résultats de TreeAnnotator pour le gène M du Coronavirus humain qui contient 43 séquences extraites de différentes espèces, en utilisant les précédents résultats de BEAST.

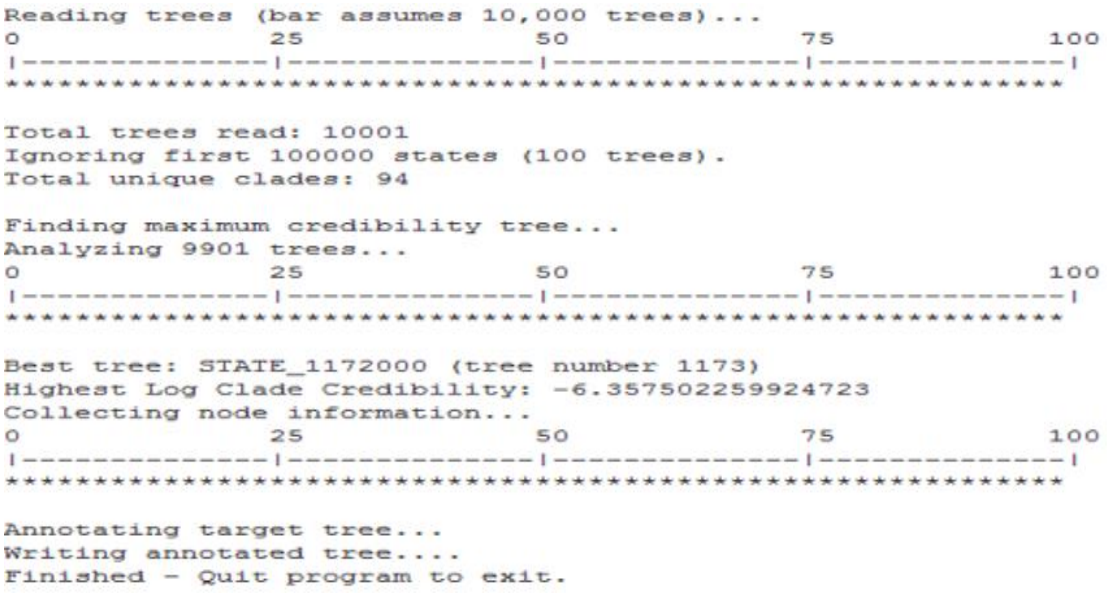

Figure 4.53 Résultat de TreeAnnotator du gène M du Coronavirus

Gène N

La figure 4.54 montre les résultats de TreeAnnotator pour le gène N du Coronavirus humain qui contient 43 séquences extraites de différentes espèces, en utilisant les précédents résultats de BEAST.

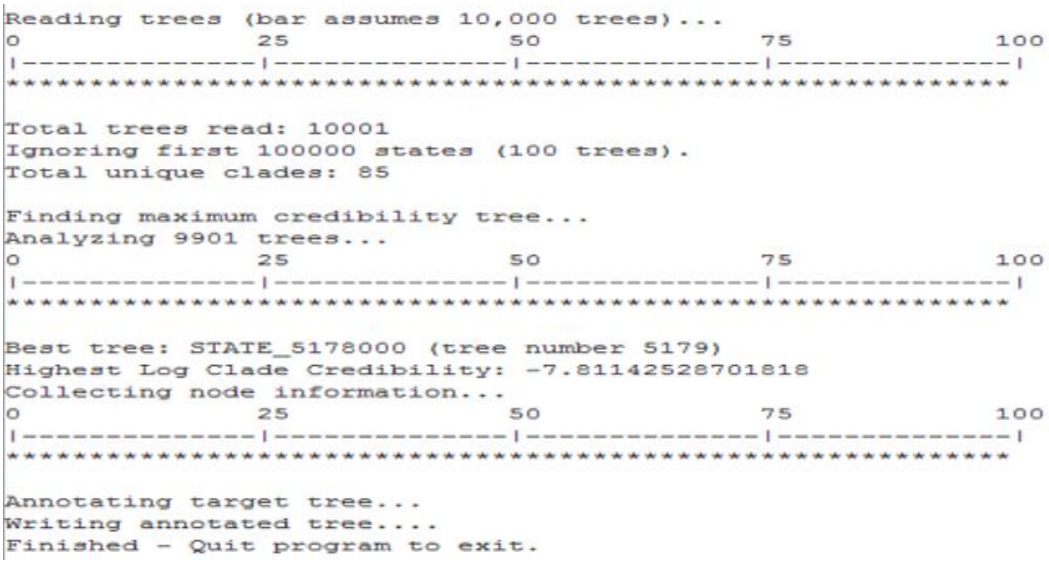

Figure 4.54 Résultat de TreeAnnotator du gène N du Coronavirus

Gène N2

La figure 4.55 montre les résultats de TreeAnnotator pour le gène N2 du Coronavirus humain qui contient 43 séquences extraites de différentes espèces, en utilisant les précédents résultats de BEAST.

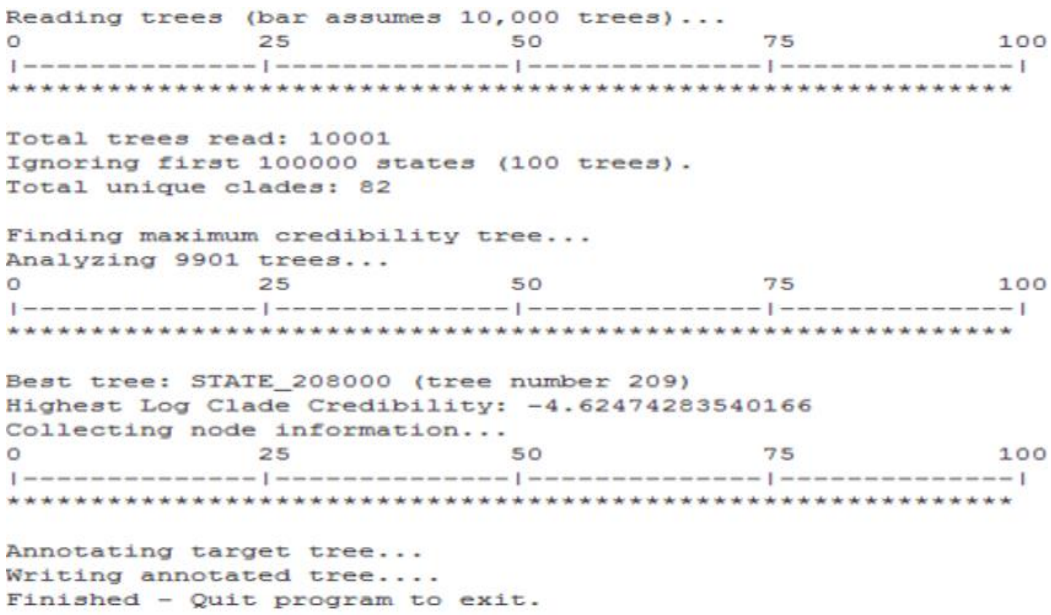

Figure 4.55 Résultat de TreeAnnotator du gène N du Coronavirus

Gène ORF1ab

La figure 4.56 montre les résultats de TreeAnnotator pour le gène ORF1ab du Coronavirus humain qui contient 43 séquences extraites de différentes espèces, en utilisant les précédents résultats de BEAST.

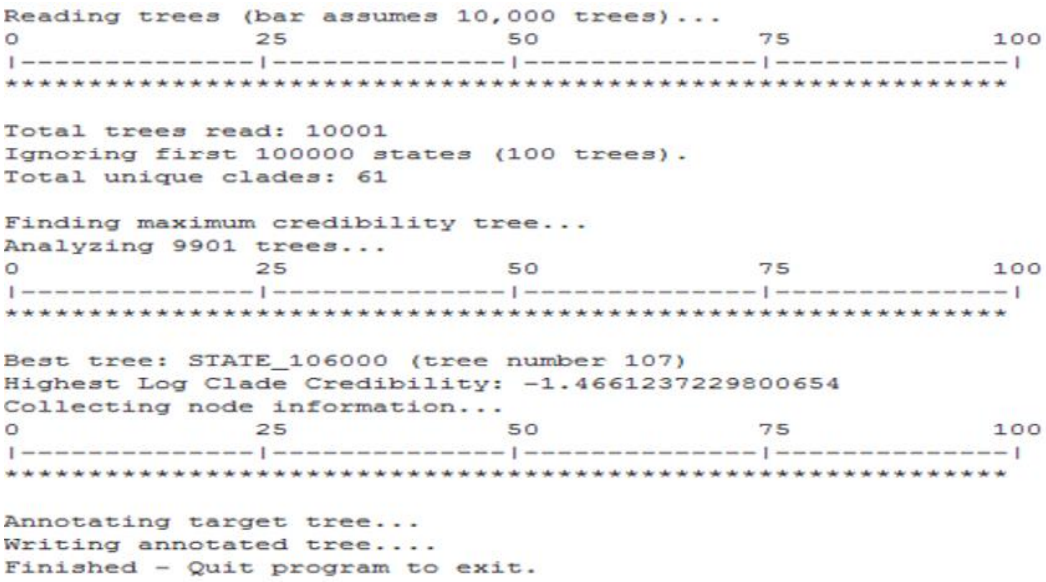

Figure 4.56 Résultat de TreeAnnotator du gène ORF1ab du Coronavirus

Gène ORF3a

La figure 4.57 montre les résultats de TreeAnnotator pour le gène ORF3a du Coronavirus humain qui contient 43 séquences extraites de différentes espèces, en utilisant les précédents résultats de BEAST.

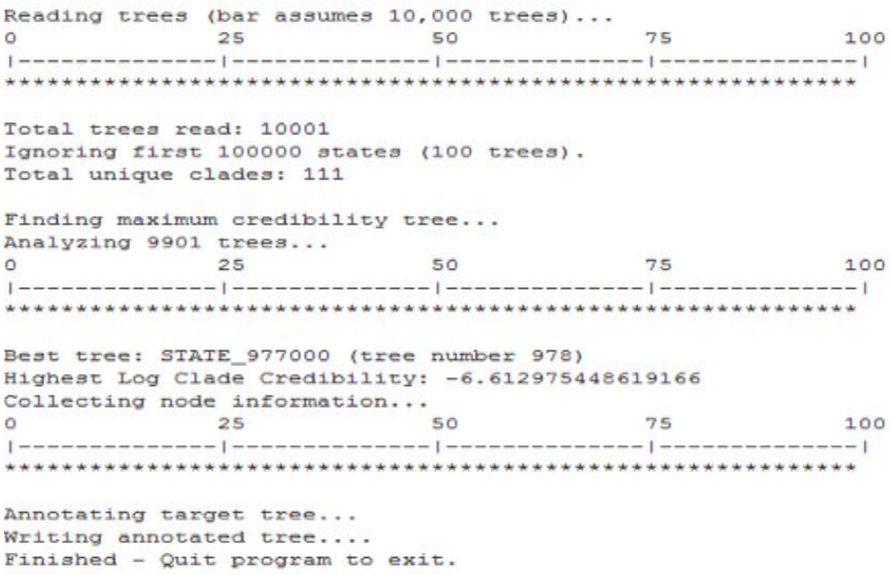

Figure 4.57 Résultat de TreeAnnotator du gène ORF3a du Coronavirus

Domaine RB

La figure 4.58 montre les résultats de TreeAnnotator pour le domaine RB du Coronavirus humain qui contient 43 séquences extraites de différentes espèces, en utilisant les précédents résultats de BEAST.

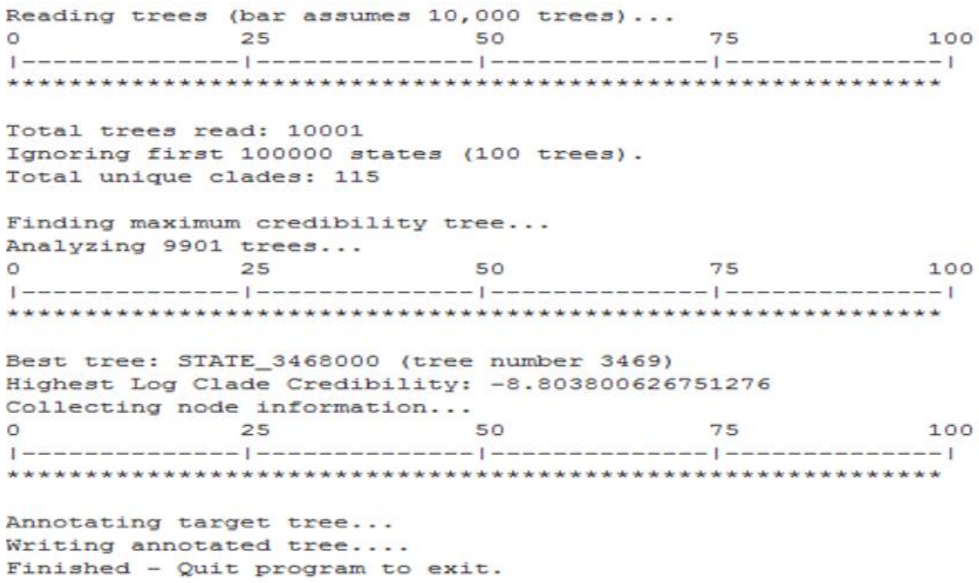

Figure 4.58 Résultat de TreeAnnotator du domaine RB du Coronavirus

Gène S

La figure 4.59 montre les résultats de TreeAnnotator pour le gène S du Coronavirus humain qui contient 43 séquences extraites de différentes espèces, en utilisant les précédents résultats de BEAST.

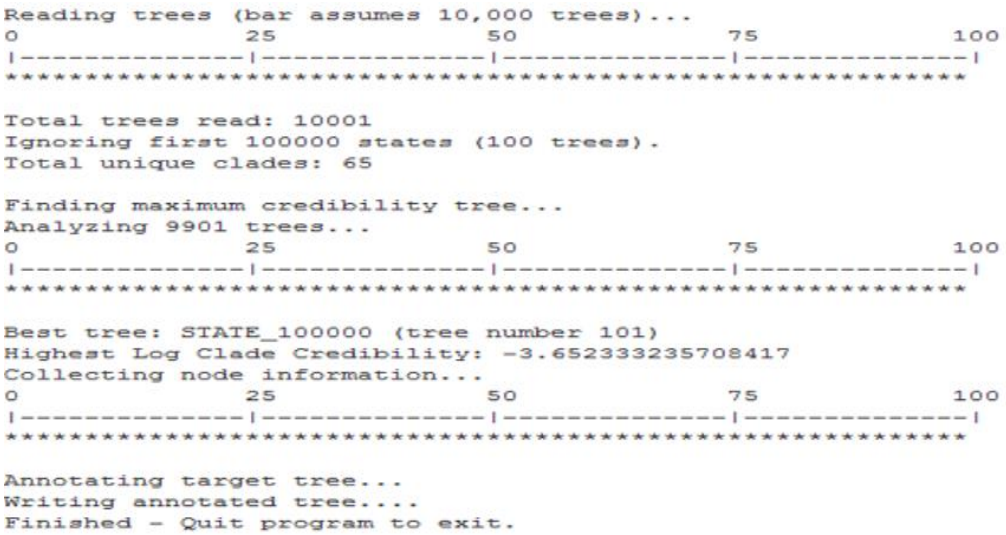

Figure 4.59 Résultat de TreeAnnotator du gène S du Coronavirus

4.4.3 Résultats de FigTree

Le but de cette étude est de présenter des arbres phylogénétiques de gènes du coronavirus, en utilisant notamment le logiciel FigTree. Les arbres ont été construits à partir de 43 séquences génétiques extraites de différentes espèces.

Les résultats obtenus permettent de visualiser les relations évolutives entre les différentes souches du SARS-Cov-2, de comprendre la diversité génétique, ainsi que les schémas de divergence évolutive. Le logiciel FigTree offre des outils graphiques avancés pour l'annotation et la représentation visuelle des arbres phylogénétiques.

Gène E

La figure 4.60 montre un arbre phylogénétique du coronavirus humain pour le gène E qui contient 43 séquences extraites de différentes espèces, en utilisant le logiciel FigTree

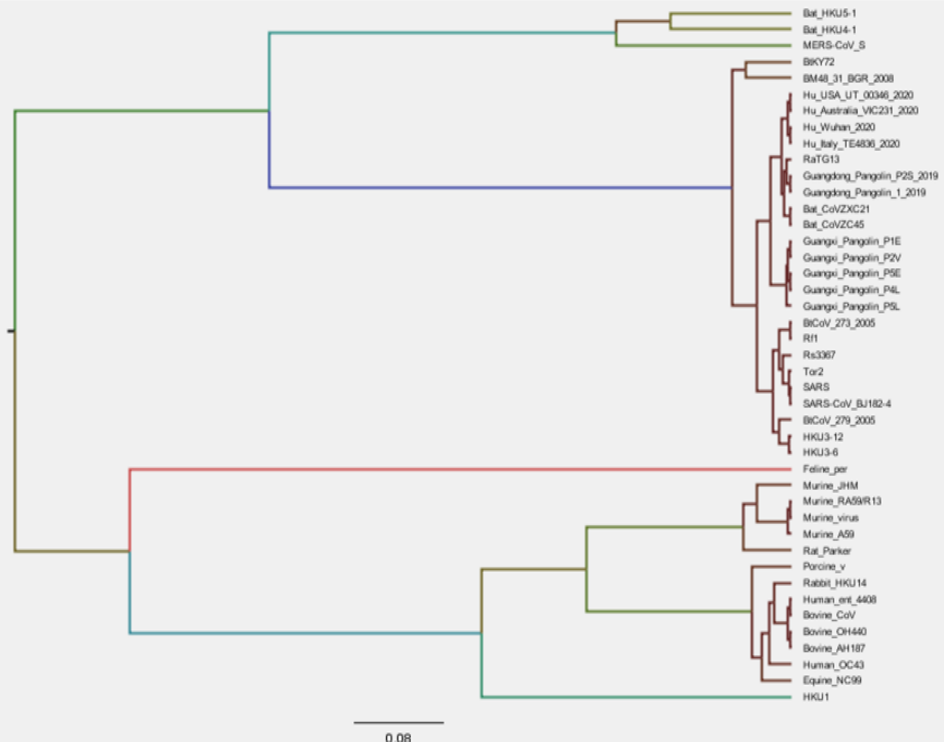

Figure 4.60 Arbre phylogénétique du coronavirus pour le gène E généré par FigTree

Gène M

La figure 4.61 montre un arbre phylogénétique du coronavirus humain pour le gène M qui contient 43 séquences extraites de différentes espèces, en utilisant le logiciel FigTree

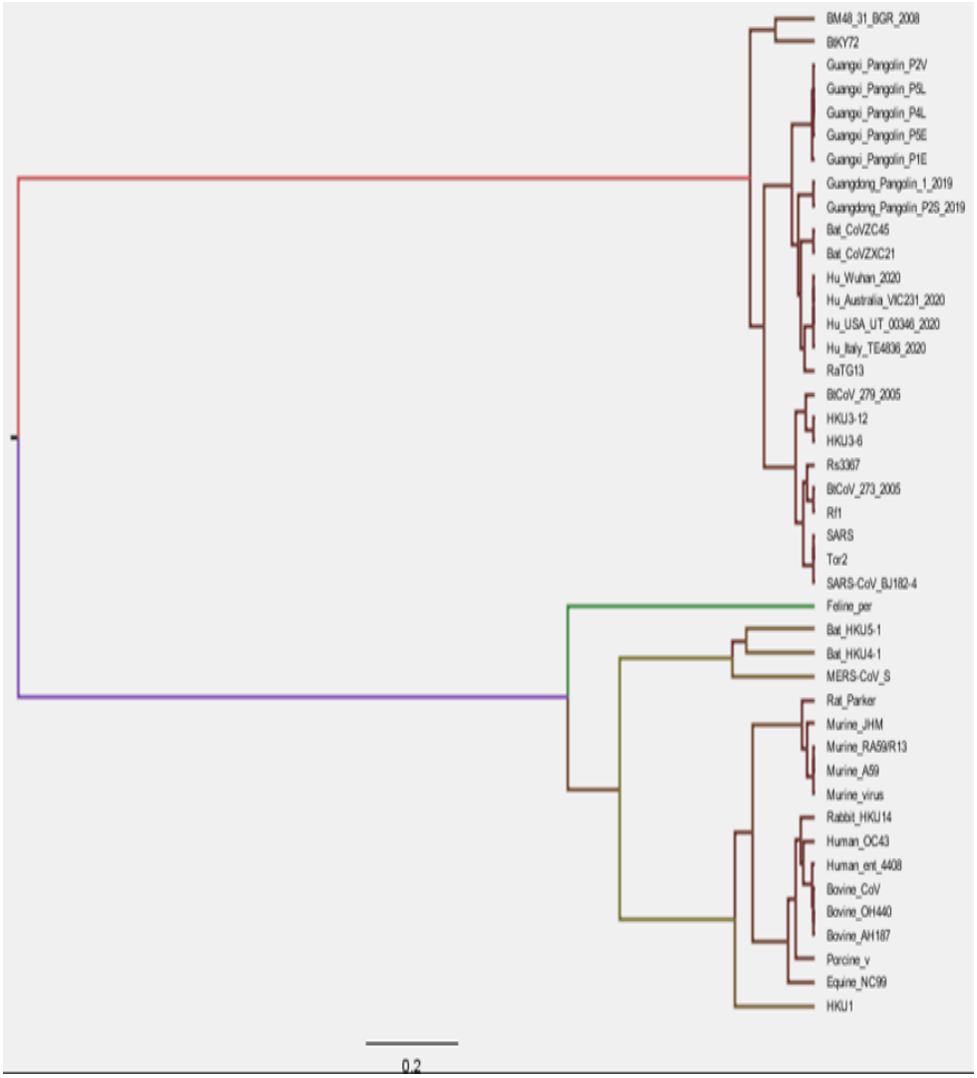

Figure 4.61 Arbre phylogénétique du coronavirus pour le gène M généré par FigTree

Gène N

La figure 4.62 montre un arbre phylogénétique du coronavirus humain pour le gène N qui contient 43 séquences extraites de différentes espèces, en utilisant le logiciel FigTree

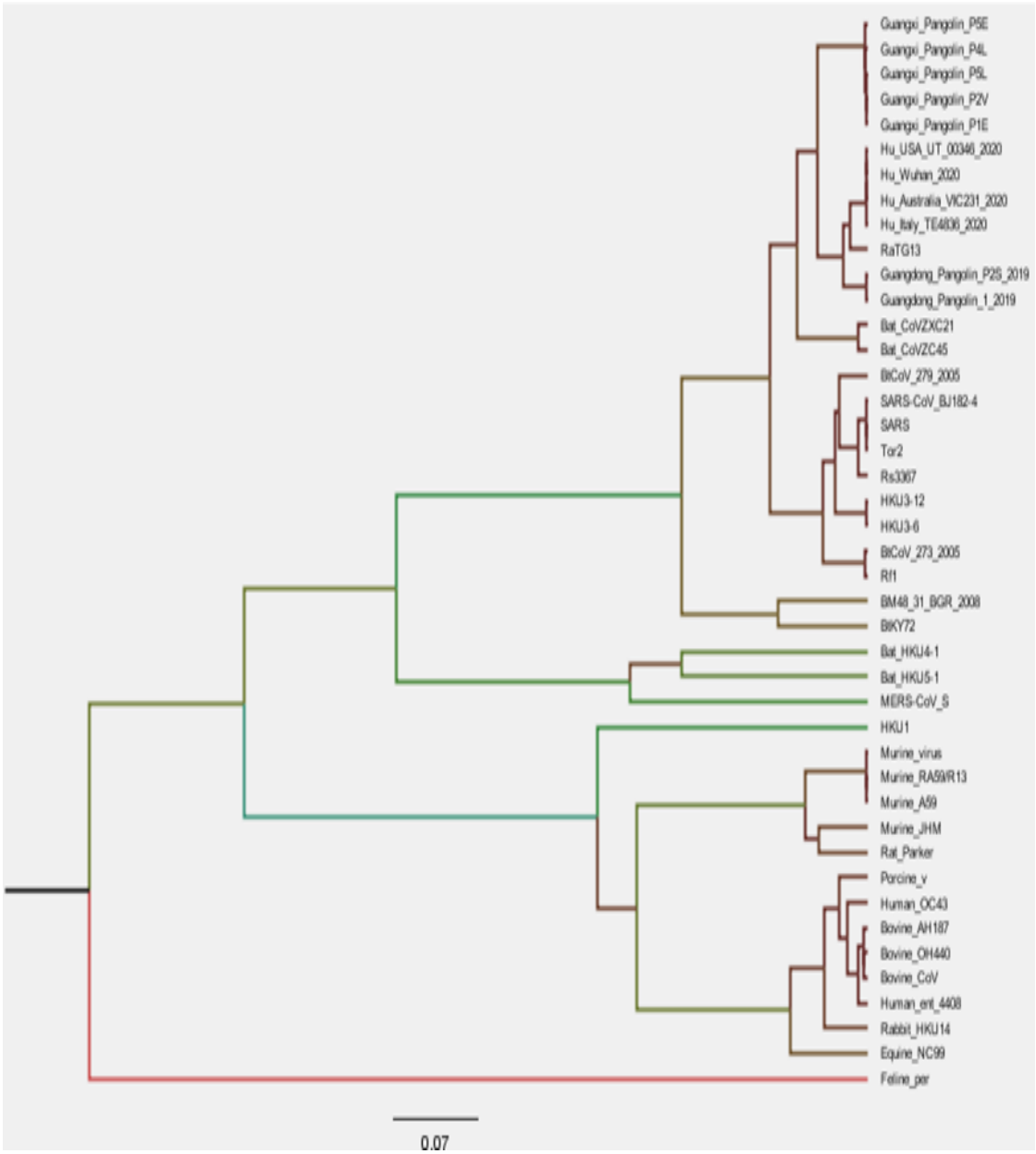

Figure 4.62 Arbre phylogénétique du coronavirus pour le gène N généré par FigTree

Gène N2

La figure 4.63 montre un arbre phylogénétique du coronavirus humain pour le gène N2 qui contient 43 séquences extraites de différentes espèces, en utilisant le logiciel FigTree

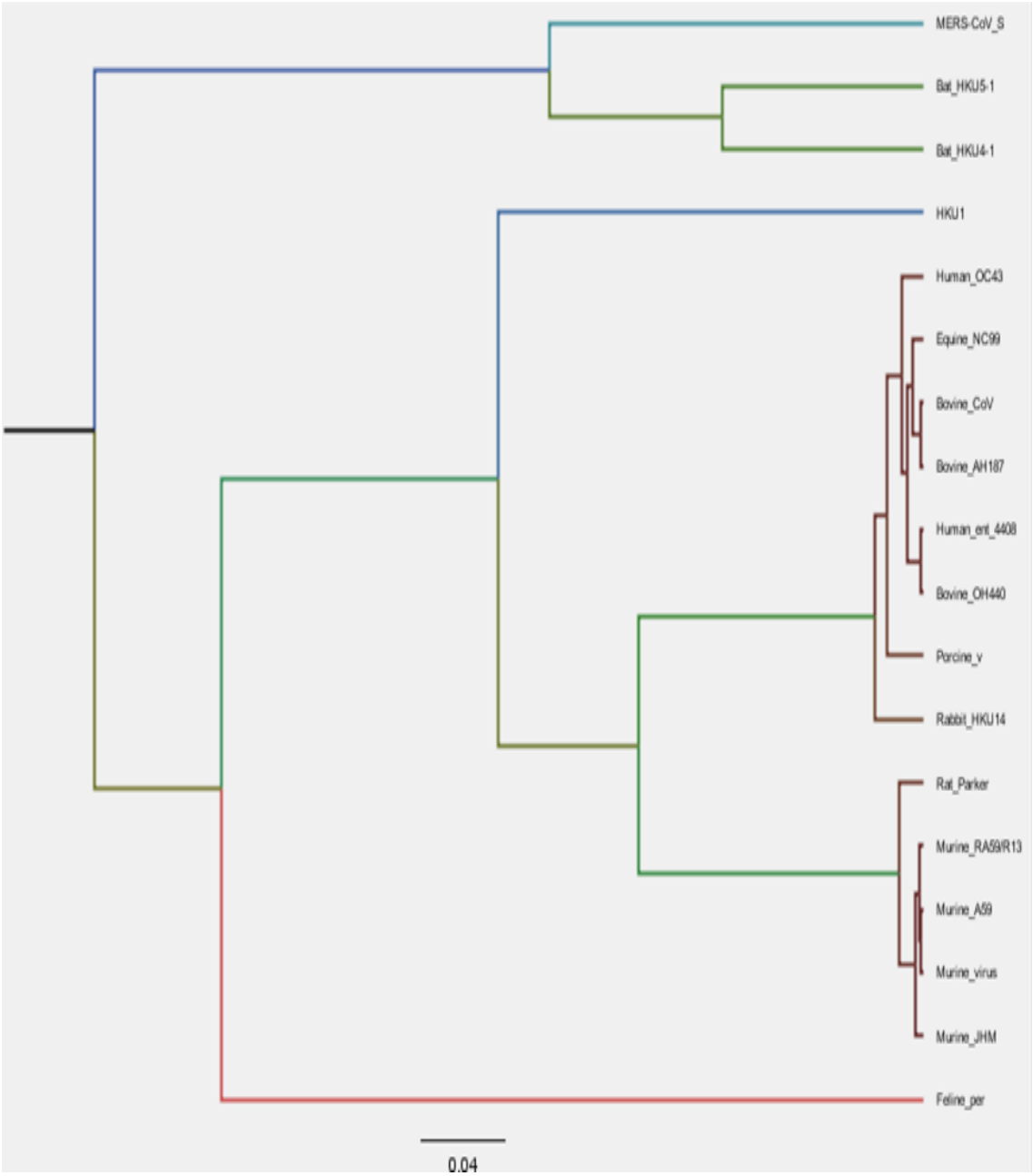

Figure 4.63 Arbre phylogénétique du coronavirus pour le gène N2 généré par FigTree

Gène ORF1ab

La figure 4.64 montre un arbre phylogénétique du coronavirus humain pour le gène ORF1ab qui contient 43 séquences extraites de différentes espèces, en utilisant le logiciel FigTree

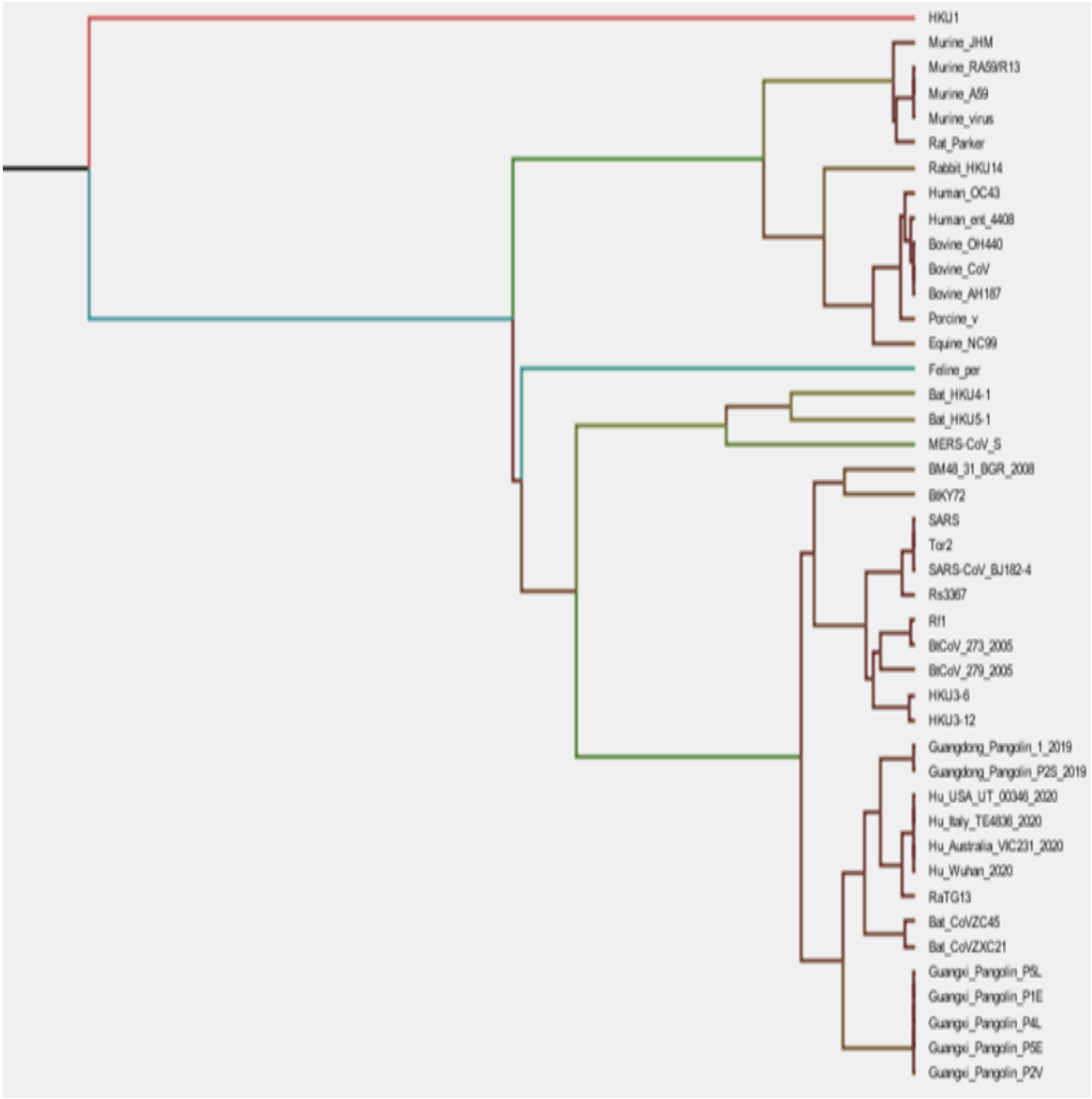

Figure 4.64 Arbre phylogénétique du coronavirus pour le gène ORF1ab généré par FigTree

Gène ORF3a

La figure 4.65 montre un arbre phylogénétique du coronavirus humain pour le gène ORF3a qui contient 43 séquences extraites de différentes espèces, en utilisant le logiciel FigTree

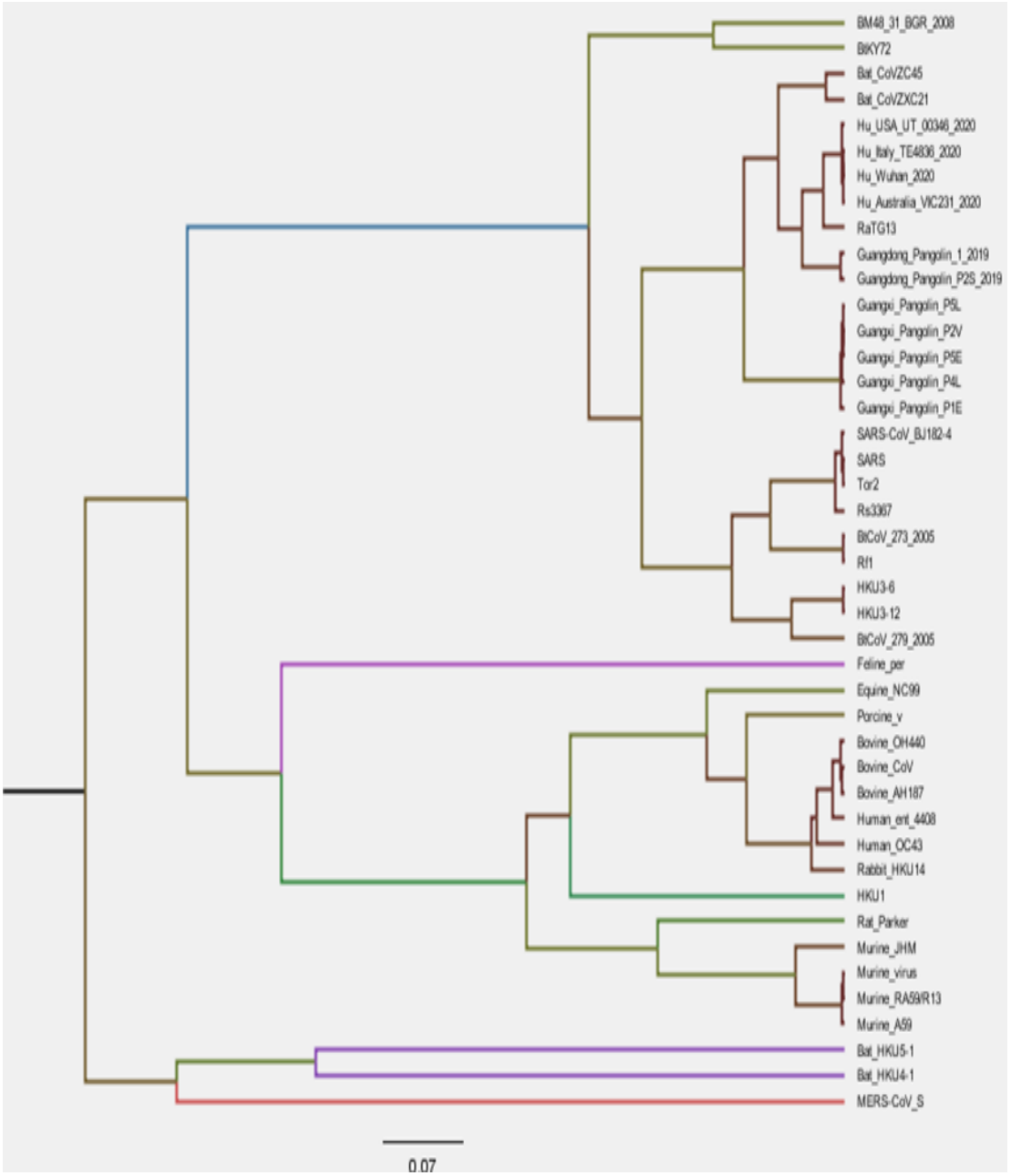

Figure 4.65 Arbre phylogénétique du coronavirus pour le gène ORF3a généré par FigTree

Domaine RB

La figure 4.66 montre un arbre phylogénétique du coronavirus humain pour le domaine RB qui contient 43 séquences extraites de différentes espèces, en utilisant le logiciel FigTree

Figure 4.66 Arbre phylogénétique du coronavirus pour le domaine RB généré par FigTree

Gène S

La figure 4.67 montre un arbre phylogénétique du coronavirus humain pour le gène S qui contient 43 séquences extraites de différentes espèces, en utilisant le logiciel FigTree

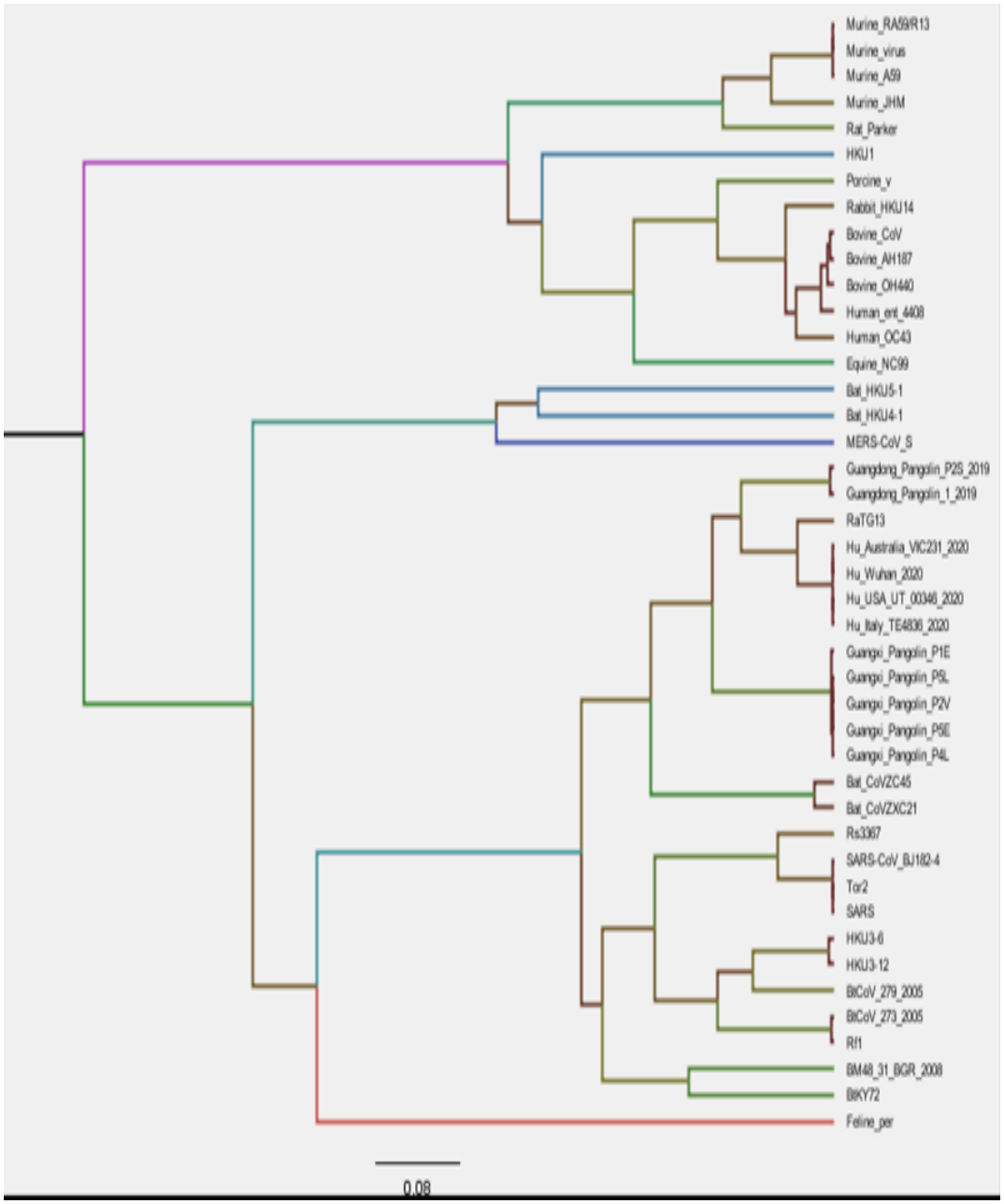

Figure 4.67 Arbre phylogénétique du coronavirus pour le gène S généré par FigTree

4.4.4 Interprétation des résultats obtenus par BEAST

Lorsque l'on interprète les arbres phylogénétiques générés par BEAST, plusieurs aspects méritent d'être pris en compte, notamment :

1— La Topologie de l'arbre : les relations de parenté entre les espèces sont représentées par la topologie de l'arbre. Les espèces regroupées sur un même nœud sont considérées comme étant étroitement apparentées les unes aux autres. Cela signifie qu'elles partagent un ancêtre commun plus récent que les espèces situées sur des nœuds différents.

2— Le Temps de divergence : les temps de divergence sont reflétés par les longueurs des branches de l'arbre. Les branches courtes indiquent des temps de divergence récents, comme c'est le cas pour les clades situés en position proche de la racine de l'arbre. À l'inverse, les branches longues correspondent à des temps de divergence plus anciens. Par exemple, les branches menant aux feuilles de l'arbre, qui regroupent des espèces ayant des homologies avec plusieurs autres clades, représentent des temps de divergence plus anciens.

3— Comparaison avec MEGA : Il est important de noter que les arbres phylogénétiques produits par BEAST présentent une cohérence remarquable avec les résultats obtenus par MEGA. Cette cohérence est observée dans la topologie de l'arbre, qui montre des relations similaires entre les espèces, ainsi que dans la représentation des temps de divergence par la longueur des branches. Les explications fournies dans l'interprétation de MEGA sont particulièrement utiles pour comprendre ces aspects de l'arbre phylogénétique.

En résumé, l'analyse des arbres phylogénétiques issus de BEAST, qui a abouti à des résultats identiques à ceux obtenus avec MEGA.

CONCLUSION

Les analyses exhaustives menées, spécifiquement la réalisation d'arbres phylogénétiques par le biais de MEGA, l'étude des transferts horizontaux de gènes entre espèces grâce à T-REX en ligne, l'évaluation des divergences génétiques par SimPlot++ ainsi que l'analyse des résultats obtenus par BEAST (particulièrement la visualisation des arbres phylogénétiques à l'aide de FigTree) ont abouti à des conclusions pertinentes :

Tout d'abord, les gênes de SARS-CoV-2 présentent une forte similarité, avec des taux élevés de ressemblances avec des gènes de coronavirus humain préalablement identifiés chez les chauves-souris RaTG13 ainsi que chez les pangolins. En revanche, ces mêmes gènes montrent des taux de similarité bien moindres avec ceux des coronavirus humains découverts chez d'autres espèces telles que le veau, le cheval, le porc, le lapin, la souris et le chat.

Nous avons trouvé que le gène E de SARS-Cov-2, qui est impliqué dans la réplication du coronavirus humain, partage une grande similarité avec les gènes de coronavirus issus des chauves-souris (RaTG13), avec un score de 75 % (voir la figure 4.29). De plus, on a noté des transferts partiels de gènes remarquables entre l'espèce humaine (SRAS, SARS-CoV_BJ182-4, Tor2) et l'espèce de chauve-souris (BM48_31_BGR_2008, BtKY72). Les transferts horizontaux partiels ont été extraits à l'aide de l'algorithme de Boc and Makarenkov (Boc & Makarenkov, 2011). En revanche, l'étude menée par Makarenkov et al. (2021) a montré des scores bootstrap médiocres du gène E ainsi qu'un seul transfert de gène trouvé.

Le gène de la protéine de la membrane M de SARS-Cov-2 jouant un rôle dans la formation de la membrane virale dévoile des similitudes avec les gènes de coronavirus humain dérivé des chauves-souris RaTG13 et les pangolins avec un score de 75 % et 70 % respectivement (Voir figure 4.33). Il y a également des transferts partiels de gènes de l'espèce du pangolin (Guangxi_Pangolin_P1E, Guangxi_Pangolin_P2V, Guangxi_Pangolin_P4L, Guangxi_Pangolin_P5E, Guangxi_Pangolin_P5L), d'humain (SARS, SARS-CoV_BJ182-4, Tor2) et de chauve-souris (HKU3-12, HKU3-6, Rf1, Rs3367, BM48_31_BGR_2008, BtCoV_273_2005, BtCoV_279_2005, BtKY7) vers l'espèce de chauve-souris (Bat_CoVZC45, Bat_CoVZXC21, RaTG1) et d'humain.

Le gène de la protéine nucléocapside N de SARS-CoV-2 essentiel à la réplication du génome viral démontre des similitudes notables avec les gènes de coronavirus humain des chauves-souris RaTG13 ainsi qu'avec ceux des pangolins avec un score de 98 % et 97 % respectivement (Voir figure 4.35). De plus, des transferts partiels ont été observés allant de l'espèce du pangolin (Guangxi_Pangolin_P4L) vers diverses souches de pangolin (Guangxi_Pangolin_P1E, Guangxi_Pangolin_P5L).

Le gène ORF1ab de SARS-CoV-2 qui se révèle hautement conservé parmi les divers isolats du coronavirus humain indique son importance capitale pour la réplication et la survie de ce dernier. Cette région ORF1ab de la chauve-souris RaTG13 avait toutes les homologies des gènes ORF1ab de l'humain Hu (Wuhan 2020, USA UT 00346 2020, Italy TE4836 2020, Australia VIC231 2020) avec un score de 99 % (Voir figure 4.37) et qui s'est déplacé horizontalement de l'espèce humaine (Hu_Wuhan_2020) et de la chauve-souris (RaTG13) vers l'espèce de chauve-souris (Bat_CoVZC45, Bat_CoVZXC21). De plus, ce gène a également été transféré horizontalement de l'espèce humaine (Hu_Wuhan_2020) et de la chauve-souris (RaTG13, Bat_CoVZC45, Bat_CoVZXC21) vers l'espèce de pangolin (Guangdong_Pangolin_1_2019).

D'autre part, le gène de la protéine ORF6 de SARS-CoV-2 impliqué dans la réplication virale partage des similitudes notables avec les gènes de coronavirus humain provenant des chauves-souris principalement la souche RaTG13 avec un score de 75 % et un transfert de l'espèce humaine Hu_Wuhan_2020 et de chauve-souris RaTG13, Bat_CoVZC45, Bat_CoVZXC21 au Guangdong_Pangolin_1_2019 a été trouvé.

Les gènes de la protéine ORF3a qui influe sur l'inflammation cellulaire dévoilent des similitudes marquées avec les gènes de coronavirus humain issu des chauves-souris et des pangolins avec un score de 75% (Voir figure 4.39) et un transfert de l'espèce de chauve-souris (Bat_CoVZC45, Bat_CoVZXC21) à l'espèce de chauve-souris (RaTG13) et de l'humain (Hu_Wuhan_2020) a été trouvé.

En ce qui concerne les gènes ORF7a, ils montrent des similarités avec les gènes de coronavirus des pangolins avec un score de 75 % et ils ont été transféré horizontalement de l'espèce de chauve-souris Bat_CoVZC45 et Bat_CoVZXC21 à l'espèce de chauve-souris RaTG13 et de l'espèce de chauve-souris Bat_CoVZC45, Bat_CoVZXC21, RaTG13 à l'espèce humaine Hu_Wuhan_2020 et enfin, un transfert de l'espèce de chauve-souris Bat_CoVZC45, Bat_CoVZXC21, RaTG13 et d'humain Hu_Wuhan_2020 au Guangdong_Pangolin_1_2019.

Le gène ORF10 du SARS-CoV-2 hautement conservé parmi les différents isolats du coronavirus humain partage une forte similitude avec les coronavirus identifiés chez les chauves-souris et les pangolins avec un score de 75 %. Il a été transféré horizontalement de l'espèce de chauve-souris Bat_CoVZXC21 à l'espèce du pangolin Guangdong_Pangolin_1_2019, Guangdong_Pangolin_P2S_2019 et il a également été transféré horizontalement de l'espèce de chauve-souris RaTG13 à l'espèce humaine Hu_Wuhan_2020

Le gène S a été transféré horizontalement de l'espèce humaine Hu_Wuhan_2020 et de la chauve-souris RaTG13, Bat_CoVZC45, Bat_CoVZXC21 à Guangdong_Pangolin_1_2019 et de l'espèce de chauve-souris Bat_CoVZC45, Bat_CoVZXC21 à l'espèce humaine Hu_Wuhan_2020 et de la chauve-souris RaTG13.

Le domaine RB a été transféré horizontalement de l'espèce du pangolin Guangdong_Pangolin_1_2019, Guangdong_Pangolin_P2S_2019 à l'espèce humaine Hu_Wuhan_2020 et de l'espèce du pangolin Guangxi_Pangolin_P2V à l'espèce du pangolin Guangxi_Pangolin_P5E.

De manière générale, l'ensemble du génome du SARS-Cov-2 à savoir son génome complet présente des taux de similarité élevés avec les gènes de coronavirus des chauves-souris avec 95 % de similitude et des pangolins avec 92 % de similitude (Voir figure 4.43). Il a été transféré horizontalement de l'espèce humaine Hu_Wuhan_2020 et de RaTG13 vers l'espèce de chauve-souris Bat_CoVZC45, Bat_CoVZXC21 ainsi qu'un transfert partiel de l'espèce humaine Hu_Wuhan_2020 et de l'espèce de chauve-souris RaTG13, Bat_CoVZC45, Bat_CoVZXC21 vers l'espèce du pangolin Guangdong_Pangolin_1_2019.

En conclusion, nos résultats confirment l'hypothèse formulée par Makarenkov et al. 2021 que le génome de SARS-Cov-2 pourrait se former comme résultat d'une recombinaison entre les génomes de RaTG13 (chauves-souris) et celui du pangolin du Guangdong. En général, nos résultats sont en accord avec ceux de Makarenkov et al. 2021, sauf les transferts horizontaux complets et partiels, trouvés pour le gène E, renforçant ainsi la crédibilité de cette étude.

Par ailleurs, nos résultats ouvrent des perspectives intéressantes pour la recherche future. Il serait bénéfique d'approfondir la compréhension des hôtes intermédiaires dans la transmission des virus, ce qui pourrait conduire à des mesures de prévention plus ciblées. Nous pensons également que l'utilisation de réseaux phylogénétiques (Layeghifard et al., 2012) et de méthodes de partitionnement, i.e. clustering, (Gondeau et al., 2012) pourrait aider à retrouver de nouvelles relations évolutives entre les différents virus. De plus, le développement continu de vaccins basés sur ces découvertes pourrait renforcer notre capacité à répondre efficacement aux futures menaces virales. Ces pistes de recherche offrent ainsi des opportunités significatives pour contribuer à l'avancement des connaissances et à la protection de la santé publique à l'échelle mondiale.

RÉFÉRENCES

Aboubakr, H. A., Sharafeldin, T. A., & Goyal, S. M. (2021). Stability of SARS-CoV-2 and other coronaviruses in the environment and on common touch surfaces and the influence of climatic conditions: A review. Transboundary and Emerging Diseases, 68(2), 296-312. https://doi.org/10.1111/tbed.13707

Acar Kirit, H., Bollback, J. P., & Lagator, M. (2022). The Role of the Environment in Horizontal Gene Transfer. Molecular Biology and Evolution, 39(11), msac220. https://doi.org/10.1093/molbev/msac220

Aiewsakun, P., & Katzourakis, A. (2016). Time-Dependent Rate Phenomenon in Viruses. Journal of Virology, 90(16), 7184-7195. https://doi.org/10.1128/JVI.00593-16

Alizon, S., & Saulnier, E. (2018). Phylodynamique des infections virales.

Amoutzias, G. D., Nikolaidis, M., Tryfonopoulou, E., Chlichlia, K., Markoulatos, P., & Oliver, S. G. (2022). The Remarkable Evolutionary Plasticity of Coronaviruses by Mutation and Recombination: Insights for the COVID-19 Pandemic and the Future Evolutionary Paths of SARS-CoV-2. Viruses, 14(1), Art. 1. https://doi.org/10.3390/v14010078

Battin, J. (2020). Pandémies : Les leçons du passé. Bulletin De L'Academie Nationale De Medecine, 204(7), 737-740. https://doi.org/10.1016/j.banm.2020.04.015

Benson, D. A., Cavanaugh, M., Clark, K., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J., & Sayers, E. W. (2012). GenBank. Nucleic acids research, 41(D1), D36-D42.

Bentley, K., & Evans, D. J. (2018). Mechanisms and consequences of positive-strand RNA virus recombination. Journal of General Virology, 99(10), 1345-1356.

Boc, A., & Makarenkov, V. (2011). Towards an accurate identification of mosaic genes and partial horizontal gene transfers. Nucleic acids research, 39(21), e144-e144.

Boc, A., Diallo, A. B., & Makarenkov, V. (2012). T-REX: a web server for inferring, validating and visualizing phylogenetic trees and networks. Nucleic acids research, 40(W1), W573-W579.

Boni, M. F., Lemey, P., Jiang, X., Lam, T. T. Y., Perry, B. W., Castoe, T. A., ... & Robertson, D. L. (2020). Evolutionary origins of the SARS-CoV-2 sarbecovirus lineage responsible for the COVID-19 pandemic. Nature microbiology, 5(11), 1408-1417.

Bonnin, T., & Lombard, J. (2019). Situer l'analyse phylogénétique entre les sciences historiques et expérimentales. Philosophia Scientae, 23-2, 131-148. https://doi.org/10.4000/philosophiascientiae.1957

Bouckaert, R., et al. (2014). BEAST 2: a software platform for Bayesian evolutionary analysis. PLoS computational biology, 10(4), e1003537.

Bouloy, M. (2002). Recombinaison chez les virus à ARN négatif : Un modèle chez les hantavirus. Virologie, 6(3). https://www.jle.com/fr/revues/vir/edocs/recombinaison_chez_les_virus_a_arn_negatif_un_modele_chez_les_hantavirus_3144/breve.phtml?tab=texte

Cardona-Ospina, J. A., Rojas-Gallardo, D. M., Garzón-Castaño, S. C., Jiménez-Posada, E. V., & Rodríguez-Morales, A. J. (2021). Phylodynamic analysis in the understanding of the current COVID-19 pandemic and its utility in vaccine and antiviral design and assessment. Human Vaccines & Immunotherapeutics, 17(8), 2437-2444. https://doi.org/10.1080/21645515.2021.1880254

Charleston, M. (2013). Phylogeny. In Brenner's Encyclopedia of Genetics (p. 324-325). Elsevier. https://doi.org/10.1016/B978-0-12-374984-0.01160-8

Corvol, P. (2021). L'envolée des publications scientifiques en temps de Covid-19—Séparer le bon grain de l'ivraie. Médecine/sciences, 37(4), Art. 4. https://doi.org/10.1051/medsci/2021039

Dallavilla, T., Bertelli, M., Morresi, A., Bushati, V., Stuppia, L., Beccari, T., ... & Marceddu, G. (2020). Bioinformatic analysis indicates that SARS-CoV-2 is unrelated to known artificial coronaviruses. European Review for Medical & Pharmacological Sciences, 24(8).

Daubin, V., & Szöllősi, G. J. (2016). Horizontal Gene Transfer and the History of Life. Cold Spring Harbor Perspectives in Biology, 8(4), a018036. https://doi.org/10.1101/cshperspect.a018036

Del Amparo, R., González-Vázquez, L. D., Rodríguez-Moure, L., Bastolla, U., & Arenas, M. (2023). Consequences of Genetic Recombination on Protein Folding Stability. Journal of Molecular Evolution, 91(1), 33-45. https://doi.org/10.1007/s00239-022-10080-2

Delaune, D., Hul, V., Karlsson, E. A., Hassanin, A., Ou, T. P., Baidaliuk, A., ... & Duong, V. (2021). A novel SARS-CoV-2 related coronavirus in bats from Cambodia. Nature communications, 12(1), 6563.

Domingo, J. L. (2021). What we know and what we need to know about the origin of SARS-CoV-2. Environmental Research, 200, 111785. https://doi.org/10.1016/j.envres.2021.111785

Doolittle, W. F. (1999). Phylogenetic Classification and the Universal Tree. Science, 284(5423), 2124-2128. https://doi.org/10.1126/science.284.5423.2124

Drummond, A. J., & Rambaut, A. (2007). BEAST: Bayesian evolutionary analysis by sampling trees. BMC evolutionary biology, 7(1), 1-8.

Drummond, A. J., Ho, S. Y. W., Phillips, M. J., & Rambaut, A. (2006). Relaxed Phylogenetics and Dating with Confidence. PLoS Biology, 4(5), e88. https://doi.org/10.1371/journal.pbio.0040088

Duchene, S., Featherstone, L., Haritopoulou-Sinanidou, M., Rambaut, A., Lemey, P., & Baele, G. (2020). Temporal signal and the phylodynamic threshold of SARS-CoV-2. Virus Evolution, 6(2), veaa061. https://doi.org/10.1093/ve/veaa061

Elie, B., & Alizon, S. (2020). Analyses génomiques et phylodynamiques du Sars-CoV-2. Revue Francophone Des Laboratoires, 2020(526), 57-62. https://doi.org/10.1016/S1773-035X(20)30314-2

Fung, S.-Y., Yuen, K.-S., Ye, Z.-W., Chan, C.-P., & Jin, D.-Y. (2020). A tug-of-war between severe acute respiratory syndrome coronavirus 2 and host antiviral defence: Lessons from other pathogenic viruses. Emerging Microbes & Infections, 9(1), 558-570. https://doi.org/10.1080/22221751.2020.1736644

Gondeau, A., Aouabed, Z., Hijri, M., Peres-Neto, P. R., & Makarenkov, V. (2019). Object weighting: a new clustering approach to deal with outliers and cluster overlap in computational biology. IEEE/ACM transactions on computational biology and bioinformatics, 18(2), 633-643.

Grandcolas, P., Daubin, V., Chave, J., Kergoat, G. J., Samadi, S., & Vignes-Lebbe, R. (2013). Systématique, Phylogénie. https://doi.org/10.13140/RG.2.1.4516.5286

Guo, Y. R., Cao, Q. D., Hong, Z. S., Tan, Y. Y., Chen, S. D., Jin, H. J., ... & Yan, Y. (2020). The origin, transmission and clinical therapies on coronavirus disease 2019 (COVID-19) outbreak–an update on the status. Military medical research, 7, 1-10.

Haeckel, E. (1866). Generelle Morphologie der Organismen: Allgemeine Grundzüge der organischen Formen-Wissenschaft, mechanisch begründet durch die von Charles Darwin reformierte Descendenz-Theorie. Band 1: Allgemeine Anatomie. Band 2: Allgemeine Entwicklungsgeschichte. de Gruyter.

Hao, P., Zhong, W., Song, S., Fan, S., & Li, X. (2020). Is SARS-CoV-2 originated from laboratory? A rebuttal to the claim of formation via laboratory recombination. Emerging microbes & infections, 9(1), 545-547.

Ho, M.-W. (2002, novembre 13). Recent Evidence Confirms Risks of Horizontal Gene Transfer. http://www.i-sis.org.uk/FSAopenmeeting.php

Ho, S. Y. W., Lanfear, R., Bromham, L., Phillips, M. J., Soubrier, J., Rodrigo, A. G., & Cooper, A. (2011). Time-dependent rates of molecular evolution. Molecular Ecology, 20(15), 3087-3101. https://doi.org/10.1111/j.1365-294X.2011.05178.x

Kapli, P., Yang, Z., & Telford, M. J. (2020). Phylogenetic tree building in the genomic age. Nature Reviews Genetics, 21(7), 428-444. https://doi.org/10.1038/s41576-020-0233-0

Keese, P. (2008). Risks from GMOs due to Horizontal Gene Transfer. Environmental Biosafety Research, 7(3), 123-149. https://doi.org/10.1051/ebr:2008014

Khailany, R. A., Safdar, M., & Ozaslan, M. (2020). Genomic characterization of a novel SARS-CoV-2. Gene Reports, 19, 100682. https://doi.org/10.1016/j.genrep.2020.100682

Kumar, S., Stecher, G., Li, M., Knyaz, C., & Tamura, K. (2018). MEGA X: Molecular Evolutionary Genetics Analysis across Computing Platforms. Molecular Biology and Evolution, 35(6), 1547-1549.

Lai, M. M. (1992). RNA recombination in animal and plant viruses. Microbiological Reviews, 56(1), 61-79.

Lam TT-Y, Jia N, Zhang Y-W, Shum MH-H, Jiang J-F, Zhu H-C, Tong Y-G, Shi Y-X, Ni X-B, Liao Y-S, Li W-J, Jiang B-G, Wei W, Yuan T-T, Zheng K, Cui X-M, Li J, Pei G-Q, Qiang X, Cheung WY-M, Li L-F, Sun F-F, Qin S, Huang J-C, Leung GM, Holmes EC, Hu Y-L, Guan Y, Cao W-C. Identifying SARS-CoV-2 related coronaviruses in Malayan pangolins. Nature. 2020;583:282–5.

Lam, T. T. Y., Jia, N., Zhang, Y. W., Shum, M. H. H., Jiang, J. F., Zhu, H. C., ... & Cao, W. C. (2020). Identifying SARS-CoV-2-related coronaviruses in Malayan pangolins. Nature, 583(7815), 282-285.

Lau, S. K., Luk, H. K., Wong, A. C., Li, K. S., Zhu, L., He, Z., ... & Woo, P. C. (2020). Possible bat origin of severe acute respiratory syndrome coronavirus 2. Emerging infectious diseases, 26(7), 1542.

Lauxmann, M. A., Santucci, N. E., & Autrán-Gómez, A. M. (2020). The SARS-CoV-2 coronavirus and the COVID-19 outbreak. International braz j urol, 46, 6-18.

Lawrence, J. G., & Ochman, H. (2002). Reconciling the many faces of lateral gene transfer. Trends in Microbiology, 10(1), 1-4. https://doi.org/10.1016/S0966-842X(01)02282-X

Layeghifard, M., Peres-Neto, P. R., & Makarenkov, V. (2012). Using directed phylogenetic networks to retrace species dispersal history. Molecular phylogenetics and evolution, 64(1), 190-197.

Lecointre, G., Le Guyader, H., & Visset, D. (2016). Classification phylogénétique du vivant (4e éd. Revue et augmentée). Belin.

Lefeuvre, C., Przyrowski, É., & Apaire-Marchais, V. (2020). Aspects virologiques et diagnostic du coronavirus Sars-CoV-2. Actualités Pharmaceutiques, 59(599), 18-23. https://doi.org/10.1016/j.actpha.2020.08.005

Lemaitre, C., Salson, M., & Touzet, H. (2022). Comment la bioinformatique a résolu le puzzle du génome du SARS-CoV-2. Interstices. https://hal.inria.fr/hal-03896532

Li, J. Y., You, Z., Wang, Q., Zhou, Z. J., Qiu, Y., Luo, R., & Ge, X. Y. (2020). The epidemic of 2019-novel-coronavirus (2019-nCoV) pneumonia and insights for emerging infectious diseases in the future. Microbes and infection, 22(2), 80-85.

Lundstrom, K. (2020). Coronavirus pandemic—therapy and vaccines. Biomedicines, 8(5), 109.

Lyons-Weiler, J. (2020). Pathogenic priming likely contributes to serious and critical illness and mortality in COVID-19 via autoimmunity. Journal of translational autoimmunity, 3, 100051.

Makarenkov, V., & Leclerc, B. (1996). Circular orders of tree metrics, and their uses for the reconstruction and fitting of phylogenetic trees. In Mathematical hierarchies and Biology (pp. 183-208).

Makarenkov V (1997). Propriétés combinatoires des distances d'arbre: Algorithmes et applications. Doctoral dissertation, Paris, EHESS.

Leclerc, B., & Makarenkov, V. (1998). On some relations between 2-trees and tree metrics. Discrete Mathematics, 192(1-3), 223-249.

Makarenkov, V., & Leclerc, B. (2000). Comparison of additive trees using circular orders. Journal of Computational Biology, 7(5), 731-744.

Makarenkov, V., & Legendre, P. (2000). Improving the additive tree representation of a dissimilarity matrix using reticulations. In Data analysis, classification, and related methods (pp. 35-40). Springer Berlin Heidelberg.

Makarenkov, V., Legendre, P., & Desdevises, Y. (2004). Modelling phylogenetic relationships using reticulated networks. Zoologica scripta, 33(1), 89-96.

Makarenkov, V., Mazoure, B., Rabusseau, G., & Legendre, P. (2021). Horizontal gene transfer and recombination analysis of SARS-CoV-2 genes helps discover its close relatives and shed light on its origin. BMC ecology and evolution, 21, 1-18.

Markov, P. V., Ghafari, M., Beer, M., Lythgoe, K., Simmonds, P., Stilianakis, N. I., & Katzourakis, A. (2023). The evolution of SARS-CoV-2. Nature Reviews Microbiology, 21(6), 361-379.

Mohamadian, M., Chiti, H., Shoghli, A., Biglari, S., Parsamanesh, N., & Esmaeilzadeh, A. (2021). COVID-19 : Virology, biology and novel laboratory diagnosis. The Journal of Gene Medicine, 23(2), e3303. https://doi.org/10.1002/jgm.3303

Ochman, H., Lawrence, J. G., & Groisman, E. A. (2000). Lateral gene transfer and the nature of bacterial innovation. Nature, 405(6784), 299-304. https://doi.org/10.1038/35012500

Pei, S., & Yau, S. S. T. (2021). Analysis of the genomic distance between bat Coronavirus RaTG13 and SARS-CoV-2 reveals multiple origins of COVID-19. Acta Mathematica Scientia, 41(3), 1017-1022.

Pellett, P. E., Mitra, S., & Holland, T. C. (2014). Basics of virology. Handbook of Clinical Neurology, 123, 45-66. https://doi.org/10.1016/B978-0-444-53488-0.00002-X

Philippe, H., & Douady, C. J. (2003). Horizontal gene transfer and phylogenetics. Current Opinion in Microbiology, 6(5), 498-505. https://doi.org/10.1016/j.mib.2003.09.008

Prabakaran P, Gan J, Feng Y, Zhu Z, Choudhry V, Xiao X, Ji X, Dimitrov DS. Structure of severe acute respiratory syndrome coronavirus receptor-binding domain complexed with neutralizing antibody. J Biol Chem. 2006;281:15829–36.

Prjibelski, A. D., Korobeynikov, A. I., & Lapidus, A. L. (2019). Sequence Analysis. In Encyclopedia of Bioinformatics and Computational Biology (p. 292-322). Elsevier. https://doi.org/10.1016/B978-0-12-809633-8.20106-4

Rambaut, A. (2016). Exploring the temporal structure of heterochronous sequences using TempEst (formerly Path-O-Gen). Virus evolution, 2(1), vew007.

Saitou, N., & Nei, M. (1987). The Neighbor-Joining Method: A New Method for Reconstructing Phylogenetic Trees. Molecular Biology and Evolution, 4(4), 406-425.

Sallard, E., Halloy, J., Casane, D., Helden, J. van, & Decroly, É. (2020). Retrouver les origines du SARS-CoV-2 dans les phylogénies de coronavirus. Médecine/sciences, 36(8-9), Art. 8-9. https://doi.org/10.1051/medsci/2020123

Samson, S., Lord, É., & Makarenkov, V. (2022). SimPlot++: a Python application for representing sequence similarity and detecting recombination. Bioinformatics, 38(11), 3118-3120.

Sanjuán, R. (2012). From Molecular Genetics to Phylodynamics : Evolutionary Relevance of Mutation Rates Across Viruses. PLoS Pathogens, 8(5), e1002685. https://doi.org/10.1371/journal.ppat.1002685

Sardon, J.-P. (2020). De la longue histoire des épidémies au Covid-19. Les Analyses de Population & Avenir, 26(8), 1-18. https://doi.org/10.3917/lap.026.0001

Segondy, M. (2020). Les coronavirus humains. Revue Francophone des Laboratoires, 2020(526), 32-39.

Sevillya, G., Adato, O., & Snir, S. (2020). Detecting horizontal gene transfer: A probabilistic approach. BMC Genomics, 21(1), 106. https://doi.org/10.1186/s12864-019-6395-5

Shu Y, McCauley J. GISAID: global initiative on sharing all influenza data–from vision to reality. Eurosurveillance. 2017;22(13):30494.

Smith, J., Johnson, A., & Brown, L.(2018). Detecting Horizontal Gene Transfer in Viruses Using T-Rex: A Bioinformatic Approach.

Song, L., Wu, S., & Tsang, A. (2018). Phylogenetic Analysis of Protein Family. In R. P. de Vries, A. Tsang, & I. V. Grigoriev (Éds.), Fungal Genomics (Vol. 1775, p. 267-275). Springer New York. https://doi.org/10.1007/978-1-4939-7804-5_21

Soucy, S. M., Huang, J., & Gogarten, J. P. (2015). Horizontal gene transfer: Building the web of life. Nature Reviews Genetics, 16(8), 472-482. https://doi.org/10.1038/nrg3962

Stapley, J., Feulner, P. G. D., Johnston, S. E., Santure, A. W., & Smadja, C. M. (2017). Recombination: The good, the bad and the variable. Philosophical Transactions of the Royal Society B: Biological Sciences, 372(1736), 20170279. https://doi.org/10.1098/rstb.2017.0279

Stavrinides, J., & Ochman, H. (2009). Phylogenetic Methods. In M. Schaechter (Éd.), Encyclopedia of Microbiology (Third Edition) (p. 247-260). Academic Press. https://doi.org/10.1016/B978-012373944-5.00272-8

Suchard, M. A., Lemey, P., Baele, G., Ayres, D. L., Drummond, A. J., & Rambaut, A. (2018). Bayesian phylogenetic and phylodynamic data integration using BEAST 1.10. Virus Evolution, 4(1), vey016.

Torres-López, J. (2020). [What is the origin of SARS-CoV-2?]. Revista Medica Del Instituto Mexicano Del Seguro Social, 58(1), 1-2.

Touzet, H., Salson, M., & Lemaitre, C. (2022). Décoder le génome : Vers la compréhension du fonctionnement du SARS-CoV-2. Interstices. https://hal.inria.fr/hal-03750389

Tratner, I. (2003). SRAS: 1. Le virus. M/S: médecine sciences, 19(8), 885-891.

Turkahia, Y., Thornlow, B., Hinrichs, A., McBroome, J., Ayala, N., Ye, C., Maio, N. D., Haussler, D., Lanfear, R., & Corbett-Detig, R. (2021). Pandemic-Scale Phylogenomics Reveals Elevated Recombination Rates in the SARS-CoV-2 Spike Region (p. 2021.08.04.455157). bioRxiv. https://doi.org/10.1101/2021.08.04.455157

Volz, E. M., Koelle, K., & Bedford, T. (2013). Viral Phylodynamics. PLoS Computational Biology, 9(3), e1002947. https://doi.org/10.1371/journal.pcbi.1002947

Wang, H., Cui, X., Cai, X., & An, T. (2022). Recombination in Positive-Strand RNA Viruses. Frontiers in Microbiology, 13. https://www.frontiersin.org/articles/10.3389/fmicb.2022.870759

White, K. A., Enjuanes, L., & Berkhout, B. (2011). RNA virus replication, transcription and recombination. RNA Biology, 8(2), 182-183. https://doi.org/10.4161/rna.8.2.15663

Willems, M., Tahiri, N., & Makarenkov, V. (2014). A new efficient algorithm for inferring explicit hybridization networks following the Neighbor-Joining principle. Journal of bioinformatics and computational biology, 12(05), 1450024.

Wong, G., Bi, Y. H., Wang, Q. H., Chen, X. W., Zhang, Z. G., & Yao, Y. G. (2020). Zoonotic origins of human coronavirus 2019 (HCoV-19/SARS-CoV-2): why is this work important? Zoological research, 41(3), 213.

World Health Organization (WHO). (2023, février 22). Coronavirus Disease (COVID-19) Situation Reports. https://www.who.int/emergencies/diseases/novel-coronavirus-2019/situation-reports

Ye, Z. W., Yuan, S., Yuen, K. S., Fung, S. Y., Chan, C. P., & Jin, D. Y. (2020). Zoonotic origins of human coronaviruses. International journal of biological sciences, 16(10), 1686.

Zhang, T., Wu, Q., & Zhang, Z. (2020). Probable pangolin origin of SARS-CoV-2 associated with the COVID-19 outbreak. Current biology, 30(7), 1346-1351.

Zhou, P., Yang, X. L., Wang, X. G., Hu, B., Zhang, L., Zhang, W., ... & Shi, Z. L. (2020). A pneumonia outbreak associated with a new coronavirus of probable bat origin. nature, 579(7798), 270-273.

Zuckerkandl, E., & Pauling, L. (1965). Evolutionary Divergence and Convergence in Proteins. In V. Bryson & H. J. Vogel (Éds.), Evolving Genes and Proteins (p. 97-166). Academic Press. https://doi.org/10.1016/B978-1-4832-2734-4.50017-6