UNIVERSITÉ DU QUÉBEC À MONTRÉAL

UN NOUVEAU LOGICIEL POUR L'ANALYSE DE SIMILARITÉ ENTRE LES SÉQUENCES GÉNÉTIQUES ET SON APPLICATION À DES DONNÉES ÉVOLUTIVES DE SARS-COV-2

MÉMOIRE

PRÉSENTÉ

COMME EXIGENCE PARTIELLE

DE LA MAÎTRISE EN INFORMATIQUE

PAR

SAMSON, STÉPHANE

DÉCEMBRE 2021

REMERCIEMENTS

Tout d'abord, j'aimerais remercier sincèrement Vladimir Makarenkov, mon directeur de recherche, pour le temps qu'il m'a consacré et ses conseils éclairés tout au long du projet. J'ai vraiment aimé faire ma maîtrise et c'est en grande partie dû à son encadrement. Merci pour m'avoir donner l'opportunité, je t'en suis reconnaissant.

J'aimerais également remercier ma copine pour son support infaillible et sa complicité. Merci Émilie d'avoir été là pour moi et pour tes encouragements quotidiens !

J'aimerais aussi remercier tous mes proches pour leur support, non seulement au cours de la maîtrise mais également dans les années passées. Je ne serais pas où je suis aujourd'hui sans vous ! Merci pour les milles et une choses que vous avez fait pour moi.

TABLE DES MATIÈRES

LISTE DES TABLEAUX vii
LISTE DES FIGURES viii
RÉSUMÉ xii
CHAPITRE I INTRODUCTION 1
1.1 Mise en contexte 1
1.2 Problématique et motivation 2
1.3 Objectifs 2
1.4 Organisation du mémoire 3
CHAPITRE II CONCEPTS PRÉLIMINAIRES 4
2.1 Variation génétique 5
2.1.1 Mutations 5
2.1.2 Recombinaisons 6
2.1.3 Transferts horizontaux de gènes 8
2.2 Modèles de substitutions 8
2.2.1 Modèle de Jukes-Cantor 10
2.2.2 Modèle de Kimura 11
2.2.3 Modèle F81 12
2.2.4 Modèle HKY85 13
2.2.5 Modèle GTR 14
2.2.6 Choix de modèles de substitutions 14
2.3 SARS-CoV-2 15
2.3.1 Génome de SARS-CoV-2 16
2.3.2 Gène S 17
2.3.3 Origine et évolution de SARS-CoV-2 18

CHAPITRE III ÉTAT DE L'ART 21
3.1 SimPlot 22
3.1.1 Données requises et préparation des groupes 22
3.1.2 Algorithme SimPlot 23
3.1.3 Algorithme Bootscan 25
3.1.4 Algorithme FindSite 27
3.2 PhiPack 28
3.2.1 Test Phi 29
3.2.2 Test NSS 30
3.2.3 Test Max-Chi 30
3.3 Autres outils de recombinaison 30
3.3.1 Recombination Analysis Tool 31
3.3.2 Recombination detection program 32
3.4 Réseaux de similarité de séquences 34
3.4.1 Types de réseaux 35
3.4.2 Bénéfices des représentations en réseaux 35
3.5 Logiciels incorporant des réseaux de similarités de séquences 39
3.5.1 PANADA 39
3.5.2 EGN 40
3.6 Beast2 41
CHAPITRE IV DÉVELOPPEMENT D'APPLICATION 44
4.1 Conception de l'application 45
4.1.1 Biopython 45
4.2 Page de création de groupes 45
4.3 Analyse SimPlot 48
4.3.1 Extraction des sous-séquences 48
4.3.2 Calcul de distances 49

4.3.3 Visualisation 50
4.3.4 Contrôle-qualité de l'analyse 51
4.4 Bootscan 56
4.5 FindSite 59
4.6 Réseaux de similarités 59
4.7 Recombinaison 62
4.7.1 PhiPack 62
4.7.2 Optimisation avec Numba 62
4.7.3 Test de proportions 64
CHAPITRE V MÉTHODOLOGIE 68
5.1 Jeu de données de 24 séquences de coronavirus du gène S 69
5.2 Jeu de données des 43 séquences de variants de SARS-CoV-2 70
5.3 Préparation des données 71
5.4 Détails d'analyse 71
CHAPITRE VI RÉSULTATS 73
6.1 Analyse du gène S 74
6.1.1 Analyse SimPlot 74
6.1.2 Analyse par l'algorithme Bootscan 75
6.1.3 Analyse FindSite 77
6.1.4 Analyse par réseaux de similarités 78
6.1.5 Analyse par test de proportions 80
6.2 Analyses des variants 81
6.2.1 Analyse du gène s des lignées Pango 81
6.2.2 Analyse du domaine RB des lignées Pango 83
6.3 Analyses phylogénétiques 85
CHAPITRE VII DISCUSSION 91
7.1 Discussion des résultats 92

7.2 Perspectives d'améliorations 93
7.3 Travaux futur sur SimPlot++ 93
CONCLUSION 95
ANNEXE A 97
ANNEXE B 99
RÉFÉRENCES 101

LISTE DES TABLEAUX

Tableau Page

2.1 Similarité des acides aminés entre SARS-CoV-2, bat-SL-CoVZXC21 et SARS-CoV au niveau des protéines structurelles (Chan *et al.*, 2020). 17

4.1 Tableau des comparaisons des vitesse d'exécution des versions du logiciel PhiPack selon les tailles d'alignements multiples 63

6.1 Tableau des résultats du test de proportions sur le jeu de données de 24 séquences du gène S. 81

A.1 Tableau des regroupements de séquences de coronavirus lors des analyses avec SimPlot++. Les numéros d'accession ainsi que les hôtes respectifs sont également présentés. 98

B.1 Tableau des lignées Pango, appellations et numéros d'accessions des 42 séquences de variants. 100

LISTE DES FIGURES

Figure Page

2.1 Représentation d'évènements de recombinaisons homologues et non-homologues entre des séquences parentales (section du haut). Les sites de recombianisons sont représentées par les lignes noires verticales. (Muslin *et al.*, 2019). 7

2.2 Représentation d'évènements de transferts horizontaux de gènes. Les différentes branches (tels 1, 2a et 2b) représentent les différentes trajectoire évolutives. (van de Guchte, 2017). 9

2.3 Matrice Q représentant le modèle de Jukes-Cantor. 10

2.4 Représentation des transitions et transversions entre les purines et pyrimidines (Lemey *et al.*, 2009). 12

2.5 Matrice Q représentant le modèle K80. 12

2.6 Matrice Q représentant le modèle F81. 13

2.7 Matrice Q représentant le modèle HKY85. 13

2.8 Matrice Q représentant le modèle GTR. 14

2.9 Représentation hiérarchique des cinq modèles de distances présentés (Lemey *et al.*, 2009). 15

2.10 Chronologie des évènements majeurs liés à SARS-CoV-2 au cours des premiers mois depuis la découverte initiale. (Hu *et al.*, 2021). 16

2.11 Représentation de la phylogénie des genres de coronavirus. (Singh et Yi, 2021) 19

2.12 Représentation du site de clivage de furine (SCF) du gène S de SARS-CoV-2 (Xia *et al.*, 2020). 20

3.1 Exemple de deux séquences représentant des histoires évolutives différentes, présentant une même distance génétique avec la séquence ancestrale. 24

3.2 Représentation du fonctionnement d'une fenêtre coulissante. . . . 25

3.3 Exemple de sorties SimPlot et Bootscan (Wu *et al.*, 2013). 26

3.4 Exemple des configurations possibles des séquences et de résultats d'analyse de sites informatifs (Robertson *et al.*, 1995). 28

3.5 Exemple de sortie d'analyse du « Recombination Analysis Tool » (Etherington *et al.*, 2005). 32

3.6 Exemple de sortie d'analyse de RDP (Martin *et al.*, 2015). 34

3.7 Représentation de différents types de réseaux. 35

3.8 Représentation de l'effet du seuil minimal de similarité sur la connectivité du réseau (Atkinson *et al.*, 2009). 37

3.9 Différentes représentations par graphes de partages de gènes entre des génomes (Corel *et al.*, 2016). 38

3.10 Exemple de graphique produit par l'entremise de PANADA (Martin *et al.*, 2013). 40

3.11 Exemple de graphique produit par l'entremise d'EGN (Halary *et al.*, 2013). 41

3.12 Exemple d'arbre consensus de Beast2 (Barido-Sottani *et al.*, 2018) 43

4.1 Page de création de groupes de SimPlot++. 48

4.2 Exemple de sortie d'analyse SimPlot par SimPlot++. 51

4.3 Représentation par heatmap des «gaps» des séquences consensus . 53

4.4 Représentation par heatmap des omissions et distances entre les séquences 55

4.5 Réimplémentation du pipeline d'analyse Bootscan à l'aide de la suite de logiciels PHYLIP. 57

4.6 Exemple de sortie d'analyse Bootscan par SimPlot++. 58

4.7 Exemple de réseau de similarité sans critères de filtre par SimPlot++. 60

6.1 Graphique SimPlot de la distance entre les groupes de séquences de CoV et le groupe SARS-CoV-2 pour le gène S. 75

6.2 Graphique Bootscan les groupes de séquences de CoV et le groupe SARS-CoV-2 pour le gène S. 76

6.3 Représentation des résultats de FindSite sur le gène S de 4 séquences de CoV. 78

6.4 Représentation en réseau des similarités entre les groupes de CoV pour le gène S 80

6.5 Graphique SimPlot de la similarité entre les groupes de variants de SARS-CoV-2 regroupés par leur lignée Pango. L'analyse a été effectuée avec une fenêtre coulissante de 200 pb, un pas de 20 pb et le modèle HKY85. 82

6.6 Heatmap des similarités entre les groupes de variants de SARS-CoV-2 regroupés par leur lignée Pango. 83

6.7 Graphique SimPlot de la distance entre les groupes de variants de SARS-CoV-2 pour le domaine RB, selon leur lignée Pango. 84

6.8 Heatmap des similarités entre les groupes de variants de SARS-CoV-2 regroupés par leur lignée Pango. 85

6.9 Reconstruction phylogénétique par Beast2 des séquences de SARS-CoV-2 et des séquences de CoV RaTG13, de pangolin de Guangdong et de Guangxi. 86

6.10 Reconstruction phylogénétique par Beast2 des séquences de SARS-CoV-2 et des séquences de CoV RaTG13, de pangolin de Guangdong et de Guangxi. Un facteur temporel de 0.27 a été ajouté pour les âges des noeuds internes. 87

6.11 Reconstruction phylogénétique par Beast2 des séquences de variants de SARS-CoV-2 et de pangolin de Guangdong. 89

6.12 Reconstruction phylogénétique par Beast2 des séquences de variants de SARS-CoV-2 et de pangolin de Guangdong. Un facteur temporel de 0.75 a été ajouté pour les âges des noeuds internes. . 90

ACRONYMES

ACE2 **E**nzyme de **C**onversion de l'**A**ngiotensine 2.

AMS **A**lignement **M**ultiple de **S**équences.

CoV **Co**rona**v**irus.

Domaine RB Domaine **R**eceptor-**B**inding.

HGT **T**ransfert **H**orizontal de **G**ènes.

MERS **S**yndrome **R**espiratoire du **M**oyen-Orient.

NSS **N**eighborhood **S**imilarity **S**core.

PHI **P**airwise **H**omoplasy **I**ndex.

RAT **R**ecombination **A**nalysis **T**ool.

RDP **R**ecombination **D**etection **P**rogram.

SCF **S**ite de **C**livage de **F**urine.

SRAS **S**yndrome **R**espiratoire de **A**igu.

RÉSUMÉ

La pandémie de SARS-CoV-2 fait partie des maladies infectieuses les plus dangereuses qui soient apparues au cours du dernier siècle. Il a été hypothétisé par le passé que les souches de coronavirus ayant mené aux épidémies de SRAS aient passées des chauves-souris à l'homme par l'entremise d'hôtes intermédiaires tels les civettes (SARS-CoV) et les chameaux (MERS-CoV). Plusieurs études ayant été publiés depuis l'apparition de ce nouveau coronavirus suggèrent que son génome présente des similarités élevées avec les CoV de certaines chauves-souris pour la majorité de ses gènes et, à certaines souches de CoV de pangolins malais pour le domaine receptor binding (RB) de la protéine S (spike protein).

Afin d'étudier plus profondément l'origine évolutive du gène S et du domaine RB de SARS-CoV-2, un nouveau logiciel, SimPlot++, inspiré du logiciel original SimPlot de Stuart C. Ray, a été développé. En plus des méthodes d'analyses offertes par SimPlot, ce nouveau logiciel a permis d'effectuer des analyses statistiques de détection d'évènements de recombinaisons, ainsi que par réseaux de similarités de séquences. Une reconstruction phylogénétique avec une composante temporelle par Beast2 a aussi été produite pour le domaine RB.

Les résultats obtenus concordent avec la littérature quant à la présence d'une région possiblement recombinée entre SARS-CoV-2 et le CoV de pangolins de Guangdong. Les arbres phylogénétiques générés suggèrent que cet évènement de recombinaison au niveau du domaine RB se serait produit au début de l'an 2018. Ces arbres suggèrent également que les premiers variants de SARS-CoV-2 seraient apparus entre novembre et décembre 2019.

MOTS-CLÉS : Recombinaison, Transferts horizontaux de gènes, SARS-CoV-2, SimPlot++, Origine évolutive

CHAPITRE I

INTRODUCTION

1.1 Mise en contexte

Le logiciel SimPlot (Lole *et al.*, 1999) est une application bioinformatique publiée en 1999, permettant la détection d'évènements de recombinaisons et de mosaïcité à partir de séquences alignées d'acides nucléiques et d'acides aminées. Ce logiciel est principalement composé de quatre fonctionnalités principales : (1) Un outil permettant de rassembler plusieurs séquences dans un même groupe et d'en produire une séquence consensus, (2) Une analyse de similarité de séquences par fenêtre coulissante nommée SimPlot, (3) une analyse phylogénétique de séquences par fenêtre coulissante nommée Bootscan et, (4) une méthode d'identification des sites informatifs entre plusieurs séquences.

SimPlot est un logiciel encore employé fréquemment, plus de deux décennies après sa publication initiale, pour l'analyse rapide de jeux de données d'origines virales, bactériennes et eucaryotes. De plus, SimPlot a joué un rôle important dans la détection d'évènements de recombinaisons chez le coronavirus SARS-CoV-2 et dans la recherche de son origine évolutive (Lam *et al.*, 2020).

Compte tenu de l'ampleur de la pandémie de SARS-CoV-2 dans la population humaine et de la hausse du nombre de variants, il serait pertinent d'étudier d'avan-

tage l'origine de ce virus par la détection d'évènements de recombinaison au niveau de son gène S et du domaine *receptor-binding* (RB). De plus, une analyse phylogénétique pourrait être effectuée afin de déterminer quand ces recombinaisons auraient pu se produire dans le passé.

1.2 Problématique et motivation

Malgré l'emploi continu du logiciel SimPlot, celui-ci n'a pas été maintenu depuis sa version 3.5.1, publiée en 2003. Cette application comporte également des limitations par son incompatibilité avec les systèmes d'exploitations MacOS et Linux, sa tendance à échouer sans messages d'erreurs durant des analyses, ainsi qu'au nombre limité d'options disponibles pour certaines fonctionnalités.

Par conséquent, il a été jugé nécessaire de développer un nouveau logiciel open-source, visant à améliorer les composantes du logiciel SimPlot original et d'y inclure de nouvelles méthodes de détection de recombinaisons.

1.3 Objectifs

Le projet consiste ainsi en trois étapes principales. Premièrement, les fonctionnalités du logiciel SimPlot original devront être redéveloppées dans un nouveau logiciel nommé SimPlot++. Ces fonctionnalités seront améliorées afin d'offrir des méthodes de calculs de distances additionnelles, des représentations visuelles de qualité « publication » ainsi que, dans le cas de SimPlot, une interface permettant d'évaluer la qualité des résultats obtenus.

Deuxièmement, des méthodes nouvelles de détection d'évènements de recombinaisons seront ajoutés à SimPlot++, à la fois par l'intégration de méthodes préexistantes telles celles du programme Phipack (Bruen *et al.*, 2006) ainsi que par

le développement de nouvelles méthodes et représentations graphiques.

Finalement, le logiciel SimPlot++ sera employé afin d'analyser des jeux de données nucléotidiques et protéiques représentant des variants de SARS-CoV-2 ainsi que des séquences de coronavirus plus éloignées, afin de mieux comprendre l'origine du virus.

1.4 Organisation du mémoire

Le mémoire est organisé en chapitre de la manière suivante. Le chapitre 2 regroupe les connaissances préalables permettant de contextualiser et détailler les concepts récurrents tout au long du mémoire tels les méthodes de distances, mutations et réseaux de similarités. Le chapitre 3 est axé sur la description des outils, logiciels et méthodes présentement disponible en lien avec les fonctionnalités de SimPlot++. Par la suite, le chapitre 4 présente en détail les nouvelles fonctionnalités de SimPlot++ ainsi que les améliorations effectuées sur les fonctionnalités du logiciel SimPlot vu au chapitre précédant. Le chapitre 5 décrit la méthodologie employée lors de la formation des jeux de données pour l'analyse des jeux de données mentionné au troisième objectif. Le chapitre 6 sert à présenter les résultats issus de ces analyses et le chapitre 7 est axé sur la discussion de ces résultats et du développement de SimPlot++. Le mémoire est par la suite finalisé par la conclusion.

CHAPITRE II

CONCEPTS PRÉLIMINAIRES

Afin de faciliter la compréhension de certains sujets abordé au cours de ce mémoire, il a été jugé nécessaire de présenter d'abord certains des concepts biologiques fondamentaux qui y sont impliqués.

À cette fin, ce chapitre permettra d'introduire et de discuter de mécanismes biologiques causant des variations génétiques, de détailler les méthodes de bases permettant de calculer la distance entre deux séquences génétiques ainsi que de présenter le virus SARS-CoV-2 au niveau génomique, particulièrement en relation avec son gène S.

2.1 Variation génétique

2.1.1 Mutations

Plusieurs mécanismes biologiques sont responsables de la modification de la composition génétique des organismes. Ces mécanismes sont la source de la grande diversité génétique observable chez le vivant. Le mécanisme le plus couramment observé est la mutation ponctuelle, qui résulte en la modification d'une paire de base de l'ADN de l'organisme. Dépendamment de la mutation en question, cette modification peut affecter l'organisme de différentes façons : (1) la mutation peut être dite silencieuse si celle-ci ne résulte pas en un codon qui cause une variation d'acide aminé lors de la synthèse protéique. (2) La mutation peut être faux-sens si elle résulte en un codon qui cause une variation d'acide aminé. (3) La mutation est dite non-sens si elle transforme un codon codant pour un acide aminé en codon stop qui cause la fin prématurée de la traduction (Lodish *et al.*, 2000).

Autres que ces trois types de mutations principales, une mutation peut également représenter une suppression ou insertion de paires de bases dans le matériel génétique. Une telle mutation peut sévèrement affecter l'organisme par le déplacement du cadre de lecture des codons lors de la traduction. Dans un tel cas, tous les codons subséquents à la mutation pourraient être lus de façon erronée (Lodish *et al.*, 2000).

La prolifération d'une mutation dans une population est hautement dépendante de son impact sur l'organisme. Il est généralement possible d'aborder l'impact d'une mutation sur l'organisme en parlant de sa « fitness » biologique, soit de « la capacité relative d'un organisme à survivre et transmettre ses gènes à la génération suivante » (traduction libre) (King *et al.*, 2007). Ce concept de fitness est un aspect important afin d'expliquer le changement de distribution des phénotypes

dans une population au fil des générations (Stuart Barker, 2009). Comme une mutation peut être qualifié de négative, neutre ou bénéfique pour l'organisme, on peut généraliser l'impact des mutations sur le fitness de l'organisme muté. Une mutation négative va nuire aux fonctions essentielles de l'organisme, et donc à sa fitness, alors qu'une mutation bénéfique va augmenter celle-ci. La mutation neutre, ne modifiant pas le fitness de l'organisme, va soit éventuellement disparaître de la population ou y être fixé (Fleischmann, 1996) (Pérez-Losada *et al.*, 2015).

Ainsi, par l'occurrence d'une ou plusieurs mutations ponctuelles bénéfiques, il est possible d'observer, au sein d'une même espèce, l'apparition de nouveaux variants par l'accroissement de leur fitness relativement aux autres membres de l'espèce (Pérez-Losada *et al.*, 2015).

2.1.2 Recombinaisons

Une autre source majeure de diversité génétique est le mécanisme de recombinaison génétique. Chez les virus, le phénomène de recombinaison a lieu lorsqu'une co-infection par deux souches virales différentes dans une même cellule hôte se produit et que ces deux souches interagissent entre eux lors de leurs réplications respectives. Les virus descendants de ces évènements de recombinaisons possèdent des génomes combinant un ou des fragments provenant des deux souches parentes (Fleischmann, 1996). Ainsi, bien que la recombinaison ne produise pas directement de la variété génétique comme dans le cas de la mutation ponctuelle, de la variété génétique est générée par le mélange de différentes souches.

Les coronavirus possèdent une prévalence envers les évènements de recombinaisons de type homologue (Rehman *et al.*, 2020). Une recombinaison homologue signifie que le site de la recombinaison est similaire chez les deux génomes impliqués, alors qu'une recombinaison non-homologue pourrait se produire à des sites différents

sur chaque brin (Pérez-Losada *et al.*, 2015).

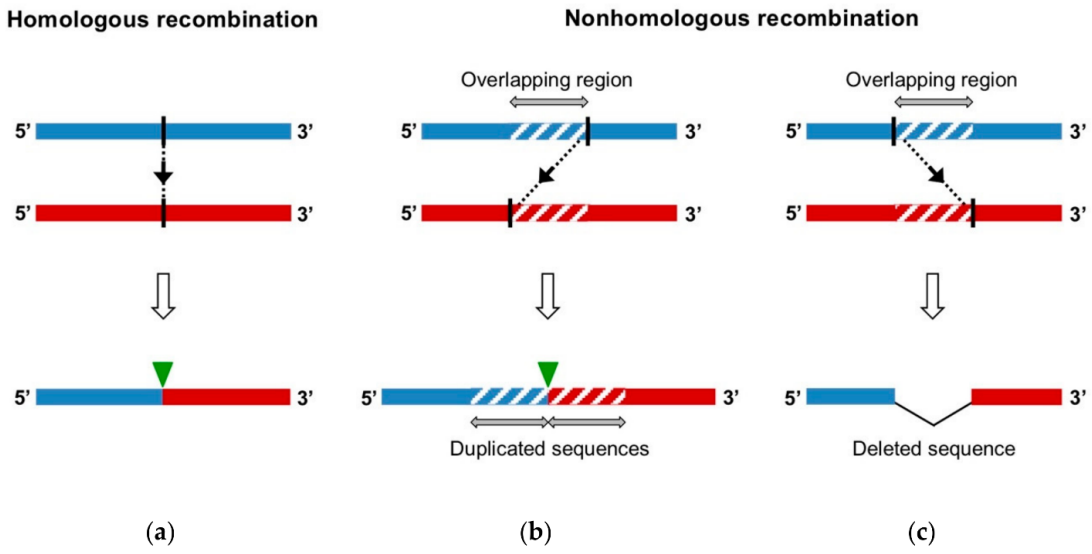


Figure 2.1: Représentation d'évènements de recombinaisons homologues et non-homologues entre des séquences parentales (section du haut). Les sites de recombianisons sont représentées par les lignes noires verticales. (Muslin *et al.*, 2019).

Cette prévalence des coronavirus envers les événements de recombinaisons homologues est une caractéristique importante lors de la détection des événements de recombinaisons chez ceux-ci. En effet, ces recombinaisons homologues mènent à l'apparition de génomes et de gènes composés de sous-régions provenant de différentes sources indépendantes (figure 2.1). Il s'agit de mosaïcité (Boni *et al.*, 2020).

En effet, les recombinaisons peuvent causer une augmentation de la transmission et de la virulence, un accroissement du nombre d'hôtes possibles, tant par une modification des vecteurs de transmissions, que par des évènements de zoonoses. De plus, les interactions entre le virus, le système immunitaire de l'hôte ainsi que la résistance aux médicaments peut être altérée (Pérez-Losada *et al.*, 2015).

2.1.3 Transferts horizontaux de gènes

Le terme « transfert horizontal de gènes » (HGT) est un terme qui représente les différents mécanismes par lesquels un transfert d'information génétique peut être effectué d'un génome à un autre sans que ceux-ci n'aient une relation parent – descendant (Un parent transmet son information génétique à ses descendants par transfert vertical de gènes) (Milner *et al.*, 2019). Les HGT peuvent être sous-divisé en deux modèles distincts : (1) Le modèle de transfert complet de gènes (modèle intergénique) où le gène d'intérêt est transféré dans son entièreté dans le génome receveur en remplaçant le gène d'origine si présent ou s'insérant dans le génome dans le cas contraire. (2) Le modèle de transfert partiel de gène (modèle intragénique) où seule une fraction du gène d'intérêt va être transféré au génome receveur (Boc et Makarenkov, 2011). Ce dernier modèle mène à la création de gènes mosaïques puisque différentes sous-régions du gène sont caractérisées par des histoires évolutives distinctes.

2.2 Modèles de substitutions

La distance génétique est une approche mathématique permettant d'évaluer la divergence entre plusieurs séquences génétiques (Nei, 1987). Bien qu'un grand nombre de tels modèles existent et sont fréquemment employés par la communauté scientifique, ceux-ci varient grandement par leur degré de complexité ainsi que par les suppositions biologiques inhérentes à chacun. Le choix d'un modèle par rapport à un autre est une décision qui doit être prise en compte en fonction du jeu de donnée et de la provenance des échantillons.

Cependant, la distance génétique entre deux séquences ne peut pas être évaluée uniquement par la proportion de positions à nucléotides égaux entre les séquence. Un tel calcul produirait une sous-estimation de la distance génétique réelle entre les

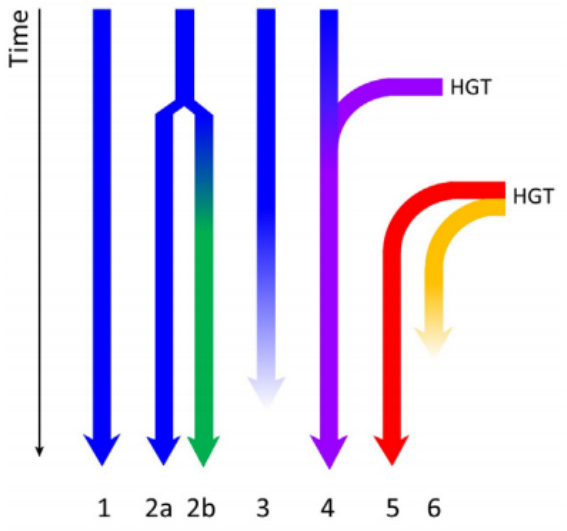


Figure 2.2: Représentation d'évènements de transferts horizontaux de gènes. Les différentes branches (tels 1, 2a et 2b) représentent les différentes trajectoire évolutives. (van de Guchte, 2017).

séquences puisque les substitutions aléatoires de nucléotides dans le temps ne sont pas prises en compte. Cette mesure ne représente donc que la distance observée entre deux séquences et non la distance génétique réelle. Il est donc nécessaire d'employer des modèles de substitutions afin de décrire les évènements génétiques aléatoires s'étant produits au cours de la divergence des séquences (Lemey *et al.*, 2009).

L'une des suppositions biologiques au centre des modèles de substitutions est la probabilité à un temps t qu'une paire de base subisse une mutation et soit substitué par une autre paire de base. Cet ensemble de probabilités est contenu dans une matrice Q, qui, dans le cas de l'ADN, représente une matrice carré 4X4 des quatre nucléotides (A, T, C et G) possibles. La matrice Q est donc la représentation matricielle du modèle de substitution, où les taux de substitutions entre chacun des nucléotides est spécifié (Jukes et Cantor, 1969).

2.2.1 Modèle de Jukes-Cantor

Le modèle de Jukes-Cantor (JC69), développé en 1969, est un modèle de substitution simple, basé uniquement sur la fréquence relative des nucléotides dans les séquences. Ce modèle suppose que chaque nucléotide a une chance égale (1/4 pour l'ADN) d'être substitué par n'importe quel autre nucléotide, produisant la matrice Q ci-dessous (figure 2.3) (Jukes et Cantor, 1969).

$$Q = \begin{pmatrix} -3/4\mu & 1/4\mu & 1/4\mu & 1/4\mu \\ 1/4\mu & -3/4\mu & 1/4\mu & 1/4\mu \\ 1/4\mu & 1/4\mu & -3/4\mu & 1/4\mu \\ 1/4\mu & 1/4\mu & 1/4\mu & -3/4\mu \end{pmatrix} \tag{2.1}$$

Figure 2.3: Matrice Q représentant le modèle de Jukes-Cantor.

Une fois la matrice Q du modèle établie, les probabilités de substitutions entre les nucléotides peuvent être calculés par l'exponentielle de la matrice Q. Par cette opération, deux équations peuvent être obtenues, soit :

$$Pii(t) = 1/4 + 3/4exp(-\mu t) \tag{2.2}$$

$$Pij(t) = 1/4 - 1/4exp(-\mu t) \tag{2.3}$$

Où P_{ii} représente la probabilité qu'un nucléotide demeure le même tout au long du temps évolutif t, P_{ij} représente la probabilité qu'un nucléotide soit substitué et μ représente le taux de substitution instantané (Lemey *et al.*, 2009).

À partir de ces deux équations, l'équation de distance génétique du modèle JC69 peut être produite :

$$d = -3/4ln(1 - 4/3P) \tag{2.4}$$

Où P est la distance observée au temps t entre les deux séquences.

Dépendamment de l'organisme d'intérêt, d'autres suppositions peuvent également être apportés aux modèles afin de représenter certaines caractéristiques biologiques. Ces suppositions seront traduites en taux de mutations instantanés dans la matrice Q du modèle.

2.2.2 Modèle de Kimura

Le modèle de Kimura (K2P ou K80) est basé sur le modèle JC69 mais y apporte un niveau de complexité supplémentaire en apportant le concept de transition et transversion. Ce concept est basé sur le principe que les acides nucléiques sont soit des purines (A et G) ou des pyrimidines (C et T). La supposition apportée au modèle K2P est que la probabilité qu'un nucléotide soit substitué par l'autre nucléotide de la même catégorie (transition) sera différente de la probabilité que ce nucléotide soit substitué par un nucléotide de l'autre catégorie (transversion) (figure 2.4). Autre que cette supposition, le modèle K2P suppose, tout comme le modèle JC69, que la fréquence relative de chaque nucléotide est de $\frac{1}{4}$ (figure 2.5) (Kimura, 1980).

Par cette matrice Q, le modèle de distance K2P peut être représenté par l'équation ci-dessous :

$$K = -1/2ln((1 - 2P - q)\sqrt{1 - 2q}) \tag{2.6}$$

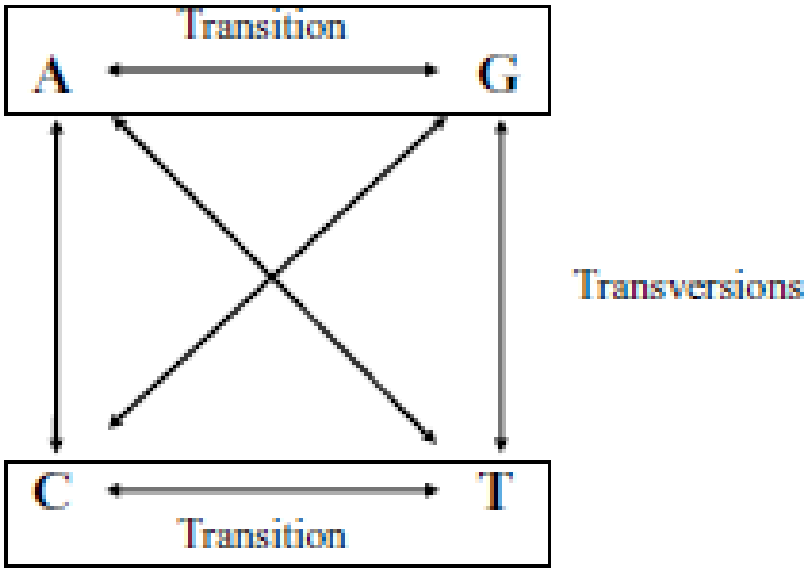


Figure 2.4: Représentation des transitions et transversions entre les purines et pyrimidines (Lemey *et al.*, 2009).

$$Q = \begin{pmatrix} * & \alpha & \beta & \Upsilon \\ \alpha & * & \Upsilon & \beta \\ \beta & \Upsilon & * & \alpha \\ \Upsilon & \beta & \alpha & * \end{pmatrix} \tag{2.5}$$

Figure 2.5: Matrice Q représentant le modèle K80.

2.2.3 Modèle F81

Le modèle F81, développé par Felsenstein en 1981, est également basé sur le modèle JC69 et en dévie uniquement par l'introduction de fréquence relative différentes pour chacun des quatre nucléotides. Ainsi, le modèle F81 diffère des modèles K2P et JC69 par le fait que les fréquences relatives ne sont pas assumé d'être $\frac{1}{4}$. Ainsi, le modèle JC69 pourrait être considéré comme un cas spécial du modèle F81

où les fréquences relatives sont effectivement égales. Ce modèle, contrairement au modèle de Kimura, ne prend pas en compte les transitions et transversions (figure 2.6) (Felsenstein, 1981).

$$Q = \begin{pmatrix} * & \pi_G & \pi_C & \pi_T \\ \pi_A & * & \pi_C & \pi_T \\ \pi_A & \pi_G & * & \pi_T \\ \pi_A & \pi_G & \pi_C & * \end{pmatrix} \tag{2.7}$$

Figure 2.6: Matrice Q représentant le modèle F81.

2.2.4 Modèle HKY85

Ce modèle, publié en 1985 par Hasegawa, Kishino et Yano, représente la fusion des modèles F81 et K2P. Celui-ci permet des fréquences relatives de nucléotides différentes de $\frac{1}{4}$ comme le modèle F81, et suppose, comme le modèle K2P, que les taux de substitutions sont différents lors des transitions et transversions (figure 2.7.) (Hasegawa *et al.*, 1985).

$$Q = \begin{pmatrix} * & \kappa\pi_G & \pi_C & \pi_T \\ \kappa\pi_A & * & \pi_C & \pi_T \\ \pi_A & \pi_G & * & \kappa\pi_T \\ \pi_A & \pi_G & \kappa\pi_C & * \end{pmatrix} \tag{2.8}$$

Figure 2.7: Matrice Q représentant le modèle HKY85.

2.2.5 Modèle GTR

Le dernier modèle présenté est le modèle de substitution GTR (« General Time-Reversible »), publié par Tavaré en 1986. Ce modèle est le plus complexe et général des modèles présentés car il intègre les concepts de transition, transversion, fréquences relatives différentes comme le modèle HKY85, mais permet également d'avoir des taux de substitutions différents pour chacune des six paires de nucléotides possibles (figure 2.8). (Tavare, 1986).

$$Q = \begin{pmatrix} * & \Upsilon_{AC}\pi_C & \Upsilon_{AG}\pi_G & \Upsilon_{AT}\pi_T \\ \Upsilon_{AC}\pi_A & * & \Upsilon_{CG}\pi_G & \Upsilon_{CT}\pi_T \\ \Upsilon_{AG}\pi_A & \Upsilon_{CG}\pi_C & * & \Upsilon_{GT}\pi_T \\ \Upsilon_{AT}\pi_A & \Upsilon_{CT}\pi_C & \Upsilon_{GT}\pi_G & * \end{pmatrix} \tag{2.9}$$

Figure 2.8: Matrice Q représentant le modèle GTR.

2.2.6 Choix de modèles de substitutions

Les modèles de substitutions représentent nécessairement des simplifications des vrais phénomènes biologiques en jeu. Les cinq modèles présentés ci-dessus sont relativement simple et peuvent, si appliqués à des jeux de données inappropriés, mener à des résultats erronés (Posada et Crandall, 2001). Il est donc important d'employer un outil d'analyse tels MEGA-X (Kumar *et al.*, 2018) ou Modeltest (Posada et Crandall, 2001) afin de déterminer les modèles les plus appropriés pour les données à analyser. La figure 2.9 ci-dessous résume les différences entre les modèles de base discutés.

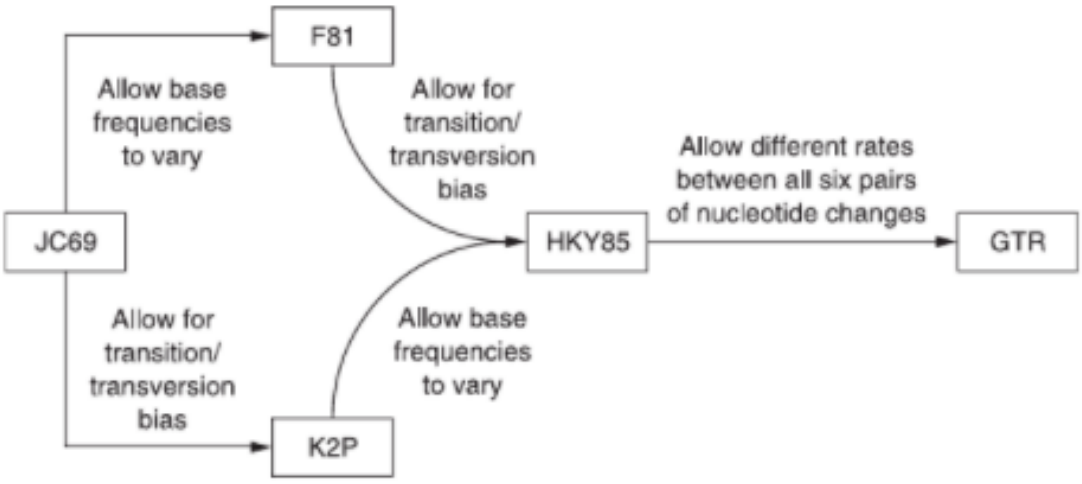


Figure 2.9: Représentation hiérarchique des cinq modèles de distances présentés (Lemey *et al.*, 2009).

2.3 SARS-CoV-2

Le SARS-CoV-2 est le troisième coronavirus hautement virulent d'origine zoonotique à faire son apparition dans la population humaine au 21ième siècle, suivant les épidémies de SRAS (syndrome respiratoire aigu sévère) de 2002 et du MERS (syndrome respiratoire du Moyen-Orient) de 2012. Ce nouveau pathogène a premièrement été détecté en décembre 2019 à Wuhan, dans la province de Hubei en Chine (Hu *et al.*, 2021). Depuis, le SARS-CoV-2 s'est propagé globalement, résultant en sa caractérisation officielle de pandémie en mars 2020 par l'Organisation Mondiale de la Santé. En août 2021, plus de 208 millions de cas de SARS-CoV-2 ont été reportés, résultant en la mort de 4.3 millions de personnes mondialement (Dong *et al.*, 2020). La figure 2.10 ci-dessous présente brièvement les évènements clés ayant marqués les débuts de la pandémie de SARS-CoV-2.

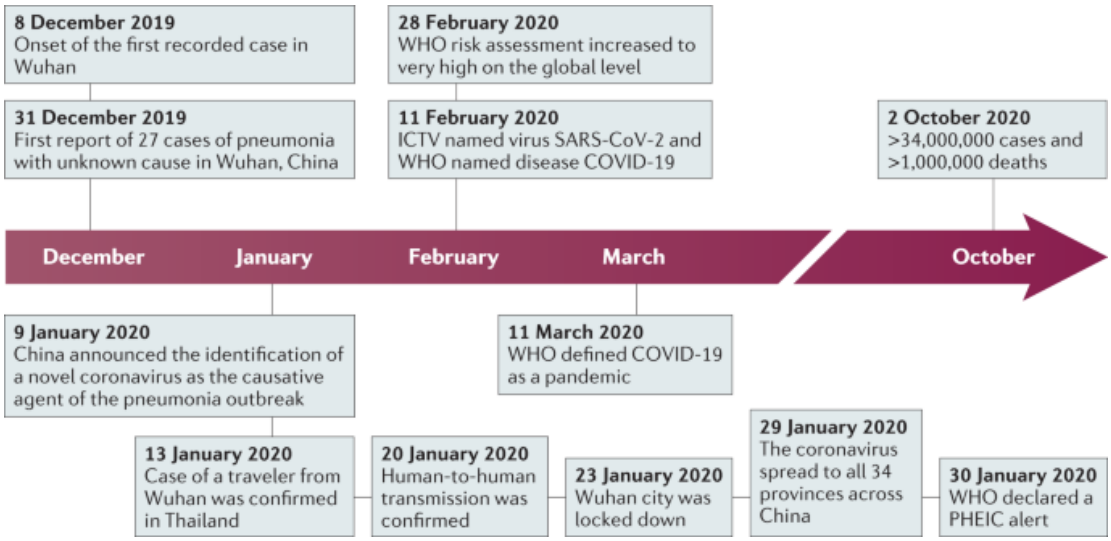


Figure 2.10: Chronologie des évènements majeurs liés à SARS-CoV-2 au cours des premiers mois depuis la découverte initiale. (Hu *et al.*, 2021).

2.3.1 Génome de SARS-CoV-2

Comme le SRAS et le MERS, le SARS-CoV-2 est un betacoronavirus (figure 2.11), caractérisé par la présence d'une enveloppe virale et possédant de l'ARN simple brin positif. Son génome est composé d'approximativement 30 000 nucléotides, codant pour 11 gènes distincts (Brant *et al.*, 2021), incluant quatre gènes structuraux : S, E, M et N. Ces quatre gènes codent respectivement pour les protéines du spicule, de l'enveloppe, de la membrane et de la nucléocapside. Cette organisation génomique est commune chez les betacoronavirus, avec lesquels SARS-CoV-2 partagent un haut niveau de similarité (Hu *et al.*, 2021).

En effet, SARS-CoV-2 présente plus de 91% de similarité avec le CoV du SRAS au niveau des acides aminés des protéines de l'enveloppe, de la membrane et de la nucléocapside, et au moins 94% de similarité lorsque comparé au génome de CoV de la chauve-souris CoVZXC21. Ainsi, des quatre protéines structurelles, seule la

protéine du spicule présente de la divergence (tableau 2.1) (Chan *et al.*, 2020).

Similarité des acides aminées (%)	2019-nCoV	2019-nCoV
	vs. bat-SL-CoVZXC21	vs. SARS-CoV
Spicule	80	76
Enveloppe	100	95
Membrane	99	91
Nucléoprotéine	94	94

Tableau 2.1: Similarité des acides aminés entre SARS-CoV-2, bat-SL-CoVZXC21 et SARS-CoV au niveau des protéines structurelles (Chan *et al.*, 2020).

2.3.2 Gène S

Le gène S de SARS-CoV-2 est responsable de la production d'une protéine de 1255 acides aminées, nommée protéine S ou protéine du spicule (« spike protein »). Cette protéine joue un rôle clé dans les mécanismes viraux de liaisons à la cellule hôte et de fusion membranaire, permettant l'infection des cellules de l'hôte et la reproduction subséquente du virus (Rota *et al.*, 2003). Cette activité virale est possible par l'interaction entre la protéine S et l'enzyme de conversion de l'angiotensine 2 (ACE2), une enzyme présente à la surface de certaines cellules chez plusieurs mammifères (tels les humains, chauve-souris, chien, chat et pangolins). La liaison entre l'ACE2 et la protéine S se produit au niveau du domaine RB (« receptor-binding ») de cette dernière et toute mutation se produisant au niveau du domaine RB a le potentiel d'affecter l'affinité entre l'ACE2 et la protéine S, résultant en une modification de la virulence (Shang *et al.*, 2020). D'autant plus, une comparaison entre le gène S du SARS-CoV et du SARS-CoV-2 suggère que le domaine RB de ce dernier présente une affinité plus élevée envers ACE2 que

le domaine RB du SARS-CoV, et ce, malgré une affinité égale ou moindre au niveau de la protéine S complète (Shang *et al.*, 2020). Tels que mentionné dans cette étude, ces résultats suggèrent que le domaine RB du SARS-CoV-2 permet une efficacité accrue d'entrée cellulaire tout en réduisant la capacité du système immunitaire à le cibler par son profil moins exposé. En plus de son importance dans les mécanismes viraux, il a été suggéré qu'un évènement de zoonose, ou saut d'espèce, par un betacoronavirus devrait nécessiter l'apparition de mutations au niveau du gène S du virus en question. Ainsi, ce gène et le domaine RB sont des régions d'intérêts marqués dans l'étude du SARS-CoV-2 (Song *et al.*, 2005).

2.3.3 Origine et évolution de SARS-CoV-2

Les analyses phylogénétiques effectuées suite au séquençage du SARS-CoV-2 ont permis de déterminer que le génome de coronavirus le plus similaire à SARS-CoV-2 est le CoV de chauve-souris *Rhinolophus affinis* RaTG13 provenant de la région de Yunnan en Chine, avec 96,2% de similarité (Guo *et al.*, 2020). La grande similarité entre les deux génomes laisse suggérer que les chauves-souris représentent potentiellement un réservoir naturel de SARS-CoV-2 mais l'absence de similarité au niveau du domaine RB ainsi que la possibilité que des évènements de recombinaisons s'y soit produites permettent d'exclure RaTG13 comme ancêtre direct de SARS-CoV-2 (Lau *et al.*, 2020).

Autre que la chauve-souris RaTG13, le génome de CoV de certains pangolins présentent également des similarités fortes avec le SARS-CoV-2. En effet, des séquences de CoV issues d'échantillons de pangolins malais dans la région de Guangdong en Chine présentent un taux de similarité très élevé avec SARS-CoV-2 en ce qui concerne les acides aminés représentant le domaine RB (Zhang *et al.*, 2020). Le CoV de pangolin de Guangdong présente un seul résidu différent au

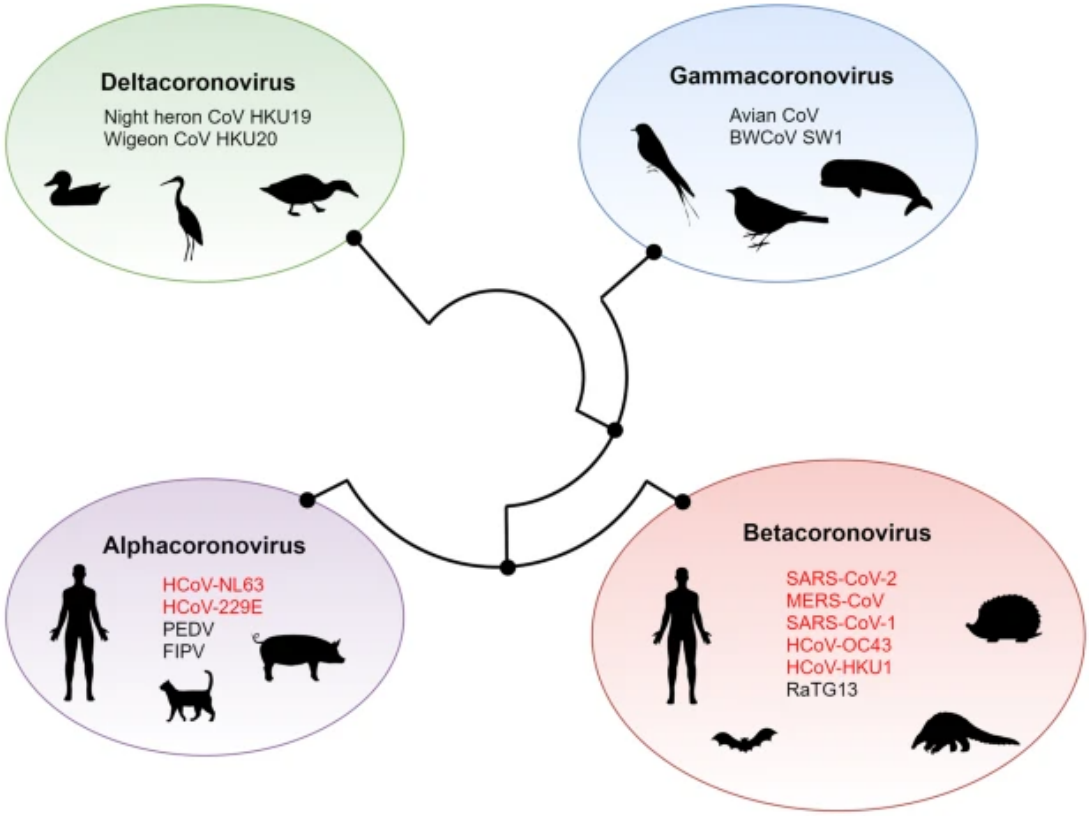


Figure 2.11: Représentation de la phylogénie des genres de coronavirus. (Singh et Yi, 2021)

niveau du motif de liaison à l'ACE2 du domaine RB (résidus 442-512), et ce résidu n'est pas l'un des 5 résidu clé impliqué directement dans la liaison avec l'ACE2. RaTG13, pour sa part, présente 14 résidus différents dans le même motif de 70 acides aminés, incluant 4 résidus parmi les 5 résidus clés. Ainsi, la proximité génétique entre SARS-CoV-2 et le CoV de pangolin de Guangdong au niveau du domaine RB et de leur affinité à l'ACE2 humaine en fait un candidat de réservoir naturel de SARS-CoV-2 (Zhang *et al.*, 2020).

Cependant, la protéine S de SARS-CoV-2 présente un site de clivage de furine (SCF) (figure 2.12) entre ces deux sous-unités S1 et S2 qui n'est pas retrouvé

ni dans les séquences de RaTG13, ni chez le CoV de pangolin de Guangdong. D'ailleurs, la présence de ce SCF est rare parmi les beta-CoV et pourrait expliquer en partie le succès de SARS-CoV-2 lors de l'infection des cellules de l'hôte (Xia *et al.*, 2020). Le SCF est effectivement connu chez les virus de l'influenza pour son rôle dans d'accroissement de la virulence(Claas *et al.*, 1998). Ainsi, le SCF de SARS-CoV-2 pourrait avoir été acquis par recombinaison avec un coronavirus éloigné.

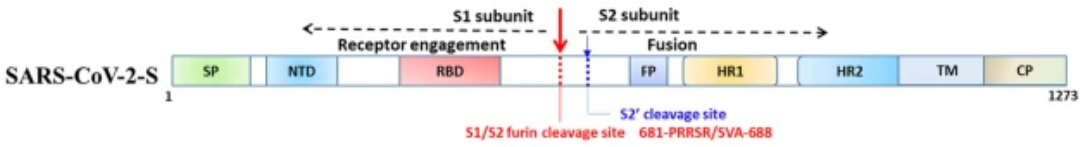


Figure 2.12: Représentation du site de clivage de furine (SCF) du gène S de SARS-CoV-2 (Xia *et al.*, 2020).

Ainsi, tels que présenté par Makarenkov et al (Makarenkov *et al.*, 2021) trois hypothèses peuvent être avancées afin d'expliquer l'évolution de SARS-CoV-2. (1) La similarité entre le domaine RB du pangolin de Guangdong et celui de SARS-CoV-2 est le résultat du phénomène d'évolution divergente à partir d'un génome proche de RaTG13, ayant favorisé des substitutions au niveau du domaine RB. Cette hypothèse suggère une absence d'évènements de recombinaisons et d'interactions dans le passé proche entre SARS-CoV-2 et les CoV de pangolins. (2) Un ou plusieurs évènements de recombinaisons complets ou partiels se sont produits entre des génomes de CoV dans une cellule hôte, résultant possiblement en la présence de gènes mosaïques chez SARS-CoV-2. (3) Les mutations et similarités observées sont le résultat du phénomène d'évolution parallèle, où les similarités génomiques entre les espèces de coronavirus ont été obtenues suite à des contraintes évolutives similaires. Une telle évolution parallèle a déjà été observée chez SARS-CoV *in vitro* (Sheahan *et al.*, 2008).

CHAPITRE III

ÉTAT DE L'ART

Plusieurs logiciels et outils ont été publiés au fil des années afin de permettre la détection d'évènements de recombinaisons. Dans ce chapitre, les principaux logiciels de ce type seront décortiqués afin de présenter leurs fonctionnements respectifs. De plus, d'autres logiciels comportant des approches intéressantes en lien avec le logiciel SimPlot++ et l'analyse subséquente des variants de SARS-CoV-2 seront discutés.

3.1 SimPlot

Le logiciel SimPlot, produit et distribué par Stuart Ray (Lole *et al.*, 1999) est un logiciel permettant la détection rapide d'évènements de recombinaisons par trois approches différentes : une analyse de distance génétique par fenêtre coulissante (aussi référée comme l'analyse SimPlot), une approche phylogénétique par la méthode de bootscan, ainsi qu'une identification des sites informatifs des séquences (FindSite). Ce logiciel a été publié initialement en 1999 et a été maintenu jusqu'à la version 3.5.1 en août 2003. Malgré l'absence de mises à jour récentes, ce logiciel est encore fréquemment employé et cité par la communauté scientifique. Par exemple, SimPlot a joué un rôle important dans l'étude de l'origine du virus du SARS-CoV-2, permettant d'établir rapidement des liens évolutifs potentiels entre ce virus et d'autres coronavirus provenant de chauve-souris et de pangolins malais (Zhang *et al.*, 2020).

3.1.1 Données requises et préparation des groupes

Le logiciel SimPlot requiert en entrée un fichier de séquences nucléotidiques ou protéiques de format Fasta, Phylip ou Nexus. Afin d'assurer la qualité de l'analyse, les séquences doivent être alignées préalablement à l'emploi de l'application. Une fois le fichier de séquences pris en charge, l'utilisateur doit manuellement regrouper les séquences du fichier en groupes de séquences évolutivement rapprochés (fortement similaires). Toutes les séquences d'un même groupe seront combinées par une méthode de construction de séquence consensus. Ainsi, si le jeu de données initial comportait 15 séquences réparties uniformément en cinq groupes de trois séquences, cinq séquences consensus seraient produites durant cette première étape du logiciel. Ces cinq séquences consensus seront utilisées pour les analyses SimPlot et phylogénétiques par bootstrap. Il est important de noter que des groupes com-

portant une seule séquence unique peuvent être formés, ne forçant pas l'utilisateur à employer l'approche des séquences consensus.

3.1.2 Algorithme SimPlot

Typiquement, une analyse de distance génétique entre deux séquences serait calculée sur l'entièreté de la longueur des séquences. Cependant, une telle analyse ne prend pas en compte le fait que certaines sous-sections d'une séquence peuvent avoir des origines évolutives différentes, comme il pourrait être le cas suite à un évènement de recombinaison. La figure 3.1 ci-dessous représente un tel exemple où la distance génétique calculée par la méthode de Jukes-Cantor entre une séquence de référence (séquence 1) et deux autres séquences sont similaires sans toutefois représenter la même histoire évolutive. Dans le cas de la séquence 2, des mutations ponctuelles ont menées à la substitution de quatre thymines (T) en adénines (A), distribués à travers l'entièreté de la séquence. Dans le cas de la séquence 3, quatre nucléotides ont également été remplacés par d'autres nucléotides, mais, contrairement à la séquence 2, ceux-ci sont regroupés afin de représenter de façon simpliste un évènement de transfert horizontal. Appliquer une méthode de calcul de distance génétique à ce scénario ne va pas révéler cette particularité entre les deux séquences puisque les résultats sont égaux (0,3041 par la méthode de Jukes-Cantor). Ce scénario est représenté à la figure 3.1 ci-dessous.

Afin d'identifier de tels évènements par calculs de distances génétiques, SimPlot emploi une méthode de calcul par fenêtres coulissantes. C'est-à-dire que contrairement à l'exemple précédant, les N premiers nucléotides de chaque séquence sont extraits et leurs distances génétiques sont calculés. Ce nombre N de nucléotide demeurera constant tout au long de l'analyse et est nommé la taille de fenêtre. Lorsque les distances de la première fenêtre sont calculées, cette fenêtre est dé-

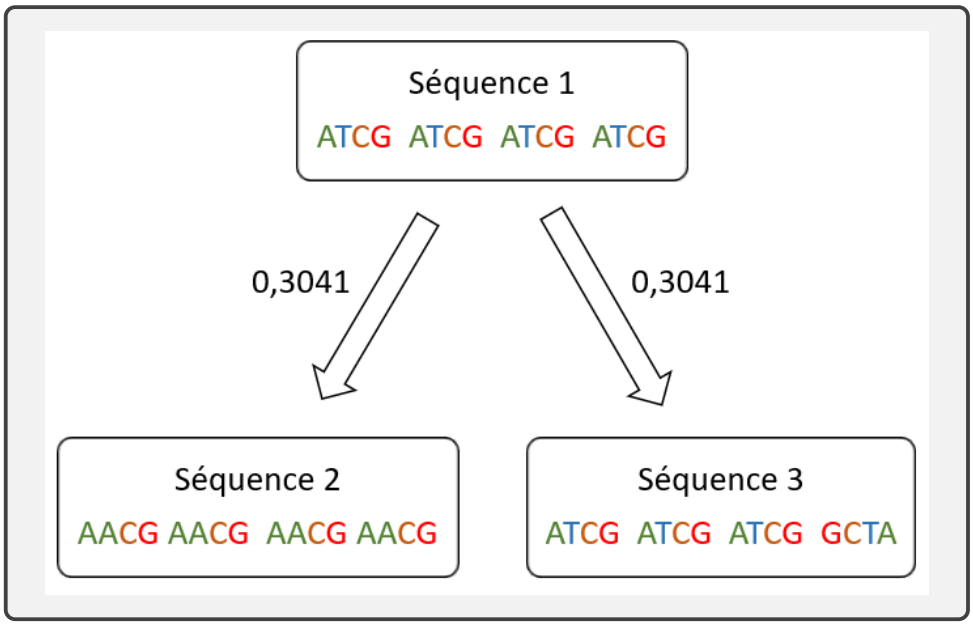


Figure 3.1: Exemple de deux séquences représentant des histoires évolutives différentes, présentant une même distance génétique avec la séquence ancestrale.

placée par un nombre M de positions, nommé le pas, et les distances entre les séquences extraites de la fenêtre sont calculés à nouveau. Ce processus est répété jusqu'à ce que la fenêtre ait traversée l'entièreté des séquences. La figure 3.2 ci-dessous représente visuellement le processus de fenêtre coulissante.

Lorsque les distances correspondants à chaque fenêtre ont été stockés, ces valeurs sont présentées à l'utilisateur sous forme de graphique en ligne représentant la similarité en fonction des positions, de chaque séquences consensus par rapport à une séquence de référence choisie par l'utilisateur. En analysant les variations de distances, il est possible d'en déduire des sites possibles où des évènements de recombinaisons ou de transferts horizontaux ont pu avoir lieu. La figure 3.3a présente un exemple de sortie de SimPlot.

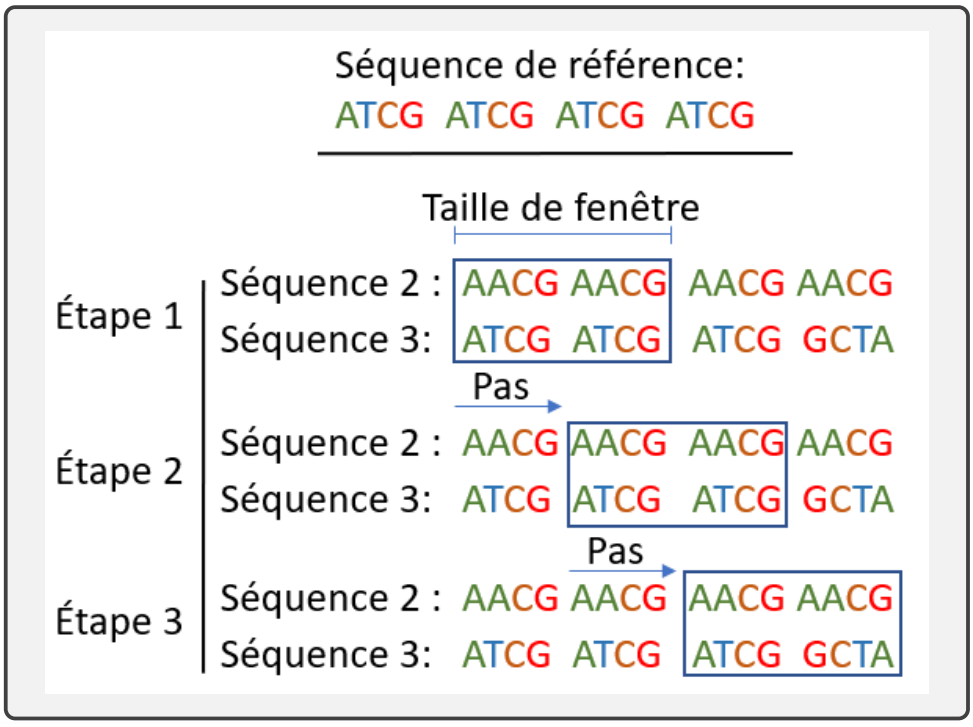


Figure 3.2: Représentation du fonctionnement d'une fenêtre coulissante.

3.1.3 Algorithme Bootscan

L'analyse Bootscan est une analyse par fenêtre coulissante qui diffère de l'analyse SimPlot par son approche phylogénétique afin d'identifier les régions d'intérêts de séquences mosaïques (Salminen *et al.*, 1995). Pour chaque fenêtre lors de l'analyse, les sous-séquences extraites sont premièrement re-échantillonnées par la méthode de bootstrap. Le jeu de données ainsi multiplié est par la suite utilisé pour générer des matrices de distances selon un modèle spécifié par l'utilisateur. Ces matrices de distances permettent de générer des arbres phylogénétiques par les méthodes de Neighbor-Joining (Saitou et Nei, 1987) et UPGMA (Sokal et Michener, 1958). Les topologies des arbres issus du bootstrap sont analysés afin de déterminer dans chacun la séquence consensus la plus similaire à la séquence de référence. Ainsi, pour chaque fenêtre d'analyse, chaque groupe autre que celle de référence

obtiendra une valeur de proximité à la séquence de référence, correspondant au pourcentage d'arbres où celui-ci était le plus similaire.

Une fois toutes les analyses de fenêtres complétées, ces résultats sont présentés à l'utilisateur à travers un graphique qui exprime les pourcentages d'arbres permutés en fonction de la position sur les séquences entières. Un exemple d'une sorte de Bootscan est présenté à la figure 3.3b ci-dessous.

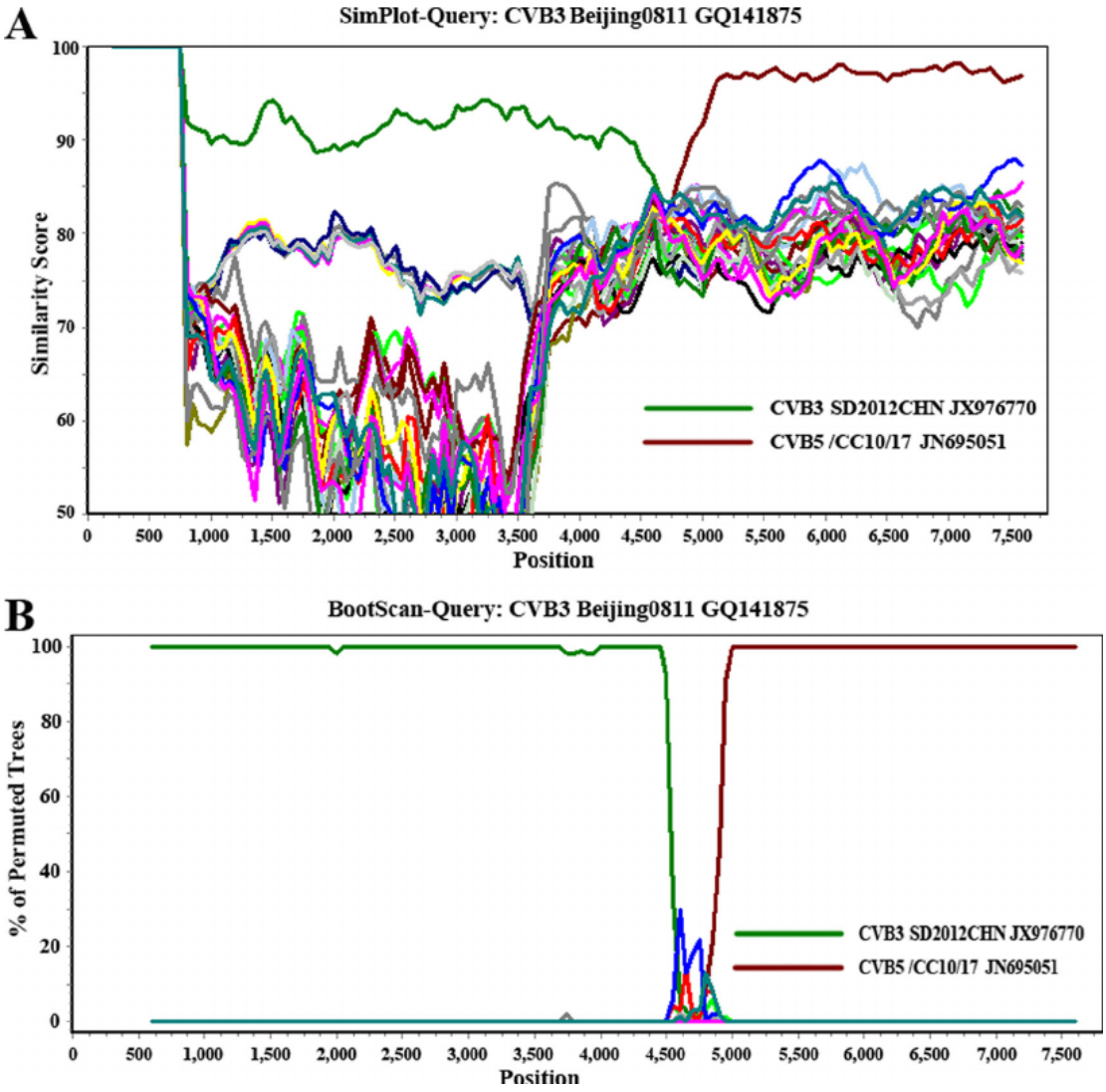


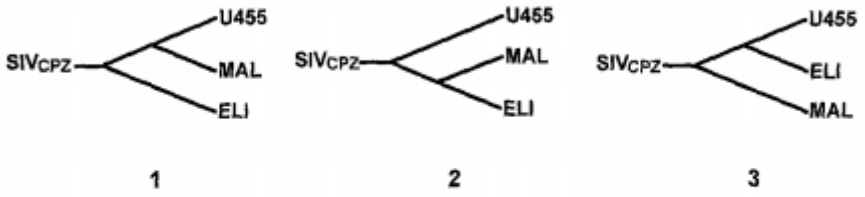
Figure 3.3: Exemple de sorties SimPlot et Bootscan (Wu *et al.*, 2013).

3.1.4 Algorithme FindSite

L'analyse FindSite est une analyse simple et rapide permettant de localiser les régions possibles de recombinaisons par l'identification de sites informatifs. La première étape de l'analyse est de sélectionner une séquence suspectée d'être issue d'un évènement de recombinaison ainsi que deux séquences provenant chacune d'une des deux lignées évolutives parentales, et une séquence autre non-relié évolutivement aux trois précédentes. Les sites informatifs seront identifiés à travers ces séquences comme étant ceux où, à une même position, deux des séquences partagent le même nucléotide, et que les deux autres séquences partagent un autre nucléotide (Robertson *et al.*, 1995).

Les sites répondants à ces critères sont considérés informatifs par la distribution de leurs positions à travers les séquences. Assumons une séquence R recombinée, ses deux séquences de lignées parentes P1 et P2, ainsi qu'une séquence externe E. Chaque site informatif détecté le long des séquences peut prendre l'une de trois configurations : la séquence R peut partager un nucléotide avec un de ses deux parents (configurations 1 et 2) ou avec le groupe externe (configuration 3). Chaque site informatif correspondant à une configuration 1 ou 2 présente une possibilité qu'un transfert horizontal s'y soit produit. Individuellement, ces sites n'offrent pas une indication forte de recombinaison mais, la présence de régions génétiques présentant une forte proportion d'une configuration 1 ou 2 par rapport à l'autres pourrait suggérer qu'un évènement de recombinaison s'y soit produit (Robertson *et al.*, 1995). La figure 3.4 ci-dessous présente les trois configurations que peuvent prendre les séquences analysées ainsi qu'un exemple de sortie standard d'analyse FindSite.

A.



B.

site	U455	MAL	ELI	SIV_{CPZ}	tree
	A	?	D		
36	G	G	A	A	1
56	G	G	A	A	1
109	G	A	A	G	2
139	A	A	C	C	1
140	G	G	C	C	1
180	G	G	A	A	1
183	A	T	A	T	3
233	A	A	G	G	1
243	C	C	T	T	1
255	G	G	A	A	1
275	T	A	A	T	2
276	G	G	A	A	1
278	A	A	G	G	1
321	A	G	A	G	3
327	C	T	T	C	2
331	C	T	T	C	2

Figure 3.4: Exemple des configurations possibles des séquences et de résultats d'analyse de sites informatifs (Robertson *et al.*, 1995).

3.2 PhiPack

Le logiciel PhiPack par Bruen et al (Bruen *et al.*, 2006) est un outil regroupant trois méthodes statistiques de détection de recombinaison. Celui-ci regroupe divers outils statistiques de détection d'évènements de recombinaison tels les tests Phi, NSS et max-Chi. Ceux-ci peuvent être exécutés avec des permutations du jeu de donnée, conférant au logiciel la capacité d'établir des valeurs de significations

(P-values) aux résultats obtenus.

3.2.1 Test Phi

Le test Phi ou « Pairwise Homoplasy Index » est une méthode statistique servant à déterminer la probabilité qu'un évènement de recombinaison ait eu lieu entre des séquences. La théorie derrière le calcul de la statistique phi repose sur le concept d'homoplasie et de compatibilité. L'homoplasie est le phénomène biologique où une similarité observée entre deux séquences d'organismes différents ne provient pas d'un ancêtre commun (Torres-Montúfar *et al.*, 2018). La compatibilité dérive de l'homoplasie par le fait que deux sites sont considérés compatibles lorsque l'hypothèse d'homoplasie entre eux est réfutée. Par le fait même, deux sites sont considérés incompatibles si une ou plusieurs homoplasies ont eu lieu (Bruen *et al.*, 2006).

Pour chaque site informatif parmi les séquences, un score d'incompatibilité des sites est évalué et ajouté à une matrice d'incompatibilité. Une fois cette matrice complétée, celle-ci permet de décrire l'histoire évolutive entre les séquences et de déterminer si des évènements de recombinaisons ont eu lieu (Bruen *et al.*, 2006). Par l'utilisation d'un test de permutation des séquences, la statistique phi peut être reproduites de multiples fois et comparée à la valeur originale du premier test afin d'obtenir une valeur-p associée à la statistique phi.

Le test Phi peut également être associé à une fenêtre coulissante sur les séquences d'intérêts. Par cette approche une statistique phi ainsi qu'un test de permutations peut être effectué sur chaque fenêtre et évaluer la mosaïcité des séquences entières.

3.2.2 Test NSS

Comme la statistique phi, le test Neighborhood Similarity Score (NSS) utilise également une matrice d'incompatibilité des sites informatifs afin d'évaluer la présence de recombinaisons dans le jeu de données. Les deux méthodes divergent par l'emploi de cette matrice. La statistique NSS est calculée par la mesure du regroupement des régions adjacentes qui sont compatibles et incompatibles (Jakobsen et Easteal, 1996). L'application du test NSS avec permutations permet également d'obtenir une valeur-p de signification du test.

3.2.3 Test Max-Chi

Le test Max-Chi est un test par fenêtre coulissante basée sur le comportement des sites polymorphiques dans les séquences, c'est-à-dire des sites où au moins une des séquences est différente des autres. Ainsi, dans chaque fenêtre d'analyse, la proportion de sites identiques et polymorphiques entre différentes régions (droites et gauche) de la fenêtre sont évaluées. Un contraste élevé des proportions entre ces régions pourrait suggérer qu'un évènement de recombinaison a eu lieu (Smith, 1992). Comme pour les autres tests, son utilisation conjointement à des permutations des sites permet de produire une valeur-p, ajoutant une valeur de signification statistique au test.

3.3 Autres outils de recombinaison

Depuis la publication de l'outils SimPlot, plusieurs autres outils de détections de recombinaisons ont été développés et publiés par la communauté scientifique. Deux de ces logiciels les plus fréquemment regroupés avec SimPlot sont ici présentés, afin de mettre de l'avant leurs particularités et distinctions avec ce dernier.

3.3.1 Recombination Analysis Tool

Le « Recombination Analysis Tool » ou RAT (Etherington *et al.*, 2005) est un outil de détection multi-plateforme de recombinaison d'alignements nucléotidiques et protéiques. Cet outil emploie une approche simple et rapide basée uniquement sur le calcul de distances par fenêtre coulissante entre les séquences afin d'identifier les régions recombinées, en se basant sur trois critères.

Le premier critère de la méthode de RAT représente le seuil maximal de similarité entre deux séquences potentiellement recombinantes, avant la région de recombinaison. Le second critère correspond au seuil minimal de similarité entre deux séquences qui identifie la présence potentielle d'un événement de recombinaison. Ces deux premiers critères représentent le « saut » de similarité entre deux séquences qui est un indicateur de recombinaison. Finalement, le troisième critère permet d'indiquer le nombre maximal de séquences pouvant participer à un même évènement de recombinaison détecté.

Les détections générées par cette méthode sont présentées à travers l'interface graphique de RAT comme présentée à la figure 3.5.

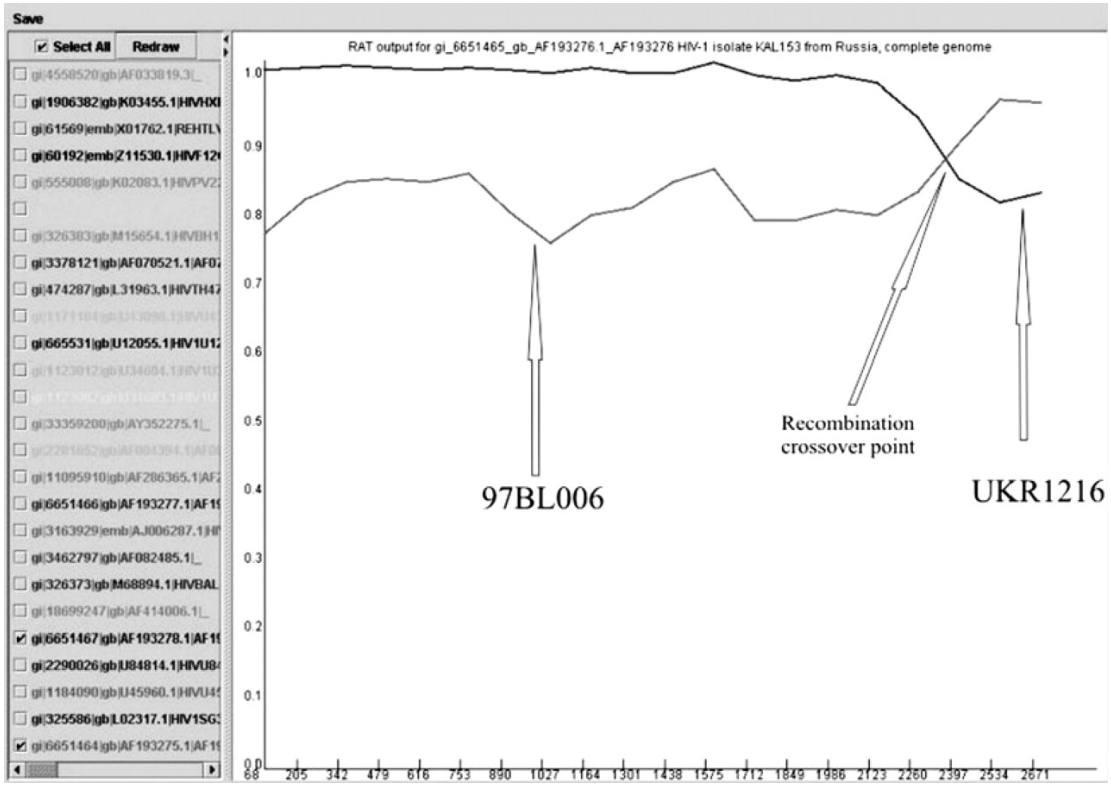


Figure 3.5: Exemple de sortie d'analyse du « Recombination Analysis Tool » (Etherington *et al.*, 2005).

3.3.2 Recombination detection program

Le « Recombination Detection Program » ou RDP est un logiciel offert sur Windows permettant également de détecter les évènements de recombinaisons. Le logiciel RDP est présentement à sa quatrième itération (RDP4), publié en 2015 (Martin *et al.*, 2015). RDP4 offre une grande sélection d'outils de détections de recombinaisons.

La méthode originale développée par (Martin et Rybicki, 2000) et présentée dans la version originale de RDP requiert un alignement multiple de plusieurs séquences et est divisible en trois étapes distinctes. Premièrement, les séquences sont regrou-

pées de façon à ce que tous les arrangements d'exactement trois séquences sont produits et que pour chacun de ces trios, les sites non-informatifs sont retirées des séquences.

Par la suite, une analyse de distance par fenêtre coulissante est effectuée sur les trios de séquences de sites informatifs afin d'identifier les régions où les deux premières séquences du trio présentent une similarité plus faible entre elles que leur similarité respective à la troisième séquence.

Finalement, la probabilité que les régions identifiées à la seconde étape soient dû à de la chance est mesurée à partir d'une équation de distribution binomiale incluant la longueur des séquences complètes et des régions recombinantes ainsi que le nombre de nucléotides en commun entre les séquences.

Une fois que tous les arrangements de trios de séquences auront été analysées, l'utilisateur doit décider d'un seuil minimal de probabilité acceptable et finalement visualiser graphiquement dans l'application RDP les régions potentiellement recombinées. Au fils des versions de RDP, d'autres méthodes de détection ont été incorporés à l'analyse afin d'employer une variété de signaux de recombinaisons différents. À partir de la somme de ces signaux, RDP4 est en mesure d'inférer le nombre minimal d'évènements de recombinaisons nécessaire pour expliquer ces signaux (Martin *et al.*, 2015). Ainsi, RDP4 permet à l'utilisateur de représenter à partir du jeu de donnée fournis des scénarios de recombinaisons complexes impliquant des multiples séquences. De plus, l'interface graphique permet à l'utilisateur d'avoir une vue d'ensemble de l'analyse et de rapidement avoir accès à l'information issue de ces multiples signaux de détections.

La figure 3.6 représente les éléments principaux d'une analyse de recombinaison par le logiciel RDP4. À travers interface, il est possible de visualiser l'alignement multiple fournie par l'utilisateur (en haut à gauche), les représentations matri-

cielles, phylogénétique et textuelles liées à l'analyse (en haut à droite), une représentation graphique des résultats statistiques liés aux signaux identifiés (en bas à gauche) et une représentation visuelle des régions recombinées (en bas à droite).

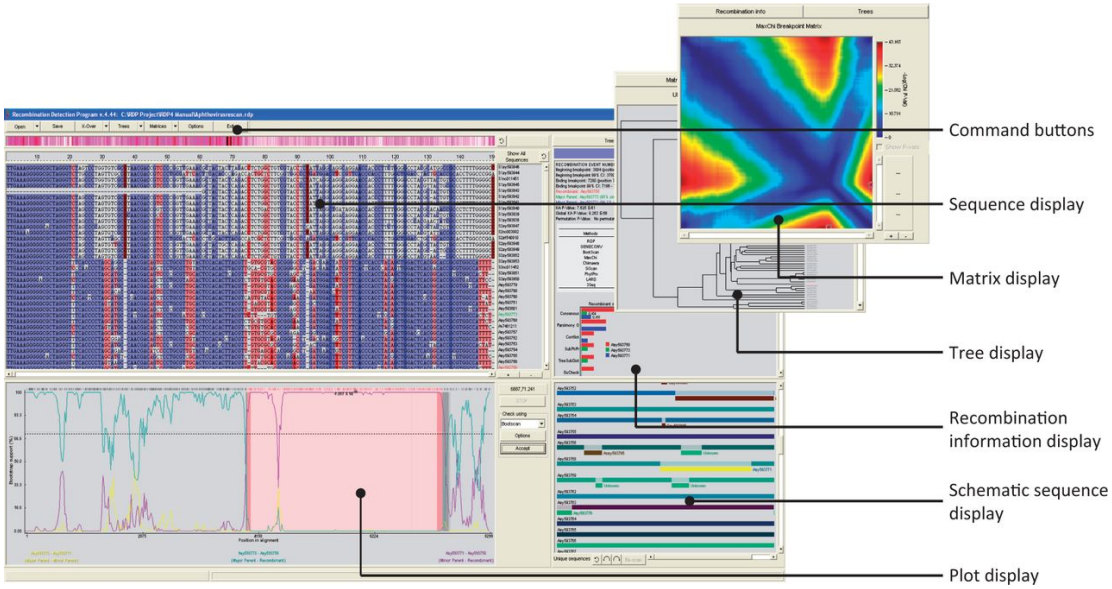


Figure 3.6: Exemple de sortie d'analyse de RDP (Martin *et al.*, 2015).

3.4 Réseaux de similarité de séquences

Plusieurs des approches présentées dans ce chapitre sont basés sur une représentation des résultats par graphes de courbes. Cette approche de représentation des résultats d'analyse permet l'inspection rapide des résultats mais présente des limitations tant qu'à la vision d'ensemble de ceux-ci. En effet, dans le cas des graphiques de similarités entre des séquences selon la position tels que proposés par les logiciels SimPlot et RAT, chaque graphique est basé sur la comparaison à une séquence de référence, ce qui limite l'agrégation et la présentation de l'ensemble des résultats en un graphique unique. De plus, ces graphes de courbes peuvent facilement devenir encombrés lorsque plus de dix courbes se croisent et se chevauchent à des mêmes positions.

3.4.1 Types de réseaux

La création d'un réseau de similarité de séquences est hautement dépendant de l'information qui doit être visualisé. Au moins une variable doit exister afin de permettre la formation d'arêtes entre les nœuds et d'ainsi développer le réseau (figure 3.7a). Dépendamment de l'analyse, le réseau peut prendre plusieurs formes. Selon les variables à représenter dans le réseau, il est possible qu'une interaction entre deux nœuds soit unidirectionnelle, tels un transfert de matériel génétique non-réciproqué. Cette relation pourrait être représentée à travers un réseau dirigé, où les arêtes présentent une direction d'interaction (figure 3.7b). Si de multiples variables sont considérées dans le réseau, un réseau multiple peut permettre de représenter de multiples arêtes entre chaque nœud (figure 3.7c)

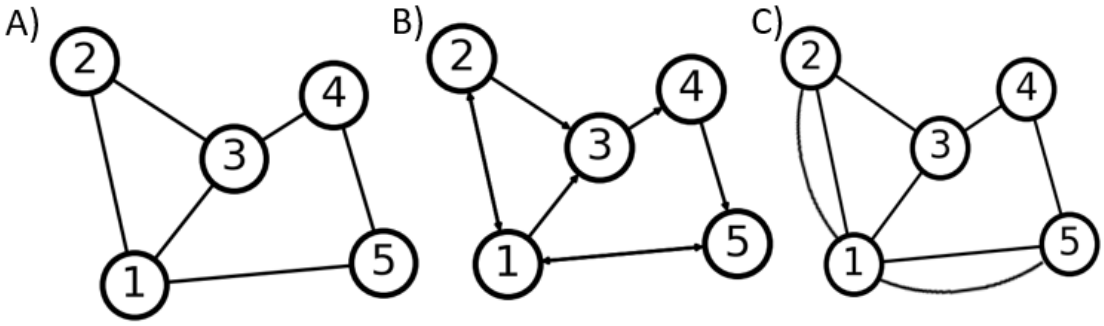


Figure 3.7: Représentation de différents types de réseaux.

3.4.2 Bénéfices des représentations en réseaux

Les réseaux de similarités de séquences peuvent ouvrir la voie à des représentations claires et concises d'interactions complexes entre divers éléments, tels des gènes et génomes, qui seraient hors de la portée d'autres types de représentations, tels par arbres phylogénétiques et graphes en courbes (Corel *et al.*, 2016).

De plus, la représentation par réseau permet d'approcher de façon moins stricte la

représentation des liens évolutifs et autres interactions entre les séquences à l'étude que d'autres représentations fréquemment employés tels les arbres phylogéniques (Corel *et al.*, 2016). Par exemple, un arbre phylogénique requiert des séquences liées évolutivement, avec ou sans groupe externe, et est difficilement capable de représenter concisément les liens évolutifs de génomes mosaïques. Les réseaux de similarités de séquences ne requièrent aucune proximité évolutive entre les séquences à l'étude, peuvent contenir de multiples groupes de séquences hautement divergentes et représenter de multiples HGT entre plusieurs séquences.

Les réseaux de similarité bénéficient également d'une capacité à être intuitivement interactifs, permettant à l'utilisateur de plus aisément explorer les résultats, mettre en évidence un ou des sous-groupes de nœuds ainsi que de restreindre la formation d'arêtes dynamiquement selon certains critères. Il s'agit donc d'une représentation pour la visualisation de résultats d'analyses de grands jeux de données, quoi que les avantages soient également visibles à petite échelle.

La figure 3.8 ci-dessous représente différentes interprétations du même jeu de donnée sous forme de réseau de similarité de séquences selon le seuil minimal de similarité requis pour la formation d'une arête entre deux nœuds.

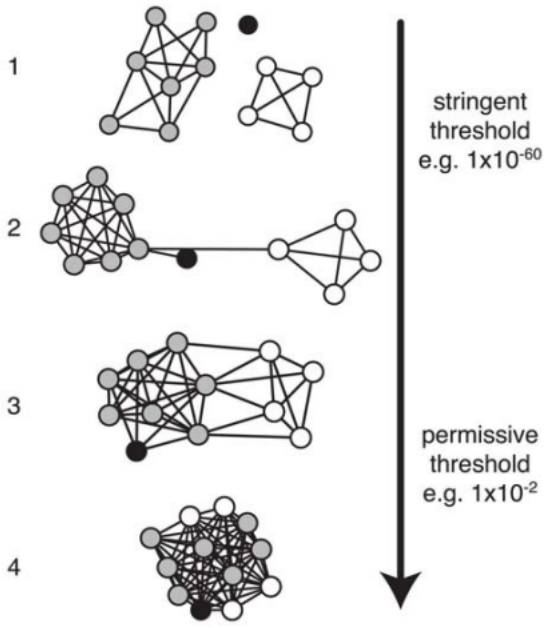


Figure 3.8: Représentation de l'effet du seuil minimal de similarité sur la connectivité du réseau (Atkinson *et al.*, 2009).

Il est également bon de noter que ce type de représentation peut prendre différents aspects selon les besoins de l'utilisateur et les hypothèses distinctes à représenter. La figure 3.9 ci-dessous présente quatre réseaux de similarités, chacun présentant une particularité :

La figure 3.9a représente un réseau de similarité de séquences où les noeuds, représentant des séquences, sont connectés par des arêtes si ceux-ci présentent une similarité élevée. Le réseau peut être séparé en groupes de composantes connectés (CC).

La figure 3.9b représente un réseau de génomes où les nœuds représentent des génomes entiers reliés par des arêtes si deux nœuds partagent au moins une famille de gènes. Dans cet exemple, les arêtes possèdent un poids identifiant le nombre de gènes partagés par chaque famille.

La figure 3.9c représente un réseau multiplexé où les nœuds représentent des génomes entiers, et les arêtes représentent les familles de gènes, distinguées par leurs couleurs respectives. Comme dans l'exemple B, les arêtes présentent un poids relatif au nombre de gènes partagés par famille.

La figure 3.9d représente un graphe biparti où les nœuds supérieurs représentent les génomes entiers et les nœuds inférieurs représentent les familles de gènes. Les nœuds inférieurs et supérieurs sont connectés par des arrêtes selon le nombre de gènes de chaque famille présente dans chaque génome. Comme dans les exemples précédents, les valeurs de poids représentent le nombre de gènes impliqués dans ce partage.

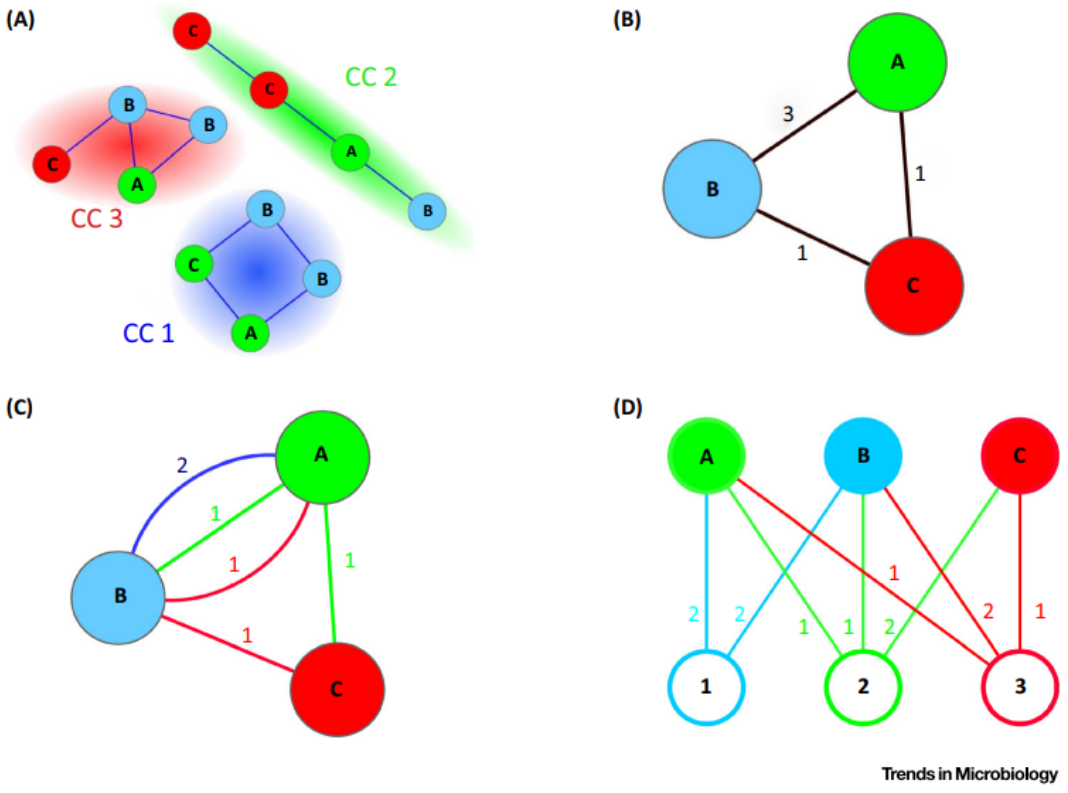


Figure 3.9: Différentes représentations par graphes de partages de gènes entre des génomes (Corel *et al.*, 2016).

3.5 Logiciels incorporant des réseaux de similarités de séquences

Plusieurs outils ont été publiés dans les dernières années permettent aux utilisateurs d'employer des réseaux de similarité de séquences lors de l'analyse des jeux de données.

3.5.1 PANADA

PANADA (Martin *et al.*, 2013) permet à l'utilisateur de produire de tels réseaux à partir de séquences protéiques, en se basant sur la qualité des alignements (pourcentage d'identité, valeur-E et longueur des alignements) entre les séquences d'acides aminées par l'entremise du logiciel BLASTALL (Altschul *et al.*, 1997). PANADA permet aussi la construction de réseaux basés sur les structures protéiques, par l'entremise d'outils tels MUSTANG (Konagurthu *et al.*, 2006) et TMalign (Zhang et Skolnick, 2005). Ci-dessous est un exemple de sortie d'analyse employant PANADA (figure 3.10).

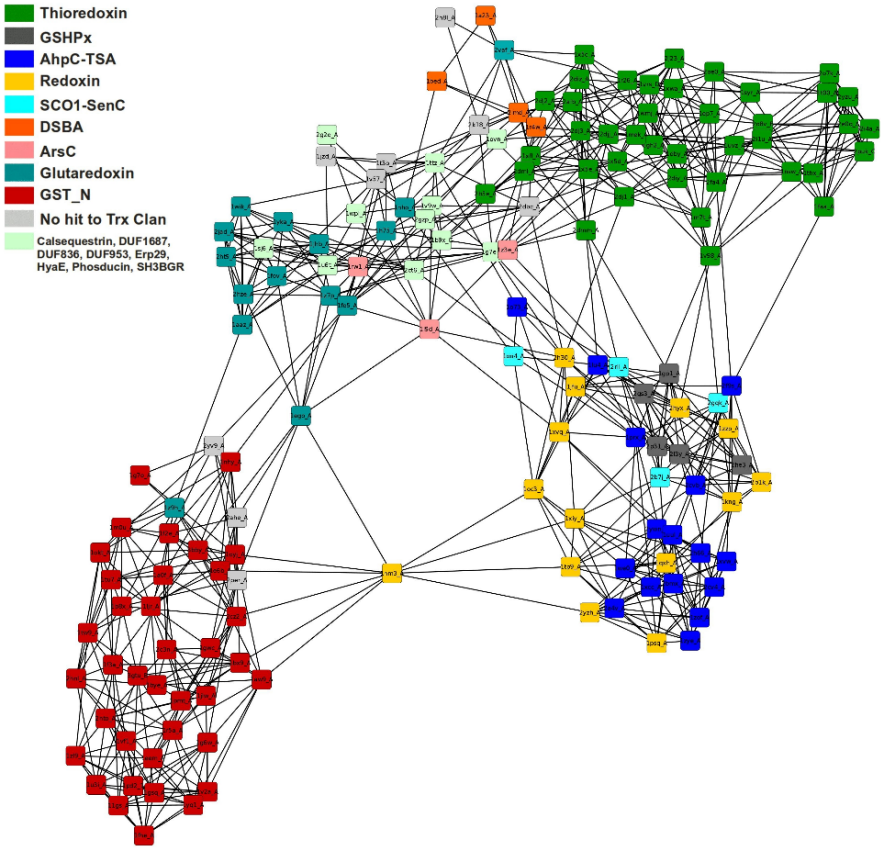


Figure 3.10: Exemple de graphique produit par l'entremise de PANADA (Martin *et al.*, 2013).

3.5.2 EGN

EGN (Halary *et al.*, 2013) permet la génération de réseaux de similarité, pour des jeux de données nucléotidiques et protéiques, basés sur des recherches BLAST ou BLAT (Altschul *et al.*, 1997). À partir des résultats de ces recherches, le réseau est produit de façon à regrouper toutes les séquences ayant une similarité élevée avec au moins une autre séquence du même groupe, et n'étant hautement similaire à aucune autre séquence hors du groupe. Le réseau résultant de ces opérations par le logiciel EGN est présenté ci-bas (figure 3.11).

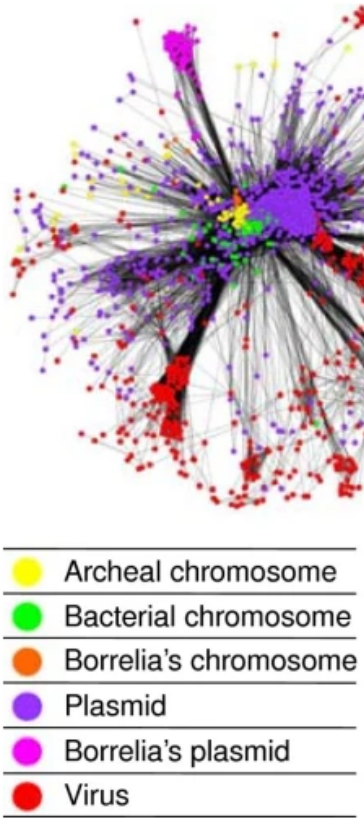


Figure 3.11: Exemple de graphique produit par l'entremise d'EGN (Halary *et al.*, 2013).

3.6 Beast2

Le logiciel Beast2 est un logiciel de reconstruction phylogénétique employant la méthode bayésienne de Monte-Carlo par chaines de Markov (MCMC). À travers ce programme, il est possible d'employer des modèles évolutifs complexes ainsi que des modèles d'horloges moléculaires afin d'inférer des scénarios évolutifs détaillés. Beast2 permet d'attribuer ces modèles spécifiquement à des sous-groupes du jeu de données afin que chaque sous-groupe soit calibré selon son comportement évolutif respectif. L'architecture de Beast2 permet à l'utilisateur de retirer les abstractions et simplifications typiquement associés aux outils simples d'inférence phylogéné-

tique et de fournir une grande flexibilité tant qu'aux spécifications des analyses (Drummond et Rambaut, 2007). De plus, Beast2 permet aux utilisateurs d'associer à chaque séquence chargée dans l'application une date numérique représentant l'âge de la séquence afin de représenter les temps d'évolution de chaque nœud des arbres phylogénétiques produit (Drummond et Rambaut, 2007).

L'emploi de la méthode MCMC consiste initialement en l'utilisation d'un arbre de départ (généré ou non par Beast2) avec des longueurs de branches aléatoires (sauf si spécifiés par l'utilisateur). De nouveaux arbres inspiré des arbres précédant sont produits et se voient assignés une probabilité relative à leur crédibilité, basé sur les paramètres (tels les antécédents, modèles de distances et horloges évolutives) de l'analyse. Éventuellement, une convergence peut être observée, favorisant un certain nombre d'arbres considérés optimaux selon leur score de crédibilité (Egan et Crandall, 2006).

Dû à l'approche par MCMC, une analyse de Beast2 peut impliquer des millions d'arbres générés, avec un échantillonnage d'arbre à chaque nombre arbitraire d'étapes. Ainsi, par exemple, une analyse de 10 000 000 d'étapes avec échantillonnage à chaque 5 000 étapes résulterait en un total de 20 000 arbres enregistrés au final (ce nombre peut être réduit en mettant de côté un pourcentage des arbres initiaux jugés « non-optimal »). Les arbres restants peuvent ensuite être condensé en un arbre consensus selon différents critères d'arbres et d'hauteur des nœuds. La figure 3.12 ci-dessous présente une visualisation typique d'arbre consensus issue de Beast2 par le logiciel FigTree V.1.4.4 (Rambaut, 2006). Dans cet exemple, un arbre consensus est représenté avec des barres au niveaux des nœuds internes représentant les intervalles les plus courts contenant 95% des probabilités postérieures (HDP 95%).

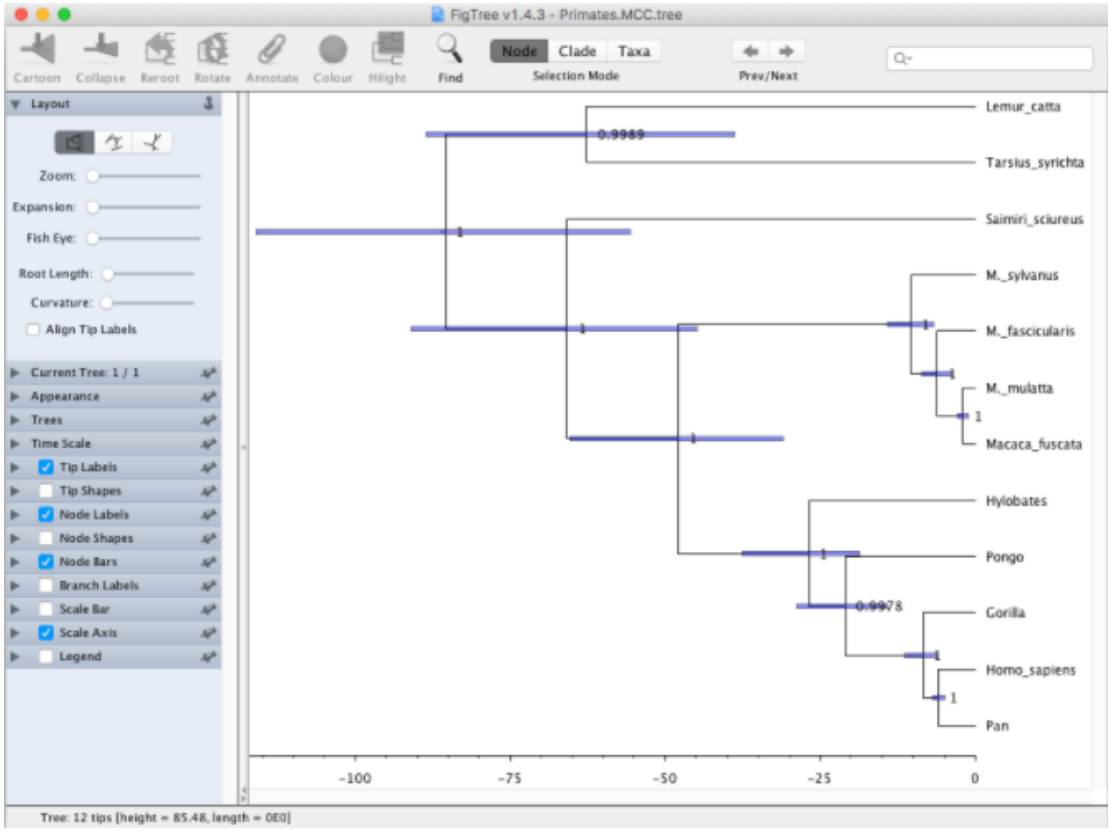


Figure 3.12: Exemple d'arbre consensus de Beast2 (Barido-Sottani *et al.*, 2018)

CHAPITRE IV

DÉVELOPPEMENT D'APPLICATION

À travers ce chapitre, les différentes analyses et composantes logiciel SimPlot++ seront abordées. Chaque section permettra de présenter chacune des six composantes principales du logiciel.

Dans le cas des méthodes du logiciel SimPlot préalablement discutées dans les chapitres précédant, l'emphase seras mise sur les améliorations apportées à chacune de celles-ci. Les nouvelles fonctionnalités seront quant à elles présentées en détail.

4.1 Conception de l'application

Afin de favoriser l'accessibilité et l'aise de modification du logiciel SimPlot++, celui-ci a été codé en langage python. Le choix de langage a également permis l'intégration facile de code provenant de quelques librairies externes, bénéficiant d'années de validation et d'améliorations par la communauté scientifique. Ces librairies sont présentées lors de leurs applications respectives, à l'exception de la librairie Biopython qui est appliquée à l'entièreté du logiciel.

4.1.1 Biopython

La librairie Biopython (Cock *et al.*, 2009) est une agglomération de divers modules bioinformatiques permettant de faciliter et standardiser l'acquisition et la manipulation de données biologiques. Dans le contexte du logiciel SimPlot++, Biopython est employé principalement afin de lire les fichiers de données génétiques (par exemple : fasta et nexus), écrire les fichiers de sorties comportant des séquences génétiques et également, de manipuler les alignements nucléotidiques et protéiques. Puisque le logiciel requiert des données déjà manipulées et non brutes, employer la librairie la plus populaire d'outils bioinformatiques de python permet d'assurer un maximum de compatibilité avec les données des utilisateurs. Outre ces applications, Biopython offre également des modèles de calculs de distances génétiques qui ont été intégré au logiciel.

4.2 Page de création de groupes

La page de création de groupes est la page centrale du logiciel, et est axée sur l'import des données et leur manipulation. Ainsi, il s'agit de la première page affichée à l'ouverture du logiciel.

Les données requises par le logiciel doivent représenter un alignement multiple de séquences nucléotidiques ou protéiques. Ces alignements multiples doivent être représentés sous un format fasta, nexus, pir, phylip, stockholm ou clustal. Un bouton d'accès rapide aux fichiers récemment ouvert par l'utilisateur est également présent pour accélérer le processus. À l'ouverture d'un fichier, un avertissement est donné à l'utilisateur si des caractères non supportés sont présents dans le jeu de données. Ces caractères sont ensuite remplacés par des tirets, représentant des « gaps » dans la séquence. Cette étape a pour but d'ajouter de la transparence quant aux manipulations effectuées sur le jeu de donnée et d'alerter l'utilisateur par rapport aux caractères non-supportés ainsi qu'à leurs positions sur les séquences.

Une fois que l'alignement multiple a été chargé, l'utilisateur peut regrouper les séquences dans des groupes créés manuellement. Ces groupes doivent consister d'un nom de groupe unique, d'au moins une séquence ainsi que d'une couleur de groupe qui sera employée lors de l'affichage des résultats. Au moins deux groupes doivent être formés pour accéder aux algorithmes SimPlot et BootScan. Toutes séquences excluent d'un groupe ne seront pas impliquées dans ces deux analyses. L'utilisateur peut sauvegarder les groupes formés à travers la création d'un nouveau fichier d'alignement multiple de format nexus afin d'éviter d'avoir à refaire cette étape dans le futur.

Finalement, cette page d'options permet également à l'utilisateur d'accéder à la fenêtre de « préférences de l'utilisateur ». Cette page permet principalement à l'utilisateur de modifier les paramètres de formation des séquences consensus issus des séquences de chaque groupe formé. Dans le logiciel SimPlot original, la formation des séquences consensus est cachée à l'utilisateur. Cela est en partie rectifié dans SimPlot++ par l'introduction d'une option pour modifier la valeur du seuil de fréquence d'un caractère à une position nécessaire pour inclure celui-ci dans la séquence consensus du groupe. De plus, il est possible de télécharger le

fichier des séquences consensus généré par le logiciel afin que l'utilisateur puisse accéder aux séquences consensus réellement analysées par les différents outils offerts. Cette option est particulièrement utile afin de valider les séquences formées et régler des problèmes potentiels liés à la qualité des résultats.

La figure 4.1 ci-dessous présente l'organisation de la page de création de groupes. Sous la partie notée « A », les boutons de chargement des fichiers, de réouverture rapide des fichiers récents, de sauvegarde des groupes formés et d'accès aux options de « préférence de l'utilisateur » sont présents. La section « B » présente les divers groupes formés par l'utilisateur. Dans l'exemple de la figure 4.1, plusieurs groupes (SARS-COV-2, Guangxi Pangolin CoV, etc.) sont déjà formés et le nombre de séquences compris dans chacun est indiqué. Le groupe SARS-CoV-2 y est actif, d'où son apparence plus foncée. La section « C » consiste en les deux boutons permettant respectivement de créer et retirer des groupes. La section « D » consiste en deux fenêtres comportant respectivement le nom des séquences incluses dans le groupe sélectionné (ici, le groupe SARS-CoV-2 composé de trois séquences), et le nom des séquences qui ne sont pas présentement dans un groupe formé. Les séquences peuvent être transférés dans ou hors d'un groupe par l'emploi des flèches séparant les deux fenêtres.

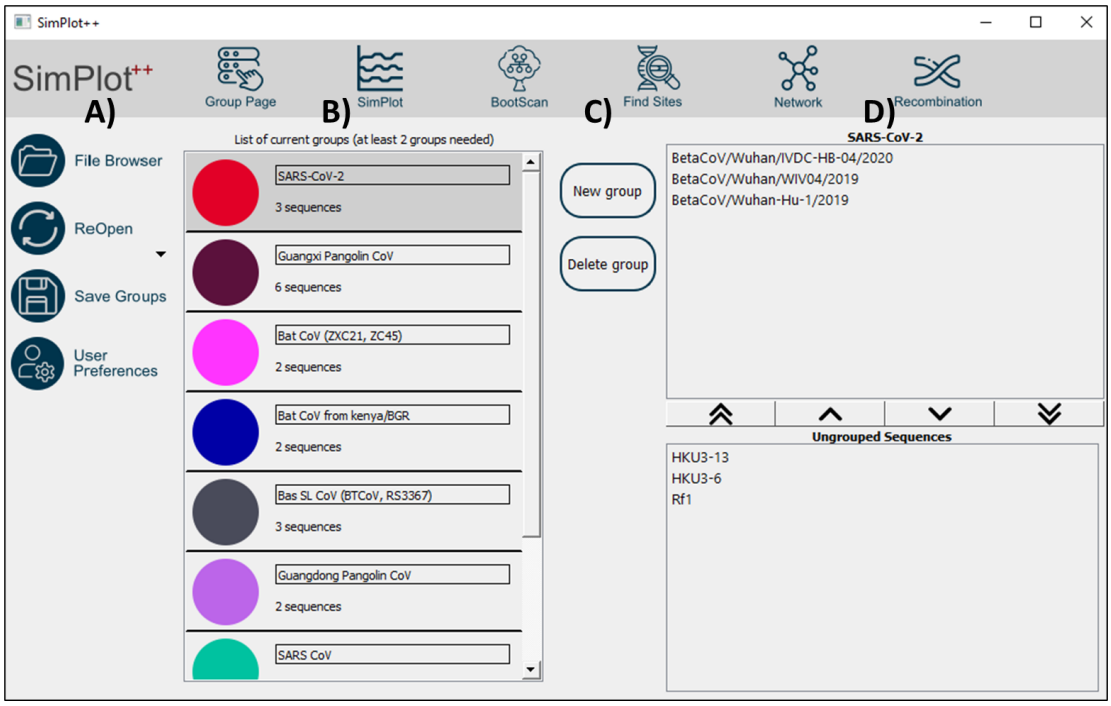


Figure 4.1: Page de création de groupes de SimPlot++.

4.3 Analyse SimPlot

4.3.1 Extraction des sous-séquences

L'analyse SimPlot débute par l'extraction des régions des séquences consensus correspondant aux positions de la fenêtre d'analyse (section 3.2). Ces sous-régions sont individuellement testées afin de vérifier si le nombre de « gaps », causés par des insertions ou des suppressions dans l'alignement multiple, est supérieur au seuil limite fourni par l'utilisateur. Si tels est le cas, la séquence en question est retirée de la fenêtre d'analyse. Puisque le comportement des modèles de distances envers les « gaps » varie (certains modèles les ignorent entièrement tels le modèle Kimura), l'utilisateur peut modifier ou retirer ce seuil maximal de « gaps » pour l'analyse.

4.3.2 Calcul de distances

Une fois les séquences validées, la distance entre chacune des séquences restantes est calculée selon le modèle choisi et les résultats sont représentés dans une matrice de distance. L'une des améliorations principales de SimPlot++ comparé aux logiciels déjà disponibles est le nombre de choix de modèles de distances disponible. Par exemple, le logiciel SimPlot original (Lole *et al.*, 1999) n'offre que quatre modèles de distances aux utilisateurs pour les séquences nucléotidiques et protéiques : Hamming, Jukes-Cantor, Kimura et F84. SimPlot++ offre 43 modèles pour les séquences nucléotidiques et 20 modèles pour les séquences protéiques. Cette grande variété de modèles comprend les modèles offerts par SimPlot et inclus les modèles les plus communs tels HKY85, GTR, JTT et WAG. Certains de ces modèles sont tirés de la librairie Cogent3 (Knight *et al.*, 2007) qui offrent une optimisation numérique des taux d'hétérogénéité du modèle selon les séquences, avec une distribution gamma. Les 5 modèles comportant cette fonctionnalité sont spécifiés dans l'application par le suffixe « -optimized » suite à leurs noms.

L'ajout de modèles de distances additionnels permet à l'utilisateur de prendre avantage de logiciels externes tels MEGA-X (Kumar *et al.*, 2018) qui offrent des outils permettant de déterminer les modèles les mieux adaptés au jeu de donnée à analyser (Posada et Crandall, 2001), tels que discuté dans la section 2.2.6. Cet ajout de modèles permet donc à l'utilisateur d'effectuer des analyses plus personnalisées envers les besoins de ses jeux de données et offre une plus grande flexibilité d'interactions avec des logiciels externes afin d'appliquer un même modèle de distance à travers de multiples types d'analyses.

Puisque l'emploi de modèles de distances plus complexes mène à des temps de calculs plus élevés, une nouvelle option de parallélisation a été développée afin de pouvoir calculer plusieurs fenêtres d'analyse à la fois. Cette option est dis-

ponible pour tous les modèles disponibles et est fortement recommandé pour les analyses utilisant les modèles optimisés de Cogent3, dû à leur temps d'exécution significativement plus élevé que ceux des autres modèles offerts.

Les matrices de distance résultantes représentent les distances entre tous les groupes de séquences. Comme l'analyse SimPlot requiert une séquence de référence, ces matrices de résultats sont utilisées afin de populer les structures de données propres à chaque séquence de référence possible.

4.3.3 Visualisation

La présentation et l'analyse des résultats SimPlot est entièrement visuelle, par l'intermédiaire d'un graphique représentant la similarité entre le groupe de référence et les autres groupes analysés selon la position sur les séquences. Afin d'offrir à l'utilisateur un maximum de liberté vis-à-vis de la manipulation des données, la librairie Matplotlib (Hunter, 2007) a été employée afin de générer les graphiques des résultats. Cette librairie est reconnue pour ses graphiques de qualité « publication » et permet donc à l'utilisateur d'employer directement le graphique de sortie du logiciel à des fins académiques. Afin d'aider l'utilisateur à modifier l'apparence du graphique, plusieurs fonctionnalités de Matplotlib ont été intégrées afin de :

1. Se déplacer sur le graphique (Zoom, déplacements cartésiens).
2. Modifier les bordures et espacements du graphique.
3. Changer les positions et textes des axes et du titre.
4. Modifier les noms, couleurs, style de lignes et de marqueurs de chaque groupe sur le graphique.
5. Sauvegarder le graphique directement sous 9 différents formats incluant PNG, PDF et JPG.

De plus, l'utilisateur peut changer de groupe de référence à travers un menu déroulant et a également l'option d'afficher les graphiques de son choix sur une nouvelle page indépendante afin de visualiser plusieurs graphiques simultanément.

Un exemple de sortie d'analyse SimPlot est présenté ci-dessous (figure 4.2).

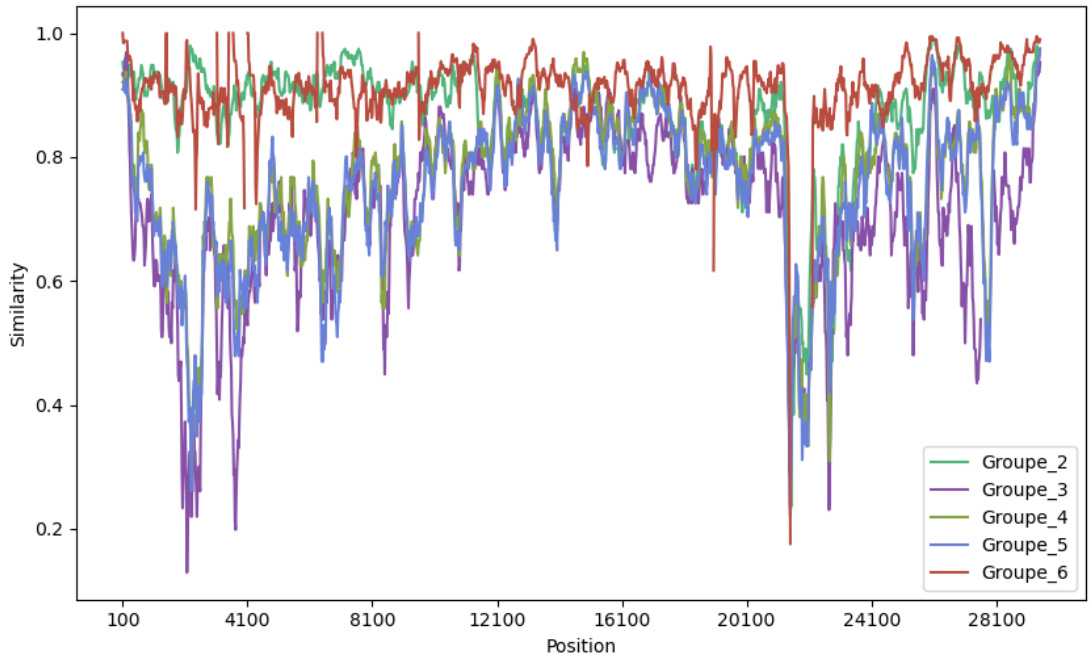


Figure 4.2: Exemple de sortie d'analyse SimPlot par SimPlot++.

4.3.4 Contrôle-qualité de l'analyse

Puisque les résultats d'une analyse SimPlot peuvent être incomplets par la présence d'une proportion de « gaps » dépassant le seuil à une certaine position ou par des distances incalculables dépendamment du modèle utilisé, il est important de donner à l'utilisateur un outil pour identifier globalement la source et l'ampleur des difficultés liés à l'analyse. Le logiciel SimPlot original a une tendance à cesser l'exécution lorsqu'une erreur de ce type est rencontré, laissant l'utilisateur sans pistes de solutions. Puisque les séquences impliquées sont généralement des

séquences consensus découlant du jeu de donnée fournie par l'utilisateur, résoudre des problèmes analytiques de ce type peut être lent.

Le logiciel SimPlot++ permet à l'utilisateur de visualiser les séquences consensus créées par le logiciel par la page de création des groupes, mais offre également une fonctionnalité nouvelle afin de générer un rapport qualité des résultats.

Ce rapport qualité est composé de quatre heatmaps représentant quatre aspects importants de l'analyse affectant sa qualité. Le premier heatmap présente, pour chaque fenêtre de chaque séquence consensus, le pourcentage de « gaps » présent. Cette représentation permet de visualiser rapidement quelles régions et séquences consensus sont problématiques et si les groupes formés devraient être réévalués.

Le second heatmap est une version simplifiée du premier, par son caractère binaire indiquant quelles fenêtres ont franchies le seuil de proportions de « gaps » permis dans l'analyse. Il a été jugé nécessaire d'inclure ces deux mesures séparément afin de maximiser la transparence du rapport en présentant à la fois les proportions brutes de « gaps » ainsi que l'inclusion ou non des sous-séquences dans l'analyse.

Ces deux premiers heatmaps sont représentés aux figures 4.3a et 4.3b. Ces figures représentent la qualité des résultats de l'analyse SimPlot présentée à la figure 4.2 précédemment. Il est possible de déterminer à travers ces deux figures que les résultats d'analyse sont partiels autour de la position 22 000 pour le groupe 5 dû à des fenêtres ayant dépassés le seuil toléré de «gaps». De plus, 5 des 6 groupes présentent des régions courtes comportant des pourcentages élevés de gaps.

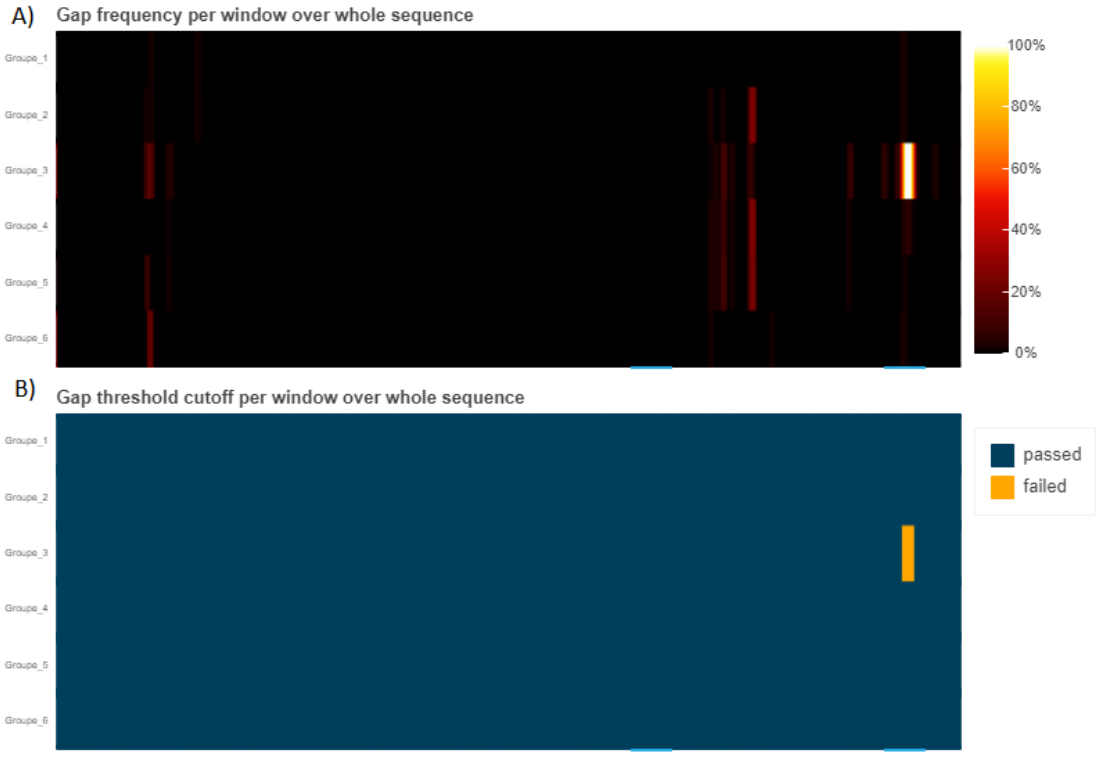


Figure 4.3: (a) Représentation par heatmap de la fréquence et distribution des «gaps» dans les séquences consensus. (b) Représentation par heatmap des régions des séquences étant au-dessus du seuil de gaps permis (en jaune).

Le troisième heatmap permet d'afficher les valeurs de distances manquantes, et donc, le niveau de complétion des résultats. Comme mentionné précédemment, certains modèles de distances peuvent être incapables de résoudre certaines mesures de distances évolutives si certains critères de similarité ne sont pas respectés. La fréquence de ces erreurs peut être amplifiée par l'utilisation de séquences évolutivement éloignées et de tailles de fenêtre courtes. Ces valeurs manquantes peuvent être difficiles à percevoir dans un graphique de courbe SimPlot dû au chevauchement des courbes. Ainsi, ce heatmap permet d'identifier exactement les positions et séquences consensus qui présentent ces erreurs.

Le dernier heatmap offert par le rapport de qualité est une représentation des résultats SimPlot sous ce format. Cette approche permet une observation alternative des résultats obtenus par l'analyse tout en identifiant clairement les données manquantes puisque contrairement à la représentation par graphique de courbes qui ignore les données manquantes, le heatmap les représentent visuellement par la couleur grise.

Les figures 4.4a et 4.4b représentent les troisièmes et quatrièmes heatmaps issus de l'analyse SimPlot de la figure 4.2. Il est possible de voir dans la figure 4.4a en jaune la région du groupe 3 dont la fréquence de « gaps » est supérieure au seuil et est donc non-calculée. Cependant, cette figure présente également sept régions du groupe 6 qui n'ont pu être calculée par le modèle JC69 dû à un problème de nature mathématique tel le logarithme d'un nombre négatif. Ces valeurs manquantes peuvent être difficiles à repérer dû au chevauchement des courbes de graphiques SimPlot. Ainsi, la figure 4.4b permet de représenter les résultats sans chevauchement, et donc, de visualiser les zones grises représentant les valeurs manquantes.

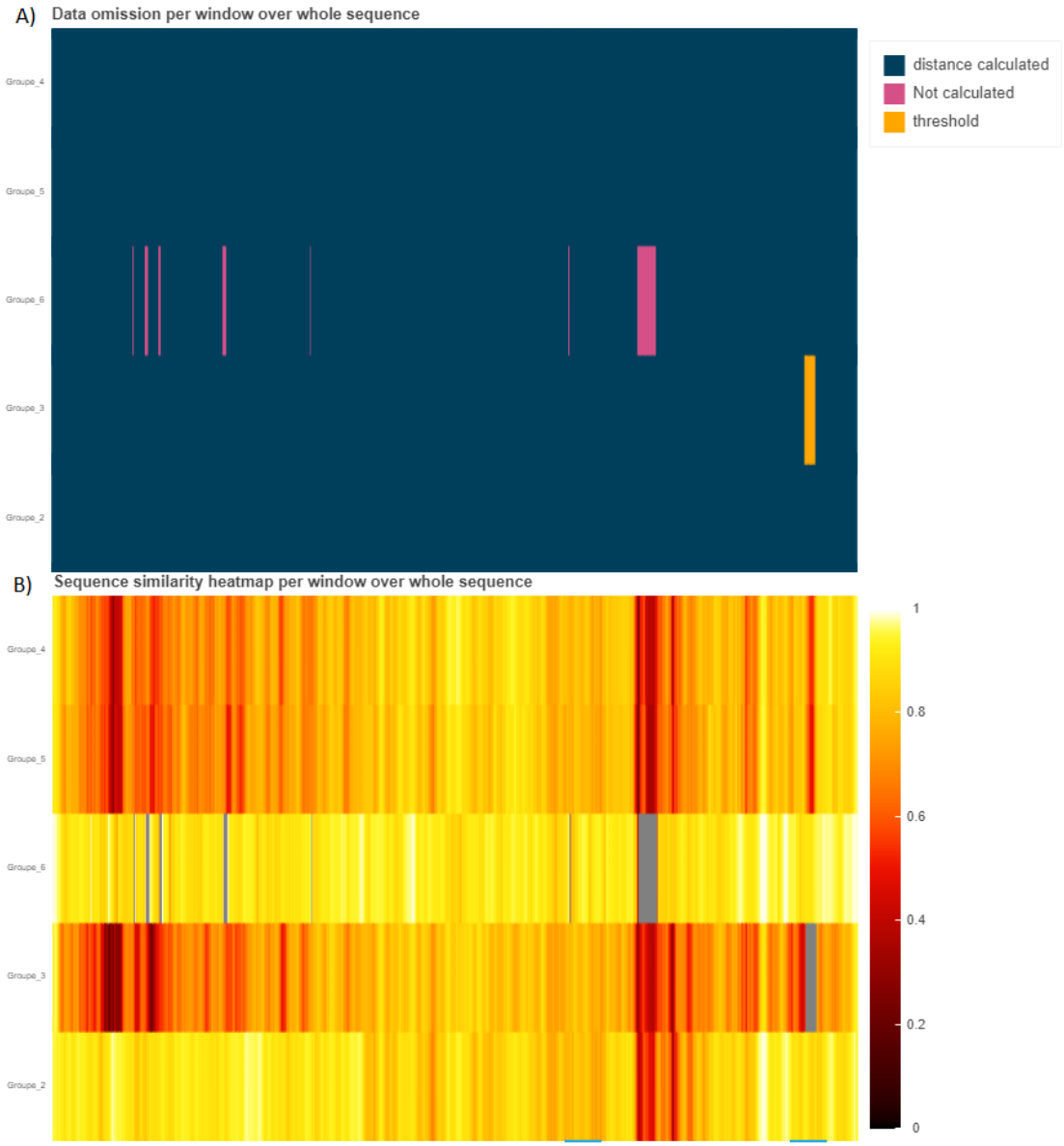


Figure 4.4: (a) Représentation par heatmap des fenêtres où les distances ont été omises ainsi que la cause. (b) Représentation par heatmap des distances entre les groupes

Ce rapport de qualité d'analyse est produit avec Bokeh (Bokeh Development Team, 2018), une librairie facilitant la conception de graphiques interactifs sur

les navigateurs web. Ainsi, le rapport interactif permet à l'utilisateur d'obtenir les données exactes composant chaque zone d'un heatmap en y déposant le curseur. Cet ajout à l'analyse SimPlot standard permet de mieux informer l'utilisateur tant qu'à la qualité des analyses produites et peut aider celui-ci à identifier et résoudre des problèmes qui pourraient survenir durant l'utilisation de cet outil.

4.4 Bootscan

La page d'analyse Bootscan permet à l'utilisateur d'employer le pipeline d'analyse de « Bootscanning » développée par Salminen et al (Salminen *et al.*, 1995). L'exécution de cette analyse est réalisée par l'emploi de quatre outils de la suite de logiciels PHYLIP (Felsenstein, 2005) employés en série. L'ensemble de l'analyse est représentée à la figure 4.5.

Tout comme l'analyse SimPlot, l'analyse Bootscan est une analyse par fenêtre coulissante. Ainsi, pour chaque déplacement de la fenêtre, la région de chaque séquence correspondante à la fenêtre consensus est extraite. Par la suite, ces sous-séquences sont employées comme données en entrée pour le logiciel SeqBoot, qui va appliquer la méthode statistique de bootstrap afin de multiplier le jeu de données aléatoirement avec une méthode de ré-échantillonnage.

Le nouveau jeu de donnée ré-échantillonné N fois est par la suite donné en entrée au logiciel DNADist afin de produire une matrice de distance entre les séquences pour chaque ré-échantillonnage effectué à l'étape précédente de bootstrap. Ces matrices de distances sont ensuite passées au logiciel neighbor afin de produire un arbre phylogénétique par la méthode « Neighbor-joining » pour chaque matrice de distance. Ces arbres sont par la suite passés au dernier logiciel de la suite PHYLIP, Consense. Ce logiciel permet de produire un arbre consensus à partir des multiples arbres en entrée et donne en sortie le nombres d'arbres où les différentes paires de

séquences ont été proximales. Cette dernière information, est extraite pour chaque paire de groupes et est stocké dans une structure de données.

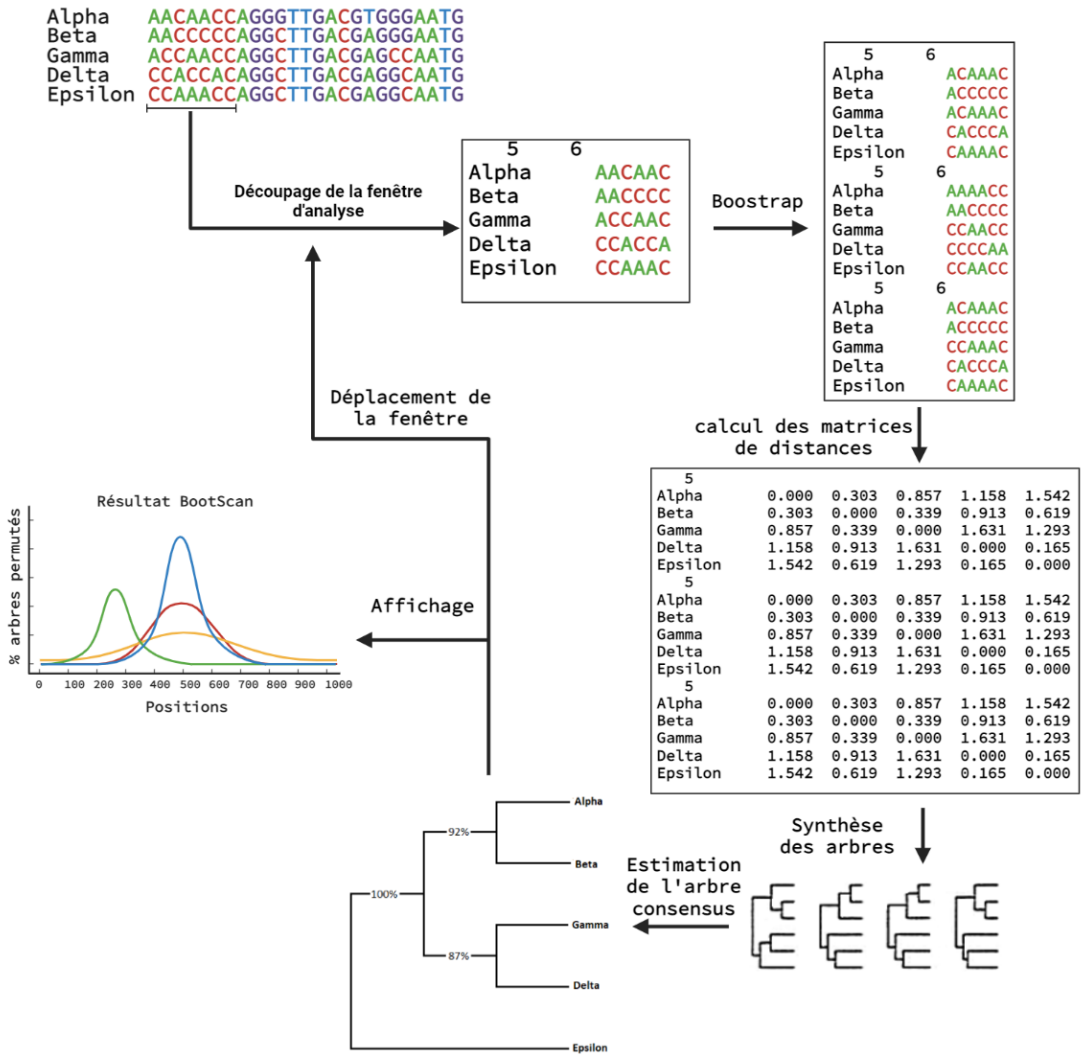


Figure 4.5: Réimplémentation du pipeline d'analyse Bootscan à l'aide de la suite de logiciels PHYLIP.

Lorsque la fenêtre d'analyse a traversée l'entièreté des séquences consensus, les pourcentages de fois où chaque groupe a été proximal au groupe de référence est calculé, pour chaque fenêtre d'analyse effectuée. Ces pourcentages sont ensuite représentés avec la librairie Matplotlib par un graphique de courbes afin de visualiser les régions recombinées. Tout comme l'analyse SimPlot, les différents outils permettant de modifier et sauvegarder les graphiques ont été incorporés à l'interface graphique du logiciel, ainsi qu'un bouton permettant de reproduire le graphique sur une nouvelle page indépendante. Une sortie typique de l'analyse Bootscan est présentée à la figure 4.6

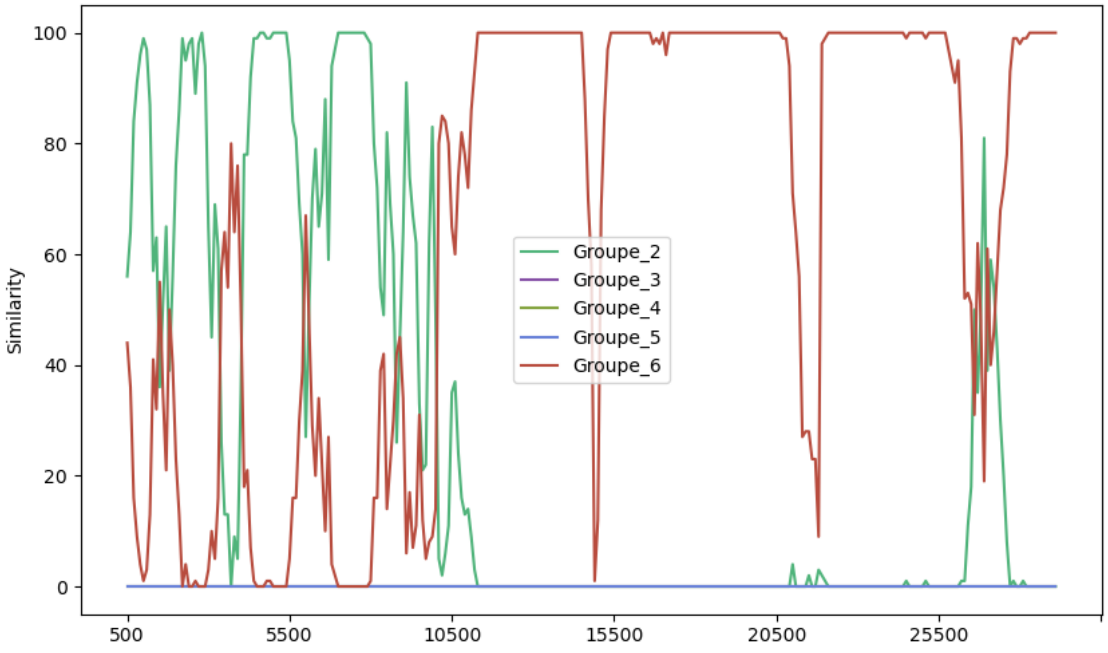


Figure 4.6: Exemple de sortie d'analyse Bootscan par SimPlot++.

4.5 FindSite

L'analyse FindSite est une méthode d'identification des sites informatifs entre quatre séquences développées par Robertson et Al (Robertson *et al.*, 1995). L'implémentation de cette approche est relativement directe, nécessitant en entrée une sélection de quatre séquences choisies par l'utilisateur. Une fois le test débutée, l'alignement de quatre séquences est analysé une position à la fois afin d'identifier les sites où deux séquences présentent un même nucléotide alors que les deux autres séquences présentent tous deux le même nucléotide différent de la première paire. Comme il existe seulement trois possibilités de regroupement entre quatre séquences, la configuration topologique de l'arbre résultant (voir la figure 3.4) est rapidement déterminée par des conditions logiques.

Les résultats sont présentés de façon textuelles à l'utilisateur de manière identique au logiciel SimPlot afin de standardiser les sorties.

4.6 Réseaux de similarités

L'analyse par réseau de similarité de SimPlot++ est une analyse basée sur les résultats de l'analyse SimPlot (Lole *et al.*, 1999). Puisque les matrices de distances de chaque fenêtre coulissante de l'analyse SimPlot représente la distance entre tous les groupes, celles-ci peuvent être utilisées afin de produire un réseau des similarités entre chaque groupe, et ceci pour chaque fenêtre.

La similarité présente entre deux groupes sur au moins une fenêtre est nommée « similarité locale », et est différente de la similarité de l'entièreté des séquences consensus, nommée « similarité globale ». Afin de distinguer ces deux types de similarités, les similarités globales sont représentées par des arêtes droites noires, alors que les similarités locales sont représentées par des arêtes courbes pointillées

de couleur rouge. Comme présenté à la figure 3.9, le réseau ainsi formé est un réseau multiplexe par sa représentation de deux couches d'informations distinctes.

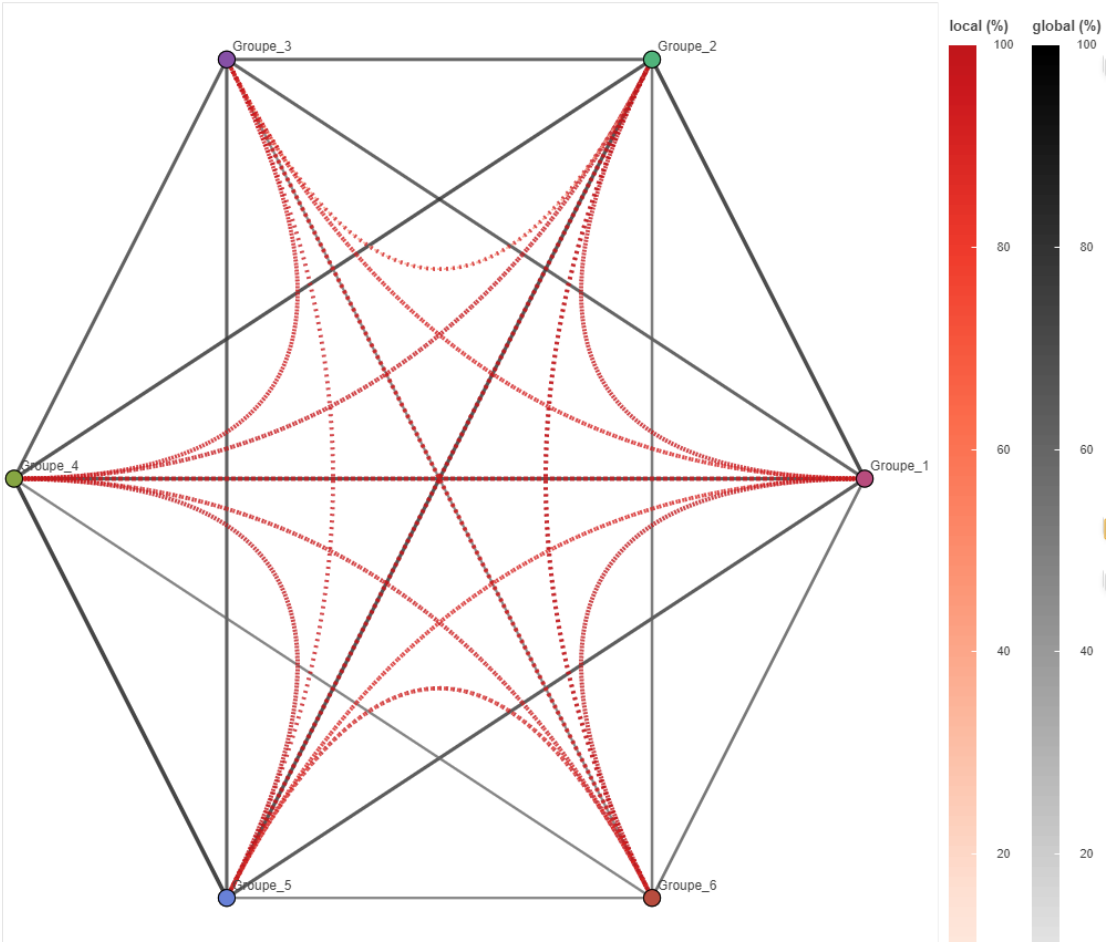


Figure 4.7: Exemple de réseau de similarité sans critères de filtre par SimPlot++.

Cependant, le réseau par défaut est trop permissif. Comme il est possible d'observer à la figure 4.7, tous les groupes sont interreliés entre eux, empêchant de distinguer les interactions d'intérêts. Plusieurs approches ont été employées afin de faciliter l'extraction des résultats d'intérêts de ce réseau. La première approche est de modifier l'aspect des différentes arêtes de similarité locales selon certains critères. À cet effet, la largeur des points des courbes pointillées est proportionnelle au pourcentage de similarité locale le plus élevé entre les deux séquences.

Ainsi, cette valeur de poids des courbes permet de représenter visuellement les groupes ayant, sur au moins une fenêtre d'analyse, une haute similarité locale entre eux. De plus, l'intensité de la couleur rouge de chaque arête pointillée est proportionnelle au pourcentage de fenêtres comportant une similarité locale sur l'ensemble des fenêtres possibles. L'intensité de la couleur noire des arêtes globales est calculée selon le pourcentage de similarité entre les deux séquences consensus. L'application de ces critères aux arêtes locales et globales permet ainsi de rapidement détecter les interactions les plus marquées entre les groupes.

Une approche additionnelle a été employée afin de permettre à l'utilisateur de filtrer les données à inclure dans le réseau, de façon dynamique et réversible. Des options ont été implémentées directement sur l'interface du réseau afin d'effectuer plusieurs opérations.

— Retirer les arêtes de similarités globales ou locales.
— Modifier le seuil de similarité minimal requis pour l'inclusion d'une donnée (seuils indépendants pour les similarité locales et globale).
— Visionner les similarités locales uniquement sur une sous-région des séquences.
— Retirer des groupes et leurs arêtes du réseau de similarité.
— Visionner sous forme de tableau les données incluses dans chaque arête du réseau.
— Visionner les résultats du test de détection de recombinaison par le test de proportion du jeu de données employé.

Le réseau ainsi que son interface est téléchargeable en format HTML et l'aspect dynamique est programmé en JavaScript, permettant à l'utilisateur de sauvegarder et partager le résultat de l'analyse tout en conservant l'aspect interactif. Le logiciel SimPlot++ n'est pas requis pour la réouverture du fichier.

4.7 Recombinaison

La page de recombinaison regroupe les différentes méthodes statistiques de détection de recombinaisons offertes par SimPlot++. Trois méthodes proviennent du logiciel PhiPack (Bruen *et al.*, 2006) et une nouvelle méthode simple et rapide a été développée afin de décrire mathématiquement les résultats des analyses SimPlot.

4.7.1 PhiPack

Puisque le logiciel PhiPack est programmé en C++ et est disponible uniquement sur Linux et MacOs, il a été jugé bénéfique de traduire les méthodes de calculs des différentes statistiques en langage python afin d'assurer une facilité accrue de maintenance de l'application ainsi que d'assurer la compatibilité du logiciel avec SimPlot++ sur les systèmes d'opération Windows.

Cette traduction du programme PhiPack de C++ vers python vient naturellement avec une réduction de la vitesse d'exécution mais, dans cette situation, cet inconvénient peut être minimisé. Comme les trois tests statistiques sont fortement structurés sous forme d'« arrays » et transformés à l'aide de calculs mathématiques simples, la librairie python Numba (Lam *et al.*, 2015) a été employée en concurrence avec NumPy (Harris *et al.*, 2020) pour accélérer le code python.

4.7.2 Optimisation avec Numba

La librairie Numba est un compilateur JIT (« Just In Time ») qui permet d'optimiser le temps d'exécution de fonctions spécifiques du code en les traduisant en langage machine lors de leur premier appel. Cette étape supplémentaire ralenti l'exécution lors de ce premier appel mais va significativement accélérer tous les

appels subséquents. Numba est donc fort avantageux dans les scénarios avec plusieurs appels de fonctions en boucles. De plus, comme Numba est appelé dans le code par l'utilisation de décorateurs, il est facile de cibler spécifiquement les fonctions et boucles qui sont computationnellement coûteuses. Numba bénéficie de (et oblige parfois) l'utilisation de NumPy, une librairie de calcul numérique scientifique qui offre des tableaux de données homogènes avec une meilleure gestion de la mémoire. Ainsi, la combinaison de Numba et NumPy peut permettre, dans certaines situations, d'approcher des vitesses d'exécutions similaires à celles d'un langage compilé (Lam *et al.*, 2015). De plus, Numba permet d'aisément paralléliser le code python en employant de multiples « threads » natifs, ignorant le GIL (« Global Interpretor Lock ») de l'interpréteur python.

Temps d'exécution (s)	472nt et 33 taxa	3765nt et 24 taxa
PhiPack	2.3	14.2
Python (base)	44.0	1411.1
Python (Numba)	10.8	58.1
Python (Numba) + parallélisation	9.4	20.3

Tableau 4.1: Tableau des comparaisons des vitesse d'exécution des versions du logiciel PhiPack selon les tailles d'alignements multiples

Comme le tableau 4.1 démontre, la version PhiPack originale est nettement plus rapide que la version python de base mais cette faiblesse peut être atténuée drastiquement par l'introduction de la libraire Numba et du compilateur JIT. Bien que la version du logiciel intégrée à SimPlot++ ne soit pas la plus rapide disponible, celle-ci est standardisée à travers toutes les plateformes et permet l'accès à cet outil sur le système d'opération Windows, qui n'en bénéficiait pas précédemment.

Également, l'interface graphique produite dans SimPlot++ permet de facilement

visualiser et modifier les paramètres d'analyse, de sauvegarder automatiquement le texte de sortie de PhiPack dans le même style que l'application originale, ainsi que de choisir d'employer le jeu de donnée brut ou les séquences consensus des groupes créés précédemment par SimPlot++.

4.7.3 Test de proportions

L'analyse par l'algorithme SimPlot est une méthode ultimement visuelle de détection des régions possiblement recombinées. Cet aspect intuitif de l'analyse est une de ces forces par sa rapidité et sa simplicité. Cependant, une analyse SimPlot comportant plusieurs groupes, sur de longues séquences génomiques, peut rapidement perdre sa simplicité. Ainsi, une méthode rapide a été développée afin d'identifier les régions génomiques où les proportions de distances entre la séquence de référence et les groupes pourrait suggérer qu'un évènement de recombinaison ait eu lieu. Pour chaque détection évaluée, la paire de groupes impliquée dans la recombinaison détectée, la région génomique où cette recombinaison ce serait produite, ainsi qu'une valeur de score relative expliqué plus bas est présentée à l'utilisateur.

Le test de proportions est une nouvelle métrique d'analyse basé sur les résultats d'une analyse SimPlot et sert d'outils mathématique pouvant aider à guider l'utilisateur vers les régions des graphiques de sortie qui montrent un potentiel de représenter des évènements de recombinaisons. Par ce fait, il s'agit d'un test indirect basé sur les proportions de distances génétiques calculées pour chaque fenêtre de l'analyse.

La structure de donnée employée pour le test correspond à un tableau où les rangées représentent les groupes de séquences autres que la séquence de référence et les colonnes représentent les positions moyennes des fenêtres d'analyses, tels qu'une fenêtre d'analyse entre les positions 0 et 200 se verrait attribuer la valeur

moyenne de position 100. Ce tableau va être rempli par les distances génétiques entre la séquence de référence et la séquence consensus des groupes, obtenus pas un modèle de distance à la discrétion de l'utilisateur.

Ce test est basé sur quelques principes décris ci-dessous, où une similarité locale représente une seule fenêtre où une séquence de référence et celle d'un groupe présentent une distance évolutive arbitrairement faible.

1. La région de recombinaison doit être l'une des plus similaires entre les deux séquences.
2. Une similarité locale élevée entre deux séquences a plus de poids si leur similarité globale est plus faible.
3. Une similarité locale élevée a plus de poids si les autres séquences consensus à la même position sont significativement plus faibles.
4. Plusieurs similarités locales consécutives augmente leur poids.

Ces quatre principes représentent les quatre étapes principales par lesquelles le score est calculé.

La première étape du test est de filtrer le tableau afin de retirer, pour chaque rangée (groupe), toutes les données de distance inférieures à la distance moyenne de cette rangée. Une distance sous la moyenne a, en accord avec le premier principe, une faible probabilité de représenter un évènement de recombinaison. Par le retrait de ces valeurs, la rapidité d'exécution de la méthode est améliorée puisqu'assumant une distribution normale des distances par rangées, 50% des distances seront retirées. Il est bon de noter également que dû à l'emploi de fenêtres coulissantes, la distance moyenne cumulée d'une rangée est différente de sa similarité globale.

La seconde étape du test est de calculer une valeur de score pour chaque distance restante dans le dataframe. Ce score est obtenu par l'équation 4.1 ci-dessous.

$$Score = \frac{\frac{d}{Distance_{globale}}}{\overline{Distance}_{position}} \tag{4.1}$$

Où :

d représente la distance entre deux séquences dans une fenêtre,

$Distance_{globale}$ représente la distance globale entre ces deux séquences,

$\overline{Distance}_{position}$ représente la distance moyenne entre la séquence de référence et les autres séquences à cette position.

Cette équation, basée sur les principes 2 et 3 présentés plus haut, permet d'ajuster les résultats bruts obtenus pour chaque position selon la similarité de leur propre séquence globalement ainsi que de la similarité entre la séquence de référence et les autres séquences analysées. Cette étape permet donc de produire un score comparable entre les positions et les séquences.

Suite à cette étape, les scores les plus élevés sont un indicateur fort des positions et groupes où la séquence consensus était plus similaire à la séquence de référence que la similarité globale entre cette séquence et la séquence de référence, en plus d'être davantage similaire à la séquence de référence que la moyenne de similarité des autres groupes à cette même position. Afin d'extraire ces valeurs d'intérêt, les scores seront filtrés afin de conserver le top 20% des scores les plus élevés à travers le tableau. La valeur de 20% a été déterminée arbitrairement afin qu'approximativement 10% du jeu de donnée brut soit représenté lors de la dernière étape, permettant plusieurs opportunités de scores d'un même groupe à des positions consécutives sans toutefois conserver un grand nombre de scores moins élevés.

Ainsi, pour chaque groupe, les scores aux positions consécutives sont additionnés afin d'ajouter du poids aux régions potentiellement mosaïques et de limiter les aberrations à des positions uniques qui pourraient se produire. Tous les scores cumulés représentant un minimum de deux positions consécutives sont gardés en mémoire avec le nom du groupe de référence, le nom du groupe d'intérêt, le score

obtenu, et les positions de début et de fin de la région identifiée. Comme l'analyse SimPlot produit N tableaux pour une analyse de N groupes (chaque groupe peut être le groupe de référence), le test de proportion est exécuté en série sur chacun de ces tableaux et les scores obtenus sont agrégés. Un filtre est employé afin de retirer les duplicatas provenant des détections inverses tels ceux d'un groupe A et B qui présentent les mêmes tendances dans les dataframes où ceux-ci sont les groupes de référence.

Au final, les 10 scores cumulés les plus élevés sont présentés à l'utilisateur, afin de guider l'analyse des résultats de SimPlot et de présenter des régions d'intérêts qui pourraient sinon être manqués.

CHAPITRE V

MÉTHODOLOGIE

Cette section du mémoire est dédiée au processus de sélection, d'acquisition et de traitement des jeux de données employés lors des analyses subséquentes. La formation des groupes employés lors des analyses SimPlot++ y sont également discutés.

Afin d'étudier l'histoire évolutive du gène S et de son domaine RB, des jeux de données appropriés ont dû être générés. Pour ce faire, des séquences génomiques de coronavirus ayant déjà été présenté comme d'intérêt dans des articles publiés traitant de l'histoire évolutive de SARS-CoV-2 ont été identifiée et téléchargés à partir de bases de données tels GISAID (Elbe et Buckland-Merrett, 2017) et GenBank (Benson *et al.*, 2013).

5.1 Jeu de données de 24 séquences de coronavirus du gène S

Afin d'identifier l'origine évolutive du gène S de SARS-CoV-2, 24 génomes de coronavirus ont été sélectionnées à partir de l'article de (Lam *et al.*, 2020). Ceux-ci représentent des coronavirus originant de chauve-souris, de pangolins ainsi que d'humains et permettront de situer SARS-CoV-2 dans le contexte élargi des betacoronavirus.

Ainsi, ce premier jeu de données consiste en 9 groupes de séquences distincts :

- 3 séquences de SARS-CoV-2 échantillonnés à Wuhan au début de la pandémie (2019-2020).
- 6 séquences de CoV isolés à partir d'échantillons de pangolins malais (*Manis Javanica*) obtenus lors d'opérations anti-contrebande dans la région de Guangxi en 2017-2018 (pangolin GX).
- 2 séquences de CoV de pangolins (*Manis Javanica*) obtenus lors d'opérations anti-contrebande dans la région de Guangdong en 2019 (pangolin GD).
- 2 séquences de CoV de chauves-souris *Rhinolophus sinicus*, originant de Chine entre 2015 et 2017 (ZXC21, ZC45).
- 3 séquences de CoV de chauve-souris *Rhinolophus* recueillis en Chine entre 2005 et 2010 (HKU3-6, HKU3-13, Rf1).

— 3 séquences de CoV de chauve-souris *Rhinolophus* recueillis en (BtCoV273, BtCoV279 et Rs3367).
— 1 séquence de CoV de chauve-souris *Rhinolophus affinis* originant de la région de Yunnan en Chine, recueillie en 2013 (RatG13).
— 2 séquences de CoV de chauve-souris *Rhinolophus* provenant du Kenya et de la Bulgarie en 2007 et 2008 (BtKY72 et BM48-31/BGR/2008).
— 2 séquences d'origine humaine (Tor2) et de civet (PCA-13).

Une représentation par tableau de ces groupes, comportant les numéros d'accessions, est disponible à l'annexe A.

5.2 Jeu de données des 43 séquences de variants de SARS-CoV-2

Afin de représenter adéquatement la diversité de variants existant dans la population humaine, la sélection des variants jugés « d'intérêt » dans le contexte de l'analyse a été effectué selon la liste fournie par le CDC (CDC, 2020) des variants classifiés comme « variants d'intérêts » et « variants préoccupants ». Ainsi, 34 séquences génomiques correspondant aux 12 lignées Pango (Rambaut *et al.*, 2020) les plus prévalentes ont été téléchargées par l'entremise de GISAID. De plus, 8 autres séquences de variants ont été sélectionnés dû à leur présence dans un article portant sur l'émergence de mutations au niveau du domaine RB (Ou *et al.*, 2021). Finalement, le génome de référence du SARS-CoV-2 a été ajoutée au jeu de données afin d'agir comme séquence de référence lors des analyses SimPlot et Bootscan.

Une représentation par tableau de ces groupes, comportant les lignées Pango, appellations communes et numéros d'accession des séquences est disponible à l'annexe B.

5.3 Préparation des données

Pour les deux jeux de données, les séquences téléchargées représentent le génome complet de chaque organisme prélevé. Ainsi, les régions génomiques correspondant au gène S ont été extraites pour chacun des génomes et alignés avec Muscle (Madeira *et al.*, 2019) avec les paramètres par défaut. Les alignements multiples de séquences (AMS) ont été peaufinés à l'aide du logiciel Gblocks (Talavera et Castresana, 2007) afin d'éliminer les positions mal alignées et améliorer la reproductibilité des analyses phylogénétiques. Les paramètres les moins stricts ont été sélectionnés pour chaque jeu de données traités.

Pour les séquences impliquées dans l'analyse du domaine RB, les régions correspondant au domaine RB ont été isolée et traduites en alignement multiple de séquences d'acides aminées à l'aide du logiciel MEGA-X (Kumar *et al.*, 2018) et alignés à nouveau avec Muscle.

5.4 Détails d'analyse

Les analyses SimPlot et Bootscan ont été effectués avec une fenêtre d'analyse de 200pb et un pas de 20pb pour les séquences du gène S. Les analyses effectuées sur le domaine RB ont été effectuées avec une fenêtre d'analyse de 20pb et un pas de 5pb. Dans tous les cas, le seuil de «gaps » par défaut de 33% a été employé. Les modèles de distances employés ont été déterminé par l'entremise de MEGA-X (Kumar *et al.*, 2018). Le modèle le plus performant offert par SimPlot++ a été sélectionné.

En ce qui concerne les analyses Beast2, les dates de prélèvements des séquences proviennent des fiches informatives associées aux numéros d'accessions. Si la date disponible pour une séquence est incomplète (jour ou mois de prélèvement man-

quant), le premier jour du mois et/ou le premier mois de l'année ont été utilisés. Les paramètres d'arbres *coalescent constant population* avec une distribution log-normale de moyenne -5 et de déviation standard de 1.25 ont été sélectionnés lors de la reconstruction phylogénétique. La sélection de ces paramètres est basée sur les ressources du site de référence «Taming the beast » (Barido-Sottani *et al.*, 2018).

CHAPITRE VI

RÉSULTATS

Ce chapitre du mémoire est axé sur la présentation et l'analyse préliminaire des résultats obtenus à travers le logiciel SimPlot++ ainsi que des logiciels autres ayant été employé lors de l'analyse.

Ainsi, les résultats liés à l'étude de l'origine du gène S de SARS-CoV-2 sont présentés en premier, suivi de l'analyse du domaine RB des variants du SARS-CoV-2. Finalement, les arbres phylogénétiques issus d'une analyse avec le logiciel BEAST2 sont présentés à la fin du chapitre.

6.1 Analyse du gène S

6.1.1 Analyse SimPlot

Une analyse SimPlot de la séquence nucléotidique du gène S a été effectuée sur le jeu de données de 24 séquences (figure 6.1). Cette analyse démontre que les génomes de SARS-CoV-2 et de RaTG13 sont fortement similaires sur la majorité du gène S, partageant une similarité globale de 92% pour ce gène. Toutefois, la similarité locale entre ces deux séquences chute jusqu'à près de 65% entre les positions 1100 et 1600. Cette région du gène S contient l'information génique correspondant au domaine RB. Alors que la similarité entre SARS-CoV-2 et RaTG13 diminue dans cette région, la similarité entre SARS-CoV-2 et la séquence consensus correspondant au groupe de CoV de pangolin de Guangdong se maintient autour de 85%. Cette interaction entre les trois groupes de séquences est un indicateur qu'un évènement de recombinaison ait pu se produire entre SARS-CoV-2 et le CoV de pangolin de Guangdong au niveau du domaine RB du gène S. Le signal de détection est d'autant plus accentué par le fait que tous les autres groupes de séquences présentent une baisse importante de similarité avec la séquence de SARS-CoV-2 sur cette région.

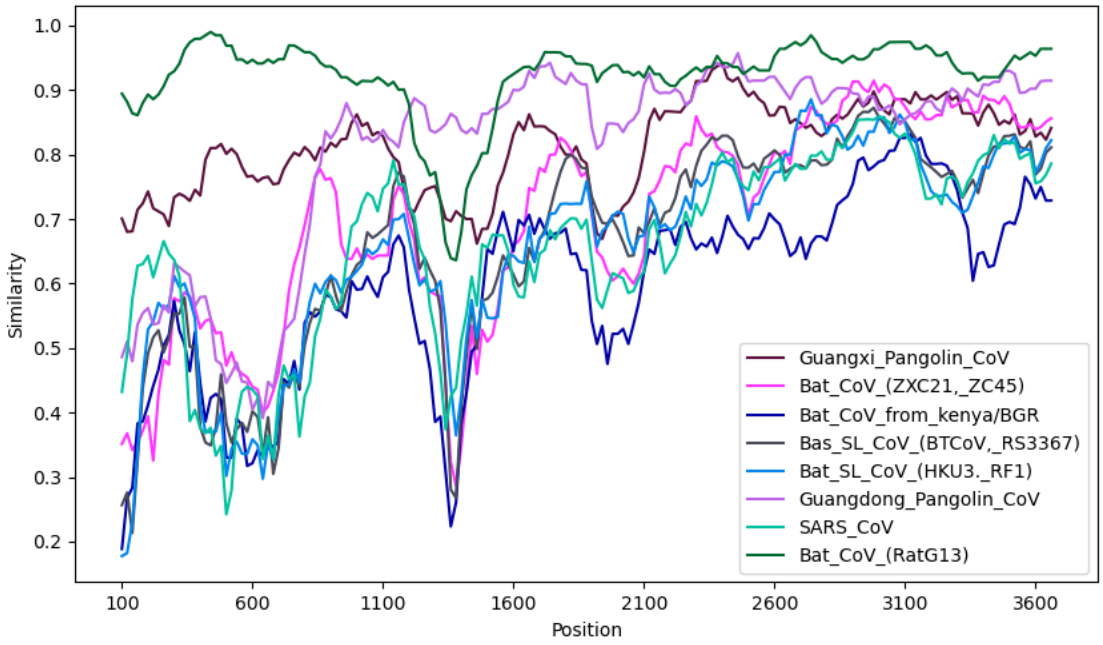


Figure 6.1: Graphique SimPlot de la distance entre les groupes de séquences de CoV et le groupe SARS-CoV-2 sur le gène S. L'analyse a été effectuée avec une fenêtre coulissante de 200 pb, un pas de 20 pb et le modèle TN93.

6.1.2 Analyse par l'algorithme Bootscan

Une analyse Bootscan a été effectuée sur le même jeu de données afin de présenter les régions potentiellement mosaïques des différents groupes. En visionnant les résultats avec SARS-CoV-2 comme référence (figure 6.2), le signal de recombinaison de l'analyse SimPlot de la figure 6.1 peut être également détecté entre les positions 1100 et 1600 du gène S. On peut y observer que le nombre d'arbres phylogénétiques où les séquences consensus de SARS-CoV-2 et RaTG13 sont proximales diminue jusqu'à atteindre 0% dans cette région, alors que le nombres d'arbres permutés où SARS-CoV-2 et le CoV de pangolin de Guangdong sont proximales augmente jusqu'à représenter 100% des arbres permutés.

Un autre signal de recombinaisons est également visible à travers cette analyse

Bootscan. Celui-ci est située autour de la position 2500, et présente une chute du pourcentage d'arbres où RaTG13 est proximale à SARS-CoV-2, alors que les groupes de pangolins de Guangdong et de Guangxi y représentent respectivement jusqu'à 80% et 20% des arbres permutés à leur valeurs les plus élevées. Ce signal est de moindre intensité que celui correspondant au domaine RB dû à sa breveté sur la séquence (une seule fenêtre d'analyse du groupe du CoV de pangolin de Guangdong est supérieure à 60%), ainsi que le fait que le signal est partagé entre les deux groupes de pangolins.

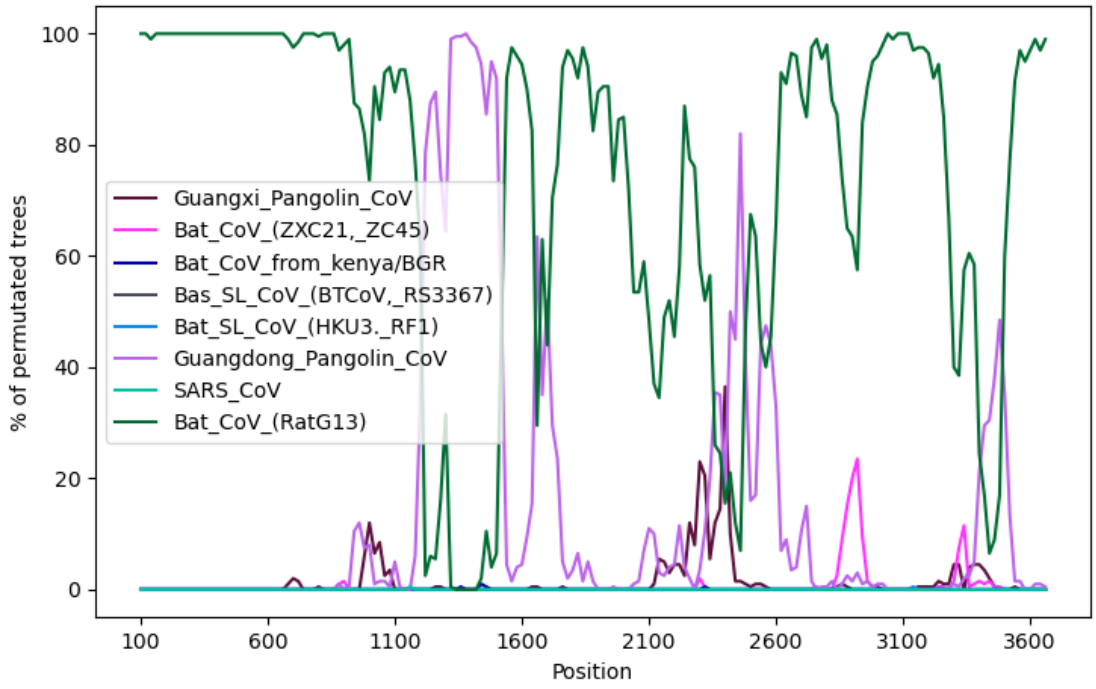


Figure 6.2: Graphique BootScan représentant pour chaque fenêtre d'analyse, le pourcentage d'arbres phylogénétiques construits où chaque séquence consensus de groupe était proximale au groupe de référence SARS-CoV-2 sur le gène S. L'analyse a été effectuée avec une fenêtre coulissante de 200pb, un pas de 20pb et 200 bootstrap. Les arbres phylogéniques sont basés sur des matrices de distances par le modèle de Kimura et générés par la méthode « Neighbour-Joining ».

6.1.3 Analyse FindSite

Une analyse de détection des sites informatifs a été effectués pour quatre séquences d'intérêts et les résultats textuels ont été manipulés afin de représenter graphiquement les résultats (figure 6.3). Comme SARS-CoV-2 a été sélectionnée comme référence, les deux séquences autres impliqués dans les signaux de recombinaisons détectés précédemment, RaTG13 et une séquence de CoV de pangolin de Guangdong (GD/P1L) ont été employés, d'autant que la séquence distante Tor2 comme groupe externe. Par la grande similarité globale entre RaTG13 et SARS-CoV-2 au niveau du gène S, il est attendu que ces deux séquences présentent un grand nombre de sites informatifs, comme présenté dans le graphique en bande de la figure 6.3. Par sa grande distance évolutive Tor2 présente peu de sites informatifs avec SARS-CoV-2 et ceux-ci sont distribués à travers l'entièreté de la séquence. Les sites informatifs impliquant GD/P1L sont également distribués à travers l'entièreté du gène mais présente une forte concentration au niveau de la région 1100-1600 identifiée précédemment, tels que présentée par le graphique à noyau de la figure 6.3. Cette forte densité représente également un signal de recombinaison entre SARS-CoV-2 et GD/P1L.

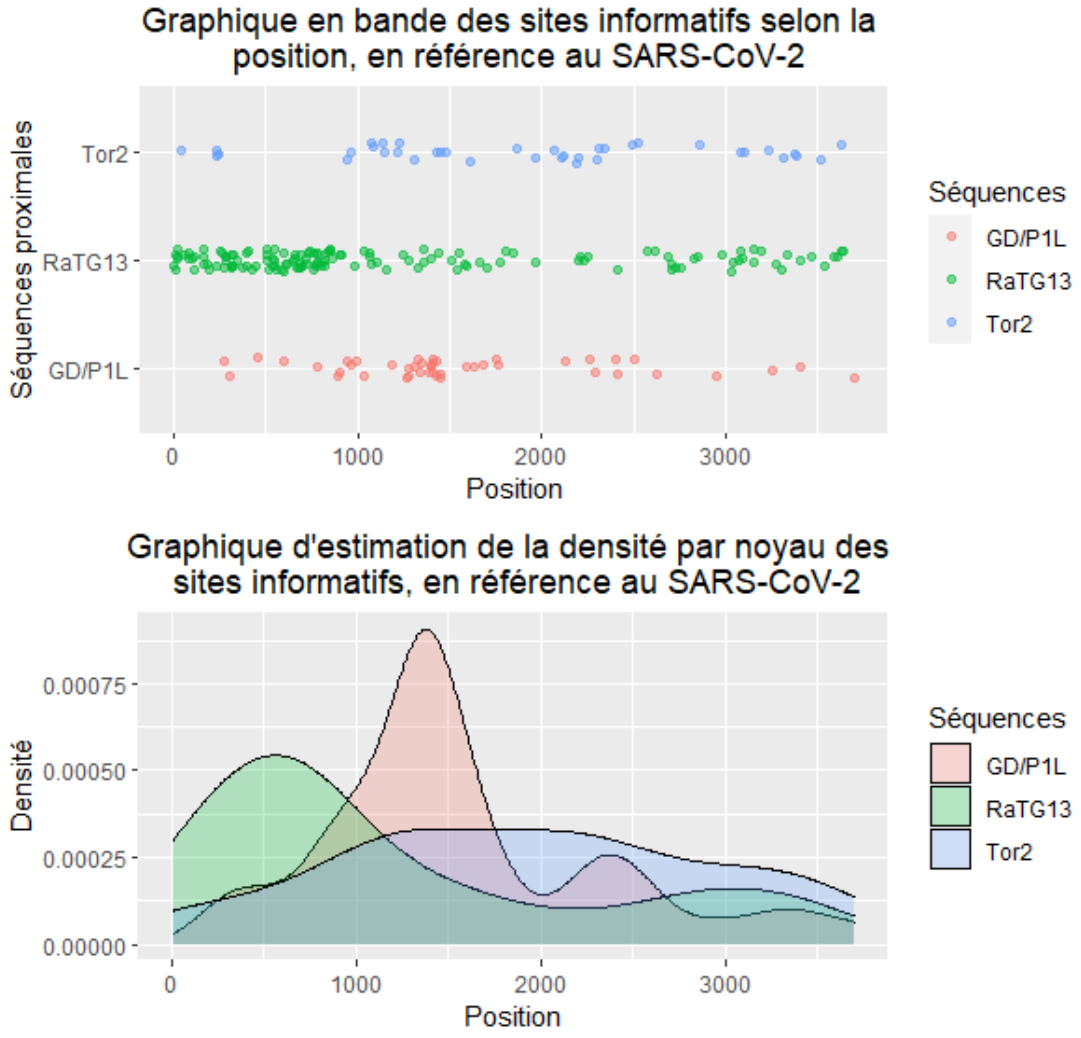


Figure 6.3: Représentation en graphique à bandes et en densité par noyau de la répartition des sites informatifs entre les séquences de SARS-CoV-2 (référence), de CoV de pangolin de Guangdong GD/P1L, de CoV de chauve-souris RaTG13 et de SARS Tor2 (groupe externe). Cette représentation est a but de visualisation et n'est pas offerte par SimPlot++

6.1.4 Analyse par réseaux de similarités

Les résultats SimPlot ont été représentés sous forme de réseaux de similarités afin de représenter globalement les résultats sans l'emploi d'un groupe de référence

(figure 6.4). Afin de représenter uniquement les similarités les plus élevées dans la région du domaine RB, le cadre de lecture a été limité aux fenêtres d'analyses comprises entre les positions 800 et 1600 du gène S. Les seuils de similarités afin de produire une arêtes entre deux groupes (noeuds) ont été établis respectivement à 88% pour les similarités locales et de 80% pour les similarité globales. À travers ce réseau, il est possible de visualiser les interactions entre les groupes dans la région génique comprenant le gène S et les signaux de recombinaisons identifiés précédemment.

Un regroupement de quatre groupes de séquences est présent dans le réseau, composé des CoV de pangolins de Guangxi et de Guangdong, ainsi que de RaTG13 et de SARS-CoV-2. Ce regroupement est attendu et est démontré par la proximité globale entre ces séquences, représenté par les arêtes noires entre eux. D'autant plus, des similarités locales (arêtes rouges) sont présentés entre SARS-CoV-2 et les groupes RaTG13 et de CoV de pangolins de Guangdong. SARS-CoV-2 et RaTG13 présentent deux similarités locales distinctes dans cette région, la première, de 95% de similarité, entre les positions 700 et 1280, et la deuxième, de 92% entre les positions 1460 et 1700. Pour le CoV de pangolin de Guangdong, une seule similarité locale est présente, de 88% entre les positions 1120 et 1620.

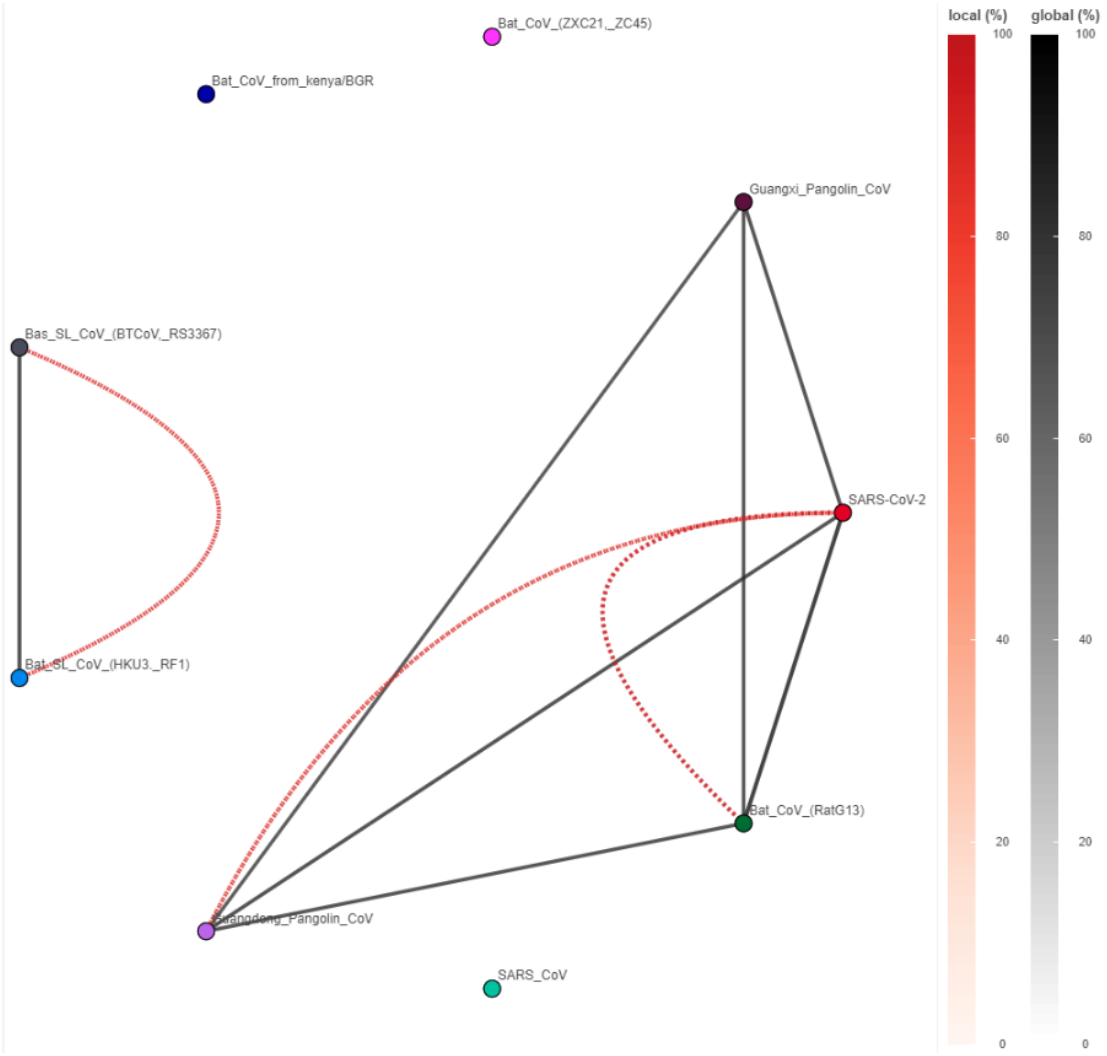


Figure 6.4: Représentation en réseau des similarités entre les groupes de séquences, représentés par les noeuds et reliés par des arêtes représentant les similarités locales (rouge) et globales (noir). Les seuils de similarités afin de produire une arêtes entre deux groupes (noeuds) ont été établis respectivement à 88% pour les similarités locales et de 80% pour les similarité globales.

6.1.5 Analyse par test de proportions

Un test de proportions a également été effectué afin de présenter les régions présentant des proportions de similarités entre les groupes analysés et le groupe de

référence pouvant représenter des évènements de recombinaisons. Les cinq régions potentiellement recombinées ayant obtenu les scores les plus élevés sont présentés dans le tableau 6.1. Le haut score obtenu entre RaTG13 et SARS-CoV-2 aux positions 280-920 est dû à la forte similarité entre ces deux séquences alors que les autres séquences consensus présentent de très faibles similarités avec RaTG13 dans cette région. La seconde région d'intérêt avec le score le plus élevé est la région 1220-1720 entre SARS-CoV-2 et le CoV de pangolin de Guangdong.

Groupe de référence	Groupe impliqué	Région détectée	Score
RaTG13	SARS-CoV-2	280-920	58.7
SARS-CoV-2	Pangolin Guangdong	1220-1720	41.2
SARS-CoV	Bat CoV (ZXC21,ZC45)	1600-2100	39.8
Bat CoV (ZXC21,ZC45)	Bat SL CoV (BTCoV, RS3367)	1020-1340	27.7
RaTG13	Pangolin Guangdong	1880-2220	26.2

Tableau 6.1: Tableau des résultats du test de proportions sur le jeu de données de 24 séquences du gène S.

6.2 Analyses des variants

6.2.1 Analyse du gène s des lignées Pango

L'analyse SimPlot effectuée sur le jeu de données des séquences nucléotidiques des différents variants de SARS-CoV-2 (figure 6.5) permet de résumer les régions mutées de chaque souche, comparativement à la séquence du gène S de référence de SARS-CoV-2 (NC045512.2). Par cette approche, les régions mutées de chaque variants peuvent être représentées graphiquement. Les régions hautement conservées, tels une courte région aux position 860-920 et une région plus grande aux positions 2260-2560 sont également visibles.

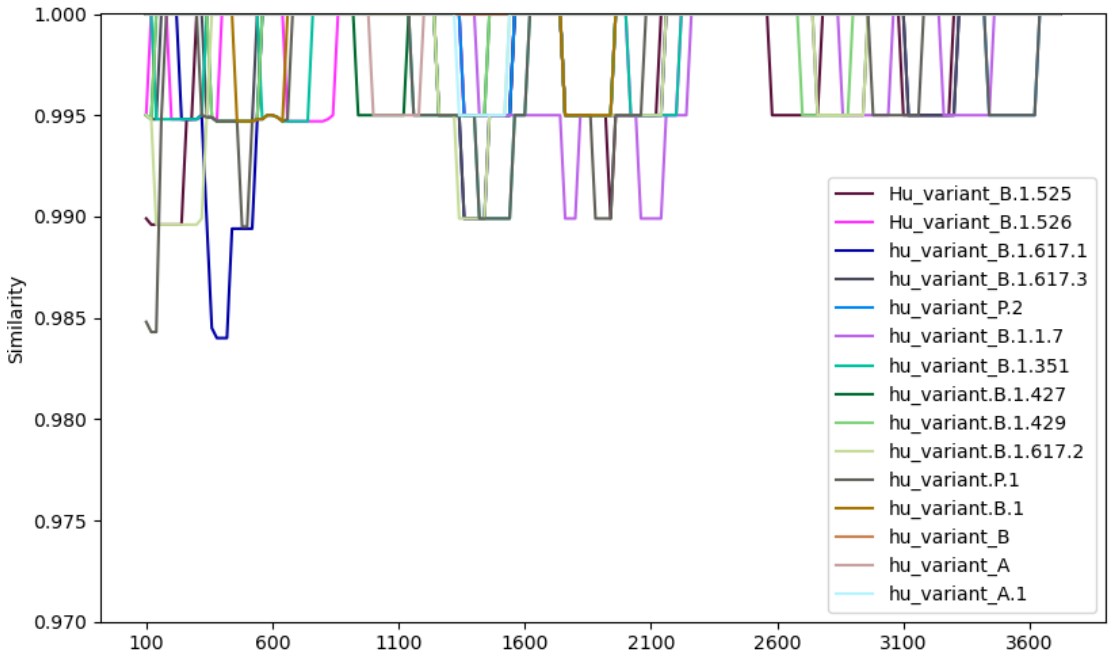


Figure 6.5: Graphique SimPlot de la similarité entre les groupes de variants de SARS-CoV-2 regroupés par leur lignée Pango. L'analyse a été effectuée avec une fenêtre coulissante de 200 pb, un pas de 20 pb et le modèle HKY85.

Les outils de contrôle-qualité inclus dans le logiciel SimPlot++ permettent également de déterminer que malgré le grand nombre d'intersections de lignes dans la figure 6.5, seul le groupe hu_variant_B (lignée Pango B) ne présente aucune mutation dans la région 1000-1600 correspondant approximativement à la région codant pour le domaine RB (figure 6.6).

Figure 6.6: Heatmap des similarités entre les groupes de variants de SARS-CoV-2 regroupés par leur lignée Pango.

6.2.2 Analyse du domaine RB des lignées Pango

Une analyse SimPlot a été effectuée sur les séquences protéiques consensus correspondant au domaine RB des différents variants de SARS-CoV-2 (figure 6.7). Afin de représenter clairement les mutations de courtes tailles, l'analyse SimPlot a été effectuée avec une fenêtre de 20 résidus et un pas de 5 résidus. Ainsi, les valeurs de distances représentées sont très sensibles aux mutations. Il est donc possible d'y observer les régions du domaine RB comportant des mutations, pour chaque lignée Pango.

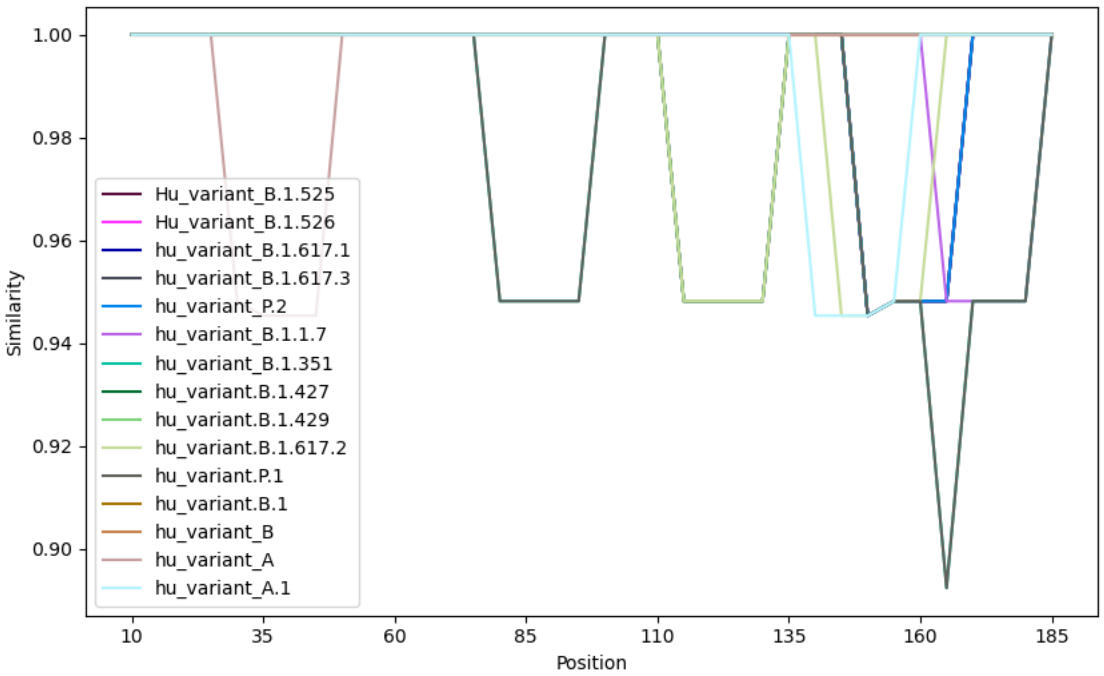


Figure 6.7: Graphique SimPlot de la similarité entre les groupes de variants du domaine RB de SARS-CoV-2 regroupés par leur lignée Pango. L'analyse a été effectuée avec une fenêtre coulissante de 20 pb, un pas de 5 pb et le modèle Kimura.

Le heatmap présenté à la figure 6.8 permet de représenter sans chevauchement de courbes les résultats obtenus à la figure 6.7.Celui-ci permet de démontrer que la majorité des mutations observées chez les variants au niveau du domaine RB sont situés autour des résidus 120 et 150.

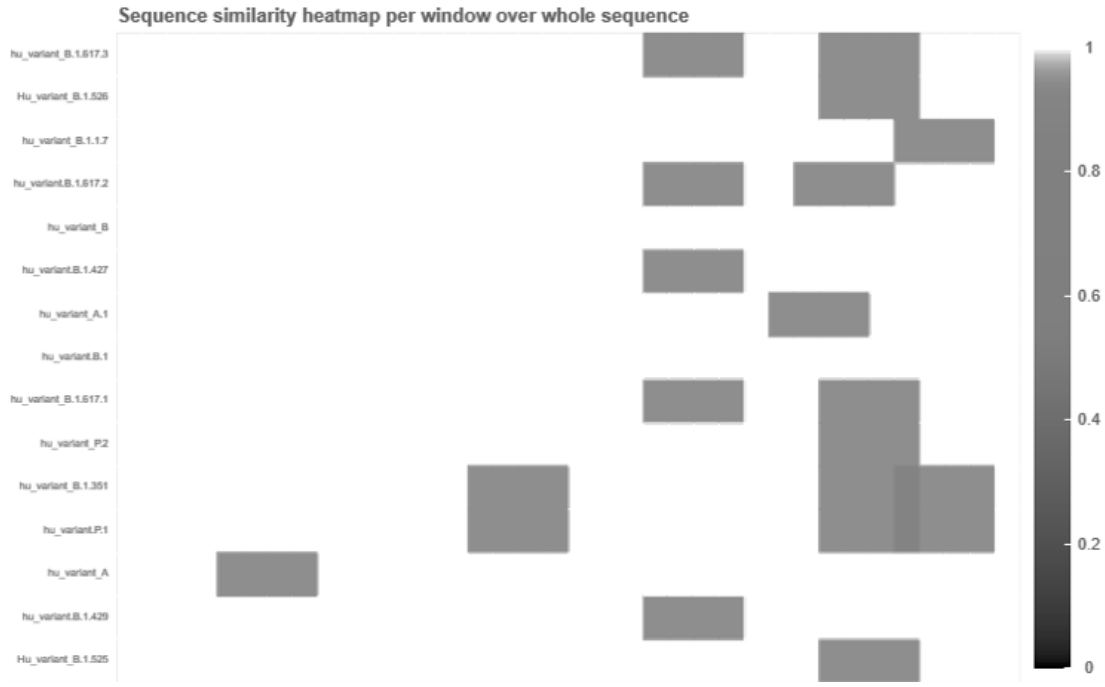


Figure 6.8: Heatmap des similarités entre les groupes de variants de SARS-CoV-2 regroupés par leur lignée Pango.

6.3 Analyses phylogénétiques

Comme présenté à la section 3.6, Beast2 est un logiciel de reconstruction phylogénétique permettant l'emploi de méthodes complexes d'horloges moléculaires et de modèles évolutifs afin de générer des arbres phylogénétiques. L'une des caractéristiques des arbres en sortie de Beast2 est l'inclusion de valeurs temporelles pour chaque noeud interne des arbres. Il est donc possible d'estimer quand deux séquences ont divergé dans le passé.

La première analyse phylogénétique consiste à représenter l'histoire évolutive des premières séquences du domaine RB de SARS-CoV-2 originant de la région de Wuhan en chine. Pour ce faire, une analyse Beast2 a été effectuée, consistant de ces séquences humaines de SARS-CoV-2, des séquences de pangolins malais

de Guangdong et Guangxi, ainsi que du CoV de chauve-souris RaTG13. Pour chacune de ces séquences, la date de prélèvement des échantillons biologiques a été incluse dans Beast2. L'analyse a été lancée avec le modèle WAG et 5 catégories gamma, le modèle d'horloge moléculaire strict et une longueur de chaîne MCMC de 10 millions d'arbres avec un prélèvement à chaque 5 000 arbres. Un *burn-in* de 5% des arbres a été retiré.

La figure 6.9 présente l'arbre consensus initial issu de cette analyse. Comme présentés dans la littérature, il est possible de voir que la séquence humaine (RBD_ref_seq) est proximale à la séquence du CoV de pangolin de Guangdong, indiquant une proximité génétique entre ces deux séquences. La séquence RaTG13 est plus distance, suivi des CoV de panglins de Guangxi. Cependant, l'horloge évolutive présentée laisse à désirer, situant l'évènement de divergence entre les séquences de SARS-CoV2 et du pangolin de Guangdong de RaTG13 en 1986.

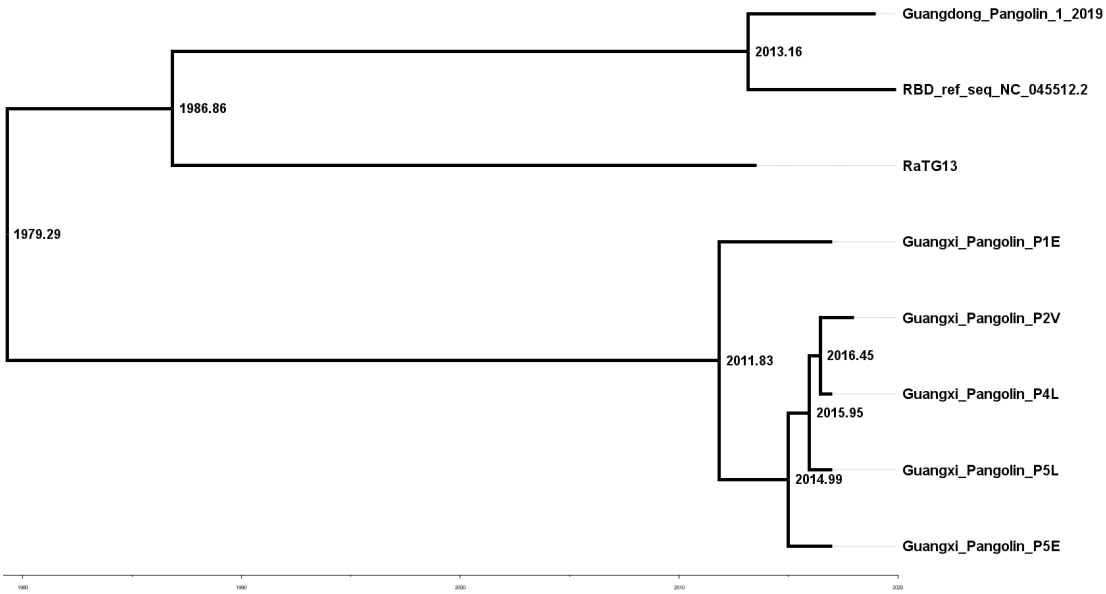


Figure 6.9: Reconstruction phylogénétique par Beast2 des séquences de SARS-CoV-2 et des séquences de CoV RaTG13, de pangolin de Guangdong et de Guangxi.

Un second arbre 6.10 a donc été généré par les logiciels de la suite Beast2 avec un facteur temporel de 0.27, afin de représenter une hypothèse plus probable de l'histoire évolutive du domaine RB de SARS-CoV-2. Cet arbre suggère que les séquences de SARS-CoV-2 et du pangolin de Guangdong aient divergées en 2018, alors que RaTG13 aient divergés vers la fin de 2011.

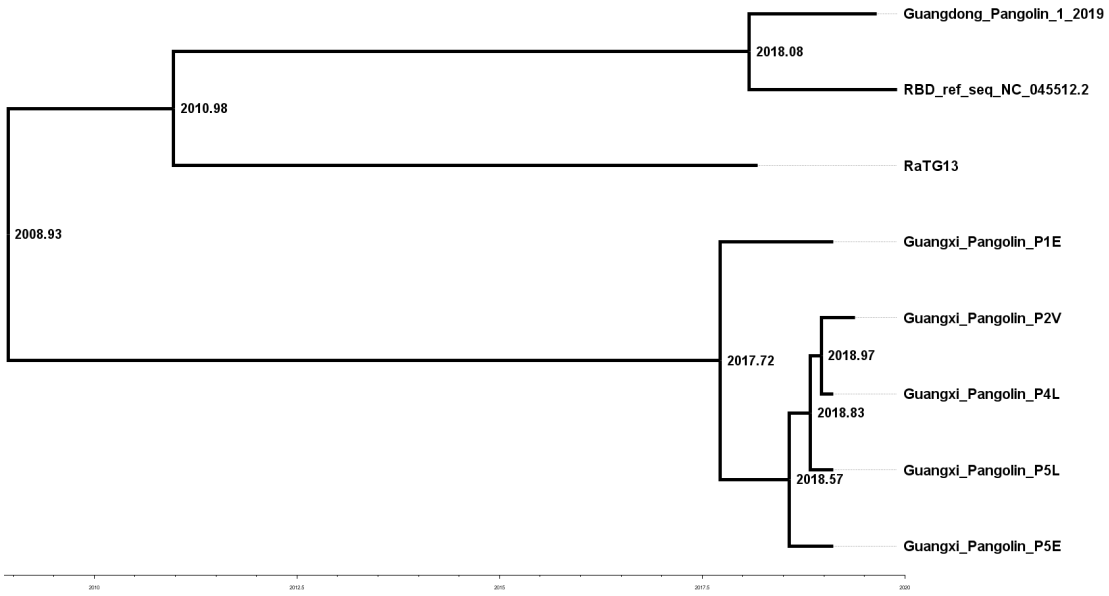


Figure 6.10: Reconstruction phylogénétique par Beast2 des séquences de SARS-CoV-2 et des séquences de CoV RaTG13, de pangolin de Guangdong et de Guangxi. Un facteur temporel de 0.27 a été ajouté pour les âges des noeuds internes.

Par la suite, une analyse Beast2 a été effectuée selon la même approche avec le jeu de données composé des séquences de pangolins de Guangdong et des variants de SARS-CoV-2, distribués selon leur lignée Pango. Comme dans l'analyse phylogénétique précédente, les dates de prélèvements des échantillons ont été employés, ainsi qu'un modèle d'évolution WAG sans paramètres gamma et le modèle d'horloge strict.

la figure 6.11 présente l'arbre consensus résultant de cette analyse. Comme attendu, la séquence de CoV de pangolin de Guangdong représente le groupe externe de l'arbre, avec un âge de divergence estimé à l'année 2017. Cependant, les hypothèses temporelles lié à cet arbre ne sont pas aussi crédible dû à des divergences entre des variants ayant lieu au début de 2019, soit un an avant que SARS-CoV-2 soit considéré pandémique.

Ainsi, comme lors de la première analyse phylogénétique, un facteur temporel de 0,75 a été appliqué à l'arbre de la figure 6.11, résultant en une nouvelle hypothèse évolutive présentée à la figure 6.12. Cette hypothèse concorde d'avantage avec l'hypothèse apportée par la figure 6.10 puisque les âges de divergences prédit entres les séquences de SARS-CoV-2 et de CoV de pangolin de Guangdong sont tous deux situé autour du début de l'année 2018. Les premiers variants de SARS-CoV-2 présentés dans cet arbre entre les lignées Pango A et B sont estimés a être apparus entre novembre 2020 et mars 2021, alors que leur échantillonnage a été effectué entre janvier et mars 2021. De plus, cette proximité entre les hypothèses temporelles des noeuds internes et les dates d'échantillonnages demeure acceptable à travers l'entièreté des séquences représentées dans cet arbre.

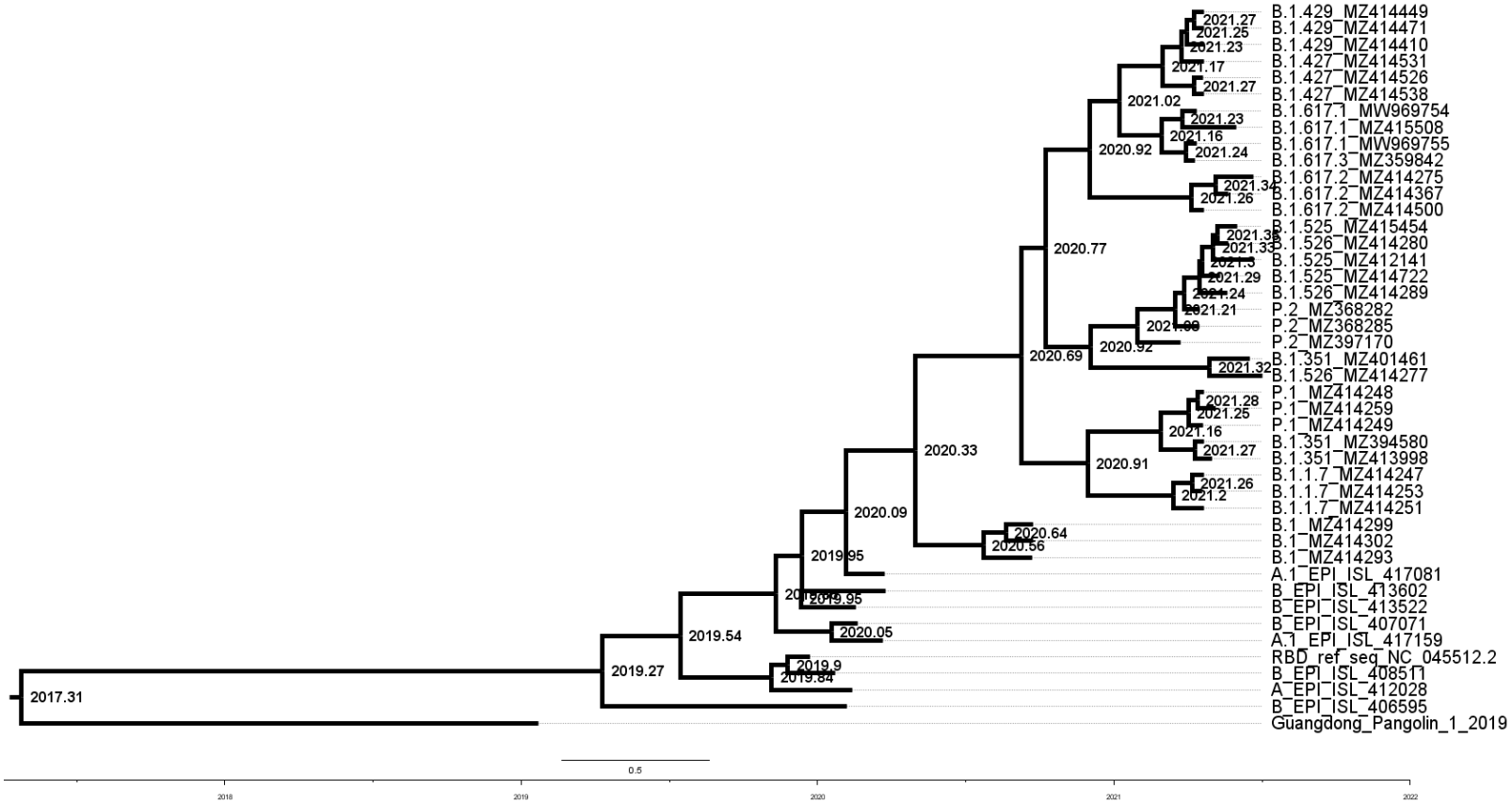


Figure 6.11: Reconstruction phylogénétique par Beast2 des séquences de variants de SARS-CoV-2 et de pangolin de Guangdong.

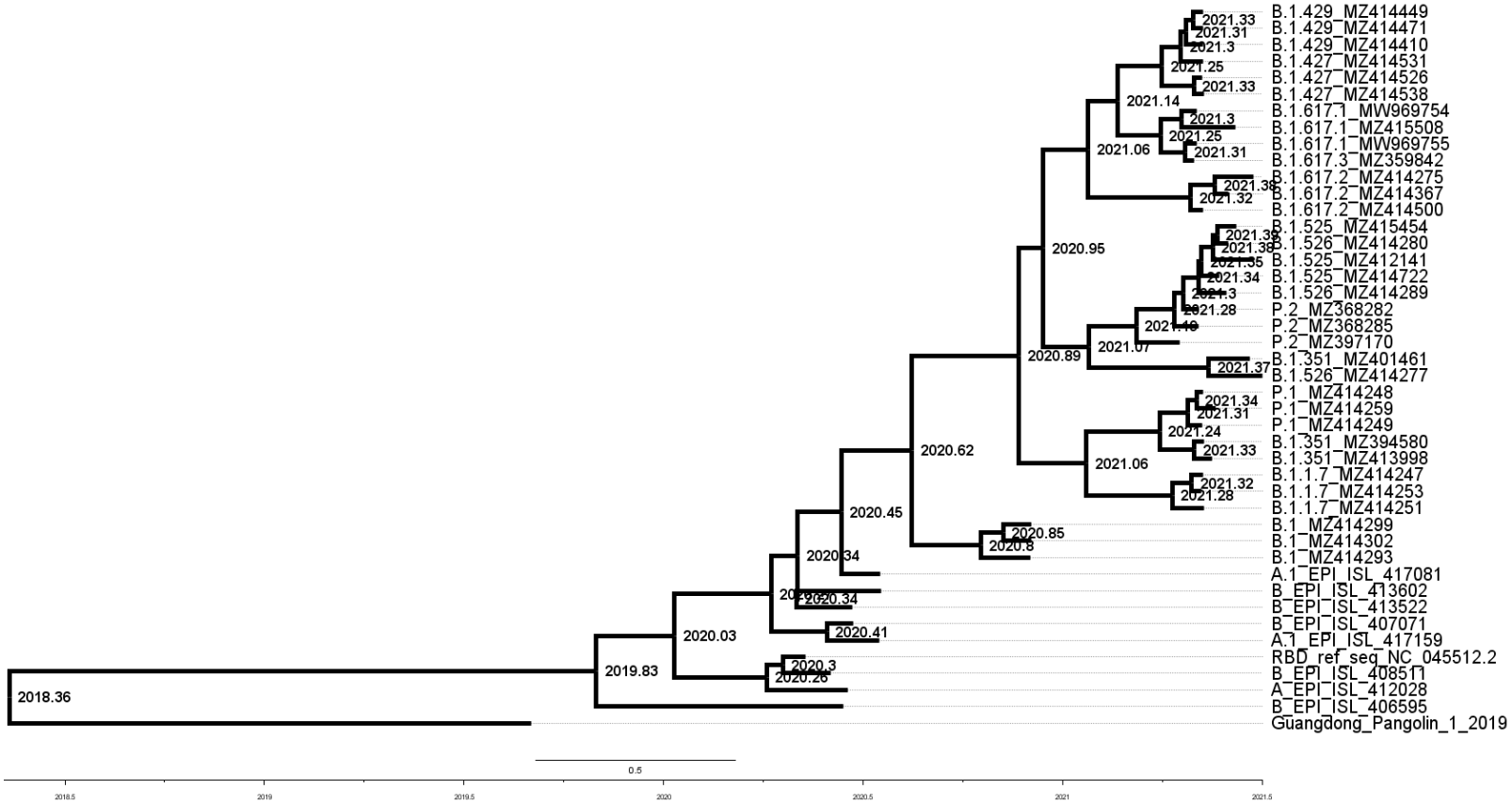


Figure 6.12: Reconstruction phylogénétique par Beast2 des séquences de variants de SARS-CoV-2 et de pangolin de Guangdong. Un facteur temporel de 0.75 a été ajouté pour les âges des noeuds internes.

CHAPITRE VII

DISCUSSION

Ce chapitre est axé sur la discussion des résultats présentés au chapitre précédent, ainsi que les situer dans un contexte biologique. Leur signification ainsi que leur concordance avec la littérature scientifique sera également abordée.

De plus, les perspectives futures d'étude de l'évolution de SARS-CoV-2 sont également discutés, ainsi que les fonctionnalités additionnelles qui pourraient être introduite au logiciel SimPlot++.

7.1 Discussion des résultats

Les analyses effectuées avec le logiciel SimPlot++ afin de détecter les évènements de recombinaisons concordent avec la théorie présenté au chapitre 2. En effet, les analyses SimPlot (figure 6.1) et Bootscan (figure 6.2) ont permis d'identifier une région potentiellement mosaïque entre les positions 1100 et 1600 du gène S. Cette région de gène est connue comme la région génique codant pour le domaine *receptor-binding* impliqué dans les mécanismes d'attachement à l'ACE2 et d'entrée cellulaire.

Selon l'hypothèse présentée à la section 2.3.3 comme quoi SARS-CoV-2 serait le résultat du phénomène d'évolution divergente à partir d'un génome proche de RaTG13 (Makarenkov *et al.*, 2021), cette région recombinée représenterait le transfert de cette région codante à partir d'une séquence de CoV provenant de pangolins malais dans un génome évolutivement proche du CoV de chauve-souris RaTG13.

Cette détection a d'autant plus été renforcée par l'identification d'une densité élevée de sites informatifs dans cette région entre les séquences de SARS-CoV-2 et du CoV de pangolin de Guangdong (figure 6.3). Ainsi, une grande proportion de sites informatifs regroupés dans une région génique est un signal suggérant qu'un évènement de recombinaison s'y soit produit (Robertson *et al.*, 1995).

De plus les tests de proportions ainsi que le réseau de similarité de séquences générés ont tous deux permis d'identifier une région recombinante dans la région du gène S correspondant au domaine RB.

Les analyses subséquentes basées sur les jeux de données de séquences protéiques représentant les domaines RB des variants de SARS-CoV-2 ont permis de visualiser les régions de ce domaine qui sont hautement conservées, ainsi que les régions

propices à l'apparition de mutations ponctuelles.

Les analyses de reconstructions phylogénétiques avec le logiciel Beast2 ont permis d'établir une hypothèse d'histoire évolutive crédible pour le domaine RB de SARS-CoV-2. Cette chronologie présente un transfert horizontal entre les séquences ancestrales à SARS-CoV-2 autour du début de l'année 2018. Celle-ci suggère également que le début de l'apparition de variants de SARS-CoV-2 a coïncidé avec l'augmentation exponentielle de la population virale dans la population humaine, soit dès la fin de 2019 et le début de 2021.

7.2 Perspectives d'améliorations

Afin d'améliorer ces résultats, il serait intéressant d'appliquer la méthodologie utilisée lors des analyses effectuées sur le gène S et le domaine RB et de les reproduire sur les autres gènes et régions géniques d'intérêts présent dans le génome de SARS-CoV-2.

Par exemple, plusieurs variants de SARS-CoV-2 présentant des taux de mortalités élevés présentent 11 régions nucléotidiques fortement conservées à travers les gènes codant pour 4 protéines spécifiques (1ab, S, M et N) (Gussow *et al.*, 2020). Assumant que ces régions aient un impact sur des mécanismes liés à la virulence, étudier leur origine évolutive pourrait approfondir les connaissances quant à l'origine de ce coronavirus.

7.3 Travaux futur sur SimPlot++

Plusieurs aspects du logiciel SimPlot++ pourraient être travaillées d'avantage afin d'accroître le nombre de fonctionnalités et de faciliter l'exploration des résultats. Ceux-ci sont regroupées ci-dessous selon la page du logiciel impliqué.

La création des groupes peut être une étape relativement lente et répétitive selon le jeu de données employé. Cela pourrait être accélérée dans une version future par une combinaison d'outils complémentaires telle la possibilité d'appliquer à un jeu de données les groupes originant d'un fichier Nexus déjà existant où ces mêmes groupes ont déjà été formé ou, de suggérer à l'utilisateur des groupes selon soit le nom des séquences ou un algorithme rapide basé sur la distance entre les séquences et un nombre de groupes à former préétabli.

L'analyse SimPlot pourrait être améliorée par la possibilité de superposer sur le graphique de courbes les résultats d'une analyse de détection de recombinaison statistique au choix tels un test Phi par fenêtre d'analyse ou un test de proportion, afin de mettre l'évidence sur les régions potentiellement mosaïques. Diverses fonctionnalités interactives pourraient également être ajoutés telle la capacité à l'utilisateur de retirer certaines courbes du graphique ou de produire un réseau de similarité directement à partir de la page SimPlot.

La page de réseau de similarité pourrait également être améliorée par l'intégration d'un algorithme de Louvain afin de pouvoir identifier des communautés présentes dans le réseau de similarité. L'intégration de graphes bipartis (tels que présenté à la figure 3.9d) comme possibilité de présentations des résultats pourrait également être explorée.

Finalement, la possibilité de produire des graphiques tels que présentés à la figure 6.3 directement dans le logiciel pourrait être bénéfique lors de l'utilisation de l'analyse FindSite, puisque les résultats sont présentés uniquement de façon textuelle dans la version courante.

CONCLUSION

À travers ce projet, il a été question de produire une application permettant de remplacer le logiciel SimPlot par l'amélioration des approches analytiques offertes par celui-ci ainsi que d'introduire de nouvelles méthodes de détection d'évènements de recombinaisons. Il était également question d'employer ce nouveau logiciel SimPlot++ afin d'étudier l'histoire évolutive du gène S et du domaine RB de SARS-CoV-2.

À travers le présent document, les améliorations apportées aux analyses SimPlot, Bootscan et Findsites ont été présentées en détails. Les nouvelles méthodes statistiques de détection de recombinaisons issue du logiciel PhiPack et développés spécifiquement pour l'analyse SimPlot (test de proportions) ont été expliqués. Une vue d'ensemble des caractéristiques de plusieurs des composantes analytiques du programme SimPlot++ qui sont des ré-implémentations de méthodes déjà existantes est présentée à la table 15.1 de (Lemey *et al.*, 2009). Il s'agit d'une ressource fort utile afin d'approfondir des concepts n'ayant pas pu être présentés dans le présent mémoire.

De plus, les différentes approches analytiques offertes par SimPlot++ ont été employées afin d'analyser deux jeux de données représentant divers génomes de coronavirus présentant des similarités avec le SARS-CoV-2, ainsi qu'un grand nombre de génomes de variants représentant les lignées Pango les plus proéminentes dans la population humaine.

Ces analyses ont permis de confirmer la présence d'une région recombinante impliquant le CoV de pangolin de Guangdong dans la région du gène S de SARS-CoV-2,

ainsi que de produire des arbres phylogénétiques par Beast2 présentant une chronologie évolutive possible de SARS-CoV-2 selon l'hypothèse que la similarité entre SARS-CoV-2, RaTG13 et le CoV de pangolin de Guangdong serait le résultat d'un ou plusieurs évènements de recombinaisons.

Ainsi, les trois objectifs présentés initialement dans le mémoire ont été accomplis. Des travaux futurs liés à SimPlot++ ainsi que l'étude de l'évolution de SARS-CoV-2 pourrait être effectués par l'apport de nouvelles fonctionnalités plus poussées concernant les réseaux de similarités de séquences. Il pourrait être fort intéressant de développer une version hybride de ces réseaux qui combinerait l'approche heuristique propre à l'analyse SimPlot à l'approche statistique offerte par les méthodes de PhiPack (test Phi, test max-chi et NSS), de façon à représenter ces résultats statistiques par le poids des arrêtes entre les noeuds (séquences). SimPlot++ pourrait également être employé dans le futur afin d'étudier l'histoire évolutive des autres régions d'intérêts de SARS-CoV-2, tels le gène M de la membrane.

Le code de SimPlot++ est stocké sur GitHub [1].

1. github.com/Stephane-S/Simplot_PlusPlus

ANNEXE A

Regroupement	Nom de séquence	Accession	Hôte
SARS-COV-2	BetaCoV/Wuhan-Hu-1/2019	NC_045512.2	Humain
	BetaCoV/Wuhan/IVDC-HB-05/2019	EPI_ISL_402121	Humain
	BetaCoV/Wuhan/IVDC-HB-04/2020	EPI_ISL_402120	Humain
Bat CoV (RatG13	Bat coronavirus RaTG13	MN996532.1	Chauve-souris
Guangdong Pangolin CoV	hCoV-19/pangolin/Guangdong/1/2019	EPI_ISL_410721	pangolin
	hCoV-19/pangolin/Guangdong/P2S/2019	EPI_ISL_410544	pangolin
Guangxi Pangolin CoV	PCoV_GX-P5E	MT040336	pangolin
	PCoV_GX-P2V	MT072864	pangolin
	PCoV_GX-P5L	MT040335	pangolin
	PCoV_GX-P1E	MT040334.1	pangolin
	PCoV_GX-P3B	MT072865	pangolin
	PCoV_GX-P4L	MT040333	pangolin
Bat CoV (ZXC21, ZC45)	bat-SL-CoVZC45	MG772933.1	Chauve-souris
	bat-SL-CoVZXC21	MG772934.1	Chauve-souris
Bat SL CoV (BTCoV, RS3367)	BtCoV/273/2005	DQ648856.1	Chauve-souris
	BtCoV/279/2005	DQ648857.1	Chauve-souris
	Bat SARS-like coronavirus Rs3367	KC881006.1	Chauve-souris
Bat SL CoV (HKU3. RF1)	Bat SARS coronavirus Rf1	DQ412042.1	Chauve-souris
	Bat SARS coronavirus HKU3-12	GQ153547.1	Chauve-souris
	Bat SARS coronavirus HKU3-6	GQ153541.1	Chauve-souris
Bat CoV from kenya/BGR	SARS-related coronavirus BtKY72	KY352407.1	Chauve-souris
	Bat coronavirus BM48-31/BGR/2008	GU190215.1	Chauve-souris
SARS CoV	Tor2	NC_004718.3	Humain
	PC4_13	AY613948.1	Civet

Tableau A.1: Tableau des regroupements de séquences de coronavirus lors des analyses avec SimPlot++. Les numéros d'accession ainsi que les hôtes respectifs sont également présentés.

ANNEXE B

Lignée Pango	Appellation	Accession
A	-	EPI_ISL_412028
A.1	-	EPI_ISL_417159
		EPI_ISL_417081
B	-	EPI_ISL_408511
		EPI_ISL_406595
		EPI_ISL_413522
		EPI_ISL_407071
B.1	-	MZ414293
		MZ414299
		MZ414302
		EPI_ISL_413602
B.1.1.7	Alpha	MZ414247
		MZ414251
		MZ414253
B.1.351	Beta	MZ413998
		MZ401461
		MZ394580
B.1.427	Epsilon	MZ414526
		MZ414531
		MZ414538
B.1.429	Epsilon	MZ414410
		MZ414449
		MZ414471
B.1.525	Eta	MZ414722
		MZ415454
		MZ412141
B.1.526	Iota	MZ414277
		MZ414280
		MZ414289
B.1.617.1	Kappa	MZ415508
		MW969754
		MW969755
B.1.617.2	Delta	MZ414275
		MZ414367
		MZ414500
B.1.617.3	-	MZ359842
P.1	Gamma	MZ414248
		MZ414249
		MZ414259
P.2	Zeta	MZ397170
		MZ368282
		MZ368285

Tableau B.1: Tableau des lignées Pango, appellations et numéros d'accessions des 42 séquences de variants.

RÉFÉRENCES

Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W. et Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST : A new generation of protein database search programs. *Nucleic Acids Research*, *25*(17), 3389–3402. http://dx.doi.org/10.1093/nar/25.17.3389

Atkinson, H. J., Morris, J. H., Ferrin, T. E. et Babbitt, P. C. (2009). Using Sequence Similarity Networks for Visualization of Relationships Across Diverse Protein Superfamilies. *PLoS ONE*, *4*(2), e4345. http://dx.doi.org/10.1371/journal.pone.0004345

Barido-Sottani, J., Bošková, V., Plessis, L. D., Kühnert, D., Magnus, C., Mitov, V., Müller, N. F., PečErska, J., Rasmussen, D. A., Zhang, C., Drummond, A. J., Heath, T. A., Pybus, O. G., Vaughan, T. G. et Stadler, T. (2018). Taming the BEAST—A Community Teaching Material Resource for BEAST 2. *Systematic Biology*, *67*(1), 170–174. http://dx.doi.org/10.1093/sysbio/syx060

Benson, D. A., Cavanaugh, M., Clark, K., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J. et Sayers, E. W. (2013). GenBank. *Nucleic Acids Research*, *41*(Database issue), D36–42. http://dx.doi.org/10.1093/nar/gks1195

Boc, A. et Makarenkov, V. (2011). Towards an accurate identification of mosaic genes and partial horizontal gene transfers. *Nucleic Acids Research*, *39*(21), e144–e144. http://dx.doi.org/10.1093/nar/gkr735

Bokeh Development Team (2018). *Bokeh : Python Library for Interactive Visualization*.

Boni, M. F., Lemey, P., Jiang, X., Lam, T. T.-Y., Perry, B. W., Castoe, T. A., Rambaut, A. et 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. http://dx.doi.org/10.1038/s41564-020-0771-4

Brant, A. C., Tian, W., Majerciak, V., Yang, W. et Zheng, Z.-M. (2021). SARS-CoV-2 : From its discovery to genome structure, transcription, and

replication. *Cell & Bioscience*, *11*(1), 136. http://dx.doi.org/10.1186/s13578-021-00643-z

Bruen, T. C., Philippe, H. et Bryant, D. (2006). A Simple and Robust Statistical Test for Detecting the Presence of Recombination. *Genetics*, *172*(4), 2665–2681. http://dx.doi.org/10.1534/genetics.105.048975

CDC (2020). Coronavirus Disease 2019 (COVID-19). https ://www.cdc.gov/coronavirus/2019-ncov/variants/variant-info.html.

Chan, J. F.-W., Kok, K.-H., Zhu, Z., Chu, H., To, K. K.-W., Yuan, S. et Yuen, K.-Y. (2020). Genomic characterization of the 2019 novel human-pathogenic coronavirus isolated from a patient with atypical pneumonia after visiting Wuhan. *Emerging Microbes & Infections*, *9*(1), 221–236. http://dx.doi.org/10.1080/22221751.2020.1719902

Claas, E. C., Osterhaus, A. D., van Beek, R., De Jong, J. C., Rimmelzwaan, G. F., Senne, D. A., Krauss, S., Shortridge, K. F. et Webster, R. G. (1998). Human influenza A H5N1 virus related to a highly pathogenic avian influenza virus. *The Lancet*, *351*(9101), 472–477. http://dx.doi.org/10.1016/S0140-6736(97)11212-0

Cock, P. J. A., Antao, T., Chang, J. T., Chapman, B. A., Cox, C. J., Dalke, A., Friedberg, I., Hamelryck, T., Kauff, F., Wilczynski, B. et de Hoon, M. J. L. (2009). Biopython : Freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics*, *25*(11), 1422–1423. http://dx.doi.org/10.1093/bioinformatics/btp163

Corel, E., Lopez, P., Méheust, R. et Bapteste, E. (2016). Network-Thinking : Graphs to Analyze Microbial Complexity and Evolution. *Trends in Microbiology*, *24*(3), 224–237. http://dx.doi.org/10.1016/j.tim.2015.12.003

Dong, E., Du, H. et Gardner, L. (2020). An interactive web-based dashboard to track COVID-19 in real time. *The Lancet Infectious Diseases*, *20*(5), 533–534. http://dx.doi.org/10.1016/S1473-3099(20)30120-1

Drummond, A. J. et Rambaut, A. (2007). BEAST : Bayesian evolutionary analysis by sampling trees. *BMC Evolutionary Biology*, *7*(1), 214. http://dx.doi.org/10.1186/1471-2148-7-214

Egan, A. et Crandall, K. (2006). Theory of Phylogenetic Estimation. In *Evolutionary Genetics : Concepts and Case Studies* 426–443. Oxford University Press.

Elbe, S. et Buckland-Merrett, G. (2017). Data, disease and diplomacy : GISAID's innovative contribution to global health. *Global challenges (Hoboken, NJ)*, *1*(1), 33–46. http://dx.doi.org/10.1002/gch2.1018

Etherington, G. J., Dicks, J. et Roberts, I. N. (2005). Recombination Analysis Tool (RAT) : A program for the high-throughput detection of recombination. *Bioinformatics*, *21*(3), 278–281. http://dx.doi.org/10.1093/bioinformatics/bth500

Felsenstein, J. (1981). Evolutionary trees from DNA sequences : A maximum likelihood approach. *Journal of Molecular Evolution*, *17*(6), 368–376. http://dx.doi.org/10.1007/BF01734359

Felsenstein, J. (2005). PHYLIP (phylogeny inference package) version 3.6. Department of Genome Sciences, University of Washington, Seattle.

Fleischmann, W. R. (1996). Viral Genetics. In S. Baron (dir.), *Medical Microbiology*. Galveston (TX) : University of Texas Medical Branch at Galveston, (fourth éd.).

Guo, Y.-R., Cao, Q.-D., Hong, Z.-S., Tan, Y.-Y., Chen, S.-D., Jin, H.-J., Tan, K.-S., Wang, D.-Y. et 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), 11. http://dx.doi.org/10.1186/s40779-020-00240-0

Gussow, A. B., Auslander, N., Faure, G., Wolf, Y. I., Zhang, F. et Koonin, E. V. (2020). Genomic determinants of pathogenicity in SARS-CoV-2 and other human coronaviruses. *Proceedings of the National Academy of Sciences*, *117*(26), 15193–15199. http://dx.doi.org/10.1073/pnas.2008176117

Halary, S., McInerney, J. O., Lopez, P. et Bapteste, E. (2013). EGN : A wizard for construction of gene and genome similarity networks. *BMC Evolutionary Biology*, *13*(1), 146. http://dx.doi.org/10.1186/1471-2148-13-146

Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N. J., Kern, R., Picus, M., Hoyer, S., van Kerkwijk, M. H., Brett, M., Haldane, A., del Río, J. F., Wiebe, M., Peterson, P., Gérard-Marchant, P., Sheppard, K., Reddy, T., Weckesser, W., Abbasi, H., Gohlke, C. et Oliphant, T. E. (2020). Array programming with NumPy. *Nature*, *585*(7825), 357–362. http://dx.doi.org/10.1038/s41586-020-2649-2

Hasegawa, M., Kishino, H. et Yano, T. (1985). Dating of the human-ape

splitting by a molecular clock of mitochondrial DNA. *Journal of Molecular Evolution*, *22*(2), 160–174. http://dx.doi.org/10.1007/BF02101694

Hu, B., Guo, H., Zhou, P. et Shi, Z.-L. (2021). Characteristics of SARS-CoV-2 and COVID-19. *Nature Reviews Microbiology*, *19*(3), 141–154. http://dx.doi.org/10.1038/s41579-020-00459-7

Hunter, J. D. (2007). Matplotlib : A 2D Graphics Environment. *Computing in Science Engineering*, *9*(3), 90–95. http://dx.doi.org/10.1109/MCSE.2007.55

Jakobsen, I. B. et Easteal, S. (1996). A program for calculating and displaying compatibility matrices as an aid in determining reticulate evolution in molecular sequences. *Bioinformatics*, *12*(4), 291–295. http://dx.doi.org/10.1093/bioinformatics/12.4.291

Jukes, T. H. et Cantor, C. R. (1969). Evolution of Protein Molecules. In *Mammalian Protein Metabolism* 21–132. Elsevier

Kimura, M. (1980). A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *Journal of Molecular Evolution*, *16*(2), 111–120. http://dx.doi.org/10.1007/BF01731581

King, R. C., Stansfield, W. D. et Mulligan, P. K. (2007). *A Dictionary of Genetics*. Oxford University Press.

Knight, R., Maxwell, P., Birmingham, A., Carnes, J., Caporaso, J. G., Easton, B. C., Eaton, M., Hamady, M., Lindsay, H., Liu, Z., Lozupone, C., McDonald, D., Robeson, M., Sammut, R., Smit, S., Wakefield, M. J., Widmann, J., Wikman, S., Wilson, S., Ying, H. et Huttley, G. A. (2007). PyCogent : A toolkit for making sense from sequence. *Genome Biology*, *8*(8), R171. http://dx.doi.org/10.1186/gb-2007-8-8-r171

Konagurthu, A. S., Whisstock, J. C., Stuckey, P. J. et Lesk, A. M. (2006). MUSTANG : A multiple structural alignment algorithm. *Proteins*, *64*(3), 559–574. http://dx.doi.org/10.1002/prot.20921

Kumar, S., Stecher, G., Li, M., Knyaz, C. et Tamura, K. (2018). MEGA X : Molecular Evolutionary Genetics Analysis across Computing Platforms. *Molecular Biology and Evolution*, *35*(6), 1547–1549. http://dx.doi.org/10.1093/molbev/msy096

Lam, S. K., Pitrou, A. et Seibert, S. (2015). Numba : A LLVM-based Python JIT compiler. Dans *Proceedings of the Second Workshop on the LLVM*

Compiler Infrastructure in HPC - LLVM '15, 1–6., Austin, Texas. ACM Press. http://dx.doi.org/10.1145/2833157.2833162

Lam, T. T.-Y., Jia, N., Zhang, Y.-W., Shum, M. H.-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, W. Y.-M., Li, L.-F., Sun, F.-F., Qin, S., Huang, J.-C., Leung, G. M., Holmes, E. C., Hu, Y.-L., Guan, Y. et Cao, W.-C. (2020). Identifying SARS-CoV-2-related coronaviruses in Malayan pangolins. *Nature*, *583*(7815), 282–285. http://dx.doi.org/10.1038/s41586-020-2169-0

Lau, S. K., Luk, H. K., Wong, A. C., Li, K. S., Zhu, L., He, Z., Fung, J., Chan, T. T., Fung, K. S. et Woo, P. C. (2020). Possible Bat Origin of Severe Acute Respiratory Syndrome Coronavirus 2. *Emerging Infectious Diseases*, *26*(7), 1542–1547. http://dx.doi.org/10.3201/eid2607.200092

Lemey, P., Salemi, M. et Vandamme, A.-M. (dir.) (2009). *The Phylogenetic Handbook : A Practical Approach to Phylogenetic Analysis and Hypothesis Testing* (second éd.). Cambridge : Cambridge University Press. http://dx.doi.org/10.1017/CBO9780511819049

Lodish, H., Berk, A., Zipursky, S. L., Matsudaira, P., Baltimore, D. et Darnell, J. (2000). Mutations : Types and Causes. *Molecular Cell Biology. 4th edition.*

Lole, K. S., Bollinger, R. C., Paranjape, R. S., Gadkari, D., Kulkarni, S. S., Novak, N. G., Ingersoll, R., Sheppard, H. W. et Ray, S. C. (1999). Full-length human immunodeficiency virus type 1 genomes from subtype C-infected seroconverters in India, with evidence of intersubtype recombination. *Journal of Virology*, *73*(1), 152–160. http://dx.doi.org/10.1128/JVI.73.1.152-160.1999

Madeira, F., Park, Y. M., Lee, J., Buso, N., Gur, T., Madhusoodanan, N., Basutkar, P., Tivey, A. R. N., Potter, S. C., Finn, R. D. et Lopez, R. (2019). The EMBL-EBI search and sequence analysis tools APIs in 2019. *Nucleic acids research*, *47*(W1), W636–W641. http://dx.doi.org/10.1093/nar/gkz268

Makarenkov, V., Mazoure, B., Rabusseau, G. et 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), 5. http://dx.doi.org/10.1186/s12862-020-01732-2

Martin, A. J. M., Walsh, I., Domenico, T. D., Mičetić, I. et Tosatto, S. C. E. (2013). PANADA : Protein Association Network Annotation, Determination

and Analysis. *PLOS ONE*, *8*(11), e78383.
http://dx.doi.org/10.1371/journal.pone.0078383

Martin, D. et Rybicki, E. (2000). RDP : Detection of recombination amongst aligned sequences. *Bioinformatics*, *16*(6), 562–563.
http://dx.doi.org/10.1093/bioinformatics/16.6.562

Martin, D. P., Murrell, B., Golden, M., Khoosal, A. et Muhire, B. (2015). RDP4 : Detection and analysis of recombination patterns in virus genomes. *Virus Evolution*, *1*(1). http://dx.doi.org/10.1093/ve/vev003

Milner, D. S., Attah, V., Cook, E., Maguire, F., Savory, F. R., Morrison, M., Müller, C. A., Foster, P. G., Talbot, N. J., Leonard, G. et Richards, T. A. (2019). Environment-dependent fitness gains can be driven by horizontal gene transfer of transporter-encoding genes. *Proceedings of the National Academy of Sciences*, *116*(12), 5613–5622.
http://dx.doi.org/10.1073/pnas.1815994116

Muslin, C., Mac Kain, A., Bessaud, M., Blondel, B. et Delpeyroux, F. (2019). Recombination in Enteroviruses, a Multi-Step Modular Evolutionary Process. *Viruses*, *11*(9), 859. http://dx.doi.org/10.3390/v11090859

Nei, M. (1987). *Chapter 9 : Genetic Distance Between Populations*. Columbia University Press.

Ou, J., Zhou, Z., Dai, R., Zhang, J., Zhao, S., Wu, X., Lan, W., Ren, Y., Cui, L., Lan, Q., Lu, L., Seto, D., Chodosh, J., Wu, J., Zhang, G. et Zhang, Q. (2021). V367F Mutation in SARS-CoV-2 Spike RBD Emerging during the Early Transmission Phase Enhances Viral Infectivity through Increased Human ACE2 Receptor Binding Affinity. *J Virol*, *95*(16), e0061721.
http://dx.doi.org/10.1128/JVI.00617-21

Pérez-Losada, M., Arenas, M., Galán, J. C., Palero, F. et González-Candelas, F. (2015). Recombination in viruses : Mechanisms, methods of study, and evolutionary consequences. *Infection, Genetics and Evolution*, *30*, 296–307.
http://dx.doi.org/10.1016/j.meegid.2014.12.022

Posada, D. et Crandall, K. A. (2001). Selecting the Best-Fit Model of Nucleotide Substitution. *S YSTEMATIC BIOLOGY*, *50*, 22.

Rambaut, A. (2006). FigTree. http ://tree.bio.ed.ac.uk/software/figtree/.

Rambaut, A., Holmes, E. C., O'Toole, Á., Hill, V., McCrone, J. T., Ruis, C., du Plessis, L. et Pybus, O. G. (2020). A dynamic nomenclature proposal for SARS-CoV-2 lineages to assist genomic epidemiology. *Nature Microbiology*,

5(11), 1403–1407. http://dx.doi.org/10.1038/s41564-020-0770-5

Rehman, Shafique, L., Ihsan, A. et Liu, Q. (2020). Evolutionary Trajectory for the Emergence of Novel Coronavirus SARS-CoV-2. *Pathogens*, *9*(3), 240. http://dx.doi.org/10.3390/pathogens9030240

Robertson, D. L., Hahn, B. H. et Sharp, P. M. (1995). Recombination in AIDS viruses. *Journal of Molecular Evolution*, *40*(3), 249–259. http://dx.doi.org/10.1007/BF00163230

Rota, P. A., Oberste, M. S., Monroe, S. S., Nix, W. A., Campagnoli, R., Icenogle, J. P., Peñaranda, S., Bankamp, B., Maher, K., Chen, M.-h., Tong, S., Tamin, A., Lowe, L., Frace, M., DeRisi, J. L., Chen, Q., Wang, D., Erdman, D. D., Peret, T. C. T., Burns, C., Ksiazek, T. G., Rollin, P. E., Sanchez, A., Liffick, S., Holloway, B., Limor, J., McCaustland, K., Olsen-Rasmussen, M., Fouchier, R., Günther, S., Osterhaus, A. D. M. E., Drosten, C., Pallansch, M. A., Anderson, L. J. et Bellini, W. J. (2003). Characterization of a Novel Coronavirus Associated with Severe Acute Respiratory Syndrome. *Science*, *300*(5624), 1394–1399. http://dx.doi.org/10.1126/science.1085952

Saitou, N. et Nei, M. (1987). The neighbor-joining method : A new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution*, *4*(4), 406–425. http://dx.doi.org/10.1093/oxfordjournals.molbev.a040454

Salminen, M. O., Carr, J. K., Burke, D. S. et McCUTCHAN, F. E. (1995). Identification of Breakpoints in Intergenotypic Recombinants of HIV Type 1 by Bootscanning. *AIDS Research and Human Retroviruses*, *11*(11), 1423–1425. http://dx.doi.org/10.1089/aid.1995.11.1423

Shang, J., Wan, Y., Luo, C., Ye, G., Geng, Q., Auerbach, A. et Li, F. (2020). Cell entry mechanisms of SARS-CoV-2. *Proceedings of the National Academy of Sciences*, *117*(21), 11727–11734. http://dx.doi.org/10.1073/pnas.2003138117

Sheahan, T., Rockx, B., Donaldson, E., Sims, A., Pickles, R., Corti, D. et Baric, R. (2008). Mechanisms of Zoonotic Severe Acute Respiratory Syndrome Coronavirus Host Range Expansion in Human Airway Epithelium. *Journal of Virology*, *82*(5), 2274–2285. http://dx.doi.org/10.1128/JVI.02041-07

Singh, D. et Yi, S. V. (2021). On the origin and evolution of SARS-CoV-2. *Experimental & Molecular Medicine*, *53*(4), 537–547. http://dx.doi.org/10.1038/s12276-021-00604-z

Smith, J. (1992). Analyzing the mosaic structure of genes. *Journal of Molecular Evolution*, *34*(2). http://dx.doi.org/10.1007/BF00182389

Sokal, R. et Michener, C. (1958). A statistical method for evaluating systematic relationships. *undefined*.

Song, H.-D., Tu, C.-C., Zhang, G.-W., Wang, S.-Y., Zheng, K., Lei, L.-C., Chen, Q.-X., Gao, Y.-W., Zhou, H.-Q., Xiang, H., Zheng, H.-J., Chern, S.-W. W., Cheng, F., Pan, C.-M., Xuan, H., Chen, S.-J., Luo, H.-M., Zhou, D.-H., Liu, Y.-F., He, J.-F., Qin, P.-Z., Li, L.-H., Ren, Y.-Q., Liang, W.-J., Yu, Y.-D., Anderson, L., Wang, M., Xu, R.-H., Wu, X.-W., Zheng, H.-Y., Chen, J.-D., Liang, G., Gao, Y., Liao, M., Fang, L., Jiang, L.-Y., Li, H., Chen, F., Di, B., He, L.-J., Lin, J.-Y., Tong, S., Kong, X., Du, L., Hao, P., Tang, H., Bernini, A., Yu, X.-J., Spiga, O., Guo, Z.-M., Pan, H.-Y., He, W.-Z., Manuguerra, J.-C., Fontanet, A., Danchin, A., Niccolai, N., Li, Y.-X., Wu, C.-I. et Zhao, G.-P. (2005). Cross-host evolution of severe acute respiratory syndrome coronavirus in palm civet and human. *Proceedings of the National Academy of Sciences*, *102*(7), 2430–2435. http://dx.doi.org/10.1073/pnas.0409608102

Stuart Barker, J. S. (2009). Defining Fitness in Natural and Domesticated Populations. In J. van der Werf, H.-U. Graser, R. Frankham, et C. Gondro (dir.), *Adaptation and Fitness in Animal Populations : Evolutionary and Breeding Perspectives on Genetic Resource Management* 3–14. Dordrecht : Springer Netherlands

Talavera, G. et Castresana, J. (2007). Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments. *Systematic Biology*, *56*(4), 564–577. http://dx.doi.org/10.1080/10635150701472164

Tavare, S. (1986). Some probabilistic and statistical problems in the analysis of DNA sequences. *Some mathematical questions in biology / DNA sequence analysis edited by Robert M. Miura*.

Torres-Montúfar, A., Borsch, T. et Ochoterena, H. (2018). When Homoplasy Is Not Homoplasy : Dissecting Trait Evolution by Contrasting Composite and Reductive Coding. *Systematic Biology*, *67*(3), 543–551. http://dx.doi.org/10.1093/sysbio/syx053

van de Guchte, M. (2017). Horizontal Gene Transfer and Ecosystem Function Dynamics. *Trends in Microbiology*, *25*(9), 699–700. http://dx.doi.org/10.1016/j.tim.2017.07.002

Wu, Z., Du, J., Zhang, T., Xue, Y., Yang, F. et Jin, Q. (2013). Recombinant

Human Coxsackievirus B3 from Children with Acute Myocarditis in China. *Journal of Clinical Microbiology*, *51*(9), 3083–3086. http://dx.doi.org/10.1128/JCM.00270-13

Xia, S., Lan, Q., Su, S., Wang, X., Xu, W., Liu, Z., Zhu, Y., Wang, Q., Lu, L. et Jiang, S. (2020). The role of furin cleavage site in SARS-CoV-2 spike protein-mediated membrane fusion in the presence or absence of trypsin. *Signal Transduction and Targeted Therapy*, *5*(1), 1–3. http://dx.doi.org/10.1038/s41392-020-0184-0

Zhang, T., Wu, Q. et Zhang, Z. (2020). Probable Pangolin Origin of SARS-CoV-2 Associated with the COVID-19 Outbreak. *Current Biology*, *30*(7), 1346–1351.e2. http://dx.doi.org/10.1016/j.cub.2020.03.022

Zhang, Y. et Skolnick, J. (2005). TM-align : A protein structure alignment algorithm based on the TM-score. *Nucleic Acids Research*, *33*(7), 2302–2309. http://dx.doi.org/10.1093/nar/gki524