Approches bioinformatiques pour la détection de l'évolution convergente dans les génomes viraux, application à la résistance aux traitements.

Marie Morel

Université Paris Cité

Institut Pasteur

Université Paris Cité

Frontières de l'Innovation en Recherche et Education - 474

Institut Pasteur – Bioinformatique Evolutive

Approches Bioinformatiques pour la Détection de l'Evolution Convergente dans les Génomes Viraux

Application à la résistance aux traitements

Par Marie MOREL

Thèse de doctorat de Génétique, Omiques, Bioinformatique et Biologie des Systèmes

Dirigée par Didier Mazel
Et par Etienne Simon-Lorière

Présentée et soutenue publiquement le 11.10.2022

Devant un jury composé de :

Didier MAZEL, Professeur, Institut Pasteur, Directeur de thèse
Etienne SIMON-LORIERE, Chargé de recherche, Institut Pasteur, Co-Directeur de thèse
Maria ANISIMOVA, Professeure, ZHAW, Rapportrice
Marie SÉMON, Maîtresse de Conférence, ENS Lyon, Rapportrice
Bastien BOUSSAU, Chargé de recherche au CNRS, Université Lyon 1, Examinateur
Pierre-Emmanuel CECCALDI, Professeur, Institut Pasteur, Examinateur
Céline SCORNAVACCA, Directrice de recherche au CNRS, Université de Montpellier, Examinatrice
Frédéric LEMOINE, Ingénieur de recherche, Institut Pasteur, Membre invité

RESUME

Approches bioinformatiques pour la détection de l'évolution convergente dans les génomes viraux. Application à la résistance aux traitements.

Mots clés : Convergence évolutive, Evolution moléculaire, Phylogénétique, Sélection, Adaptation, VIH, VHC, Mutations de résistance, Variants mineurs.

Les génomes des virus à ARN présentent un taux d'évolution parmi les plus élevés, ce qui leur permet de s'adapter rapidement en cas de changement d'environnement. Au sein de leurs hôtes, ces virus existent généralement sous la forme d'une population de mutants dont les génomes sont distincts, bien qu'apparentés. Cette caractéristique présente un avantage évolutif important car parmi la multitude de variants, il y a de plus grande chance que certains soient adaptés à un nouvel environnement. Par exemple, les virus à ARN sont connus pour être capables de changer d'hôte, d'échapper à la reconnaissance du système immunitaire ou de résister aux traitements antiviraux. Ainsi, les virus à ARN sont surreprésentés parmi les agents pathogènes émergents et ré-émergents (anciennement connus mais dont l'incidence augmente de manière inattendue), et leur évolution rapide contribue probablement à leur risque d'émergence élevé, ce qui constitue un défi majeur pour leur contrôle.

La résistance aux traitements antiviraux peut s'avérer un véritable problème dans le contrôle de certains virus pour lesquels il n'existe pas de vaccin. D'une part, certains traitements peuvent devenir inefficaces, ce qui se caractérise par la réplication du virus et pouvant entrainer le décès de l'hôte. D'autre part, des patients traités peuvent avoir une charge virale suffisante pour transmettre des virus déjà résistants, réduisant ainsi l'éventail des thérapies envisageables en première intention.

Durant cette thèse j'ai exploré sous deux angles différents la question de l'adaptation des virus aux traitements antiviraux : 1) l'identification de convergence moléculaire (émergence répétée et indépendante d'une même mutation) dans de grands jeux de données de séquences (plusieurs milliers de séquences), et 2) l'étude des variations génomiques au sein de populations virales soumises à un traitement antiviral. Pour cela, j'ai développé une méthode de détection de convergence évolutive dans les séquences protéiques en suivant l'hypothèse que la présence de mutations de convergence était un indicateur des pressions de sélection sur les séquences virales. J'ai également étudié l'évolution de la diversité virale intra-hôte lorsque les virus sont soumis à un traitement antiviral. Les organismes étudiés pour ce projet sont le virus de l'immunodéficience humaine (VIH) et le virus de l'hépatite C (VHC), tous deux responsables d'affections longue durée, traitées par traitements antiviraux et responsables de millions d'infections à travers le monde.

ABSTRACT

Bioinformatics approaches for detecting convergent evolution in viral genomes. Application to treatment resistance.

Key words: Evolutionary convergence, Molecular evolution, Phylogenetics, Selection, Adaptation, HIV, HCV, Resistance mutations, Minor variants.

The genomes of RNA viruses have some of the highest rates of evolution, allowing them to adapt rapidly to changing environments. Within their hosts, these viruses generally exist as a population of mutants with distinct, but related genomes. This is an important evolutionary advantage because, among the multitude of variants, there is a higher chance that some will be adapted to a new environment. For example, RNA viruses are known to be able to switch hosts, evade recognition by the immune system or resist antiviral treatments. Thus, RNA viruses are over-represented among emerging and re-emerging pathogens (previously known but unexpectedly increasing in incidence), and their rapid evolution probably contributes to their high risk of emergence, representing a major challenge for their control.

Resistance to antiviral treatments can be a real problem in the control of some viruses for which there is no vaccine. On the one hand, some treatments can become ineffective, which is characterised by the replication of the virus and can lead to the death of the host. On the other hand, treated patients may have a sufficient viral load to transmit already resistant viruses, thus reducing the range of possible first-line therapies.

During this thesis, I explored the question of viral adaptation to antiviral treatments from two different angles: 1) the identification of molecular convergence (repeated and independent emergence of the same mutation) in large sequence datasets (several thousand sequences), and 2) the study of genomic variations within viral populations subjected to antiviral treatment. For this purpose, I developed a method for detecting evolutionary convergence in protein sequences under the assumption that the presence of convergent mutations was an indicator of selection pressures on viral sequences. I also studied the evolution of intra-host viral diversity when viruses are subjected to antiviral treatment. The organisms studied for this project are the human immunodeficiency virus (HIV) and the hepatitis C virus (HCV), both of which are responsible for long-term illnesses, treated with antiviral therapy, and responsible for millions of infections worldwide.

REMERCIEMENTS

SOMMAIRE

Résumé i
Abstract ii
Remerciements iii
Liste des figures vii
Liste des tableaux ix
Liste des annexes ix
1 Introduction Générale 1
1.1 Généralités sur les virus 1
1.1.1 Définition et historique 1
1.1.2 Evolution et propagation des virus 5
1.2 Outils et méthodes bioinformatiques pour l'étude de l'évolution moléculaire 12
1.2.1 Production et analyse des séquences virales 12
1.2.2 Analyse comparative des génomes viraux 16
1.2.3 Modèles d'évolution 17
1.2.4 Inférence phylogénétique 21
1.2.5 Reconstruction de caractères ancestraux 26
1.2.6 Sélection positive ou négative 27
1.3 VIH, VHC et mutations de résistance 30
1.3.1 Le VIH 30
1.3.2 Le VHC 38
2 Etude de la convergence évolutive chez les virus 42
2.1 La convergence évolutive : un signe d'adaptation ? 42
2.1.1 Quand l'évolution se répète 42
2.1.2 Les bases génétiques des convergences phénotypiques. 43
2.1.3 Comment expliquer que l'évolution se répète ? 45
2.1.4 Convergences moléculaires de premier plan et d'arrière-plan 49
2.2 La convergence moléculaire est un phénomène répandu chez les virus 50
2.2.1 La nature des virus facilite l'émergence de convergences moléculaires 50
2.2.2 Convergences observées en conditions expérimentales 50
2.2.3 Convergences observées en milieu naturel 51
2.3 Quelles méthodes permettent d'étudier la convergence au niveau moléculaire 55
2.3.1 Recherche manuelle des mutations 55
2.3.2 Association phénotype-génotype 56
2.3.3 Sélection positive directionnelle 56

2.3.4 Topologies et autres signaux phylogénétiques 57
2.3.5 Détection de mutations spécifiques aux espèces cibles 59
2.3.6 Nécessité d'une nouvelle méthode 61
3 Développement d'une méthode pour détecter la convergence moléculaire 63
3.1 Une approche par simulation et corrélation 63
3.1.1 Des simulations pour estimer le nombre d'émergence d'une mutation 64
3.1.2 Une mesure de corrélation pour estimer si une mutation émerge indépendamment du phénotype convergent ou non. 66
3.1.3 Implémentation dans le logiciel ConDor 66
3.2 Article : Accurate detection of Convergent Mutations in Protein Alignments with ConDor 66
3.3 Analyses complémentaires 92
3.3.1 Performances sur simulations 92
3.3.2 Sensibilité de la composante émergence 96
3.3.3 Sélection positive 100
3.4 Limites : étude du SARS-CoV-2 102
3.5 Conclusions 104
3.6 Perspectives 105
4 Etude de la résistance aux traitements chez le VHC 106
4.1 Présentation des données et du problème 106
4.2 Article : Genomic variations associated with drug resistance in HCV genotype 6 infected patients failing DAA-based therapy. 107
4.3 Conclusions et Perspectives 120
5 Conclusion Générale 121
6 Annexes 123
Matériel supplémentaire de l'article "Genomic variations associated with drug resistance in HCV genotype 6 infected patients failing DAA-based therapy". 123
Bibliographie 130
Table des Matières 165

LISTE DES FIGURES

Figure 1 : taux de mutations par site en fonction de la taille du génome chez les virus et d'autres entités biologiques. 6
Figure 2 : Composition du génome d'un recombinant CRF02_AG du VIH-1. 8
Figure 3 : Etapes de réassortiment du virus de la grippe jusqu'à obtenir la souche H1N1 épidémique. 9
Figure 4 : Exemple de visualisation d'un mapping de lectures sur un génome de référence. 14
Figure 5 : Alignement en acides aminés. 16
Figure 6 : Mutations cachées à un site pendant un temps t. 17
Figure 7 : Transitions et transversions entre les nucléotides. 19
Figure 8 : Représentation d'un arbre phylogénétique binaire et enraciné. 22
Figure 9: Principe du bootstrap de Felsenstein par rééchantillonnage des sites d'un alignement. 26
Figure 10 : Phylogénie des différents clades du VIH-1 et de plusieurs SIV. 31
Figure 11 : Répartition dans le monde des principaux sous-types et formes recombinantes du VIH-1. 33
Figure 12 : Illustration d'un virion du VIH-1 avec les différentes protéines le composant. 34
Figure 13 : Composition du génome du VIH-1 d'après le génome de référence HXB2. 34
Figure 14 : Cycle de réplication du VIH-1 montrant les mécanismes d'action des différentes classes d'antiviraux. 35
Figure 15 : Représentation simplifiée du génome du VHC et des 10 protéines exprimées. 40
Figure 16: Exemples de morphologies convergentes : ailes d'insecte, de chauve-souris et d'oiseau. 43
Figure 17 : Différents types de mutations de convergence moléculaire 44
Figure 18 : Arbre phylogénétique de la PEPC chez les graminées avec les positions sous sélection positive en lien avec le métabolisme en C4. 45
Figure 19 : Fréquences d'équilibre des acides aminés selon différents modèle d'évolution des protéines. 47
Figure 20 : Matrice de transition BLOSUM62 représentée sous la forme de « heatmap ». 48
Figure 21 : Schéma illustrant l'influence de la pléiotropie sur la sélection de mutations. 48
Figure 22: Représentation de convergence moléculaire pouvant être observée à partir de données expérimentales. 55
Figure 23: Phylogénie d'espèce et phylogénies alternatives proposées par Parker et al pour tester la présence de convergence dans les gènes d'animaux écholocateurs. 58
Figure 24: Principe de détection des mutations convergentes spécifiques aux espèces cibles. 60
Figure 25: Probabilités associées aux états des nœuds enfants N_1 et N_2, lors de l'évolution de l'état au nœud parent N_0. 64
Figure 26: Comparaison entre la distribution du nombre attendu d'émergences sur 10'000 simulations et le nombre observé à la position 123 d'un alignement VIH-1 de la reverse transcriptase. 65
Figure 27: distributions bivariées par paires de plusieurs statistiques pour les DRMs détectées et les mutations convergentes candidates sur le jeu de données vraies du VIH-1. 97
Figure 28 : distributions bivariées par paires de plusieurs statistiques pour les DRMs détectées et non détectées sur le jeu de données vraies du VIH-1. 98
Figure 29: Mutations détectées par la composante émergence, triées par nombre d'émergence. 99

Figure 30: Comparaison de la fréquence des nucléotides dans les vraies données et dans les simulations après différents temps d'évolution. 103
Figure 31 : Provenance des différents échantillons plasmatiques des patients infectées avec le génotype 6 du VHC. 106

LISTE DES TABLEAUX

Tableau 1 : Classification de Baltimore des virus 4
Tableau 2 : Matrices de taux pour différents modèles markoviens d'évolution de séquences ADN. 20
Tableau 3 : Exemples de convergences moléculaires. Les positions suivies d'une étoile indiquent des confirmations expérimentales de l'effet des mutations sur le phénotype. 44
Tableau 4: Performances de PC, FADE, ConDor et ses sous composantes sur données synthétiques. 93
Tableau 5: résultats de FEL, FUBAR et MEME sur la détection des DRMs dans le jeu de données réelles du VIH-1. 100

LISTE DES ANNEXES

Table S1: characteristics of viral genomes from patients with TF. Caractéristiques des génomes viraux provenant de patients ayant connu un échec du traitement.

Table S2: characteristics of viral genomes from patients successfully treated. Caractéristiques des génomes viraux provenant de patients traités avec succès.

Table S3: Mutations found by Geno2Pheno on samples HCV24(B) and HCV03(B). Mutations trouvées par Geno2Pheno sur les échantillons HCV24(B) et HCV03(B)

Table S4: Polymorphisms found at baseline on resistance-associated positions in NS5A for successfully treated patients or patients with TF. Polymorphismes trouvés avant traitement sur les positions associées à de la résistance dans la NS5A pour les patients traités avec succès ou les patients en échec thérapeutique.

Table S5: Polymorphisms found at baseline on resistance-associated positions in NS5B for successfully treated patients or patients with TF. Polymorphismes trouvés avant traitement sur les positions associées à de la résistance dans la NS5B pour les patients traités avec succès ou les patients en échec thérapeutique.

Figure S1: Heatmap of the pairwise distance between newly sequenced genomes from patients infected with HCV GT6 and annotated subtypes from GT6. Matrice de la distance par paire entre les génomes nouvellement séquencés de patients infectés par le VHC GT6 et les sous-types annotés du GT6.

Figure S2: Diversity index by samples at pre-treatment and post-treatment. Indice de diversité des échantillons avant et après traitement.

1 INTRODUCTION GÉNÉRALE

Au cours de cette thèse, je me suis intéressée aux mutations de résistance chez les virus, et en particulier aux moyens de les détecter. Les mutations de résistance surviennent lorsque les virus sont soumis à un ou plusieurs traitements antiviraux, et se traduisent par des changements dans la séquence du génome virale. Ces mutations permettent aux virus d'échapper au traitement, c'est-à-dire de continuer à se répliquer malgré la présence du traitement. Cette forme d'adaptation est permise par un certain nombre de propriétés (populations de grande taille, taux de réplication élevé, temps de génération court et taux de mutation élevé) inhérentes aux virus, des entités biologiques simples mais capable d'évoluer et de s'adapter rapidement.

Au cours de ce chapitre je définirai l'objet d'étude de ces travaux, les virus, en insistant dans un premier temps sur les différents mécanismes qui permettent leur évolution rapide et leur propagation à grande échelle. Ensuite je décrirai différentes approches computationnelles qui permettent de quantifier et de détecter les signes d'adaptation chez les virus. Enfin, je m'intéresserai de manière plus précise aux deux principaux organismes étudiés dans cette thèse : le virus de l'immunodéficience humaine (VIH), et le virus de l'hépatite C (VHC). Nous verrons que ces deux virus arborent des mutations de résistance au traitement antiviraux pouvant menacer les opérations de santé publique visant à leur contrôle.

1.1 GENERALITES SUR LES VIRUS

1.1.1 Définition et historique

Définir les virus de manière simple et générale n'est pas un exercice évident tant la découverte de nouveaux virus nous amène à reconsidérer régulièrement leurs caractéristiques. Pourtant nous avons tous une intuition de ce qu'est un virus et d'autant plus ces deux dernières années en raison du contexte sanitaire. Le plus simplement, nous pouvons définir les virus comme des agents infectieux de petite taille (microscopique) et parasites obligatoires dans la mesure où ils ont besoin de la machinerie d'une cellule hôte infectée pour se répliquer.

1.1.1.1 Découverte des virus

Les manifestations liées aux virus étaient connues bien avant que ces derniers ne soient identifiés et il a fallu attendre la fin du dix-neuvième siècle pour que l'on appréhende les premières caractéristiques biologiques des virus. En effet, le vaccin contre la rage a été mis au point par Louis Pasteur et ses collaborateurs en 1885 avant même que les chercheurs considèrent l'existence d'agents infectieux différents des « microbes » qui avaient été découverts grâce à la microscopie optique. A cette époque-là, le terme virus existait pourtant déjà mais désignait n'importe quel pathogène (bactérie, parasite, virus).

En 1892, le chercheur russe Dmitri Ivanovsky (Ivanowsky 1892), décrit la présence d'un agent infectieux contenu dans la sève des plants de tabac atteints de la mosaïque du tabac. Cet agent infectieux n'est pas retenu par les filtres de Chamberland alors utilisés pour filtrer les bactéries et autres microbes. Pour autant, Dmitri Ivanovsky ne conclut pas à la découverte d'un type d'agent infectieux différent des bactéries. Il pense qu'il peut s'agir d'un très petit microbe ou d'une toxine, d'un poison sécrété par la plante. En 1898, le hollandais Martinus Beijerinck (Beijerinck 1898)

démontre que cet agent infectieux n'est pas de nature bactérienne ni une toxine mais ce qu'il appelle un *Contagium vivum fluidum* (« germe vivant soluble »). Parallèlement, en Allemagne, Friedrich Loeffler et Paul Frosch (Loeffler et Frosch 1898) montrent l'origine virale de la fièvre aphteuse en 1898, là encore un germe capable de passer au travers des filtres bactériens. On notera d'ailleurs, non sans une certaine ironie, que Louis Pasteur décédé en 1895, ne connaitra jamais la découverte des virus en tant que tels.

Le début du vingtième siècle voit ainsi l'essor de la découverte de nombreux virus, dont la caractéristique principale est alors leur filtrabilité, c'est-à-dire leur capacité à passer au travers d'un filtre de Chamberland, mais on ne connait alors ni leur taille, ni leur forme. Le développement de nouvelles technologies a permis l'essor de la recherche en virologie. En 1935, Wendell Stanley (Stanley 1935) réussi à purifier et cristalliser le virus de la mosaïque du tabac ce qui permettra par la suite de l'étudier par diffraction aux rayons X (Bawden et al. 1936) et de déterminer qu'il est composé de protéines et d'acides nucléiques. De même, le développement de la microscopie électronique permet d'observer les premiers virus (Kausche, Pfankuch, et Ruska 1939; Nagler et Rake 1948; Van Rooyen et Scott 1948; Reagan et Brueckner 1952) et de comprendre leur morphologie et donc de commencer à les classifier (Almeida 1963; Almeida et Waterson 1970). La virologie est devenue en quelques dizaines d'années une discipline à part entière au sein de la microbiologie.

1.1.1.2 Définition des virus

Il faudra attendre 1957, pour qu'André Lwoff (1957) propose la première véritable définition des virus.

> *Les virus pourraient donc être définis comme : des entités strictement intracellulaires et potentiellement pathogènes possédant une phase infectieuse, et (1) ne possédant qu'un seul type d'acide nucléique, (2) se multipliant sous la forme de leur matériel génétique, (3) incapables de croître et de subir une fission binaire, (4) dépourvus de système de Lipmann [incapables de produire de l'ATP].*

Pendant longtemps la définition de Lwoff a peu évolué et l'on définit communément les virus comme des agents infectieux de très petite taille (de l'ordre de la dizaine voire centaine de nanomètres), dont l'information génétique se présente sous la forme d'un seul type d'acide nucléique (ADN ou ARN) contenu dans une capside protéique et qui ne peuvent se répliquer qu'en parasitant une cellule.

Cette définition porte toutefois à débat, notamment depuis la découverte en 2003, d'un virus géant infectant des amibes. Il s'agit d'Acanthamoeba polyphaga Mimivirus (La Scola et al. 2003) qui présente une taille supérieure à 400nm (soit deux fois la taille maximale alors communément admise pour un virus (Lwoff 1957; Sandaa et al. 2001)) et un génome de plus de 1.2 millions de paires de bases codant 911 protéines. La notion de virus géants ou « girus » apparait et d'autres virus géants sont découverts peu après (Boyer et al. 2009 ; Philippe et al. 2013). Outre la taille de ces virus c'est aussi la présence de gènes codant pour des fonctions cellulaires (production de protéines, transport de molécules, réparation de l'ADN) (Raoult et al. 2004; Arslan et al. 2011) qui interroge et relance le débat sur les questions « les virus sont-ils vivants ? » et « les virus sont-ils des organismes ?». Ces questions ne sont pas récentes puisqu'elles figuraient déjà dans l'article de Lwoff en 1957 (Lwoff 1957). Les virus n'étant pas capable de se répliquer seuls et d'assurer des

fonctions métaboliques, ils peuvent finalement être considérés comme des particules inertes en dehors de leur hôte. De la même manière, les virus ne sont pas représentés sur l'arbre du vivant divisé en 3 domaines (procaryotes, eucaryotes, archées) quoique Didier Raoult et Patrick Forterre proposent, en 2008 (Raoult et Forterre 2008), un nouvel arbre prenant en compte les virus, ainsi qu'une nouvelle définition des virus :

> *« Virus : organisme codant pour une capside, composé de protéines et d'acides nucléiques, qui s'auto-assemble en une nucléocapside et utilise un organisme codant pour des ribosomes pour l'achèvement de son cycle de vie. »*

Cet article recevra une réponse de la part de Roland Wolkowicz et Moselio Schaechte (2008). Selon eux, la définition de Raoult et Forterre omet la "caractéristique principale de ce qui fait un virus un virus" : il se désagrège et perd son intégrité physique, sa progéniture se reconstituant après réplication à partir des parties nouvellement synthétisées. C'est donc la caractéristique de désintégration et de reconstitution, uniquement vraie pour les virus qui serait essentielle à leur définition.

Depuis, d'autres découvertes ont poussé les chercheurs à revoir la manière de définir un virus. Sans en faire une liste exhaustive, on peut citer l'existence de virus qui infectent d'autres virus (La Scola et al. 2008), les Pandoravirus et leurs 2500 gènes dont la plupart sont inconnus (Philippe et al. 2013), de virus sans capside (Kanhayuwa et al. 2015)… Finalement, notre vision des virus est encore récente et notre compréhension de ces entités évoluera sans doute avec les futures découvertes et progrès. Nous avons d'ailleurs un aperçu très limité de l'ensemble des virus, avec un peu plus de 10'000 génomes de référence recensés à ce jour (Lefkowitz et al. 2018), alors qu'on estime le nombre de particules virales sur terre à 10^{31} (Hendrix et al. 1999; Mushegian 2020), le nombre de virus différents chez les mammifères à un minimum de $3x10^5$ (Anthony et al. 2013) et que 10^5 nouveaux virus à ARN viennent d'être découverts (Edgar et al. 2022).

1.1.1.3 Classification des virus

Afin de mieux étudier les virus et étant donné leur grand nombre et leur très grande diversité, des efforts ont été fait pour les classer. Leur organisation n'est pas évidente car contrairement aux organismes du domaine du vivant, une origine unique avec un même ancêtre commun des virus n'est pas vérifiée. Une classification phylogénétique de tous les virus, basée sur des similarités de séquences homologues n'est donc pas possible.

Une première classification est tout d'abord proposée pour classer les virus selon leur morphologie (Almeida 1963) mais elle est très vite remplacée par la classification de Baltimore (Tableau 1) qui permet de grouper les virus selon l'organisation de leur génome et leur mode de réplication. En effet, les virus infectent une cellule hôte pour se multiplier et des mécanismes différents existent pour parvenir à cette fin. La classification de Baltimore est subdivisée en 7 groupes selon si le génome viral est constitué d'ARN ou d'ADN, simple ou double brin et à sens positif ou négatif.

ADN/ARN	Groupe	Définition	Enzyme pour la réplication	Exemples de virus
Virus à ADN	Groupe I	Virus à ADN double brin	Polymérase de la cellule hôte	Virus de la variole
	Groupe II	Virus à ADN simple brin	Polymérase de la cellule hôte	
Virus à ARN	Groupe III	Virus à ARN double brin	RdRp	Rotavirus
	Groupe IV	Virus à ARN simple brin à polarité positive	RdRp	Dengue, SARS-CoV-2, Virus Hépatite C
	Groupe V	Virus à ARN simple brin à polarité négative	RdRp	Virus de la rage, virus de la grippe
Rétrovirus	Groupe VI	Rétrovirus à ARN simple brin	Reverse transcriptase	VIH, HTLV
	Groupe VII	Rétrovirus à ADN double brin		VHB

Tableau 1 : Classification de Baltimore des virus

Outre la classification de Baltimore, les virus sont organisés par l'ICTV[1] en différents ordres, familles, sous-familles, genres et espèces. Une même espèce de virus est définie selon un groupe monophylétique de virus qui partagent des propriétés distinctes des autres espèces de virus selon plusieurs critères. Ces critères peuvent être la structure de la capside, l'existence ou non d'une enveloppe, les types de protéines exprimées, les hôtes infectés, le pouvoir pathogène et surtout la similarité des séquences génétiques. La classification actualisée de 2021 organise ainsi un peu plus de 10'000 espèces de virus.

Au-delà de l'espèce virale, la classification et la nomenclature des virus n'est pas universellement définie, et chaque famille de virus a ses propres spécificités. Par exemple le VIH-1 (une des deux espèces de VIH) est divisé en groupes M, N, O et P, le groupe M étant ensuite subdivisé en neuf sous-types ou clades (A-D, F-H, J et K) (Robertson et al. 2000). De même, le virus de l'hépatite C est divisé en 9 génotypes et 80 sous-types (D. B. Smith et al. 2014; Charlotte Hedskog et al. 2019). Ces subdivisions servent à marquer la variabilité génétique et la divergence d'une espèce virale par l'acquisition de mutations par exemple. De la même manière on peut lire dans (Kuhn et al. 2013) qu'il « n'existe pas de définition universellement acceptée des termes "souche", "variant" et "isolat" dans la communauté scientifique ». Toutefois il est d'usage de désigner une souche virale comme un variant viral génétiquement stable qui diffère du virus naturel de référence en ce qu'elle provoque un phénotype d'infection significativement différent et observable (type de maladie différent, infection d'un type d'hôte différent, transmission par des moyens différents, etc.) (Kuhn et al. 2013). Un variant désignant alors un virus légèrement différent du virus de référence avec quelques mutations de différence par exemple.

De même qu'il existe des systèmes de classification différents pour des virus différents, on peut trouver des classifications différentes au sein d'un même virus. Le SARS-CoV-2 en est un bon exemple avec plusieurs systèmes de nomenclature dont aucun n'est pour l'instant

[1] International committee of taxonomy of viruses.

universellement accepté. D'ailleurs l'OMS a annoncé au 31 mai 2021 nommer les variants du SARS-CoV-2 par une lettre grecque[2] sans que ce système ne remplace les noms scientifiques utilisés. Parmi les différents systèmes on peut citer le système « d'année et lettre » utilisé par Nexstrain (Hadfield et al. 2018), les clades GISAID définis par des mutations essentielles ou encore le système dynamique des lignées Pango qui se concentre sur les variants circulant activement (Rambaut et al. 2020). Ces différentes notations permettent d'ailleurs de ne pas nommer les virus et leurs lignages à partir d'indications géographiques ce qui peut être source de discriminations[3]. Etant donné la spécificité de la pandémie de SARS-CoV-2 et le suivi des variants en temps réel, les différents variants sont nommés avant de savoir s'ils vont se répandre au sein de la population. Afin de savoir quels variants sont les plus à risque il existe également une classification en variant préoccupant (VOC), variant à suivre (VOI) et variant sous surveillance (VUM)[4].

Il est fort probable que la classification des virus soit encore amenée à évoluer au gré de la recherche. D'ailleurs des liens phylogénétiques commencent à être mis en évidence entre différents virus à ARN. On suppose qu'ils proviennent d'un même ancêtre commun duquel ils auraient hérité l'enzyme nécessaire à leur réplication, l'ARN polymérase ARN dépendante (RdRp pour *RNA dependent RNA polymerase*) (Wolf et al. 2018; Kuhn et al. 2019). Même si les analyses pour parvenir à cette conclusion sont controversées (Edward C. Holmes et Duchêne 2019), les virus ARN appartiennent désormais à un même règne « *Riboviria* », comme groupe monophylétique probable au sein de l'ICTV (Walker et al. 2019). Pourtant les virus ARN recoupent 3 classes différentes de la classification de Baltimore, et de nouvelles formes de classification commencent à être suggérées pour résoudre ces incohérences (Koonin et al. 2020).

1.1.2 Evolution et propagation des virus

Si les virus sont aussi difficiles à classifier, c'est d'une part parce que nous ne connaissons qu'une petite partie d'entre eux, mais aussi parce qu'il est difficile de comparer les séquences virales entre elles en raison de leur grande diversité. En effet, les virus évoluent et accumulent au cours de leur cycle de réplication des changements au niveau de leur séquence ADN ou ARN. Après un certain temps d'évolution (disons des millions d'années), si suffisamment de changements ont eu lieu entre deux séquences, nous risquons de ne plus être capables de révéler leur similitude.

Ces modifications de leur génome jouent un rôle important dans l'évolution des virus car elles peuvent impacter la structure des protéines encodées et ainsi modifier la transmissibilité des virus, leur pathogénicité, la reconnaissance par les anticorps, etc. (Fitch et al. 1991; Fleischmann 1996; Leslie et al. 2004; Timm et al. 2004; Tsetsarkin et al. 2007; Parvez and Parveen 2017). Au cours du temps, différents variants peuvent évoluer à partir d'une même souche virale. Ils se différencient alors par les changements au sein de leur génome. Favorisés par les pressions de sélection, les variants les plus adaptés à un environnement donné deviennent majoritaires par l'effet de la sélection naturelle. On a pu observer ce phénomène en temps réel avec les variants Delta ou Omicron du SARS-CoV-2 qui présentaient des taux de transmission plus élevés que les

[2] https://www.who.int/news/item/31-05-2021-who-announces-simple-easy-to-say-labels-for-sars-cov-2-variants-of-interest-and-concern
[3] https://asm.org/Articles/2021/May/Why-Scientists-Should-Not-Name-Diseases-After-Plac
[4] https://www.who.int/fr/activities/tracking-SARS-CoV-2-variants

autres variants et qui sont rapidement devenus majoritaires dans de nombreux pays (Elliott et al. 2021; Karim and Karim 2021; Tegally et al. 2021; Callaway 2022; Sun et al. 2022).

Les virus évoluent majoritairement selon trois mécanismes : les mutations, la recombinaison et le réassortiment.

1.1.2.1 Les mutations comme moteurs de l'évolution virale.

Une mutation survient au niveau du génome lorsqu'un nucléotide est remplacé par un autre, ou lors de phénomène d'insertion (une ou plusieurs bases sont ajoutées) ou de délétion (suppression d'une ou plusieurs bases). Ce phénomène survient notamment lors de l'étape de réplication lorsque la molécule d'ARN ou d'ADN est copiée dans la cellule hôte. En effet, les enzymes responsables de la réplication peuvent faire des erreurs et insérer des mutations. Il existe trois grands types d'enzymes impliquées dans ce processus chez les virus : Les polymérases classiques (ARN et ADN polymérases ADN-dépendante), les ARN polymérases ARN dépendante (RdRps) et les transcriptase inverse (RT pour *reverse transcriptase*). Ces trois types d'enzymes ne possèdent pas la même fidélité et induisent plus ou moins d'erreurs, ce qui a des répercussions sur l'évolution des génomes viraux. Ainsi, les virus à ADN qui utilisent des polymérases classiques pour se répliquer ont en moyenne des taux de mutation plus faibles que les virus à ARN ou les rétrovirus comme illustré en Figure 1 (Elena et Sanjuán 2005; Duffy, Shackelton, et Holmes 2008; Sanjuán et al. 2010). Cette meilleure fidélité s'explique notamment par la présence d'un système de relecture des polymérases classiques qui permet de corriger certaines erreurs (Garcia-Diaz et Bebenek 2007).

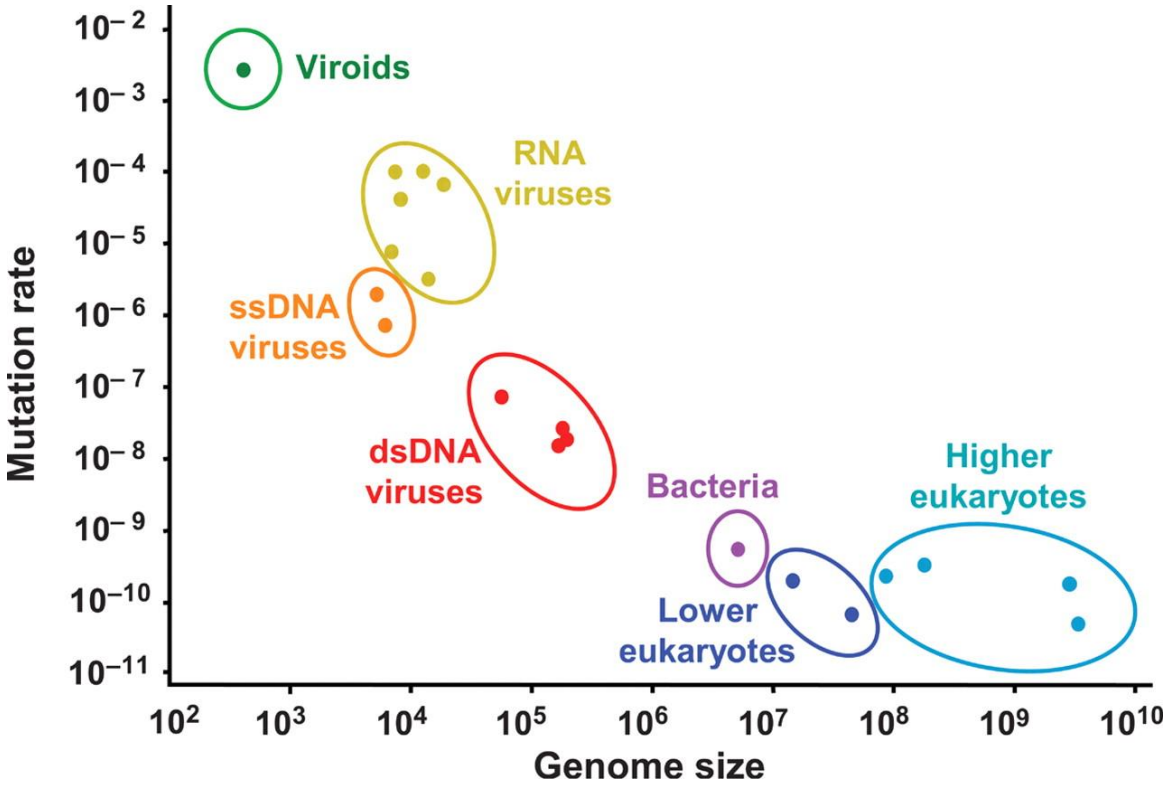


Figure 1 : taux de mutations par site en fonction de la taille du génome chez les virus et d'autres entités biologiques. Source : (Gago et al. 2009)

La RdRp ne possède pas de système de relecture et donc de rectification des erreurs. On estime que ces enzymes font environ 1 erreur toutes les 10'000 bases (Elena et Sanjuán 2005). C'est-à-

dire qu'à chaque génération, un virus ayant un génome de 10'000 bases (moyenne des virus à ARN) va accumuler au moins une mutation. Les virus ARN possèdent ainsi le taux de mutation le plus élevé parmi les virus. Les coronavirus sont une exception parmi les virus à ARN car ils sembleraient posséder un système de relecture grâce à une exonucléase et un taux de mutation plus faible que d'autres virus à ARN (Minskaia et al. 2006; Denison et al. 2011; Robson et al. 2020). Il existe d'ailleurs un lien de corrélation élevé entre la taille des génomes et le taux de mutation, comme on peut le voir sur la Figure 1. C'est sans doute grâce à ce système de relecture que les coronavirus possèdent parmi les plus longs génomes des virus à ARN avec environ 30'000 bases (Snijder et al. 2003; Gorbalenya et al. 2006).

La transcriptase inverse des rétrovirus ne possède pas non plus de système de relecture. La fidélité des RT est tout de même plus élevée que celle des RdRps, de l'ordre de 0.1-0.2 mutations par génome et par réplication dans le cas du VIH ou du virus de l'hépatite B (Drake 1993; Drake et al. 1998). Le taux de mutation du VIH varie lors de son cycle de réplication et est par exemple moins élevé lorsqu'il est intégré à l'ADN de son hôte puisqu'il utilise alors le système de réplication de l'hôte qui lui a une fonction de correction (Abram et al. 2010). En dehors des erreurs lors de la réplication, au sein de l'hôte, le virus est soumis aux pressions exercées par la réponse immunitaire de l'hôte, ce qui peut entrainer de nombreuses mutations via par exemple des mécanismes de déamination (Alizon et Fraser 2013; Cuevas et al. 2015).

Les mutations surviennent de manière aléatoire le long du génome et toutes ne vont pas avoir un impact sur le virus. Selon leur effet, on peut distinguer 3 types de mutations : les mutations létales ou délétères, les mutations neutres et les mutations avantageuses. Une mutation aura un impact sur le virus notamment si elle affecte ses capacités à se répliquer et à engendrer une descendance dans un environnement donné. Le degré d'adaptation d'un virus à son environnement est couramment défini en anglais par la 'fitness' du virus (Domingo et Holland 1997; Wargo et Kurath 2012). Nous continuerons à employer cet anglicisme dans la suite du manuscrit.

Les mutations délétères et létales disparaissent généralement rapidement car un virus possédant ce type de mutations ne va pas pouvoir se répliquer ou beaucoup moins efficacement et va donc être remplacé au sein de la population virale. Les mutations neutres ou silencieuses concernent les mutations qui ne changent pas la séquence en acides aminés de la protéine (en raison de la redondance du code génétique) ou alors qui n'induisent pas de changement important au niveau de la conformation des protéines du virus. En d'autres termes, ces mutations n'impactent pas la *fitness* du virus. Elles ne vont donc être ni sélectionnées ni remplacées mais rester au sein de la population virale. Enfin dans de rares cas, les mutations peuvent induire des changements dans la structure des protéines virales et vont se révéler avantageuses. Des mutations avantageuses vont par exemple permettre au virus de pénétrer plus efficacement dans une cellule ou d'infecter un nouveau type de cellule, échapper au système immunitaire de l'hôte, permettre un changement d'hôte, échapper aux traitements antiviraux, etc. Ces mutations sont relativement rares mais peuvent être sélectionnées dans un type d'environnement donné sous l'action de la sélection naturelle, puis disparaitre lorsque l'environnement change ou bien perdurer et donner lieu à l'émergence de nouveaux variants (Desai et Fisher 2007; Loewe et Hill 2010; Duffy 2018).

Le fort taux de mutation des virus conduit la plupart du temps à des mutations délétères et limite la taille des génomes des virus à ARN. Il semble d'ailleurs que le taux de mutation des virus à ARN soit optimal. En effet, il a été démontré in vitro qu'en augmentant artificiellement le taux de

mutation de différents virus à ARN, la population s'effondre par un processus appelé mutagenèse létale (Eigen 2002; Bull, Sanjuán, et Wilke 2007).

1.1.2.2 L'importance de la recombinaison dans l'émergence de nouvelles souches

La recombinaison constitue le deuxième mécanisme d'évolution des génomes viraux. Elle consiste en un échange de segments entiers du génome (Pérez-Losada et al. 2015). Elle peut survenir entre plusieurs virus qu'ils soient de la même espèce ou non. Des échanges de gènes peuvent également avoir lieu avec le génome de la cellule hôte (Filée, Pouget, et Chandler 2008). La recombinaison est plus fréquente chez les virus à ADN et les rétrovirus (notamment lorsqu'ils sont sous leur forme ADN) car les enzymes de l'hôte sont capables de catalyser la re-ligation (soudure) de l'ADN après la cassure et l'échange de gènes. Pour qu'il y ait recombinaison entre plusieurs virus, il faut qu'une même cellule soit infectée par plusieurs virus différents. Si les virus entrainent des infections transitoires aigües, les co-infections sont plus rares de même que les chances de recombinaison. Dans le cas où les virus recombinants sont d'espèces différentes ou de sous-types différents, ce mécanisme permet d'assimiler très rapidement beaucoup d'information génétique conduisant à un véritable « saut » évolutif et l'acquisition potentielle de nouvelles fonctions (Simon-Loriere et Holmes 2011).

Le VIH-1 se trouve souvent sous la forme de recombinants en échangeant des morceaux de son génome avec des virus d'autres sous-types du VIH-1. Ces recombinants se retrouvent majoritairement dans des zones géographiques où des virus de différents sous-types coexistent et peuvent coinfecter un même individu. La Figure 2, illustre un virus recombinant du VIH-1 issu de la recombinaison entre le sous type A1 et le sous type G. On constate qu'il existe plusieurs sites de cassure au niveau des différents gènes du VIH-1. Ce sont donc des morceaux importants de gènes qui peuvent être échangés.

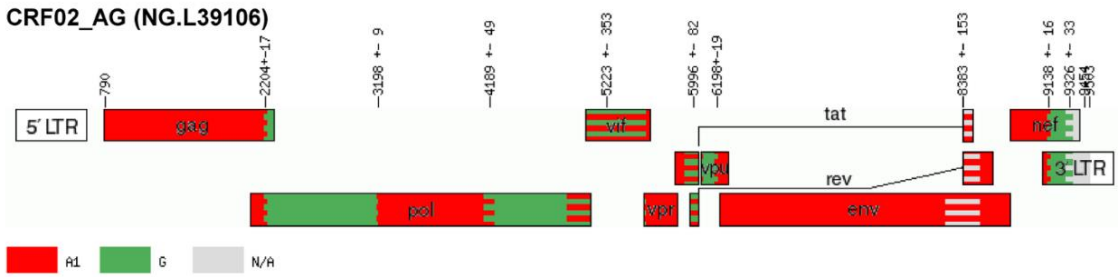


Figure 2 : Composition du génome d'un recombinant CRF02_AG du VIH-1.
Source : (Yebra et al. 2018)

Chez les virus ARN la recombinaison est beaucoup plus rare mais existe, notamment chez les Picornaviridae et les Coronaviridae (auquel appartient le SARS-CoV-2) (Simon-Loriere et Holmes 2011). D'ailleurs des formes recombinantes du SARS-CoV-2 entre les variants Delta et Omicron ont été mis en évidence récemment (Montagutelli et al. 2022).

1.1.2.3 Le réassortiment chez les virus segmentés.

Certains virus présentent des génomes segmentés, c'est-à-dire que l'information génétique n'est pas contenue sur une seule molécule d'acide nucléique mais segmentée sur plusieurs brins qui encodent des protéines différentes. Les virus à génomes segmentés peuvent échanger facilement différents brins de leur génome, sans cassure du génome viral, c'est ce qu'on appelle le

réassortiment. Si plusieurs variants d'un virus infectent une même cellule, les brins peuvent être échangés et les particules virales libérées par la cellule infectée vont alors présenter un génome mosaïque réassorti. Le virus de la grippe A, qui est composé de huit segments, est connu pour ses nombreux réassortiments notamment car des virus de différents sous-types peuvent infecter un même réservoir animal. Les porcs sont connus pour être des « points chauds » de réassortiment pour le virus de la grippe A car ils sont sensibles aux virus de la grippe aviaire, humaine et porcine et peuvent donc être infectés concomitamment par des souches de différentes espèces. Si cela arrive, il est possible d'observer des réassortiments entre ces virus et voir émerger un nouveau type de virus. C'est notamment ce qui s'est produit en 2009 lorsqu'un virus de la grippe A H1N1 contenant des gènes de virus porcins, aviaires et humains est apparu au printemps 2009 et a été transmis à la population humaine (Shinde et al. 2009; Trifonov et al. 2009). Ce phénomène est illustré Figure 3 où l'on voit que des segments provenant de 4 souches de grippe différentes ont abouti à la formation du virus de la grippe H1N1 et augmenté le potentiel épidémique de ce virus au sein de la population humaine.

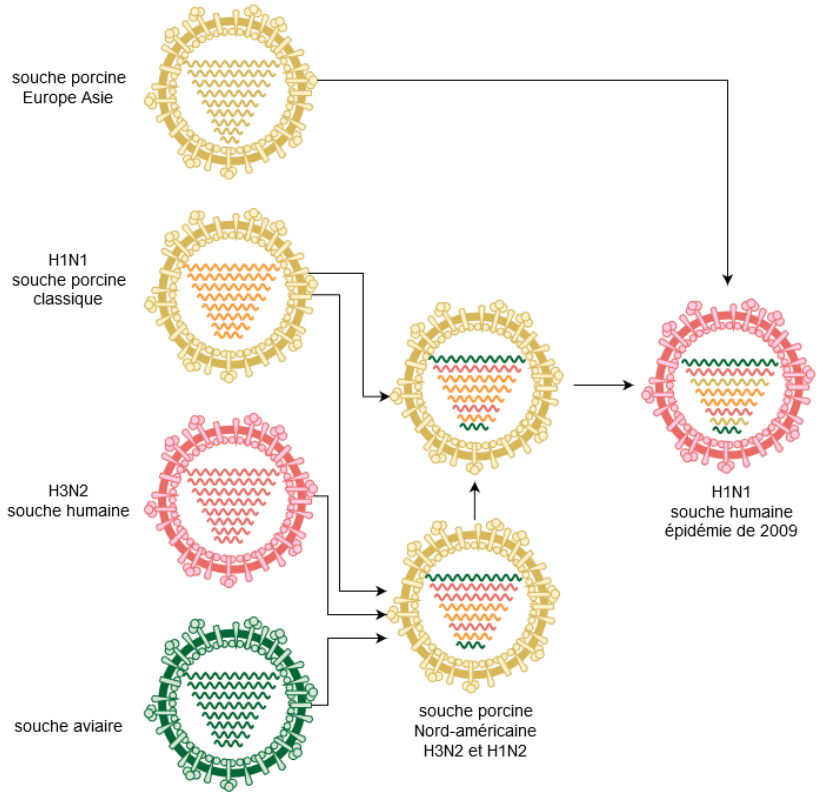


Figure 3 : Etapes de réassortiment du virus de la grippe jusqu'à obtenir la souche H1N1 épidémique. Inspirée de (Trifonov, Khiabanian, et Rabadan 2009)

1.1.2.4 La population virale

Le taux de mutation élevé observé chez certains virus, associé avec un temps de génération court (quelques heures), une taille de population importante et des taux de réplication élevés conduit ces virus à exister généralement sous la forme d'une population de mutants dont les séquences génomiques sont génétiquement apparentées mais distinctes. Ainsi, on ne peut pas correctement représenter un virus par une seule séquence nucléotidique. Au terme du séquençage la séquence qui est obtenue pour décrire le virus est une séquence consensus qui résume à chaque position le nucléotide majoritaire.

La population virale présente un avantage évolutif important car elle est plus résiliente/adaptée (*fit* en anglais) qu'un groupe génétiquement homogène. Au sein d'un hôte, la taille de la population virale peut être extrêmement grande, avec des milliers de virions produits chaque jour au sein d'un hôte (Perelson et al. 1996). Ainsi, parmi la multitude de variants présents, il y a une probabilité non nulle que certains soient adaptés à un nouvel environnement voire à un nouvel hôte.

En raison de leur capacité d'adaptation, les virus à ARN sont sur-représentés parmi les maladies émergentes et ré-émergentes. Ebola, la COVID-19, la rage, la dengue, l'hépatite C, etc. sont autant de maladies causées par des virus à ARN. On remarquera d'ailleurs que certaines de ces maladies émergent par zoonose c'est-à-dire que le virus change d'organisme hôte en « sautant » d'un animal à l'homme.

1.1.2.5 Les facteurs qui influencent l'évolution virale

Au cours de leur cycle de réplication, les virus sont soumis à toute forme de pression de sélection qui vont façonner leur évolution et favoriser l'augmentation en fréquence de certaines mutations au sein de la population virale. Au sein d'un hôte, un virus est ciblé par la défense immunitaire dans ce que l'on désigne parfois comme « une course à l'armement » dans laquelle les deux systèmes cherchent à se dépasser. L'immunité intra-hôte peut donc façonner l'évolution d'un virus en favorisant les variants moins bien reconnus par les cellules de la défense immunitaire : on parle alors d'échappement immunitaire (Lucas et al. 2001; Leslie et al. 2004; Bhattacharya et al. 2007). A l'inverse, dans le cas de personnes immunodéprimées, le système immunitaire de l'hôte va exercer peu de pression sur l'agent infectieux qui pourra alors subsister dans l'hôte et se répliquer pendant de longues périodes (Kaiser et al. 2006; Pinsky et al. 2010; Dunn et al. 2015; Choi et al. 2020). Dans certains cas, l'évolution d'un virus dans un hôte où la réponse immunitaire est amoindrie permet la multiplication rapide de ce virus augmentant les chances d'apparition de variants particulièrement adaptés à l'hôte et potentiellement hautement transmissible (Corey et al. 2021; Weigang et al. 2021).

Un grand nombre de maladies émergentes sont le résultat de zoonoses ce qui implique que le virus a pu s'adapter à un nouveau type d'hôte et est parvenu à passer la barrière inter-espèce (T. Kuiken et al. 2006). Afin de changer d'hôte, les particules virales doivent pouvoir entrer dans les cellules du nouvel hôte, recruter sa machine cellulaire, se répliquer et générer des virions capables à leur tour d'infecter des cellules et de se répliquer. Ces différentes étapes sont autant de pression de sélection qui influencent la transmissibilité d'un virus. Ces changements d'environnements entrainent également des changements dans la composition de la population virale. En effet, seul un nombre restreint de variants sont capables d'entrer dans les cellules des différents hôtes et de se répliquer chez chaque type d'hôte et la diversité virale (nombre de variants distincts au sein de la population) va alors chuter : on parle de goulot d'étranglement de la population virale (Hongye Li et Roossinck 2004; Zwart et Elena 2015). Ces goulots d'étranglement surviennent aussi dans le cas de virus qui se propagent via des vecteurs d'autres espèces (moustiques et autres insectes dans le cas de virus de plantes) (Weaver et al. 2021) ou bien lors d'infection entre différents individus d'une même espèce (Gutiérrez, Michalakis, et Blanc 2012; da Silva et al. 2017). Cette réduction importante de la diversité virale et de la taille de population rend moins efficace la sélection naturelle au profit de la dérive génétique (des concepts que nous expliquerons ultérieurement). De manière similaire, lorsqu'un virus s'adapte durablement à un nouvel hôte de

nouvelles pressions s'exercent et peuvent favoriser l'apparition de nouveaux variants (Tamuri et al. 2009; Longdon et al. 2014; Troupin et al. 2016; Longdon et al. 2018) tout en augmentant éventuellement la fitness du virus.

Les traitements antiviraux peuvent également avoir une influence importante sur l'évolution des virus. Les antiviraux sont des molécules qui inhibent le fonctionnement de certaines protéines virales dans le but de freiner voire stopper la réplication virale. Lors de l'administration d'un antiviral, les variants sauvages (*wild-type*) sensibles aux traitements vont rapidement disparaitre. En revanche, certaines mutations peuvent permettre au virus d'échapper (de résister) aux antiviraux. L'apparition de ces mutations de résistance peut entrainer un échec thérapeutique se traduisant par une multiplication du virus chez un patient traité. Pour limiter l'apparition des mutations de résistance, on administre les antiviraux en polythérapie (trithérapie pour le VIH par exemple) avec différents antiviraux qui ciblent différents mécanismes de la réplication du virus. La plupart des antiviraux développés à ce jour ciblent le VIH et le VHC, deux virus pour lesquels des mutations de résistance ont été observées (Cheng et al. 2016; Clutter et al. 2016; Jacobson 2016).

1.2 OUTILS ET METHODES BIOINFORMATIQUES POUR L'ETUDE DE L'EVOLUTION MOLECULAIRE

En décrivant différents aspects de la biologie des virus et de leur évolution, nous avons vu dans la première partie de cette introduction que les virus évoluaient avec des taux de mutation élevés et que leurs séquences génomiques renfermaient de nombreuses informations sur cette évolution. Dans cette partie, je vais décrire les méthodes qui permettent de comparer les séquences virales et d'étudier leur évolution.

Cette partie n'a pas vocation à introduire de manière exhaustive l'ensemble des outils bioinformatiques permettant d'étudier ou de mesurer l'évolution moléculaire. Je présenterai ici un contexte général des outils et concepts utilisés dans ma thèse et notamment des méthodes de phylogénie. La plupart des méthodes présentées ici ne sont d'ailleurs pas réservées à l'étude des virus.

L'étude de l'évolution moléculaire s'intéresse aux changements des séquences nucléotidiques (ADN, ARN) ou protéiques (acides aminés) au cours du temps. Dans ce cadre, de nombreuses méthodes bioinformatiques permettent de mesurer la fréquence de ces changements (vitesse d'évolution), leurs causes (pression de sélection), et leurs éventuelles conséquences (changements phénotypiques) etc. En raison de leur évolution rapide, les virus sont des objets d'étude idéaux pour quantifier et comprendre l'évolution moléculaire. La phylogénie moléculaire occupe une place privilégiée dans ce domaine car elle permet d'inférer l'histoire évolutive des séquences étudiées et leurs relations de parenté.

Afin d'étudier l'évolution des virus, il est nécessaire d'accéder à leur génome via des technologies de séquençage. Ces séquenceurs produisent de grandes quantités de données brutes qu'il faut ensuite analyser via des workflows bioinformatiques afin de reconstruire les génomes complets. Par la suite, il est possible de comparer ces génomes entre eux afin d'étudier leur histoire évolutive.

Je présenterai dans une première partie la manière dont les séquences sont produites par les séquenceurs et les méthodes de reconstruction des génomes complets à partir de ces données brutes. Puis j'introduirai les méthodes qui permettent de comparer ces génomes et d'inférer leur histoire évolutive.

1.2.1 Production et analyse des séquences virales

1.2.1.1 Séquençage et production des données.

L'étude de l'évolution à l'échelle moléculaire, nécessite d'avoir accès aux séquences nucléotidiques des organismes à étudier. L'obtention de ces séquences se fait par le séquençage des composants élémentaires des séquences, les bases azotées : adénine, thymine (ou uracile pour l'ARN), guanine et cytosine (A, T (ou U), C et G). Pour cela, le séquençage permet de déterminer la séquence des bases (A, C, G, T, U) constituant le génome viral. La grande majorité des technologies de séquençage fonctionnent pour l'ADN. Dans le cas des virus ARN on passe par l'ADN complémentaire pour leur séquençage.

Il existe plusieurs technologies de séquençage, la première étant le séquençage Sanger mis au point en 1977 (Sanger, Nicklen, et Coulson 1977). Cette technologie a été utilisée pour séquencer

le premier génome humain en 2003 après 10 ans de travail et un investissement de plusieurs milliards de dollars (https://www.genome.gov/human-genome-project/Completion-FAQ). Les technologies de séquençage ont beaucoup évolué et aujourd'hui le séquençage dit à haut débit ou séquençage de nouvelle génération (NGS) permet de séquencer un génome humain complet pour moins de 1000 dollars en quelques jours (Goodwin, McPherson, et McCombie 2016).

Les technologies actuelles de séquençage se différencient par la taille des « lectures » (séquences individuelles lues par le séquenceur), le taux d'erreur de ces lectures et les méthodes pour obtenir ces lectures. Les technologies comme Illumina ou IonTorrent vont produire des lectures courtes (au maximum entre 150 et 300 nucléotides)[5] avec un taux d'erreur faible (de l'ordre de 1/1000), alors que les technologies comme PacBio (Eid et al. 2009) et Nanopore (Clarke et al. 2009) vont produire des lectures longues (entre 5000 et 25000 nucléotides en moyenne pour PacBio et atteignant le million de bases pour Nanopore (Payne et al. 2019)), avec un taux d'erreur plus élevé mais pouvant toutefois être réduit (Jain et al. 2018; Wenger et al. 2019; Sahlin et Medvedev 2021). Pour l'assemblage de génome ou le séquençage d'organismes non modèles, les lectures longues présentent un avantage car elles permettent de reconstruire avec moins d'ambiguïté les régions répétées (notamment les télomères ou les centromères), et faciliter l'assemblage de grands génomes.

Dans le cas des virus, qui ont des génomes courts, on utilise surtout des technologies avec des tailles de lectures courtes comme Illumina ou IonTorrent. Ces technologies permettent également des profondeurs de séquençage plus importantes (Quiñones-Mateu et al. 2014; Houldcroft, Beale, et Breuer 2017) et des coûts moins élevés notamment grâce au séquençage de plusieurs échantillons en même temps (*multiplexing ;* (Arul et Robinson 2019)). La profondeur du séquençage correspond au nombre moyen de lectures uniques couvrant chaque base nucléique et est exprimée en X. En séquençant en profondeur, on peut s'intéresser aux variants mineurs afin de détecter des mutations rares ou minoritaires dans la population virale mais qui pourraient devenir avantageuses dans certaines conditions. L'étude des variants mineurs permet également de calculer la diversité génétique d'un échantillon. Des logiciels permettent de quantifier la proportion de variants mineurs et déterminer s'il s'agit d'erreurs de séquençage ou non (Grubaugh et al. 2019).

En raison de sa petite taille, de son prix d'acquisition réduit et de sa facilité de transport, la technologie Nanopore MinION est aussi utilisée pour le séquençage de virus sur le terrain dans le cadre de suivi épidémique comme lors de l'épidémie d'Ebola en Afrique de l'Ouest en 2014-2015 (Hoenen et al. 2016; Quick et al. 2016). Les données de suivi épidémique peuvent être utilisées pour guider les mesures de contrôle, en mesurant la vitesse de propagation de l'épidémie, surveillant les signatures d'adaptation à l'hôte, identifiant de possibles cibles thérapeutiques, etc. (Gardy, Loman, et Rambaut 2015). Cette technologie est également utilisée afin de reconstruire des génomes du virus de la grippe et identifier rapidement de possibles réassortiments (King et al. 2020). Par ailleurs, en réponse à la pandémie de SARS-CoV-2 et grâce à son coup réduit d'acquisition, de nombreux laboratoires et pays, utilisent les technologies Nanopore pour le séquençage des génomes du SARS-CoV-2[6] et ce notamment grâce au protocole ARTIC (Lambisia et al. 2022).

[5] https://www.illumina.com/systems/sequencing-platforms.html
[6] https://nanoporetech.com/covid-19/community-timeline

1.2.1.2 Obtention d'une séquence consensus

Les lectures en sortie de séquençage correspondent à des petits fragments du génome (ou des séquences) séquencé, qu'il faut réordonner et fusionner par des méthodes bioinformatiques pour reconstruire un génome complet. Cette étape peut se faire soit par mapping en alignant les lectures sur une séquence de référence, soit par assemblage de novo (avec ou sans séquence de référence).

On utilise le mapping lorsque l'on a déjà une idée de l'organisme que l'on séquence, et que l'on s'attend à des séquences proches. On peut alors utiliser une séquence connue et de bonne qualité de cet organisme comme référence, sur laquelle on va aligner l'ensemble des lectures. La séquence consensus sera ensuite reconstruite en listant les différences (mutations, insertions, délétions) majoritaires (affectant plus de la moitié des lectures aux positions concernées) que l'on trouve dans l'échantillon. Une correspondance imparfaite entre une lecture et la référence peut indiquer la présence de mutations qui peuvent refléter une réalité biologique mais aussi des erreurs de séquençage. Afin de déterminer l'origine des mutations, on peut utiliser la couverture, qui correspond au nombre moyen de lectures qui s'alignent sur, ou "couvrent", les positions sur le génome de référence. Pour désigner la couverture d'un seul nucléotide on parle de profondeur de couverture (Sims et al. 2014). Plus la profondeur de séquençage est importante à une position, plus on aura confiance que l'observation de variants à cette position reflète une réalité biologique.

Dans le cas des virus et notamment des virus à ARN, les échantillons séquencés reflètent la diversité de la population virale. Certains variants sont minoritaires dans la population et d'autres sont majoritaires. La séquence consensus représente la séquence des nucléotides les plus fréquents à chaque position. La Figure 4, représente une visualisation, adaptée du logiciel IGV (Robinson et al. 2011), d'un résultat d'assemblage par mapping des lectures sur un génome de référence.

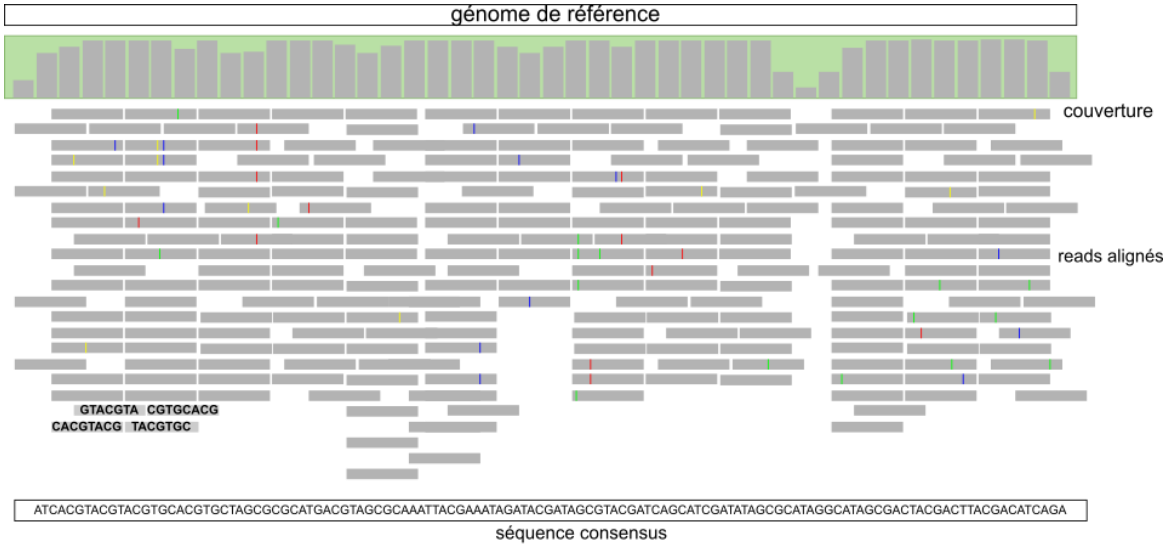


Figure 4 : Exemple de visualisation d'un mapping de lectures sur un génome de référence.
Les rectangles gris horizontaux représentent les lectures de séquençage qui ont été 'mappé' sur le génome de référence. Les traits colorés verticaux sur les lectures représentent les variations par rapport au génome de référence.

L'assemblage de novo permet de reconstruire des séquences complètes sans a priori sur l'organisme séquencé. On peut ainsi reconstruire des génomes d'organismes encore inconnus et ne présentant aucune homologie avec des espèces déjà séquencées (Yooseph et al. 2007). Cette

approche est particulièrement utile en métagénomique où l'on cherche à analyser l'ensemble des microorganismes d'un échantillon environnemental.

Une fois les génomes reconstruits, il est possible de les annoter en identifiant les différents gènes, leur rôle biologique, proposer une traduction des séquences nucléotidiques en acides aminés, etc.

1.2.1.3 Les bases de données génomiques

Les nouveaux génomes peuvent ensuite être partagés publiquement en les déposant sur des bases de données. En particulier, une collaboration internationale de plus de 30 ans existe pour le partage des données de séquence nucléotidiques (*International Nucleotide Sequence Database Collaboration* (INSDC; http://www.insdc.org/)). Cette infrastructure permet, grâce à des mises à jour quotidiennes, le partage des données et métadonnées de trois bases de données ADN : la Banque de données ADN du Japon (DDBJ) (Tateno et al. 2002), les Archives Européennes de Nucléotides (ENA) (Cummins et al. 2022) et GenBank (Benson et al. 2013). La base de séquence ADN GenBank du Centre National de l'Information Biologique au Etats-Unis est la plus importante dans le domaine. Elle contient la majorité des séquences ADN publiques annotées et sa taille augmente de manière exponentielle, avec 237'520'318 séquences déposées en avril 2022 et un doublement du nombre de bases nucléotidiques tous les 18 mois environ[7].

Ces bases de données sont très généralistes et contiennent des séquences provenant de tout type d'organismes. On trouve également des bases de données spécialisées dans les virus comme Vipr (Pickett et al. 2012), ou la base de données de pathogènes de Los Alamos pour le VIH (C. Kuiken, Korber, et Shafer 2003), le VHC ou Ebola. Ces bases collectent et annotent des séquences virales et permettent de les interroger selon certains critères spécifiques aux virus (données de patients, traitements, données immunologiques, etc.) ce qui facilite les analyses. Une autre base importante pour le VIH est la base de données de Stanford qui répertorie les séquences associées au développement de résistance (Rhee 2003).

De la même manière, il existe des bases de données spécialisées dans le stockage des séquences de protéines accompagnées de leur description dont des informations sur leur repliement et leur structure, d'annotations fonctionnelles, etc. Les bases de données UniProt (The UniProt Consortium 2021), SwissProt et TrEMBL (Bairoch et Apweiler 2000) en sont un exemple.

Les bases de données permettent de maintenir durablement les données et les métadonnées associées et de les partager à l'ensemble de la communauté scientifique. Elles jouent ainsi un rôle essentiel dans le mouvement de libre accès pour la littérature académique (https://www.budapestopenaccessinitiative.org/) et dans les efforts réalisés de la part de la communauté scientifique afin que les données soient FAIR (« trouvables, accessibles, interopérables et réutilisables ») (Wilkinson et al. 2016). On pourra par exemple citer l'effort international qui a été fait dans le séquençage et le partage des données SARS-CoV-2 avec plus de 10'369'509 séquences déposées sur la plateforme GISAID (https://www.gisaid.org/) au moment de l'écriture de cette introduction. Il reste toutefois des efforts à faire dans ce sens, et des critiques

[7] https://www.ncbi.nlm.nih.gov/genbank/statistics/

existent sur l'accès des données de la part de GISAID ainsi que sur un manque de transparence de la plateforme[8].

1.2.2 Analyse comparative des génomes viraux

1.2.2.1 Comparaison de séquences et homologie

Une fois les séquences produites par séquençage ou téléchargées à partir de bases de données, on peut les comparer entre elles afin d'identifier leurs similarités. La recherche de similarité est souvent utilisée pour trouver des régions ou des séquences potentiellement homologues, c'est-à-dire héritées d'un ancêtre commun. L'homologie est un concept fondamental en génomique comparative pour plusieurs raisons. Tout d'abord, si des séquences ont une origine commune elles partagent sans doute des propriétés en commun (fonction par exemple). Ainsi, on peut annoter des fonctions hypothétiques d'une séquence inconnue à partir d'une séquence préexistante si elles présentent de fortes similarités. Trouver des séquences homologues chez différentes espèces permet également de déterminer qu'elles ont une origine commune et donc des liens de parenté que l'on va pouvoir reconstruire. L'hypothèse sous-jacente est que plus les séquences évoluent, plus elles accumulent des mutations. En estimant le nombre de mutations entre deux séquences on peut alors avoir une idée du temps qui les sépare.

1.2.2.2 Alignement

La recherche de similarités entre séquences passe par une première étape qui est l'alignement. En effet, au cours de l'évolution, les séquences peuvent subir des insertions ou délétions, qui vont modifier leur longueur, et rendre les séquences incomparables directement. L'alignement de séquences est un moyen de comparer des séquences, afin d'identifier les positions homologues. Cela peut se faire pour deux séquences dans le cas de l'alignement par paires, ou plus dans le cas de l'alignement multiple. Lors de l'alignement, on ajoute des espaces (gaps) aux positions où des indels ont eu lieu. Ainsi les régions similaires sont superposées (ou alignées). Le résultat d'un alignement multiple est une matrice dont les colonnes sont constituées des caractères homologues (dérivés d'un ancêtre commun), et les lignes correspondent aux séquences alignées. La Figure 5 représente un alignement multiple de séquences protéiques. La première position est ainsi occupée par le seul acide aminé leucine "L". La longueur de l'alignement correspond au nombre total de positions alignées, qui est supérieur ou égal à la longueur de la plus longue séquence avant alignement.

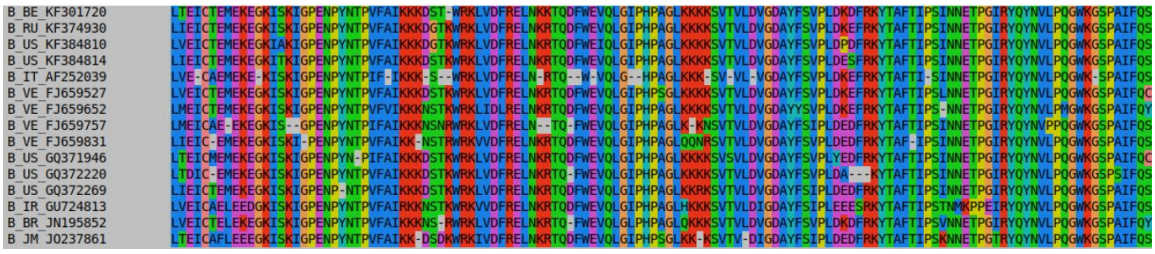


Figure 5 : Alignement en acides aminés.
Les identifiants des séquences sont donnés à gauche. Les couleurs correspondent à des propriétés biochimiques proches. Les traits d'union représentent les gaps introduit dans les séquences lors de l'alignement.

[8] https://www.science.org/content/article/critics-decry-access-transparency-issues-key-trove-coronavirus-sequences

Les alignements peuvent se faire à partir de séquences nucléotidiques ou protéiques. Pour des séquences très proches on favorisera des alignements de séquences nucléotidiques qui permettent une plus grande finesse (redondance du code génétique) alors qu'avec des séquences plus éloignées, si l'on a des séquences codantes, les similarités seront plus évidentes en acides aminés (moins de saturation à de grandes distances évolutives).

1.2.2.3 Algorithmes d'alignement

De nombreux algorithmes ont été conçus pour aligner des séquences. Nous pouvons distinguer d'une part l'alignement par paires, et d'autre part l'alignement de séquences multiples (≥3 séquences). Les algorithmes d'alignement par paires cherchent à trouver l'alignement qui maximise le score de correspondance entre deux séquences par programmation dynamique. Ils peuvent se faire à l'échelle de la séquence complète (Needleman et Wunsch 1970) (alignement global) ou alors chercher des alignements locaux sur une ou des sous-parties de la séquence (Smith et Waterman 1981). Ces algorithmes sont dits optimaux dans le sens où ils donnent la meilleure solution possible.

Ces algorithmes ne fonctionnent plus pour l'alignement de séquences multiples car ils deviennent trop coûteux en temps de calcul. Les stratégies utilisées pour l'alignement de séquences multiples utilisent différentes heuristiques (par exemple l'alignement progressif (Feng et Doolittle 1987); la transformée de Fourier rapide, FFT ou le modèle de Markov caché, HMM) et permettent d'aligner plusieurs centaines voire plusieurs milliers de séquences. Parmi les logiciels d'alignement multiple les plus connus on peut citer Clustal Omega (Thompson, Higgins, et Gibson 1994; Sievers et Higgins 2021), MAFFT (Katoh et al. 2002; Katoh et Standley 2013), MUSCLE (Edgar 2004), ou T-COFFEE (I. M. Wallace et al. 2006).

1.2.3 Modèles d'évolution

A partir de séquences alignées, on peut estimer leur distance évolutive c'est-à-dire définir le nombre moyen de mutations par site s'étant produites depuis que ces séquences ont divergé de leur ancêtre commun (Perrière et Brochier-Armanet 2010). Cette estimation se fait grâce à des modèles d'évolution que je vais décrire ici de manière succincte. De plus amples détails peuvent être trouvés dans les ouvrages suivants (Gascuel 2005; Gascuel et Steel 2007).

1.2.3.1 Distance observée ou p-distance

La méthode la plus simple consiste à compter le nombre de mutations observées entre deux séquences alignées. Ainsi, pour des séquences alignées sur *n* positions et qui présentent *m* mutations on obtient $p = \frac{m}{n}$. Cette méthode approxime bien la distance évolutive sur des temps très courts, mais sur des temps plus longs, elle la sous-estime. En effet, sur une même position, plusieurs mutations peuvent avoir eu lieu, comme illustré Figure 6. Dans certains cas, après des temps de divergence longs, si suffisamment de mutations se sont accumulées on peut assister à un phénomène de saturation. Le nombre de mutations entre deux séquences n'est alors plus un indicateur de leur temps de divergence.

Figure 6 : Mutations cachées à un site pendant un temps t.
Les mutations successives vers T, G et A ne sont pas détectables en comparant les séquences à t=0 puis à t.

Les processus évolutifs sont complexes et il est impossible de reconstruire avec certitude l'historique des substitutions d'une séquence ancestrale aux séquences que nous observons aujourd'hui. Des hypothèses et des modèles ont été proposés pour donner une meilleure approximation de la probabilité pour un site d'évoluer d'un caractère $x \rightarrow y$ au cours du temps t.

1.2.3.2 Modèles de Markov en temps continu.

1.2.3.2.1 Définition et caractéristiques

Dans les modèles dits de Markov, plusieurs hypothèses sous-tendent l'apparition des mutations. Tout d'abord on suppose que l'évolution est continue dans le temps et sans mémoire. Cela signifie que les mutations sont des évènements aléatoires et l'information pour la prédiction des états futurs est entièrement contenue dans les états présents. Ainsi le processus de mutations suit une chaine de Markov en temps continu. Un modèle d'évolution peut alors être représenté selon une matrice de taux Q stable dans le temps. Cette matrice donne les taux de transitions entre deux états x et y. Dans le cas de substitution nucléotidique entre les bases (états) ACGT cette matrice peut être représentée comme suit :

$$\begin{pmatrix} -\mu(a\pi_C + b\pi_G + c\pi_T) & a\mu\pi_C & b\mu\pi_G & c\mu\pi_T \\ g\mu\pi_A & -\mu(g\pi_A + d\pi_G + e\pi_T) & d\mu\pi_G & e\mu\pi_T \\ h\mu\pi_A & i\mu\pi_C & -\mu(h\pi_A + i\pi_C + f\pi_T) & f\mu\pi_T \\ j\mu\pi_A & k\mu\pi_C & l\mu\pi_G & -\mu(j\pi_A + k\pi_C + l\pi_G) \end{pmatrix}$$

Équation 1: Matrice des taux instantanés Q

Les lignes et les colonnes suivent l'ordre ACGT et, par exemple, le deuxième terme de la première ligne représente le taux de substitution de A vers C. Ce taux est égal à la fréquence à l'équilibre de C, π_C, que multiplie le taux de substitution instantané global μ que multiplie le facteur a qui correspond à un taux décrivant la quantité de substitutions de A vers C en comparaison avec les autres substitutions. Les éléments sur la diagonale sont calculés afin que chacune des lignes somme sur zéro. Les matrices de taux instantanés sont normalisées par le terme μ afin que les taux de substitution par site correspondent à la probabilité d'observer une mutation par unité de temps.

Les modèles de substitution décrits par cette matrice Q s'accompagnent d'autres propriétés décrivant le modèle markovien :

- Hypothèse d'indépendance : chaque position de la séquence évolue de manière indépendante et une mutation à une position n'influence pas les mutations à d'autres positions. Chaque position peut donc être étudiée séparément.
- Hypothèse d'homogénéité : les taux de substitution ne changent pas au cours du temps.
- Hypothèse d'uniformité : tous les sites d'une séquence suivent le même processus et les probabilités de substitution sont donc les mêmes pour tous les sites. Cela signifie donc que les sites évoluent à la même vitesse.
- Hypothèse de stationnarité : les fréquences relatives de A, C, G et T, $\pi_A, \pi_C, \pi_G, \pi_T$ sont à l'équilibre. Cela permet de définir à priori les probabilités stationnaires de A, C, G et T qui correspondent à la fréquence qu'on attendrait après un temps d'évolution infini.
- Hypothèse de réversibilité : la quantité de substitution d'un état i vers un état j est égale à la quantité de substitution de l'état j vers l'état i. Ainsi $a = g, b = h, c = i, d = j, e = k$ et $f = g$ dans la matrice Q.

Finalement, la probabilité de changement d'une base vers une autre au cours d'un temps t peut être calculée à partir de la matrice Q selon la formule suivante :

$$P(t) = \exp(Qt)$$

1.2.3.2.2 Modèles markoviens pour l'évolution de séquences ADN

Selon les valeurs attribuées aux taux de transition entre nucléotides a, b, c, d, e, f, et les fréquences à l'équilibre, différents modèles ont été définis et sont répertoriés dans le Tableau 2. Le modèle le plus simple est le modèle JC69 de Jukes et Cantor (1969) qui suppose l'égalité entre chacun des taux de transition ainsi que l'égalité des fréquences à l'équilibre. Le seul paramètre de la matrice Q_{JC69} est donc μ et les valeurs de la matrice sont toutes égales à 0.25. Une amélioration de ce modèle est le modèle F81 (Felsenstein 1981) qui suppose toujours des taux de transition égaux entre nucléotides mais les fréquences à l'équilibre de chaque nucléotide sont différentes. Le facteur K présent dans les modèles K80 (M. Kimura 1980) et HKY (Hasegawa, Kishino, et Yano 1985) prend en compte la différence entre les transitions et les transversions en considérant respectivement des fréquences à l'équilibre égales ou non. En effet, comme représenté Figure 7, un changement entre une purine et une pyrimidine (transversion) a un impact plus important que des changements entre deux purines ou deux pyrimidines (transition). Le modèle le plus complexe est le modèle GTR (Tavaré et Miura 1986) qui autorise des taux de transition entre nucléotides différents pour chaque base en plus de prendre en compte les fréquences à l'équilibre de chaque base.

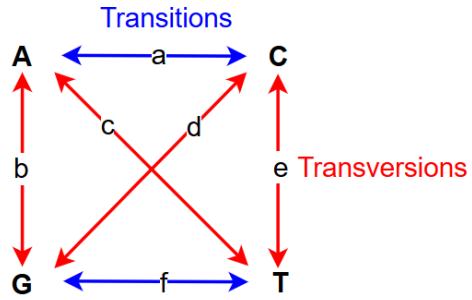


Figure 7 : Transitions et transversions entre les nucléotides.

Modèle	Matrice
JC69 (Jukes et Cantor 1969)	$Q_{JC69} = \frac{1}{\mu}\begin{pmatrix} - & 0.25 & 0.25 & 0.25 \\ 0.25 & - & 0.25 & 0.25 \\ 0.25 & 0.25 & - & 0.25 \\ 0.25 & 0.25 & 0.25 & - \end{pmatrix}$
K80 (M. Kimura 1980)	$Q_{K80} = \frac{1}{\mu}\begin{pmatrix} - & 0.25 & 0.25k & 0.25 \\ 0.25 & - & 0.25 & 0.25k \\ 0.25k & 0.25 & - & 0.25 \\ 0.25 & 0.25k & 0.25 & - \end{pmatrix}$
F81 (Felsenstein 1981)	$Q_{F81} = \frac{1}{\mu}\begin{pmatrix} - & \pi_C & \pi_G & \pi_T \\ \pi_A & - & \pi_G & \pi_T \\ \pi_A & \pi_C & - & \pi_T \\ \pi_A & \pi_C & \pi_G & - \end{pmatrix}$

HKY (Hasegawa, Kishino, et Yano 1985)	$Q_{HKY} = \frac{1}{\mu}\begin{pmatrix} - & \pi_C & k\pi_G & \pi_T \\ \pi_A & - & \pi_G & k\pi_T \\ k\pi_A & \pi_C & - & \pi_T \\ \pi_A & k\pi_C & \pi_G & - \end{pmatrix}$
GTR (Tavaré et Miura 1986)	$Q_{GTR} = \frac{1}{\mu}\begin{pmatrix} - & a\pi_C & b\pi_G & c\pi_T \\ a\pi_A & - & d\pi_G & e\pi_T \\ b\pi_A & d\pi_C & - & f\pi_T \\ c\pi_A & e\pi_C & f\pi_G & - \end{pmatrix}$

Tableau 2 : Matrices de taux pour différents modèles markoviens d'évolution de séquences ADN.

1.2.3.2.3 Modèles markoviens pour l'évolution de séquences protéiques

Les modèles présentés précédemment sont des modèles pensés pour l'évolution de séquences nucléotidique. Pour étudier l'évolution de l'ADN, il existe également des modèles de codons qui prennent en compte des taux de substitution différents en fonction de la position du codon ou des mutations synonymes (silencieuse) et non-synonymes (avec remplacement de l'acide aminé) (Muse et Gaut 1994; Goldman et Yang 1994). Au niveau protéique, avec 20 acides aminés, le nombre de paramètres à estimer est trop important pour conceptualiser un modèle théorique. Les modèles d'évolution des séquences protéiques sont donc le plus souvent empiriques.

Certaines matrices sont des modèles généraux comme les patrices PAM (Dayhoff, Schwartz, et Orcutt 1978) et BLOSUM (Henikoff et Henikoff 1992) et ont été calculées à partir d'observations sur un grand nombre de protéines. Par exemple, Les matrices PAM sont des matrices de taille 20 x 20 où chaque ligne et chaque colonne représente un des 20 acides aminés. La matrice contient un score de vraisemblance pour chaque substitution d'un acide aminé par un autre pour un temps d'évolution donné. Il existe plusieurs matrices PAM pour différents intervalles de temps d'évolution des séquences protéiques. Ces matrices ont été introduites par Margareth Dayhoff et ont été calculées sur 1572 mutations observées au sein de 71 familles de protéines étroitement liées et partageant au moins 85% de similarité.

D'autres modèles sont spécifiques et sont calculés spécialement sur un organisme donné ou une protéine. C'est le cas par exemple du modèle HIVb (Nickle et al. 2007) que j'ai beaucoup utilisé pour l'étude du VIH dans cette thèse. Ce modèle a été construit à partir de 8 jeux de données VIH et un total de 7189 substitutions entre acides aminés.

Les modèles sont donnés sous la forme de la matrice d'échangeabilité symétrique R. Elle contient les taux de substitution $R_{x\leftrightarrow y}$ entre les différents états (nucléotides ou acides aminés) ainsi que le vecteur des fréquences à l'équilibre Π. La matrice de substitution Q peut en effet être déterminée à partir de R et des fréquences de chaque caractère (fréquences à l'équilibre, calculées empiriquement ou optimisées à partir de l'alignement) :

$$Q_{x,y} = \pi_y R_{x\leftrightarrow y} \; for \; x \neq y$$
$$Q_{x,x} = -\sum Q_{xy}.$$

1.2.3.3 Modèles de mélange pour l'évolution de séquences protéiques

Les modèles markoviens d'évolution des séquences protéiques reposent sur une matrice de substitution unique pour représenter les différents processus d'évolution d'une protéine. Cependant, l'évolution de chaque site répond à des mécanismes différents influencés par la structure des protéines, leur exposition aux différents solvants, le code génétique… Ainsi la

plupart du temps, seul un petit nombre d'acides aminés différents sont possibles à un site donné. Cette réalité biologique n'est pas prise en compte par les modèles markoviens car ils sont sans mémoire et vont donc pouvoir favoriser un plus grand nombre d'acides aminés possibles à une position.

Les modèles de mélange sont une alternative aux modèles markoviens à matrice unique et permettent de modéliser l'évolution des sites selon différentes matrices d'échangeabilité ou différents profils. Ainsi, les modèles de mélange EX2 ou EX3 (Le, Lartillot, et Gascuel 2008) sont composés respectivement de deux ou trois matrices selon l'accessibilité des acides aminés au solvant. Les matrices CAT (Le, Gascuel, et Lartillot 2008) quant à elles, représentent différents vecteurs de probabilités stationnaires (ou fréquences à l'équilibre) des 20 acides aminés. Ces vecteurs de probabilités ont été estimés sur de vraies données et sont regroupés en matrices de 10 à 60 profils. Par exemple, la matrice C10 (de taille 10 x 20) contient 10 vecteurs de probabilités stationnaires pour chacun des 20 acides aminés. Pour un profil donné (un des vecteurs de probabilités), la probabilité d'observer un acide aminé *b*, après un temps *t=l* et sachant qu'à *t=0* l'acide aminé était *a* est donnée par :

$$P(b|a,l) = e^{-l}\delta_{ab} + \left(1 - e^{-l}\pi_b\right),$$

Avec δ_{ab} le symbole de Kronecker (vaut 1 si a = b et 0 sinon) et π_b la fréquence à l'équilibre de b dans le profil π.

1.2.3.4 Vitesses d'évolution

Afin d'être utilisables, les modèles markoviens ou les modèles de mélange font des hypothèses fortes. On sait par exemple que tous les sites n'évoluent pas à la même vitesse ou que des mutations à certains sites influencent des sites voisins (conformation 3D, etc.). Des améliorations sont donc proposées afin d'augmenter la fiabilité des modèles. On peut notamment modéliser les taux d'évolution des sites (leur vitesse) grâce à une loi gamma (Γ), que l'on va discrétiser, en définissant des catégories (4 par exemple). Ainsi, lors de l'inférence, on va pouvoir associer une catégorie de la loi gamma à chaque site. La forme de cette distribution est paramétrée par le facteur de forme α qui modélise l'hétérogénéité entre sites (paramètre qui sera également optimisé lors de l'inférence). Il est également possible de rajouter une classe invariante pour les sites constants. Le modèle FreeRate (Susko et al. 2003; Soubrier et al. 2012) repose sur le même principe de taux variables selon les sites, par catégories, mais en relâchant la contrainte d'une distribution gamma des taux. Les taux d'évolution correspondant à chacune des catégories sont alors directement estimés des données.

1.2.4 Inférence phylogénétique

Les modèles d'évolution permettent d'estimer la distance évolutive entre plusieurs séquences. A partir de données moléculaires et de leur distance évolutive, il est possible de reconstruire un arbre phylogénétique qui représente les relations de parenté entre séquences, gènes ou espèces. La reconstruction d'une phylogénie est un processus complexe qui fait appel à différentes méthodes d'inférence phylogénétique (Felsenstein 2003; Delsuc et Douzery 2004a; 2004b; Lemey, Salemi, et Vandamme 2009) que je vais brièvement introduire ici.

1.2.4.1 Arbre phylogénétique

En théorie des graphes, un arbre est un graphe connexe et acyclique, c'est-à-dire un ensemble de sommets (nœuds) connectés par des arrêtes (branches) de telle sorte que toute paire de nœuds

est reliée exactement par un chemin. En biologie, cette représentation est utilisée pour représenter les relations de parenté entre divers organismes (espèces, groupe d'espèces, gènes, individus) ou taxons. On parle alors d'arbre phylogénétique (ou phylogénie), tel que représenté Figure 8 : Représentation d'un arbre phylogénétique binaire et enraciné. Les feuilles (nœuds externes) représentent les taxons étudiés (espèces ou séquences actuelles) et les nœuds internes, les ancêtres communs, non observés. Un arbre phylogénétique peut être orienté lorsqu'il est enraciné, c'est-à-dire qu'un nœud racine est créé et représente l'ancêtre commun à toutes les feuilles de l'arbre. L'arbre représente alors l'évolution orientée depuis la racine vers les feuilles.

La longueur des branches de l'arbre peut être optimisée à partir des données et d'un modèle d'évolution, et représente une distance évolutive entre les nœuds (en nombre de substitutions par site), un pourcentage de divergence ou le temps écoulé. Nous étudions ici uniquement des arbres dont la longueur des branches représente la distance évolutive et qui sont dit binaires (chaque nœud interne est relié à exactement 3 voisins).

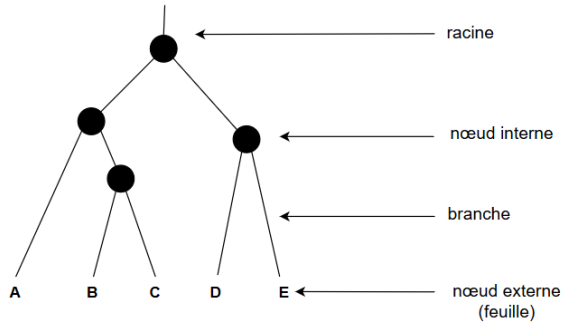


Figure 8 : Représentation d'un arbre phylogénétique binaire et enraciné

Un arbre peut être enraciné en utilisant un groupe externe, c'est à dire un groupe monophylétique de séquences ou d'organismes distants (mais pas trop) d'un point de vue évolutif des objets d'études entre eux. Le nœud à la racine sera alors le nœud séparant les deux groupes : le groupe externe et le groupe étudié.

La manière dont les feuilles et les nœuds internes d'un arbre sont connectés définit la topologie de l'arbre. Le nombre de topologies différentes pour un arbre binaire non enraciné de *n* feuilles (n>2) est égal à :

$$\frac{(2n-5)!}{2^{n-3}(n-3)!}.$$

Cela signifie que pour 10 feuilles il existe plus de 2 millions de topologies alternatives et pour 55 feuilles on dépasse le nombre d'atomes dans l'univers, estimé à 10^{80}. Il est alors impossible de tester toutes les phylogénies possibles à la recherche de la topologie optimale. Des heuristiques appropriées qui n'explorent qu'une petite fraction de cet espace de recherche doivent être appliquées avec pour conséquence qu'elles peuvent fournir des solutions sous-optimales. Le choix d'une méthode qui propose un équilibre entre un temps de calcul court et une bonne précision est donc primordial.

Deux grandes familles de méthodes d'inférence existent : les méthodes de distance et les méthodes basées sur les caractères. Elles se distinguent entre autres par le temps de calcul nécessaire, leur biais, le type de données utilisé, et l'approche algorithmique utilisée.

1.2.4.2 Méthodes de distance

Les méthodes de distance cherchent à trouver l'arbre qui représente au mieux la matrice des distances évolutives entre paires de séquences. Pour cela, les distances sont calculées selon les modèles présentés précédemment. Cette matrice de distance est ensuite utilisée pour inférer une topologie et les longueurs de branches. Les principales méthodes de distance pour reconstruire la topologie sont les méthodes de clustering et les méthodes de minimum évolution.

Les méthodes de clustering (comme UPGMA) fonctionnent par étape en groupant les séquences ou groupes de séquences (aussi appelés unités taxonomiques opérationnelles ; OTUs) les plus proches entre elles, c'est-à-dire celles pour lesquelles la distance génétique est la plus faible. Lorsque deux OTUs sont groupés ils deviennent un nouvel OTU simple. Les OTUs sont ainsi groupés jusqu'à ce qu'il n'en reste plus que deux.

En raison de certaines limitations des méthodes de clustering, des algorithmes dits d'arbres de distance additive ont été développés : les méthodes de minimum d'évolution (« A Simple Method for Estimating and Testing Minimum-Evolution Trees » 1992) et de Neighbour-Joining (Saitou et Nei 1987; Studier et Keppler 1988). En minimum d'évolution, l'arbre considéré comme optimal est celui qui minimise la somme des longueurs des branches. Pour chaque topologie, les longueurs de branches sont optimisées à partir de la matrice de distance. Comme il n'est pas possible de parcourir l'ensemble de toutes les topologies, des heuristiques comme le Neighbour-Joining sont utilisées. A partir d'une topologie en étoile (toutes les feuilles sont reliées par un seul nœud), l'algorithme cherche tout d'abord le couple de taxon avec la plus faible distance génétique (calculée à partir d'une matrice modifiée de la matrice de distance par paires). Ces taxons sont groupés en un nouveau nœud. Une nouvelle matrice de distance par paires est alors calculée entre les taxons restants et ce nouveau nœud. L'algorithme est ensuite répété. D'autres version de Neighbour-Joining ont été proposées, notamment en suivant le principe de minimum d'évolution équilibré (Desper et Gascuel 2002; Gascuel et Steel 2006), de Relaxed Neighbour-Joining (Evans, Sheneman, et Foster 2006) ou encore de réarrangements topologiques tels que les NNI, SPR (FASTME ; (Desper et Gascuel 2004; Lefort, Desper, et Gascuel 2015).

Les méthodes de distance sont très rapides et sont souvent utilisées pour reconstruire de très grands arbres (plusieurs milliers de feuilles) ou comme première approximation d'un arbre avant d'appliquer d'autres méthodes d'inférence phylogénétique.

1.2.4.3 Méthodes basées sur les caractères

Contrairement aux méthodes de distance, les méthodes de caractères n'utilisent pas la matrice de distance par paire de séquence mais peuvent utiliser l'information de n'importe quel caractère discret (trait morphologique, séquences, propriétés physiologiques, information géographique). Lors de l'analyse de séquences, toutes les séquences sont comparées simultanément pour un site de l'alignement et chaque site est analysé de manière indépendante. Un site de l'alignement est alors un caractère qui prend différents états qui sont les nucléotides ou les acides aminés à ce site. L'autre avantage des méthodes de caractères est qu'elles conservent l'information des états de chaque taxon de sorte qu'elles peuvent être utilisées pour de la reconstruction des caractères

ancestraux. Ces méthodes se basent sur un score (ou critère d'optimalité) pour l'évaluation des arbres et cherchent l'arbre avec le meilleur score. Le score utilisé correspond au nombre de changements pour les méthodes de parcimonie, le log de la vraisemblance en maximum de vraisemblance. Dans les approches bayésiennes, au lieu de chercher à optimiser le score d'un arbre, on cherche une distribution d'arbres qui maximise un critère d'optimalité. A partir de cette distribution on peut ensuite trouver l'arbre (ou plusieurs arbres) avec la meilleure probabilité postérieure.

Méthodes de parcimonie : on cherche l'arbre (ou les arbres) avec le moins de changements évolutifs (changement d'état d'un caractère) qui permet d'expliquer les données (Kluge et Farris 1969; Fitch 1971). Dans le cas de données moléculaires, on cherche l'arbre qui présente le moins de substitutions totales. La manière la plus naïve d'identifier l'arbre le plus parcimonieux consiste à considérer successivement chaque arbre possible et à rechercher l'arbre ayant le plus petit score. Comme expliqué précédemment, cette stratégie n'est pas applicable dès que les jeux de données dépassent une dizaine de taxons. Parmi les heuristiques pour rechercher l'arbre optimal, une classe de méthode consiste à réarranger un premier arbre non optimal par des perturbations de type permutation de branches. Ces permutations peuvent être des échanges du voisin le plus proche entre 4 taxons (*nearest neighbour interchange* ; NNI), retrait et réinsertion de sous-arbres (*subtree pruning and regraffting* ; SPR) ou bissection et reconnexion d'arbres (tree bisection and reconnection ; TBR). Après chaque réarrangement, le score de l'arbre obtenu est calculé et chaque arbre avec un meilleur score remplace le précédent jusqu'à ce que tous les réarrangements soient testés. Toutefois ces heuristiques sont sensibles aux optimums locaux et on ne peut pas garantir que la ou les topologies trouvées soient les plus parcimonieuses parmi l'ensemble de tous les arbres possibles. Dans le cas où plusieurs topologies présentent le même score, un arbre consensus qui représente le scénario majoritaire est calculé. En raison de certains biais (attraction des longues branches notamment) et de l'absence de longueur de branches, les méthodes de parcimonie sont peu utilisées pour la reconstruction d'arbres phylogénétiques. Elles sont toutefois utilisées dans la reconstruction des états ancestraux que je détaillerai par la suite.

Maximum de vraisemblance : Les méthodes de maximum de vraisemblance sont des méthodes probabilistes fondées sur le concept de vraisemblance qui correspond à la probabilité conditionnelle d'observer des données sous une certaine hypothèse. Appliquées à la phylogénie les méthodes de vraisemblance cherchent à maximiser la vraisemblance de l'arbre *L(T)* qui correspond à la probabilité d'obtenir les données D (par exemple un alignement de séquences) considérant un arbre *T* (défini par une topologie et des longueurs de branches) :

$$L(T) = P(D|T).$$

Etant donné que les sites sont considérés indépendants, la vraisemblance d'un arbre est calculée comme le produit des vraisemblances de tous les sites. Comme la vraisemblance est généralement extrêmement faible, elle est souvent exprimée sous la forme d'un logarithme :

$$log(L(T)) = \sum_{i=1}^{n} \log\big(P(d_i|T)\big),$$

avec d_i les données à la position *i* et *n* la longueur de l'alignement. En théorie, le meilleur arbre est celui qui a la meilleure vraisemblance, mais pour obtenir cet arbre l'algorithme doit parcourir toutes les topologies possibles, ce qui est impossible. En pratique, un premier arbre est construit avec une méthode rapide (distance ou parcimonie) et est utilisé comme point de départ. Des réarrangements locaux de cet arbre sont ensuite effectués pour explorer l'espace des arbres. Les

paramètres optimisés au cours de l'inférence de l'arbre sont : ses longueurs de branches, sa topologie, ainsi que les paramètres du modèle d'évolution des séquences (défini par la matrice de taux instantané Q ; voir section modèles d'évolution). Le maximum de vraisemblance a une très bonne précision mais présente des temps de calculs relativement longs en comparaison des autres méthodes. La méthode de maximum de vraisemblance est celle que j'ai utilisée au cours des travaux de cette thèse pour l'inférence des arbres.

Méthodes Bayésiennes : L'approche bayésienne est également une méthode probabiliste qui, comme son nom l'indique, repose sur une application du théorème de Bayes :

$$P(H|D) = \frac{P(D|H)P(H)}{P(D)}.$$

Ce théorème permet de calculer la probabilité postérieure d'une hypothèse *H* sachant les données *D :* $P(H|D)$. D'après le théorème de Bayes, cette probabilité postérieure est fonction de la vraisemblance $P(D|H)$ et de la probabilité à priori de l'hypothèse $P(H)$.
Appliqué à un arbre *T*, celui-ci permet de déterminer la probabilité postérieure (calculée à posteriori) d'un arbre phylogénétique à partir de probabilités définies à priori d'observer un arbre sachant les données D. Il s'énonce ainsi :

$$P(T|D) = \frac{P(D|T)P(T)}{P(D)},$$

avec $P(D|T)$ la fonction de vraisemblance, $P(T)$ les probabilités à priori de l'arbre et $P(D)$, la probabilité des données. Comme précédemment la probabilité de l'arbre peut être exprimée comme la probabilité de sa topologie, de ses longueurs de branches ainsi que des paramètres du modèle d'évolution. Ainsi, le calcul de la probabilité postérieure d'un arbre nécessite de parcourir l'espace des topologies, des longueurs de branches et des paramètres du modèle d'évolution ce qui n'est pas atteignable. Dans les méthodes bayésiennes, la probabilité postérieure d'un arbre est donc estimée par des algorithmes d'échantillonnage de Monte Carlo en chaîne de Markov (MCMC). L'idée est que l'on peut estimer une distribution de probabilités en échantillonnant à intervalle régulier dans un espace multidimensionnel de ses paramètres. Plus le nombre d'échantillonnage sera important et plus l'estimation de la distribution sera proche de la réalité. La fréquence à laquelle un arbre a été visité lors du parcours de la chaine de Markov donne l'estimation de sa probabilité postérieure.

1.2.4.4 Robustesse des phylogénies

Une fois avoir reconstruit une phylogénie, il peut être utile d'estimer sa robustesse c'est-à-dire évaluer à quel point on a « confiance » dans les clades observés. Une méthode largement répandue est la méthode du bootstrap qui à l'origine permet d'évaluer la précision de la plupart des estimations statistiques (Efron 1979; Efron, Halloran, et Holmes 1996). Une adaptation aux arbres phylogénétiques a ensuite été proposée par Felsenstein (Felsenstein 1985) et consiste à reconstruire des arbres à partir de rééchantillonnages avec remise des sites de l'alignement.

Plus précisément, le procédé consiste à tirer de manière aléatoire avec remise les sites de l'alignement pour obtenir un alignement bootstrap de la taille de l'alignement original (voir Figure 9). Ainsi certaines positions seront tirées plusieurs fois et d'autres jamais. Un grand nombre d'alignements bootstrap sont générés (entre 100 et 1000 le plus souvent) et un arbre est reconstruit à chaque tirage, en utilisant la même méthode que l'inférence originale. Pour chaque branche de l'arbre original, on regarde le nombre d'arbres bootstrap dans lesquels elle est

retrouvée à l'identique. On estime qu'une branche retrouvée dans 70-80% des tirages est robuste (Hillis et Bull 1993). Cependant plus il y a de taxons et plus la probabilité de retrouver à l'identique une branche dans les réplicas bootstrap diminue (notamment pour les branches internes). Une alternative au bootstrap a ainsi été proposée (Lemoine et al. 2018), pour laquelle « la présence de branches inférées dans les réplicas est mesurée à l'aide d'une distance de "transfert" graduelle plutôt que l'indice binaire de présence ou d'absence utilisé dans la version originale de Felsenstein ». Cette distance de transfert quantifie le nombre de réarrangements (taxons à déplacer ou à retirer) à effectuer pour retrouver une branche à l'identique dans les réplicas bootstrap. Une branche supportée à 100% ne signifie pas que la branche est nécessairement vraie, seulement qu'elle est robuste parce que beaucoup d'information dans l'alignement la corrobore.

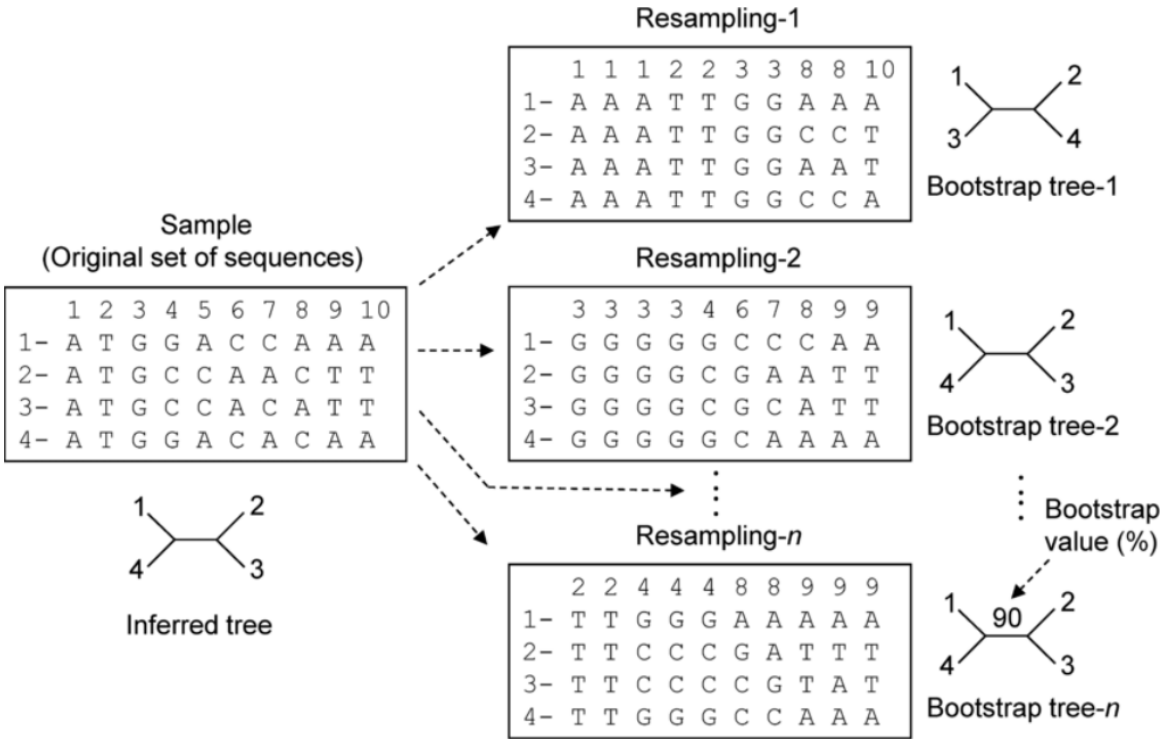


Figure 9: Principe du bootstrap de Felsenstein par rééchantillonnage des sites d'un alignement. Source : (Ateto 2014)

1.2.5 Reconstruction de caractères ancestraux

Une fois reconstruites, les phylogénies permettent d'étudier les relations de parenté entre les objets d'étude mais pas uniquement. Dans le cas d'une phylogénie enracinée, on connait le sens de l'évolution. Cela signifie qu'à partir des données de l'alignement, on peut remonter dans l'arbre nœud après nœud et inférer les états ancestraux des caractères étudiés. On peut ainsi reconstruire les séquences ancestrales (ou d'autres caractères comme des zones géographiques ou des traits phénotypiques) jusqu'à la racine de l'arbre. Ce principe repose toutefois sur l'hypothèse que la phylogénie inférée est exacte et si des incertitudes sont présentes dans l'arbre cela impactera la reconstruction ancestrale.

La reconstruction des caractères ancestraux peut être utile pour identifier un réservoir animal dans le cas d'une zoonose virale (Drexler et al. 2012), reconstruire le chemin géographique lors de

la propagation d'une épidémie (Wallace et al. 2007; Lemey et al. 2014; Dudas et al. 2017), etc. La reconstruction ancestrale est aussi utilisée dans la recherche de convergence que je détaillerai de manière plus complète dans le chapitre suivant.

Différentes méthodes permettent de reconstruire les états ancestraux et sont conjointes aux modèles d'inférence phylogénétiques : les méthodes de parcimonie, le maximum de vraisemblance et les méthodes bayésiennes.

Comme pour l'inférence phylogénétique, le maximum de parcimonie ne repose pas sur un modèle statistique de l'évolution, et ne tient donc pas compte des longueurs de branches et dans sa version la plus simple, tous les changements de caractères sont considérés équiprobables. Cette méthode est donc très rapide mais peu précise. Les états ancestraux reconstruits sont donc ceux qui impliquent le moins de changements. Il peut arriver que plusieurs scénarios soient à égalité et les états ancestraux sont alors ambigus. Des algorithmes (Maddison et Maddison 2000), que je ne détaillerai pas ici, permettent de résoudre une partie de ces ambiguïtés.

Dans les méthodes de maximum de vraisemblance, la probabilité de chacun des états du caractère reconstruit à chaque nœud est calculée en fonction d'un modèle d'évolution spécifique. Les probabilités étant dépendantes des longueurs de branches et des états observés, elles doivent être calculées à chaque nœud et ces méthodes sont plus coûteuses en temps de calcul que la parcimonie. Le plus souvent, les méthodes de maximum de vraisemblance calculent : 1) les probabilités marginales de chacun des nœuds étant dans chacun des états possibles (Felsenstein 2003) ou 2) la reconstruction jointe d'un scénario unique avec la probabilité postérieure la plus élevée (Pupko et al. 2000). Avec l'approche marginale (1), les résultats peuvent être difficiles à interpréter (plusieurs probabilités à chaque nœuds) alors que la méthode jointe (2) ne reflète pas l'incertitude de certaines reconstructions. Des méthodes intermédiaires ont ainsi été proposées comme l'approximation des probabilités marginales postérieures (MPPA ; Ishikawa et al. 2019), qui permet, de prédire un état unique aux nœuds avec une faible incertitude, et plusieurs états dans les régions incertaines (généralement autour de la racine).

Les méthodes précédentes considèrent que la phylogénie reconstruite est exacte et ne prennent donc pas en compte de possibles incertitudes. L'inférence bayésienne présente l'avantage de relier la probabilité conditionnelle d'un état à la vraisemblance de l'arbre, ainsi qu'au degré d'incertitude associé à cet arbre (Ronquist 2004; Joy et al. 2016). Toutefois, le temps d'inférence des états ancestraux est plus long que pour les autres méthodes.

1.2.6 Sélection positive ou négative

Comme nous l'avons vu précédemment, les mutations apparaissent de manière aléatoire le long du génome, certaines étant avantageuses, d'autres neutres ou délétères. Ces mutations pourront ensuite être fixées dans une population ou éliminées. Par l'effet de la sélection naturelle, les mutations avantageuses tendent à être sélectionnées et fixées dans la population (sélection positive) alors qu'au contraire les mutations délétères tendent à être éliminées (sélection négative). Les mutations neutres ne devraient pas être affectées par la sélection. La dérive génétique est un processus différent dans le sens où les fréquences des mutations neutres, délétères et avantageuses fluctuent aléatoirement dans une population au cours du temps. Dans

une population réelle, la sélection naturelle et la dérive génétique influent sur les fréquences alléliques.

Les mutations au niveau des acides nucléiques peuvent être distinguées en deux classes. En raison de la redondance du code génétique, certaines mutations n'entrainent pas de changement de l'acide aminé encodé. C'est ce qu'on appelle les mutations synonymes (ou silencieuses). A l'inverse, les mutations non-synonymes (ou remplacement) induisent un changement de l'acide aminé encodé par le codon muté. Par leur possible impact au niveau des séquences protéiques, les mutations non-synonymes ont plus de chance de se révéler avantageuses ou délétères et seront plus sensibles à la sélection. En considérant que les mutations surviennent de manière aléatoire, sans pression de sélection, on devrait observer une même proportion de mutations synonymes et non-synonymes. Sous sélection négative ou neutre, les mutations non-synonymes délétères devraient s'accumuler moins vite que les mutations synonymes. A l'inverse, sous sélection positive, les mutations non-synonymes avantageuses seront favorisées par rapport aux mutations synonymes. Ainsi, l'étude du taux relatif des mutations non-synonymes (dN ou β) et synonymes (dS ou α) est un outil puissant pour caractériser les pressions de sélection qui s'exercent sur des séquences génétiques codantes.

Le ration ω=dN/dS est une mesure usuelle des pressions de sélection (Motoo Kimura 1977). Si ce ratio est supérieur à 1 cela signifie que des pressions de sélection positives sont à l'œuvre, s'il est inférieur à 1 des pressions négatives, et égal à 1, de l'évolution neutre. Toutefois, ω = 1 peut aussi cacher une alternance de sélection négative et positive dans le temps.

Le terme sélection positive peut être employé pour désigner deux processus évolutifs : la sélection directionnelle et la sélection de diversification (*diversifying selection*). La sélection directionnelle s'applique quand, à une position, des mutations vers un acide aminé en particulier vont être sélectionnées. Ce phénomène survient par exemple quand un virus est soumis au même traitement antiviral et que les mêmes mutations vont être sélectionnées (Frost et al. 2000). On parle de sélection de diversification quand plusieurs acides aminés vont être maintenus dans la population. Cela est notamment le cas chez certains virus quand des positions sont la cible du système immunitaire (Moore et al. 2002) et que le virus évolue par la sélection de différents variants d'échappement immunitaires.

Différentes méthodes implémentées dans différents logiciels permettent d'estimer le ratio ω=dN/dS à partir de différents modèles de substitutions de codons. Dans l'hypothèse la plus simple et la plus globale, les taux de mutations synonymes et non-synonymes ne varient pas entre les sites ou le long de l'arbre (Goldman et Yang 1994). Le ratio est donc homogène selon les différents sites ou les différentes branches. En réalité, tous les sites ne sont pas soumis aux mêmes pressions de sélection et seulement certains sites sont sous sélection positive. Des modèles de codons permettant de faire varier ω selon les différents sites ont ainsi été proposés (Nielsen et Yang 1998; Y. Suzuki et Gojobori 1999; Z. Yang 2000; Yoshiyuki Suzuki 2004; Pond et Muse 2005). De la même manière, les pressions de sélection ne sont probablement pas constantes au cours du temps (et donc selon les différentes branches). Des modèles complexes permettent ainsi de faire varier le ratio ω à la fois entre les sites et au sein des branches (Ziheng Yang et Nielsen 2002; Zhang, Nielsen, et Yang 2005; Pond et al. 2006; Murrell, Wertheim, et al. 2012).

Parmi les implémentations logicielles pour estimer la sélection, on peut citer PAML (Yang 1997; Yang 2007) avec notamment le programme CODEML. Une autre implémentation des modèles de substitutions de codons est le package HYPHY (Pond, Frost, et Muse 2005) et sa version en ligne DATAMONKEY (Pond et Frost 2005). J'ai utilisé dans cette thèse différents programmes du package HYPHY notamment en raison de leur rapidité d'exécution sur de grands jeux de données.

1.3 VIH, VHC ET MUTATIONS DE RESISTANCE

Dans les chapitres suivants de cette thèse je me suis particulièrement intéressée à l'étude du VIH et du VHC afin d'identifier des mutations de résistance, préalablement connues ou non.

Ces deux virus présentent des taux de mutation élevés, et circulent sous la forme de nuages de variants. Ils sont tous les deux responsables de millions d'infections à travers le monde, et il n'existe pas de vaccin pour les traiter. Le recours aux traitements antiviraux est donc à ce jour la seule indication thérapeutique.

1.3.1 Le VIH

Le VIH est l'agent responsable du syndrome d'immunodéficience acquise ou SIDA qui entraîne chez les patients infectés un affaiblissement du système immunitaire et une vulnérabilité accrue aux infections opportunistes. D'après ONUSIDA[9], on estime que depuis le début de l'épidémie de sida, le VIH a infecté 79 millions de personnes et causé le décès d'environ 36 millions de personnes. Bien que l'on observe un ralentissement de l'épidémie depuis un pic mondial d'infection en 1997, encore 37 millions de personnes vivaient avec le VIH en 2020, dont 1,5 millions de personnes nouvellement infectées. Même si ce chiffre est le plus bas enregistré depuis 1990, il reste loin de l'objectif fixé par l'OMS des 500'000 nouvelles infections par an (Organisation mondiale de la Santé 2016). Le VIH a par ailleurs causé le décès de 680'000 personnes en 2020. L'épidémie est aujourd'hui majoritairement présente en Afrique subsaharienne qui concentre 67% des personnes vivant avec le VIH. En France, on compte un peu plus de 6'000 personnes nouvellement infectées chaque année, dont la moitié provient de personnes nées à l'étranger.

1.3.1.1 Découverte du VIH et manifestations cliniques

Le SIDA a été cliniquement décrit pour la première fois en 1981 (Centers for Disease Control (CDC) 1981) mais la présence du VIH chez l'homme est antérieure. Des analyses rétrospectives ont permis d'identifier le VIH dans des échantillons prélevés chez des patients dès 1959 au Congo et dans les années 1970 en Norvège où un marin aurait développé le SIDA en 1966 après avoir été infecté lors d'un voyage au Cameroun en 1961 et 1962 (Lindboe et al. 1986; Land et al. 1988).

Le VIH a été pour la première fois isolé et séquencé en 1983 par Luc Montagnier et Françoise Barré-Sinoussi à l'Institut Pasteur en pensant qu'il pouvait être l'agent causatif du SIDA (Barre-Sinoussi et al. 1983). Ils le nommèrent lymphanodenopathy virus (LAV). En 1984, Robert Gallo et son équipe identifiaient également le VIH, nommé alors en anglais human T-lymphotropic virus type 3 (HTLV-III) (Popovic et al. 1984). Après s'être rendu compte que le LAV et le HTLV-III étaient le même virus, le nom VIH fut décidé par une commission internationale en 1986. La même année, un autre type de VIH, également responsable du SIDA (nommé VIH-2), est découvert en Afrique de l'Ouest (Clavel et al. 1986). Le VIH découvert en 1983 et responsable de la pandémie mondiale est donc le VIH-1.

Le VIH est un rétrovirus à ARN du genre des lentivirus (« lent ») avec une période d'incubation longue. Il infecte des cellules du système immunitaire et préférentiellement les lymphocytes T CD4+ et des macrophages. Il se transmet par différents fluides corporels tels que le sang, le lait,

[9] https://www.unaids.org/fr/resources/fact-sheet

les sécrétions vaginales ainsi que le sperme. Une infection au VIH se déroule en 3 phases : la phase de primo-infection, la phase asymptomatique et la phase SIDA. La phase de primo-infection a lieu dans les premières semaines suivant une infection au VIH, durant laquelle des symptômes analogues à la grippe peuvent survenir. Ensuite, vient une phase asymptomatique parfois longue de plusieurs années, durant laquelle les personnes infectées sont contagieuses, avec un système immunitaire se dégradant progressivement mais sans manifestation clinique grave. Sans traitement, l'état de santé de la personne se dégrade, entrainant des symptômes tels que perte de poids, fièvre, toux etc. jusqu'à entrer en phase SIDA, 2 à 15 ans après l'infection. La destruction des cellules immunitaires empêche en effet l'organisme de se défendre normalement, laissant apparaître des maladies opportunistes causées par des bactéries, champignons, parasites et certains cancers.

1.3.1.2 Diversité génétique du VIH

A ce jour, le VIH est subdivisé en deux types distincts, le VIH-1 et le VIH-2. Le VIH-1 est le plus répandu géographiquement et représente la grande majorité des infections, alors que le VIH-2 est majoritairement présent en Afrique de l'Ouest avec des taux de transmission plus faibles. Les VIH-1 et VIH-2 ont été transmis des primates à l'homme au cours de plusieurs transmissions zoonotiques depuis des virus de l'immunodéficience simienne (VIS). Ces transmissions indépendantes ont engendré l'émergence de différentes lignées de VIH : les groupes M, N, O et P du VIH-1 et les groupes A à H du VIH-2 (Hemelaar 2012).

Des études de phylogénétique moléculaire (voir Figure 10) montrent que les groupes M et N du VIH-1 ont émergé indépendamment de VIS infectant des chimpanzés de l'espèce *Pan troglodytes troglodytes* alors que les groupes O et P sont plus proches de VIS de gorilles (eux même dérivés de l'espèce *Pan troglodytes troglodytes)* (Takehisa et al. 2009). Les groupes A à H du VIH-2 sont apparentés à des VIS trouvés chez le singe vert mangabey. Les VIH-1 et VIH-2 ont donc des origines géographiques différentes.

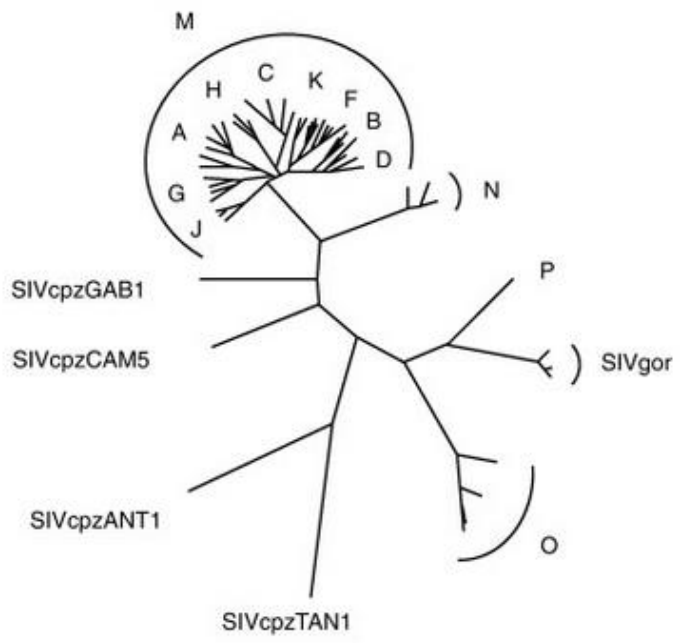


Figure 10 : Phylogénie des différents clades du VIH-1 et de plusieurs SIV.
Source : (Hemelaar 2012).

Au sein du VIH-1, les différents groupes ne présentent pas la même répartition et infectent des tailles de population très variables. Les groupes N, O et P ont peu diffusés en dehors du Cameroun et infectent un nombre restreint d'individus. Le groupe P, a été par exemple identifié uniquement chez quelques personnes au Cameroun (Simon et al. 1998; Vallari et al. 2011; Alessandri-Gradt et al. 2018). Le groupe M représente plus de 99% des infections du VIH-1. Originaire également du Cameroun, il se serait diffusé mondialement par l'actuelle Kinshasa dans la république démocratique du Congo (Vidal et al. 2000; Keele et al. 2006). On estime que trois facteurs principaux ont permis la propagation exponentielle du VIH-1 : l'utilisation de seringues infectées notamment dans les centres de santé (Marx, Alcabes, et Drucker 2001) , la prostitution (Faria et al. 2014; Pépin 2021) et le commerce de poches de plasma dont certaines contaminées (Volkow, Lopez, et Torres 1997).

Au cours de sa diffusion, le groupe M a évolué et présente la plus importante diversité génétique au sein du VIH-1. Il est subdivisé en 9 sous-types connus (A-D, F-H, J et K). Le sous-type C est le sous-type avec la plus forte prévalence et représente 46% des infections mondiales. Comme on peut le voir Figure 11, il est majoritaire en Inde, Afrique du Sud et dans la corne de l'Afrique. Le sous-type B est majoritairement trouvé en Europe occidentale, Amérique et Australie. Alors qu'il ne représente que 10% des infections mondiales, la majorité des études cliniques sur le VIH-1 sont sur ce sous-type. La région géographique avec la plus grande diversité génétique est l'Afrique subsaharienne et notamment le Cameroun et la république démocratique du Congo d'où le VIH-1 est originaire (McCutchan 2006; Kilmarx 2009; Hemelaar 2012).

En plus des sous-types principaux du VIH-1, on trouve aussi de nombreux virus recombinants avec une incidence variable selon la zone géographique (Simmonds et al. 1991; Hemelaar et al. 2020). Si un même virus recombinant est séquencé chez au moins trois patients dont les infections sont indépendantes, et que ce virus est donc devenu une souche épidémique on parle de forme recombinante circulante (CRF pour *circulating recombinant form*) (Robertson et al. 2000).

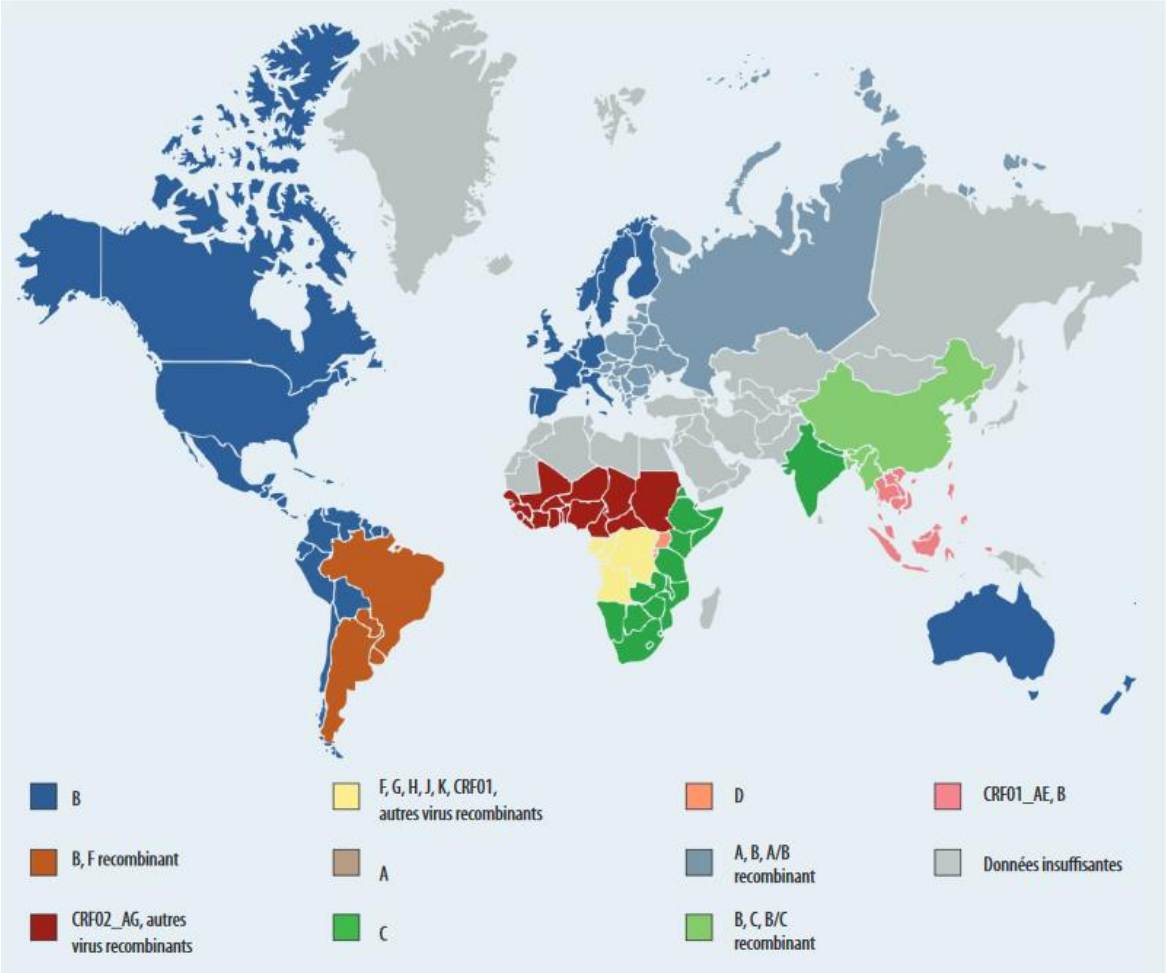


Figure 11 : Répartition dans le monde des principaux sous-types et formes recombinantes du VIH-1.
Source : https://www.edimark.fr/Front/frontpost/getfiles/17903.pdf

1.3.1.3 Structure et génome

Le VIH, représenté Figure 12, est un virus enveloppé, son enveloppe étant constituée de fragments membranaires de la cellule infectée associée à des glycoprotéines virales (gp120 et gp41) nécessaires à l'entrée du virus dans la cellule. L'enveloppe protège une capside composée d'un assemblage de la protéine p24. La capside renferme deux copies d'un brin d'ARN simple brin à polarité positive ainsi que des enzymes essentielles à la réplication du VIH : la transcriptase inverse, l'intégrase et la protéase. Entre l'enveloppe et la capside se trouve une matrice de protéines virales p17 qui assure l'intégrité du virion (« Human Immunodeficiency Virus (HIV) » 2016).

Le génome du VIH est long d'environ 9719 nucléotides (Wain-Hobson et al. 1985; Ratner et al. 1985). La Figure 13 illustre qu'il est composé d'une zone codante flanquée de deux zones non codantes aux extrémités 5' et 3' qui forment les séquences terminales longues répétées (LTR pour *long terminal repeat*). Ces séquences répétées sont essentielles à l'intégration du génome viral dans l'ADN de l'hôte et à sa transcription.

La partie codante du génome est composé de 9 gènes : *gag*, *pol*, *env*, *tat*, *rev*, *nef*, *vpr*, *vif* et *vpu* (ou *vpx* pour le VIH-2) qui codent 15 protéines (Frankel et Young 1998). Les gènes *gag*, *pol* et *env* sont des gènes de structure, communs à tous les rétrovirus. Les autres gènes codent des protéines

régulatrices ou accessoires. L'existence d'un dixième gène, *asp,* codant une protéine antisens et associée au développement épidémique du VIH-1 a été démontré par des méthodes bioinformatiques (Cassan et al. 2016).

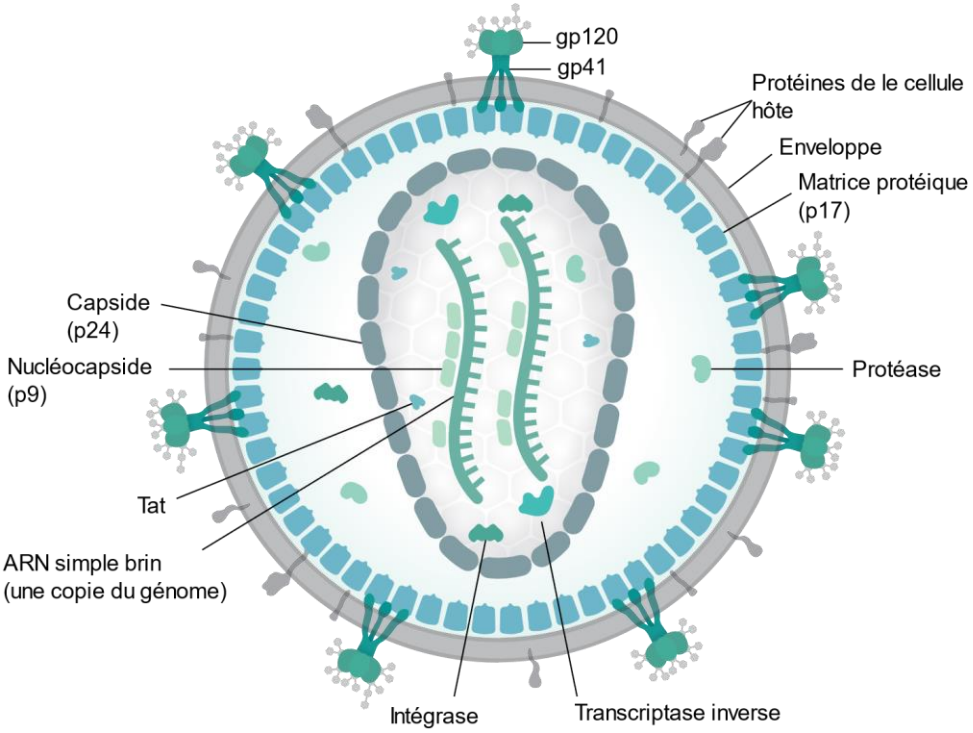


Figure 12 : Illustration d'un virion du VIH-1 avec les différentes protéines le composant.
Adaptée de Thomas Splettstoesser (www.scistyle.com) – sous licence commune CC BY-SA 4.0.

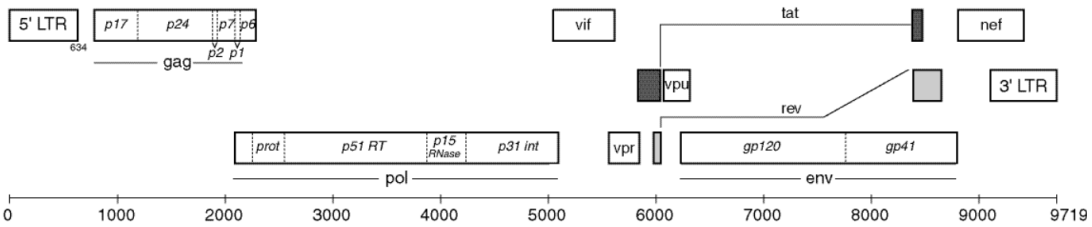


Figure 13 : Composition du génome du VIH-1 d'après le génome de référence HXB2.
Les différents gènes et les protéines codées sont représentés selon les différents cadres de lecture (1er en haut et 3eme en bas). La barre d'échelle représente les positions nucléotidiques le long du génome.
Source : https://www.hiv.lanl.gov/content/sequence/HIV/MAP/landmark.html

1.3.1.4 Le gène *pol*

Le gène *pol* est d'abord traduit sous la forme d'une polyprotéine Gag-Pol car les gènes gag et pol sont chevauchant. Cette polyprotéine contient des protéines non encore matures : protéase, reverse transcriptase, RNase et intégrase. C'est la protéase une fois associée en dimère qui va venir cliver les différents points de cassure de la polyprotéine et ainsi permettre de produire les protéines matures fonctionnelles (Brik et Wong 2003). Cette association en dimère de la protéase est réalisée par le précurseur de la protéase lui-même par un procédé dit d'auto-traitement (Louis, Clore, et Gronenborn 1999).

Le gène *pol* code 3 protéines : La reverse transcriptase, l'intégrase et la protéase. La reverse transcriptase convertit (rétrotranscrit) l'ARN viral simple brin en ADN double brin. L'ADN rétrotranscrit est ensuite intégré par l'intégrase à l'ADN de la cellule-hôte. L'ADN viral peut soit rester dormant dans le noyau, soit être transcrit en ARNm et traduit par la cellule hôte en polyprotéine Gag-Pol, elle-même clivée par la protéase, etc.

Ces 3 enzymes sont spécifiques du virus et jouent un rôle essentiel dans sa réplication. En raison de leurs spécificités elles sont la cible de nombreuses molécules antivirales.

1.3.1.5 Mécanismes d'action des antiviraux

Il n'existe à ce jour pas de traitement qui permette d'éliminer durablement et totalement le VIH de l'organisme d'un patient infecté. Les antiviraux (ou antirétroviraux) du VIH permettent de réduire la charge virale durablement à des niveaux où le virus n'est plus détectable dans le sang (<50 copies par ml) (Vergidis, Falagas, et Hamer 2009; Collaboration 2010). Il a par ailleurs été démontré qu'un patient traité efficacement par thérapie antirétrovirale avec une charge virale en dessous des seuils de détection ne peut plus transmettre le virus lors de rapports sexuels même non protégés (Donnell et al. 2010; Cohen et al. 2011; Rodger et al. 2019).

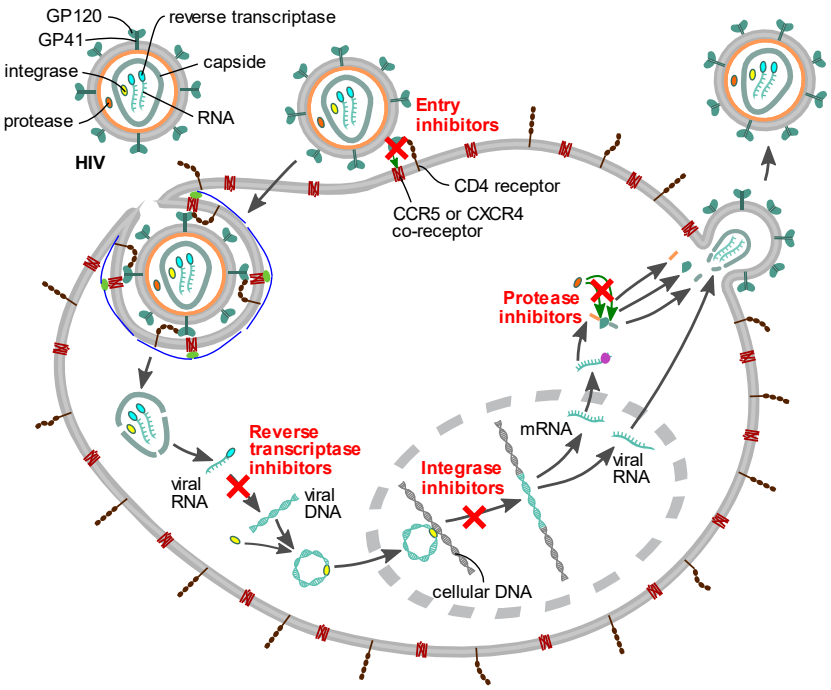


Figure 14 : Cycle de réplication du VIH-1 montrant les mécanismes d'action des différentes classes d'antiviraux. Source : Thomas Splettstoesser

Il existe différentes classes d'antirétroviraux en fonction de la phase où ils interviennent dans le cycle de réplication du VIH représenté Figure 14:

- Avant l'entrée du virus dans la cellule hôte, les inhibiteurs de fusion, et les inhibiteurs d'entrée vont empêcher le virus de se lier aux récepteurs CCR5 et CD4 et de fusionner avec la membrane de la cellule. Ils peuvent pour cela se fixer aux protéines de l'enveloppe virale ou aux récepteurs cellulaires. La liaison au récepteur CCR5 est une fixation non

compétitive qui va entrainer une modification conformationnelle empêchant la liaison de la gp120.

- Les inhibiteurs nucléosidiques de la transcriptase inverse (NRTI pour *nucleosidic reverse transcriptase inhibitor*) sont des analogues nucléosidiques et nucléotidiques qui inhibent la transcription inverse. Ils entrent en compétition avec le substrat de cette enzyme et empêchent l'élongation du brin d'ADN rétrotranscrit. Ils sont intégrés au brin d'ADN lors de la transcription et empêchent l'incorporation de nouvelle bases azotées.
- Les inhibiteurs non nucléosidiques de la transcriptase inverse (NNRTI pour *non nucleosidic reverse transcriptase inhibitor*) sont des inhibiteurs non compétitifs. Ils se lient près du site actif de la transcriptase inverse, changeant sa conformation et gênant la manipulation des nucléotides, rendant l'enzyme inefficace.
- Les inhibiteurs de la protéase empêchent le clivage des polyprotéines Gag et Gag-Pol et donc la formation des peptides matures. Les particules virales produites sont alors immatures et non-infectieuses.
- Les inhibiteurs de l'intégrase bloquent l'intégration de la molécule d'ADN viral dans le génome de la cellule cible.

Le premier antirétroviral à avoir été approuvé est un NRTI, l'azidothymidine (AZT) en 1987 (Fischl et al. 1987). Cependant dès 1989, des études ont commencé à observer une diminution de la sensibilité du virus chez des patients traités à l'AZT depuis plusieurs mois (Larder, Darby, et Richman 1989). Par la suite, d'autres molécules se sont avérées moins efficaces quelques mois après leur introduction, synonyme d'un échappement du virus au traitement et donc un échec thérapeutique jusqu'au décès du patient. Ce phénomène de résistance est lié à l'apparition de mutations dans les protéines ciblées et qui rendent les inhibiteurs inefficaces (Meyer et al. 1999). L'usage des antirétroviraux à cette époque était donc très limité et on attendait que le nombre de cellules de l'immunité atteigne un seuil critique pour les administrer.

En 1996, les traitements modernes voient le jour avec la trithérapie (Palella et al. 1998) qui est une combinaison de trois antirétroviraux incluant une nouvelle classe de molécules, les inhibiteurs de la protéase. En ciblant différentes étapes du cycle du VIH, les trithérapies sont beaucoup moins sensibles aux mutations de résistance. En effet, la probabilité qu'apparaissent dans un même virion 3 mutations simultanées qui protègent des différents antirétroviraux est très faible. Il existe aujourd'hui une quarantaine de molécules antirétrovirales qui sont administrées en trithérapies dès le début de l'infection[10]. Ces traitements ont permis de sauver des millions de personnes mais ne sont pas encore accessibles à toutes les personnes séropositives notamment dans certains pays d'Afrique de l'Ouest (UNAIDS, 2016).

1.3.1.6 Mutations de résistance aux traitements

Le VIH est un des virus avec un des taux de mutation les plus élevé et même si les mutations avantageuses sont très rares, elles peuvent survenir et conduire à la résistance du virus aux traitements antirétroviraux. Toutes les molécules antirétrovirales sont susceptibles de perdre en efficacité voire devenir inactive en raison de mutations de résistance. La résistance aux antirétroviraux est un phénomène préoccupant pour plusieurs raisons. Tout d'abord, en cas d'échec thérapeutique, le patient traité redevient contagieux et peut alors propager des variants

[10] https://hivinfo.nih.gov/understanding-hiv/fact-sheets/fda-approved-hiv-medicines

résistants du virus limitant alors le choix des molécules prescrites chez les nouveaux patients. De plus, le virus se multipliant à nouveau, l'état de santé du patient peut se détériorer jusqu'à la phase SIDA (François Clavel et Hance 2004; Pennings 2013; World Health Organization, Global Fund, et US Centers for Disease Control and Prevention 2017).

La résistance survient lorsque des variants dans la population virale comportant des mutations sur les gènes ciblés par les antirétroviraux (apparues par hasard) sont sélectionnés : mutations sur la protéase, l'intégrase, la transcriptase inverse ou des protéines d'enveloppe. Elles sont donc différentes en fonction du type de molécule antivirale utilisée (Clutter et al. 2016).

Les mutations de résistance peuvent changer la conformation des protéines virales, les rendant souvent moins efficaces ou moins « fit » (Quiñones-Mateu et Arts 2002). Elles sont donc souvent accompagnées de mutations compensatoires ou accessoires qui rétablissent l'efficacité des protéines mutées (Nijhuis, Deeks, et Boucher 2001). Certaines mutations de résistance surviennent seules et suffisent à limiter l'efficacité des antirétroviraux (Turner, Brenner, et Wainberg 2003). Chez certains patients on observe une association de plusieurs mutations de résistance (Marcelin et al. 2005, 2019; Wensing et al. 2019; Boyer et al. 2022).

Dans ma thèse, je me suis particulièrement intéressée aux mutations de résistance de la transcriptase inverse. Elles sont de deux types : les mutations de résistance aux NRTI et les mutations de résistances aux NNRTI. Dans le premier cas, ces mutations empêchent les inhibiteurs nucléosidiques de se fixer à l'ADN en cours de transcription soit en l'excisant une fois intégré (mutations TAM pour thymidine analog mutations en anglais) soit en favorisant l'incorporation de nucléosides et nucléotides naturels. Les TAMs classiques (M41L, D67N, K70R, L210W, T215Y/F et K219Q/E) ont été découvertes rapidement après l'utilisation de l'AZT en monothérapie. Les autres mutations majeures aux NRTI et non TAMs sont les mutations M184V/I, K65R, L74V, Y115F et Q151M. Dans le second cas, les mutations de résistance empêchent la fixation des NNRTI à une poche de la transcriptase inverse proche du site actif. Ces mutations se trouvent aux positions 100, 101, 103, 106, 138, 181, 188, 190 et 230.

1.3.1.7 Implications des mutations de résistance

Les mutations de résistance peuvent survenir en réponse au traitement, notamment si l'observance n'est pas régulière. Dans certains pays d'Afrique subsaharienne l'accès au traitement peut être très difficile, avec des ruptures de stock fréquentes. Ces arrêts de traitement favorisent le développement de mutations de résistance (Ssemwanga et al. 2015).

On observe aussi des cas de résistance chez des personnes non traitées et n'ayant jamais été traitées. On estime que des résistances pré-traitement sont présentes chez 10% des personnes traitées pour la première fois. Notamment des mutations de résistance sont trouvées chez des enfants lors de transmission de la mère à l'enfant. Des sondages effectués en Afrique subsaharienne ont montré que la moitié des enfants diagnostiqués comme porteur du VIH portaient un virus résistant aux NNRTI classiques[11].

Lorsqu'un patient est détecté comme positif au VIH, un dépistage d'éventuelles mutations de résistance est réalisé par séquençage. Des algorithmes ont d'ailleurs été développé afin de tester les implications phénotypiques des mutations sur les différents génotypes du VIH-1 comme

[11] https://www.afro.who.int/news/who-unveils-plan-tackle-rising-hiv-drug-resistance-africa

Geno2pheno (Beerenwinkel et al. 2003). En fonction des résistances détectées, les combinaisons de molécules antirétrovirales sont adaptées pour la trithérapie. Les mutations de résistances ont été extensivement étudiées et sont recensées notamment dans la base de données de résistance du VIH, développée et maintenue par l'université de Stanford (Rhee 2003).

1.3.2 Le VHC

Le VHC est un agent viral responsable de plusieurs maladies chroniques hépatiques. Après infection, le virus devient persistant chez 75 à 85% des malades et infecte aujourd'hui environ 58 millions de personnes à travers le monde d'après l'organisation mondiale de la santé[12] (OMS). Le VHC constitue un problème de santé publique majeur, causant environ 1.5 millions de nouvelles infections et le décès de 290'000 personnes en 2019. Récemment, le développement de traitements antiviraux a permis l'élimination du VHC chez plus de 95% des patients traités (Falade-Nwulia et al. 2017). Grâce à ces traitements, l'OMS a fixé l'objectif de réduire les infections au VHC de 90% et de diminuer le nombre de décès de 60% par rapport à 2015 d'ici 2030 (World Health Organization 2016).

1.3.2.1 Découverte du VHC et manifestations cliniques

Les hépatites sont des inflammations des cellules du foie pouvant avoir plusieurs causes dont des infections virales. Le VHC n'est pas le seul virus pouvant causer des hépatites et avant son identification, les virus de l'hépatite A (Feinstone, Kapikian, et Purcell 1973) et de l'hépatite B (Blumberg et Alter 1965) avaient déjà été identifiés.

En 1975, des patients suivis pour des hépatites post transfusionnelles se révèlent négatifs aux virus de l'hépatite A et B laissant penser à l'existence d'un autre agent infectieux encore inconnu (Feinstone et al. 1975). En 1978, l'agent responsable des hépatites dites « non-A, non-B » était démontré comme étant probablement un virus (Tabor et al. 1978). Le VHC sera finalement identifié comme l'agent causatif des hépatites non-A non-B, alors renommées hépatites C, en 1989 (Choo et al. 1989; Kuo et al. 1989). Il s'agit d'un virus de la famille Flaviviridae et du genre des *Hepacivirus* au sein duquel se trouvent des espèces virales infectant plusieurs vertébrés (canidés, chevaux, chauve-souris, rats, certains primates…).

Le VHC peut causer des infections aigües ainsi que des infections chroniques. La plupart des infections aigües sont asymptomatiques et guérissent spontanément dans les 6 mois chez 25% des personnes infectées (Marcellin 1999; Orland, Wright, et Cooper 2001). Cette phase asymptomatique rend difficile l'estimation du nombre exact de personnes infectées et favorise la propagation du virus. Pour le reste des personnes infectées, la maladie évolue dans une forme chronique pouvant causer cirrhose du foie voire cancer hépatique. Ce virus est transmis par voie sanguine, notamment lors de transfusion de produits sanguins non testés, matériel d'injection non stérilisé, usage de drogues par injection…

1.3.2.2 Diversité génétique du VHC

Le VHC présente une grande variabilité génétique et compte 8 différents génotypes caractérisés ainsi qu'une subdivision en plus de 90 sous-types (https://talk.ictvonline.org/ictv_wikis/flaviviridae/w/sg_flavi/56/hcv-classification, 2019). Des

[12] https://www.who.int/news-room/fact-sheets/detail/hepatitis-c

souches de génotypes différents ne partagent pas plus de 65-70% de similarité et des séquences appartenant au même sous-type partagent plus de 85% de similarité (D. B. Smith et al. 2014; C. Li et al. 2015; Charlotte Hedskog et al. 2019). L'annotation des sous-types se fait par l'ajout d'une lettre (par ordre alphabétique) après le numéro du génotype.

La prévalence des différents génotypes et sous-types varie selon les régions du monde et seule une petite fraction de sous-types (1a, 1b, 3a et 2a) est responsable de la grande majorité des infections. Ces sous-types dits « épidémiques » sont répandus sans distinction géographique alors que d'autres sous-types dits endémiques sont restreints à des régions particulières où ils représentent localement la majeure partie des infections. Ainsi, le génotype 4 est majoritairement présent en Afrique centrale et au Moyen Orient, le génotype 5 en Afrique du Sud et le génotype 6 en Asie du Sud-Est (Gower et al. 2014; Petruzziello et al. 2016; Blach et al. 2017; Shenge, Odaibo, et Olaleye 2019).

La propagation des sous-types épidémiques est récente (19ème siècle) et serait due à des contaminations parentérales (usage de seringues non désinfectées, transferts de sang contaminé…) (Peter Simmonds 2013).

Des analyses de datation moléculaire estiment une origine commune des différents génotypes du VHC il y a 3000 ans, les plus vieux génotypes se trouvant en Asie. Cela est concordant avec la grande diversité de génotypes et de sous-types observés en Asie, le génotype 6, endémique en Asie du Sud-Est, étant le plus variable avec 31 sous-types reconnus à ce jour (Forni et al. 2018). Les origines exactes du VHC sont encore floues mais on suppose qu'il provient d'une infection intra-espèce unique et qu'il s'est ensuite diversifié au sein de la population humaine. En effet, des hepacivirus ont été trouvés chez plusieurs mammifères (Drexler et al. 2013; Quan et al. 2013; Gemaque et al. 2014; Baechlein et al. 2015; El-Attar et al. 2015), les plus proches du VHC se trouvant chez les équidés et les canidés (Pybus et Thézé 2016). Ces derniers auraient toutefois émergé il y a environ 1000 ans (Forni et al. 2018) et l'organisme à partir duquel le VHC a émergé reste donc à déterminer.

1.3.2.3 Structure et Génome

Le VHC est un virus à ARN simple brin de polarité positive. Il possède une enveloppe lipidique qui entoure sa capside icosaédrique. Des protéines d'enveloppe ancrées dans la structure lipidique assurent la reconnaissance des particules virales par les récepteurs CD81 présents dans certaines cellules du foie. Au sein de la capside se trouve le génome à ARN, d'environ 9600 nucléotides (Choo et al. 1989). Il code une polyprotéine de 3000 acides aminés qui est ensuite clivée en dix protéines comme représenté Figure 15 : la protéine de capside C, les protéines d'enveloppe E1 et E2, et des protéines non structurales impliquées dans la production des virions (p7, NS2) ou qui assurent les fonctions enzymatiques utiles au cycle viral (NS3, NS4A, NS4B, NS5A et NS5B) (Chevaliez et Pawlotsky 2006; Moradpour et Penin 2013). En plus de la partie codante, on trouve les régions 5' et 3' non codantes qui sont fortement conservées et qui jouent un rôle dans la reconnaissance de l'ARN viral par les ribosomes (Jubin 2001) ainsi que dans sa traduction et sa transcription (Yi et Lemon 2003).

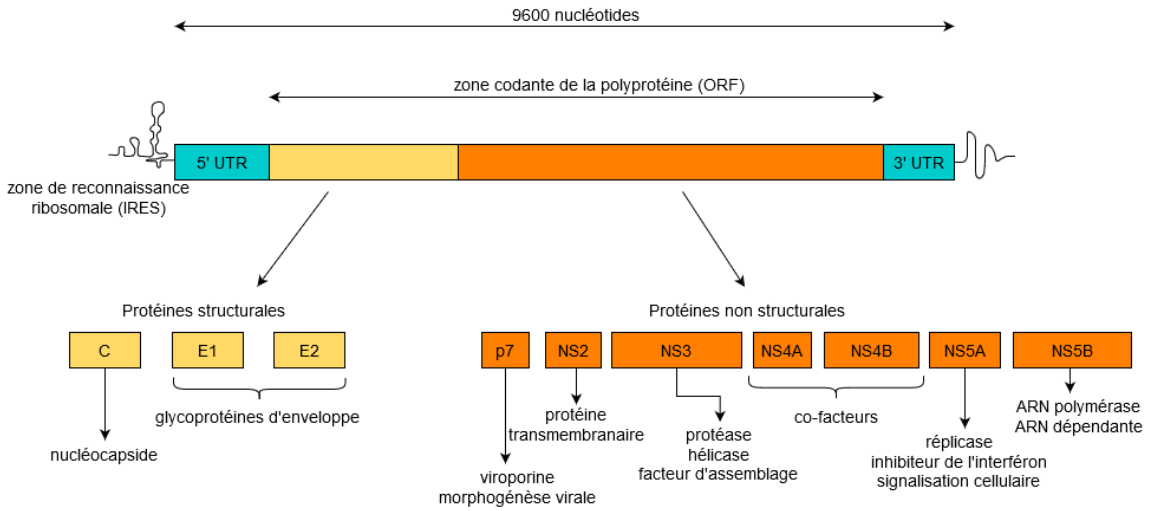


Figure 15 : Représentation simplifiée du génome du VHC et des 10 protéines exprimées.

1.3.2.4 Mécanismes d'action des antiviraux et résistance

Depuis 2011, le VHC est traité par des antiviraux à action directe (AADs) qui ont prouvé une efficacité supérieure à 95% chez la plupart des génotypes (Falade-Nwulia et al. 2017) notamment grâce à l'utilisation du Sofosbuvir depuis 2013 (Lawitz et al. 2013). Cette efficacité est mesurée par la réponse virologique soutenue plusieurs semaines après arrêt du traitement. Cela signifie que le virus a été éliminé de l'organisme et que les patients sont guéris, contrairement au VIH pour lequel il n'existe pas de cure. Les AADs agissent sur plusieurs étapes dans le cycle de réplication du VHC, en particulier en ciblant les protéines NS3, NS5A et NS5B. Le VHC évolue sous la forme d'une population virale et des mutations de résistance peuvent survenir sur chacune des protéines ciblées par les traitements. Elles peuvent alors causer un échappement du virus au traitement et un échec thérapeutique (Lontok et al. 2015; Wyles et Luetkemeyer 2017; Sorbo et al. 2018; Sarrazin 2021). L'utilisation des AADs en polythérapie réduit considérablement l'impact des mutations de résistance et des patients qui présentent des mutations de résistance avant traitement peuvent tout de même atteindre une réponse virologique soutenue. Au cours de ma thèse je me suis intéressée aux mutations de résistance des protéines NS5A et NS5B.

1.3.2.4.1 Inhibiteurs de NS5A et résistance

La protéine NS5A est une phosphoprotéine virale dont le rôle complet au sein de la réplication du VHC n'est pas encore totalement élucidé (Macdonald et Harris 2004; Fridell et al. 2011). On sait toutefois qu'elle est indispensable dans la régulation de la réplication du virus, son assemblage et sa sortie des cellules. Il existe plusieurs inhibiteurs de la NS5A qui ont prouvé un fort pouvoir antiviral ainsi qu'une efficacité clinique très élevée notamment en association avec d'autres antiviraux. Alors qu'une substitution est usuellement considérée comme conférant de la résistance si la dose curative est multipliée par 2.5, certaines mutations sur la NS5A peuvent créer un changement du facteur de résistance par 100. Sur le génotype 1, il s'agit de mutations aux positions 28,30,31 et 93 (Lontok et al. 2015; Sorbo et al. 2018; Perales et al. 2018). Il a été démontré que les mutations de résistance de la NS5A pouvaient persister au sein de la population virale intra-hôte jusqu'à plusieurs années après arrêt du traitement, pouvant affecter les chances d'un retraitement en cas d'échec (Wyles, Mangia, et al. 2017).

Les trois prochains paragraphes présentent un bref état des lieux des mutations de résistance chez le VHC. Je recommande aux lecteurs la lecture de ces revues (Lontok et al. 2015; Sorbo et al. 2018; Perales et al. 2018) pour des informations plus complètes.

1.3.2.4.2 Inhibiteurs de la polymérase et résistance

La protéine NS5B est la RdRp du VHC. Comme pour le VIH, il existe des inhibiteurs nucléos(t)idiques et non nucléos(t)idiques. Le Sofosbuvir est un analogue nucléos(t)idique qui cause un arrêt prématuré de la séquence nucléotidique. Le site actif de la NS5B est très conservé, ce qui fait que le Sofosbuvir fonctionne pour tous les génotypes. On dit qu'il a une efficacité pan-génotypique (Wu et al. 2019). L'autre inhibiteur de la NS5B, le Dasabuvir est un inhibiteur non compétitif qui se fixe sur des sites allostériques de la protéine (Kati et al. 2015). Cette fixation change la conformation de NS5B et empêche son activité polymérase. Il existe peu de mutations de résistance au Sofosbuvir sur la protéine NS5B en raison de leur impact sur l'activité polymérase (Lawitz et al. 2013; Sarrazin et al. 2016). La substitution S282T est la plus connue mais disparait rapidement en cas d'arrêt du traitement (C. Hedskog et al. 2015). Les mutations de résistance aux inhibiteurs non nucléos(t)idiques sont plus fréquentes.

Avant la découverte des AADs, le VHC était traité par ribavirine et interférons qui amélioraient la réponse virale dans la moitié des cas (Manns et al. 2001; Fried et al. 2002). Le développement des AADs, curatifs à 95% a créé un véritable élan dans le traitement du VHC. Cependant, l'objectif fixé par l'OMS de supprimer le VHC comme menace pour la santé d'ici 2030 est encore loin d'être atteint. En effet, en 2017, seul 20% des personnes infectées dans les pays à faible et moyen revenus sont diagnostiquées et parmi elles, l'accès au traitement reste difficile notamment en raison de prix parfois élevés[13].

[13] https://www.who.int/publications/i/item/9789240019003 page 7

2 ETUDE DE LA CONVERGENCE EVOLUTIVE CHEZ LES VIRUS

Dans le chapitre précédent, nous avons vu que les virus présentaient des taux de mutations sans équivalent chez les autres organismes, leur permettant de s'adapter rapidement à des contraintes environnementales fortes, jusqu'à même pouvoir échapper dans certains cas aux traitements antiviraux. Cette rapidité d'adaptation en fait un modèle de prédilection pour étudier les mécanismes sous-jacents aux différents processus évolutifs. De plus, grâce au séquençage à haut débit, il est possible de séquencer et suivre l'évolution des virus en temps réel que ce soit dans des conditions expérimentales ou réelles.

Parmi les mécanismes liés à l'adaptation, je me suis particulièrement intéressée ici à la convergence évolutive, qui est définie de manière générale comme l'acquisition indépendante de traits similaires dans des lignées distinctes au cours de l'évolution. La convergence évolutive est particulièrement intéressante car elle peut se traduire chez les virus directement par les mutations de résistance. En effet les mêmes mutations peuvent apparaitre indépendamment chez des virus soumis à une même contrainte : le traitement antiviral.

Dans ce chapitre, je définirai plus précisément la convergence évolutive, en me concentrant particulièrement sur la convergence moléculaire, qui nous intéresse spécifiquement dans le cas des virus. J'illustrerai mes propos avec plusieurs exemples de convergence. Je décrirai ensuite quelles méthodes ont été conçues pour détecter de la convergence évolutive puis finalement j'aborderai leurs limites et la nécessité d'une nouvelle méthode adaptée au cas des virus.

2.1 LA CONVERGENCE EVOLUTIVE : UN SIGNE D'ADAPTATION ?

La convergence évolutive a d'abord été identifiée au niveau phénotypique chez les animaux. Afin de mieux comprendre ses mécanismes, éloignons-nous quelque peu de l'étude des virus pour nous concentrer sur des organismes plus complexes et plus facilement observables.

2.1.1 Quand l'évolution se répète

Théorisée par Charles Darwin en 1859, la théorie de la sélection naturelle est le « processus qui, sous l'effet de conditions naturelles, assure la survie des êtres et des espèces les plus aptes à lutter pour leur existence dans un milieu donné et entraîne l'élimination plus ou moins complète de ceux qui sont moins bien adaptés »[14]. Lorsque deux espèces soumises à des mêmes pressions de sélection développent un phénotype/une fonction similaire/analogue de manière indépendante (sans qu'il ne soit hérité), on parle de convergence évolutive. La convergence évolutive peut alors être interprétée comme un signe d'adaptation en réponse à des contraintes environnementales fortes qui ont façonnées de la même manière des espèces différentes (Losos 2011; Stayton 2015; Storz 2016).

Il existe de multiples exemples où l'évolution se répète pour répondre à ce qui semble une même contrainte. Le cas d'école par excellence est sans doute l'apparition indépendante du vol chez les oiseaux, les chauves-souris et les insectes avec le développement de structures similaires : les ailes

[14] https://www.cnrtl.fr/definition/sélection

comme représentées Figure 16. Des cas de convergence sont aussi observées pour l'adaptation en milieu aquatique avec l'évolution progressive de morphologies hydrodynamiques chez les cétacés, les otaries, les poissons ou les oiseaux manchots. Pourtant ils ne partagent pas un même ancêtre commun qui leur aurait transmis cette morphologie en ogive permettant de réduire la force de trainée de l'eau. Ces deux exemples de convergences observées au niveau morphologique sont souvent cités car ils permettent facilement d'appréhender le concept de convergence. Pour autant, la convergence peut se retrouver sur un large éventail de traits et de caractéristiques, et peut concerner tout type d'organismes. J'aborderai dans la suite de ce chapitre des cas de convergence avec des modifications du métabolisme, l'acquisition de nouveaux organes ou de nouvelles fonctionnalités etc…

Figure 16: Exemples de morphologies convergentes : ailes d'insecte, de chauve-souris et d'oiseau.

Partant du principe que l'évolution pouvait se répéter, des auteurs se sont également interrogés sur la possibilité de prédire l'évolution (Stern, Orgogozo, et Rausher 2008; Stern et Orgogozo 2009; Bridgham 2016; Agrawal 2017). Si la réponse à cette question est hors de portée à l'heure actuelle, elle a le mérite de nous amener à mieux comprendre les chemins évolutifs et les mécanismes qui façonnent les espèces. En plus de satisfaire notre curiosité, cette question présente également un intérêt pour l'étude de l'évolution des virus. En effet, à l'heure où les virus émergents et ré-émergents posent des problèmes de santé publique à l'échelle mondiale, anticiper leur évolution est essentiel pour ajuster au mieux les mesures visant à les contrôler.

2.1.2 Les bases génétiques des convergences phénotypiques.

Grâce aux technologies de séquençage et à l'étude des données moléculaires, la compréhension des mécanismes qui sous-tendent les observations phénotypiques a été facilitée. En effet, les caractères morphologiques étant contrôlés par l'expression des gènes, la convergence phénotypique pourrait trouver son origine au niveau moléculaire. En étudiant les bases moléculaires de la convergence phénotypique, les chercheurs ont montré qu'un phénotype convergent pouvait être régit par : i) des gènes différents (Wittkopp et al. 2003; Fry et al. 2009; Steiner et al. 2009; Sauter et al. 2009; Kluge et al. 2014), ii) des groupes de gènes similaires (Larter et al. 2018), iii) plusieurs mutations au niveau du même gène (Zhang 2003; Li et al. 2008), voire iv) une ou plusieurs mutations ponctuelles identiques au sein du même gène (voir références Tableau 3). C'est ce dernier cas auquel je vais m'intéresser ici et pour lequel quelques exemples sont listés dans le tableau 3 ci-dessous.

Lorsque des mutations se produisent vers un même acide aminé (ou dans certains cas des acides aminés similaires (Rey et al. 2018; Besnard et al. 2009)) et se fixent de manière indépendante dans des lignées différentes, on parle de convergence moléculaire. Il en existe 3 types : les mutations parallèles, convergentes et les réversions. Comme illustré Figure 17, les mutations parallèles dérivent d'un même état ancestral, les mutations convergentes dérivent d'états ancestraux

différents et les réversions sont un retour vers un état ancestral perdu (Zhang et Kumar 1997). Certains considèrent que les mutations convergentes sont des meilleures preuves d'adaptation car elles sont moins probables que les mutations parallèles ou les réversions (J. Zhang et Kumar 1997; J. Zhang 2006). Cette distinction a sans doute peu d'importance (voir justification dans Arendt et Reznick (2008)) et je me suis intéressée dans mes travaux à l'étude des trois types de mutations. Je parlerai alors indistinctement de convergence moléculaire ou de mutations convergentes.

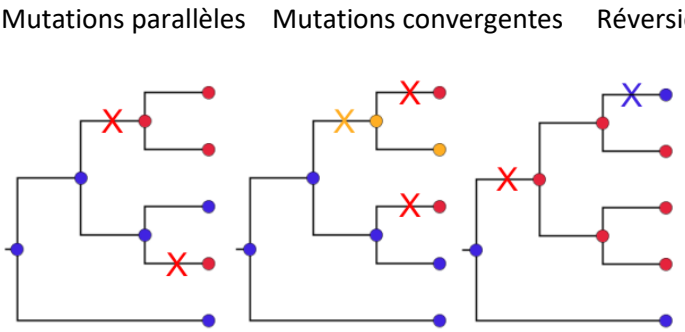


Figure 17 : Différents types de mutations de convergence moléculaire

Phénotype	Gène(s)	Positions concernées	Espèces concernées	Référence(s)
Animaux fermenteurs	Lysozyme stomachal	75 et 87	Langur, vache, hoazin	(J. Zhang et Kumar 1997)
Régime alimentaire à base de feuilles	RNases pancréatiques	4*, 6* et 39*	Colobes et langurs	(J. Zhang 2006)
Echolocation	Prestine	7*, 384*, 392* et 497*	Chauve-souris et cétacés	(Li et al. 2008; Li et al. 2010; Liu et al. 2010; Davies et al. 2012; Liu et al. 2014; 2018)
Vision en faible luminosité	Rhodopsine (RH1)	83*, 122*, 261* et 292*	Poissons	(Sugawara et al. 2005; Yokoyama et al. 2008; Yokoyama, Yang, et Starmer 2008)
Résistance insecticides	Pompe à sodium-potassium	111 et 122	Insectes se nourrissant de plantes toxiques	(Zhen et al. 2012; Dobler et al. 2012)
Métabolisme en C4	Phosphoénolpyruvate carboxylase	572, 665, 733, 761 et 780*	Graminées et carex	(Christin et al. 2007; Besnard et al. 2009)
Adaptation à l'hypoxie	Hémoglobine (sous unité β)	13 et 83	Colibris	(Projecto-Garcia et al. 2013)
Résistance aux toxines (hétérosides cardiotoniques)	Pompe à sodium-potassium	111 et 119	Insectes, amphibiens, reptiles et mammifères	(Ujvari et al. 2015)

Tableau 3 : Exemples de convergences moléculaires. Les positions suivies d'une étoile indiquent des confirmations expérimentales de l'effet des mutations sur le phénotype.

Un exemple de convergence moléculaire est illustré Figure 18, issue d'une étude de (Christin, Weinreich, et Besnard 2010). Dans cette étude, les auteurs se sont intéressés à l'émergence récurrente du métabolisme en C4 chez des graminées et ont cherché des mutations convergentes dans la phosphoénolpyruvate carboxylase (PEPC), une enzyme indispensable au métabolisme en C4. La gauche de la figure représente un arbre phylogénétique de plusieurs espèces de graminées dont les clades avec le métabolisme en C4 sont représentés en rouge. Les espèces avec le phénotype ancestral ont un métabolisme en C3 et sont représentées en noir. Des positions particulières de la PEPC sont associées à cette phylogénie. Ces positions ont été choisies car elles possèdent des mutations convergentes, corrélées avec le métabolisme en C4. Par exemple, à la position 572, l'acide aminé ancestral (en blanc) était un acide glutamique (E) et a été remplacé chez plusieurs des clades convergents par une glutamine (Q) surlignée en rouge. La position 780 est assez remarquable car tous les clades convergents possèdent une sérine (S) alors que les clades avec le phénotype ancestral ont tous une alanine (A). Une sérine serait donc apparue 8 fois indépendamment au cours de l'évolution des graminées à la position 780, à chaque fois dans les espèces avec le phénotype convergent et uniquement celles-ci (du moins sur ces données). Des expériences de mutagénèse ont d'ailleurs révélé que des mutations à la position 780 altéraient les propriétés catalytiques de la PEPC (Svensson, Bläsing, et Westhoff 2003).

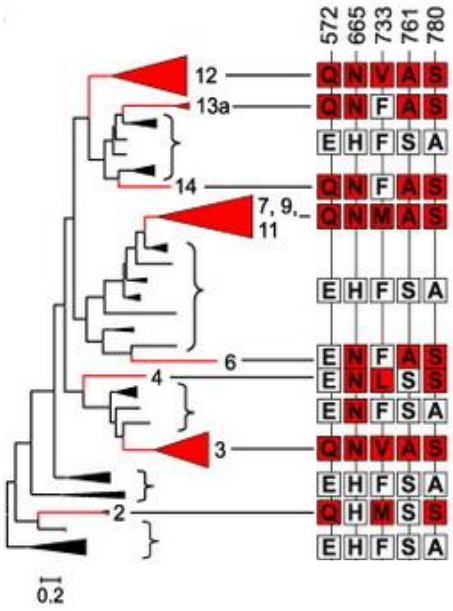


Figure 18 : Arbre phylogénétique de la PEPC chez les graminées avec les positions sous sélection positive en lien avec le métabolisme en C4.
Figure issue de (Christin, Weinreich, et Besnard 2010) Les gènes spécifiques du métabolisme en C4 sont en rouge dans l'arbre. Les lignées C4 sont numérotées selon (Christin et al. 2008). Les barres d'échelle représentent les substitutions attendues par site. Les acides aminés sous évolution convergente supposée sont surlignés en rouge.

2.1.3 Comment expliquer que l'évolution se répète ?

L'observation de convergence moléculaire a suscité de nombreux débats et interrogations, et notamment au sujet des facteurs qui expliquent l'émergence répétée de certaines mutations. Deux points de vue opposent un aspect mutationniste et sélectionniste de l'évolution (Lenormand, Chevin, et Bataillon 2016). Dans la vision sélectionniste, le principal moteur de l'évolution serait la sélection naturelle. Selon cette hypothèse, seules certaines mutations parmi

une multitude possible vont permettre de conduire au phénotype optimal pour une pression de sélection donnée. Par sélection naturelle, ces mutations particulières seront sélectionnées, pouvant ainsi potentiellement conduire à des convergences. La vision mutationniste considère qu'aux pressions de sélection s'ajoutent des contraintes sur l'émergence de mutations. Le nombre possible de mutations à une position donnée est en fait limité et joue donc un rôle dans l'apparition d'un caractère. Ainsi la principale force évolutive serait les mutations. En d'autres termes selon la vision mutationniste, des traits à première vue convergents pourraient émerger sans présence de contraintes environnementales mais en raison de forces s'appliquant sur les mutations (Nei 2005; Stoltzfus 2006). En réalité, les deux processus sont sans doute à l'œuvre, et observer des convergences moléculaires n'est pas suffisant pour invoquer un phénomène adaptatif à l'œuvre. Il ne fait aucun doute que des pressions fortes peuvent conduire à des mutations convergentes mais la réciproque n'est pas toujours vraie.

2.1.3.1 Pressions de sélection et sélection naturelle

La principale raison invoquée lors de l'étude de la convergence moléculaire est la sélection naturelle (Losos 2011). Dans l'hypothèse où plusieurs mutations sont équiprobables à une position donnée, si une mutation conduit à un phénotype beaucoup plus avantageux dans l'environnement alors les organismes porteurs de cette mutation engendreront sans doute plus de descendance. Cet effet conduit à la fixation de cette mutation dans la population et à terme l'adaptation de l'espèce. Plus les pressions de sélection seront fortes et plus une mutation avantageuse pour l'espèce aura tendance à se fixer rapidement dans la population. Si une même mutation a des effets similaires dans deux lignées distinctes, en présence de contraintes suffisamment fortes, la sélection naturelle conduit ainsi à de la convergence moléculaire.

Si l'effet de la sélection naturelle est bien accepté à des niveaux supérieurs d'organisation biologique, il n'est pas simple de démontrer qu'une mutation convergente est le résultat de la sélection naturelle. De nombreux autres effets peuvent influencer la probabilité de fixation d'une mutation plutôt qu'une autre.

2.1.3.2 Biais de mutation

Les mutations des séquences protéiques sont le résultat de mutations au niveau nucléotidique. Comme nous l'avons vu en introduction, les taux de substitution entre les différentes bases azotées (A, C, G et T ou U) ne sont pas les même et les transitions sont plus fréquentes que les transversions. Par exemple, dans les génomes de mammifères on trouve des taux importants pour les mutations C -> T et G -> A (qui sont des transitions). De plus, ce biais peut être accentué par la présence de dinucléotides CpG (un nucléotide cytosine suivi par une guanine) qui jouent notamment un rôle dans l'expression des gènes chez les eucaryotes (Deaton et Bird 2011). Les cytosines de ces dinucléotides peuvent être méthylées ce qui facilite leur mutation en thymines et augmente d'autant plus le taux de transitions C -> T. Il existe donc des biais de mutation au niveau des bases nucléotidique qui, en fonction du code génétique, peuvent conduire à l'usage de certains codons plutôt qu'à d'autres et conduire *in fine* à des mutations convergentes (Storz 2016; Stoltzfus et McCandlish 2017). Storz et al. (2019) ont d'ailleurs démontré un cas d'évolution convergente causée par des mutations dans les dinucléotides CpG dans l'hémoglobine de certains vertébrés adapté à la respiration en haute altitude.

Ces biais se révèlent d'ailleurs dans la composition en acides aminés des génomes et tous les acides aminés n'ont pas la même probabilité d'apparition. Différents modèles ont ainsi estimé,

sur de grands jeux de données, les fréquences d'équilibre des acides aminés (c'est-à-dire les fréquences des acides aminés après un temps d'évolution infini). On peut voir par exemple sur la Figure 19 que la leucine (L) est un des acides aminés les plus fréquents, que ce soit pour un modèle généraliste estimé sur 49'697 séquences protéiques (LG (Le et Gascuel 2008)), un modèle estimé sur 1'025 protéines du VIH (HIVb (Nickle et al. 2007)) ou un modèle estimé sur des protéines mitochondriales (MtVER (Le, Dang, et Le 2017)). A l'inverse les acides aminés cystéine (C) et tryptophane (W) sont les plus rares et l'on s'attend à moins de mutations vers ces acides aminés.

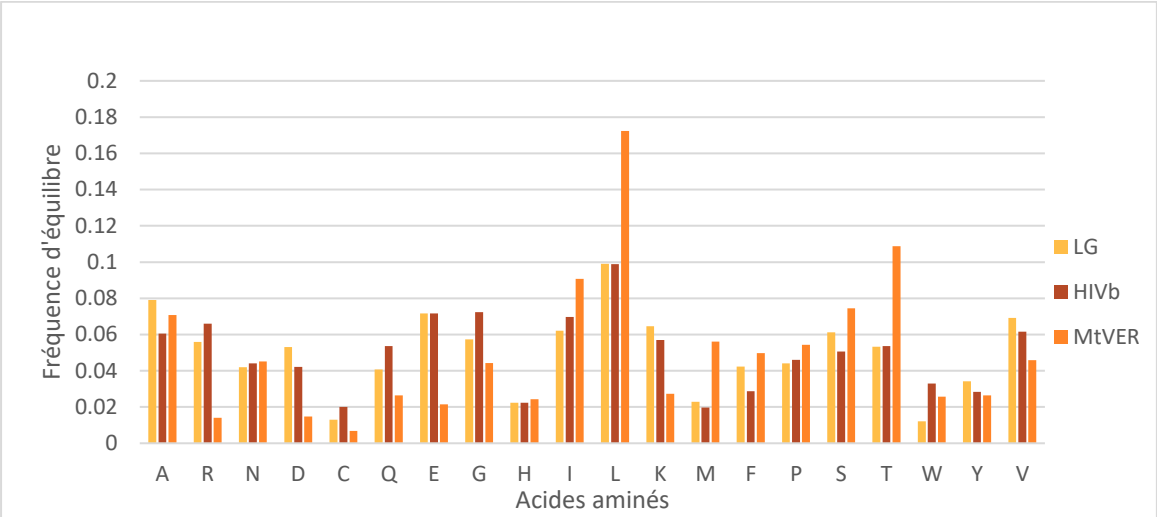


Figure 19 : Fréquences d'équilibre des acides aminés selon différents modèle d'évolution des protéines.

De plus, les acides aminés n'ont pas la même échangeabilité c'est-à-dire que les taux de substitutions entre acides aminés sont différents. Par exemple les acides aminés valine (V) et isoleucine (I) sont hautement interchangeables. Les matrices BLOSUM (que nous avions évoquées dans l'Introduction générale), comme celle Figure 20, constituent une bonne représentation de ce phénomène. Dans cette figure, une valeur élevée indique une plus grande interchangeabilité des acides aminés correspondants. On retrouve l'interchangeabilité entre l'isoleucine et la valine avec un score de 3 (en clair sur la figure). On remarque que le tryptophane (W) qui est rare a principalement des scores négatifs : il est peu substituable avec les autres acides aminés. Ces différents taux s'expliquent entre autres par des propriétés physico-chimiques similaires entre les acides aminés facilement interchangeables mais aussi par des biais de substitutions au niveau de la séquence nucléotidique.

2.1.3.3 Biais de fixation, pléiotropie et épistasie

Une mutation est dite pléiotrope si elle affecte plusieurs phénotypes (Sivakumaran et al. 2011). En effet, selon leurs propriétés physico-chimiques, certains acides aminés peuvent altérer plusieurs aspects d'une protéine : fonction, repliement, solubilité, etc. Ainsi une mutation qui augmente un aspect de la fonction d'une protéine peut simultanément altérer d'autres fonctions. Comme illustré Figure 21, parmi plusieurs mutations qui peuvent avoir un effet équivalent sur le phénotype, on s'attend à ce que celles qui sont majoritairement sélectionnées sont celles avec le moins d'effet pléiotropique délétère (Chevin, Martin, et Lenormand 2010; Stern 2013; Storz 2016).

De la même manière, certains acides aminés à des positions différentes peuvent interagir au sein d'une protéine, donnant ainsi lieu à des interactions épistatiques (l'effet phénotypique des

différentes mutations combinées ne peut être prédit par la somme des effets individuels de chaque mutation). L'épistasie pourrait alors jouer un rôle dans la probabilité de convergence moléculaire. En particulier, l'épistasie peut réduire le nombre de mutations possibles et entrainer de la sélection purificatrice (Storz 2016).

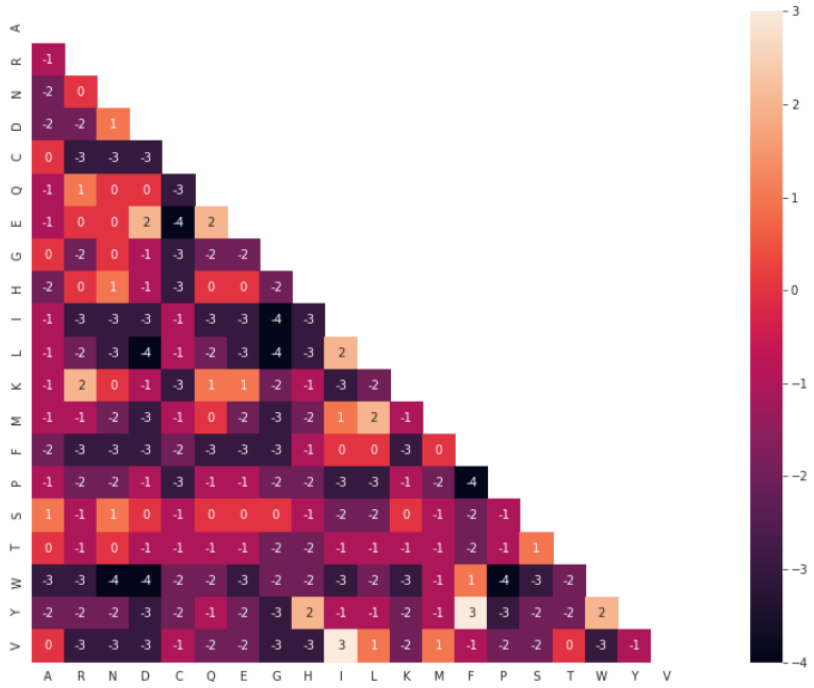


Figure 20 : Matrice de transition BLOSUM62 représentée sous la forme de « heatmap ».
En abscisse et en ordonnée sont représentés les 20 acides aminés. Les chiffres indiquent les taux d'échangeabilité entre acides aminés. Plus le chiffre est grand, plus l'échangeabilité est élevée. Les couleurs correspondent aux différents chiffres. Les taux sur la diagonale ne sont pas représentés.

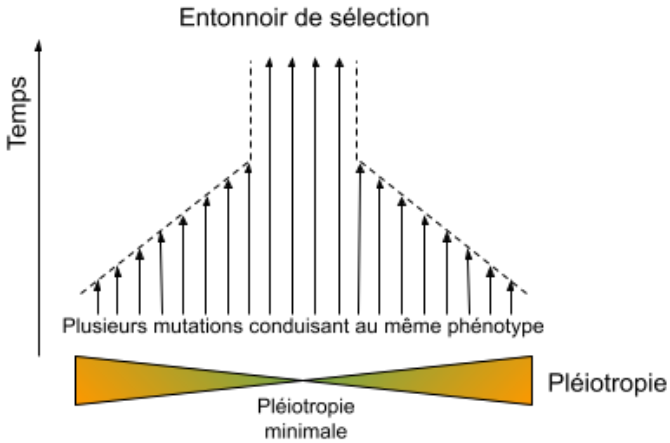


Figure 21 : Schéma illustrant l'influence de la pléiotropie sur la sélection de mutations.
Inspiré de (Gompel et Prud'homme 2009).

2.1.3.4 Taille de population et taux d'évolution

Il a également été démontré (Bailey et al. 2017) que la probabilité de convergence était déterminée par la taille de population, c'est-à-dire le nombre d'individus qui composent une espèce. En effet, l'efficacité de la sélection est d'autant plus élevée que le nombre d'individu

participant à la diversité génétique est élevé. Une grande taille de population permet en effet d'augmenter les chances qu'une mutation avantageuse se fixe par sélection naturelle. A la sélection naturelle, s'ajoute l'effet de la dérive génétique qui concerne le changement aléatoire des fréquences alléliques sans que ces changements ne soient causés par des pressions environnementales. A l'inverse de la sélection naturelle, plus le nombre d'individu qui compose une population est faible plus l'effet de la dérive génétique sera important.

Un autre facteur important dans l'émergence de convergence est le taux d'évolution. Ce taux d'évolution est régi par des facteurs intrinsèques (taux de mutation élevé en raison des erreurs de la polymérase) et à l'échelle d'un organisme ou d'une population, par le contexte génétique ou l'environnement. Plus le taux d'évolution est élevé, plus les séquences génétiques évoluent rapidement dans un même intervalle de temps et plus il y aura la possibilité qu'une mutation avantageuse se répande dans la population.

2.1.4 Convergences moléculaires de premier plan et d'arrière-plan

Comme nous venons de le voir, l'émergence de convergences moléculaires dépend de facteurs multiples et ne correspond pas nécessairement à des pressions de sélection particulières. Comment alors différencier la convergence moléculaire qui sous-tend un phénotype convergent (que l'on suppose être un signe d'adaptation à une contrainte donnée) plutôt que tous les autres types de convergence ? En reprenant les termes employés par Rey et al. (2019), l'étude des bases moléculaires des convergences phénotypiques revient à distinguer la « convergence de premier plan » de la « convergence d'arrière-plan ». Dans le premier cas, cela correspond à des mutations corrélées au phénotype convergent, et dans le second à de la convergence neutre ou non-adaptive ainsi qu'aux mutations corrélées à un phénotype différent de celui étudié. Il convient de rappeler ici que dans les travaux de cette thèse je n'ai pas cherché à déterminer si un phénotype était convergent ou non ni à quelle pression de sélection il correspondait. J'ai étudié ici les bases moléculaires de phénotypes convergents préalablement caractérisés. Des approches ont été développées pour explorer la question de la convergence phénotypique (Ingram et Mahler 2013; Arbuckle, Bennett, et Speed 2014; Arbuckle et Speed 2016; Stayton 2015), que je ne détaillerai pas dans ce manuscrit.

Il est difficile de connaître a priori le degré attendu de convergence d'arrière-plan. Des auteurs se sont intéressés à cette question voyant que l'observation de convergences moléculaires était souvent interprétée comme un signe d'adaptation dès lors que les espèces étudiées partageaient un phénotype convergent. Ils ont montré que, par hasard, les niveaux de convergence d'arrière-plan étaient les plus élevés sur les sites à évolution rapide, entre les acides aminés les plus échangeables et entre des espèces proches. Ainsi les taux de convergence d'arrière-plan ont tendance à décroitre à mesure que les espèces divergent (Goldstein et al. 2015; Zou et Zhang 2015b).

2.2 LA CONVERGENCE MOLECULAIRE EST UN PHENOMENE REPANDU CHEZ LES VIRUS

2.2.1 La nature des virus facilite l'émergence de convergences moléculaires

Chez les eucaryotes supérieurs, on suppose qu'il est relativement rare qu'un phénotype convergent soit le résultat de mutations identiques acquises indépendamment dans des gènes orthologues (Losos 2011). En effet, leur génome est grand et un phénotype est souvent le résultat de cascades moléculaires complexes où interviennent l'expression de plusieurs gènes. On imagine donc qu'un large éventail de possibilités au niveau moléculaire puissent être à l'origine d'un même phénotype convergent.

Chez les virus et en particulier les virus à ARN, un grand nombre de conditions sont remplies pour augmenter les chances d'observer des convergences moléculaires. Tout d'abord leurs génomes sont petits et fortement contraints et les mutations avantageuses sont plutôt rares. De plus, lorsque l'on parle de convergence évolutive chez les virus on parle de convergence chez une même espèce de virus. Les organismes étudiés sont donc relativement peu divergents ce qui augmente la probabilité de convergences qu'elles soient de premier ou d'arrière-plan. Par ailleurs, la taille de population chez les virus est très importante et ils possèdent des taux d'évolution parmi les plus élevés.

De nombreux cas de convergences moléculaires ont été observés chez les virus en conditions expérimentales mais aussi en milieu naturel lors de transmission inter-espèces, en réponse à des traitements antiviraux ou en réponse au système immunitaire de l'hôte. Comme précédemment, la convergence chez les virus s'observe à différents niveaux, de la mutation ponctuelle à des changements génétiques plus larges (réarrangements, délétions de plusieurs nucléotides, etc.). Je m'intéresserai toutefois uniquement ici à la convergence résultant de mutations ponctuelles.

2.2.2 Convergences observées en conditions expérimentales

Les virus évoluant rapidement et leur génome pouvant être séquencé facilement, il est possible de suivre la dynamique d'apparition et de perte des mutations et de tester des hypothèses évolutives. Des auteurs se sont ainsi intéressés à l'évolution des virus en conditions expérimentales soit pour tester des hypothèses évolutives généralistes soit pour mieux comprendre l'évolution des virus.

Une des premières expériences pour estimer la fréquence de l'évolution convergente dans l'évolution des virus a été conduite par Bull et al. (1997). Ils ont ainsi cultivé pendant 11 jours plusieurs réplicas de phages ϕX174 à haute température sur deux espèces bactériennes différentes (Escherichia coli et Salmonella typhimurium). Après séquençage, ils ont observé des mutations convergentes spécifiques à l'hôte utilisé et d'autres partagées par les lignées qui avaient évolué dans des hôtes différents. Ils ont également montré que le niveau de convergence conduisait à des erreurs dans la reconstruction phylogénétique des lignées de phage. La survenue de convergence moléculaire était déjà connue chez les virus dans le cas par exemple de résistance au traitement (Boucher et al. 1992; Kellam et al. 1994; Borman, Paulous, et Clavel 1996) ou de pression sélective de l'hôte (Holmes et al. 1992). Pour autant on ne connaissait pas l'étendue de la convergence moléculaire dans des conditions génériques de culture.

Wichman et al. (1999) et Miller et al. (2016) ont également étudié la convergence évolutive chez des populations de phages en répliquant leur évolution dans diverses conditions. Dans leur expérience, Wichman et al. (1999) ont par exemple trouvé que 96% des mutations qu'ils observaient étaient adaptatives dont la moitié étaient des convergences.

D'autres travaux ont permis de mettre en évidence l'importance de certains facteurs comme l'épistasie ou la taille de la population virale dans la probabilité d'émergence de mutations convergentes (Cuevas, Elena, et Moya 2002; Bailey et al. 2017). De même, il a été possible d'identifier les différents niveaux biologiques où la convergence peut survenir (J. R. Meyer et al. 2012; Miller et al. 2016; Sackman et al. 2017). Les travaux de Sackman et al. (2017) ont par ailleurs permis de mettre en évidence que ce n'étaient pas toujours les mutations les plus bénéfiques qui étaient sélectionnées et que les biais mutationnels jouaient un rôle important dans l'émergence de convergence. Plus récemment, Bertels et al. (2019) ont testé la convergence du VIH-1 dans une expérience d'évolution à long-terme (315 jours) dans deux cultures de lymphocytes T humains. Cela leur a permis d'estimer la fréquence de la convergence évolutive dans un environnement constant. Là encore, ils ont trouvé des niveaux élevés de convergence, avec 62% des mutations devenues majoritaires à l'issue de l'expérience dans une lignée qui apparaissent également dans l'autre lignée.

Par ailleurs, l'étude de virus en conditions expérimentales a permis de montrer que des virus soumis à des mêmes changements d'hôtes présentaient des mutations convergentes (Wichman et al. 1999; Liang, Lee, et Wong 2002; Remold, Rambaut, et Turner 2008; Bedhomme, Lafforgue, et Elena 2012; Longdon et al. 2018). Pour confirmer ce propos, il a été montré dans une culture de phages, qu'un retour vers l'hôte ancestral s'accompagnait de la perte de ces mutations (Crill, Wichman, et Bull 2000). L'expérience menée par Liang, Lee, et Wong (2002) illustre un cas remarquable de coévolution convergente où le passage successif d'un virus de la tache cerclée chlorotique de l'hibiscus sur un nouvel hôte conduit de manière répétée à des mutations aux 8 mêmes positions.

Ces expériences où l'on répète l'évolution de mêmes populations virales sont également un moyen de tester la fameuse hypothèse de Gould (1989) sur la répétabilité de l'évolution. Selon Gould, s'il était possible de « rembobiner la cassette de l'évolution » à son point de départ, le résultat obtenu en rejouant l'évolution des espèces serait complètement différent de ce que l'on observe à présent. Cette hypothèse s'oppose donc à une possible répétabilité de l'évolution. A partir des exemples précédents, on peut toutefois affirmer qu'à l'échelle moléculaire, cette hypothèse n'est pas vérifiée. Pour autant, les virus étudiés dans ces expériences sont peu divergents et si l'on devait remonter au dernier ancêtre commun universel, il est impossible de savoir ce que l'on obtiendrait après des milliards d'années d'évolution et de spéciation…

2.2.3 Convergences observées en milieu naturel

En conditions naturelles, les virus sont régulièrement soumis à des changements d'environnement, ce qui engendre des pressions sélectives favorisant l'apparition de convergences. Des convergences moléculaires ont par exemple été observées lors de l'adaptation à de nouvelles espèces d'hôtes, pour échapper au système immunitaire, ou encore dans le cas de traitements antiviraux.

2.2.3.1 Adaptation à de nouvelles espèces hôtes

Certains virus sont capables de s'adapter à une nouvelle espèce d'hôte lors de ce qu'on appelle un changement d'hôte (Longdon et al. 2014). Lorsqu'un virus passe d'un animal à l'humain, on parle de zoonose. De nombreuses épidémies sont le résultat de zoonoses virales comme pour le VIH-1 qui prend son origine chez les chimpanzés (Hemelaar 2012), la souche H1N1 de grippe également appelée grippe aviaire dont le réservoir animal se trouve chez les oiseaux (Webby et Webster 2001) ou encore le SARS-CoV-2 probablement hérité d'une chauve-souris (Edward C. Holmes et al. 2021; Zhukova et al. 2021). Lors de l'infection d'un nouvel hôte, certaines mutations peuvent être sélectionnées en lien avec le nouvel hôte, si elles favorisent par exemple l'entrée dans les cellules ou permettant d'échapper à la réponse immunitaire. Les évènements de changement d'hôtes sont donc souvent accompagnés de modifications des génomes viraux témoignant de leur adaptation au nouvel hôte. Les possibilités de mutations étant limitées chez les virus, ces adaptations peuvent être convergentes.

Troupin et al. (2016) ont ainsi identifié des mutations convergentes en lien avec l'adaptation du virus de la rage des canidés aux blaireaux-furets, dont deux sont suspectées d'avoir facilité l'adaptation du virus à son nouvel hôte. Ce cas n'est pas isolé et d'autres exemples existent notamment chez le VIH-1 (Wain et al. 2007; Bertels, Metzner, et Regoes 2021), le virus de la grippe aviaire H5N1 (Steel et al. 2009) ou chez le variant H7N9 de la grippe (Xiang et al. 2018). De la même manière, un cas de convergence chez le virus du chikungunya a permis son adaptation à une nouvelle espèce de moustique comme vecteur de propagation (Tsetsarkin et al. 2007; Vignuzzi et Higgs 2017) augmentant ainsi l'aire géographique de présence du virus. Le séquençage massif du SARS-CoV-2 a permis de suivre son évolution en temps réel lors de son adaptation à l'homme, révélant ainsi des cas intéressants de convergence. Plusieurs des variants d'intérêts sont porteurs de mutations ayant émergé indépendamment. C'est le cas notamment de mutations dans la protéine Spike favorisant l'attachement au récepteur ACE2-RBD et l'entrée du virus dans les cellules (N439K, N501Y, E484K ou S477N) (Peacock et al. 2021; Martin et al. 2021).

De nombreux autres exemples de convergence lors de changement d'hôte existent, notamment à des niveaux d'organisation biologique supérieurs aux mutations (Longdon et al. 2014; Gutierrez, Escalera-Zamudio, et Pybus 2019).

2.2.3.2 Echappement au système immunitaire de l'hôte

Lors de l'infection, les virus doivent faire face à la réponse immunitaire de l'hôte. Chez les mammifères, cela implique deux mécanismes, d'une part la réponse immunitaire innée, et d'autre part la réponse immunitaire acquise. Lors de la réponse immunitaire innée, les cellules infectées par un virus sécrètent des interférons qui activent la production de protéines antivirales dans les cellules voisines, ralentissant ainsi la réplication virale (Tosi 2005). De plus, les cellules infectées sont reconnues par les lymphocytes NK (ou cellules tueuses naturelles) qui les éliminent grâce à leur activité cytotoxique. Ces cellules tueuses entrainent par ailleurs l'activation de la réponse immunitaire acquise. La réponse immunitaire acquise ou adaptative fait intervenir les lymphocytes B (responsables de la fabrication d'anticorps) et les lymphocytes T cytotoxiques (LTC) (responsables de l'immunité cellulaire). Les LTC en particulier jouent un rôle important dans la défense contre les virus en reconnaissant les cellules infectées. En effet, celles-ci portent au niveau de leurs antigènes des épitopes qui témoignent de la présence d'un corps étranger.

Les virus sont donc confrontés à tout un éventail de mécanismes défensifs de la part de l'hôte qui peuvent être contournés si des mutations d'échappement apparaissent, permettant alors aux virus de poursuivre leur réplication. Les stratégies d'échappement immunitaire chez les virus sont nombreuses (Ploegh 1998; Simmons, Willberg, et Paul 2013; Ye et al. 2013; Beachboard et Horner 2016; Nelemans et Kikkert 2019) et nous aborderons ici le cas des mutations d'échappement immunitaire.

Des mutations d'échappement ont été constatées dans des génomes du VIH-1, permettant d'échapper aux LTC notamment par des mutations survenant dans les épitopes reconnus par les LTCs (Borrow et al. 1997; Leslie et al. 2004; Bhattacharya et al. 2007; Hashimoto et al. 2010). Une liste des mutations d'échappement a d'ailleurs été proposée par Arcia et al. (2017) selon les différentes molécules de surface portées par les LTCs. Des mécanismes similaires ont été observés chez le VHC (Bowen et Walker 2005; Ray et al. 2005; Gaudieri et al. 2006; Walker et al. 2016; Salimi Alizei et al. 2021). Le VHC dispose d'ailleurs de nombreux mécanismes pour échapper au système immunitaire (Thimme, Lohmann, et Weber 2006), ce qui entraîne des infections chroniques chez 75% (55-85%) des patients infectés (Micallef, Kaldor, et Dore 2006). Les autres hépatites virales fonctionnent d'ailleurs selon un principe similaire (Salimi Alizei et al. 2021). De la même manière, la mutation E484K dans la spike du SARS-CoV-2 a émergé de nombreuses fois et a été démontrée comme permettant d'échapper au système immunitaire (Harvey et al. 2021). On peut également citer l'exemple du virus de la grippe où quelques mutations, notamment dans l'hémagglutinine (glycoprotéine antigénique), permettent au virus d'échapper au système immunitaire (Kaverin et al. 2002; Smith et al. 2004). C'est le cas par exemple des mutations A125T, A151T et L217Q sur l'hémagglutinine de la souche H7N9 (Chang et al. 2020).

Les mutations d'échappement sont parfois accompagnées de mutations compensatrices, qui compensent le phénotype potentiellement moins "adapté" en restaurant la « viabilité » du virus (Peyerl et al. 2004; Friedrich et al. 2004). La transmission de virus avec des mutations d'échappement chez des patients ne possédant pas les mêmes anticorps, entraine parfois la réversion de certaines de ces mutations démontrant qu'elles sont surtout avantageuses dans un environnement donné (Leslie et al. 2004; Bowen et Walker 2005). L'étude des mutations d'échappement est importante car au-delà du système immunitaire, elles peuvent également permettre d'échapper à l'immunité conférée par la vaccination (Novella, Domingo, et Holland 1995).

2.2.3.3 Résistance

L'un des cas de convergence le mieux caractérisé est sans doute la résistance aux antiviraux. La résistance chez le VIH et le VHC a été abordée dans le chapitre introductif de cette thèse et nous avons vu qu'il existait une grande variété de mutations de résistance, classées selon les gènes ciblés par les antiviraux ou le traitement administré (Clutter et al. 2016). Les mutations de résistance au traitement (DRMs pour *drug resistance mutations*) sont effectivement des mutations convergentes puisqu'elles apparaissent de manière répétée et indépendante en réponse à une contrainte forte (le traitement). D'une certaine manière, les mutations compensatrices (ou accompagnatrices ou accessoires) qui permettent de rétablir la *fitness* du virus après l'émergence de DRMs sont également des mutations convergentes.

Au début des années 1990, si des premiers cas de résistance chez les VIH-1 étaient rapportés (Larder, Darby, et Richman 1989; St. Clair et al. 1991), le lien avec la convergence évolutive au

niveau moléculaire n'était pas encore clairement établi (Doolittle Russell F. 1994). Crandall et al. (1999) remarquèrent, en suivant l'évolution de séquences de VIH chez 8 patients avant administration du traitement puis plus de 59 semaines après, que le virus chez 5 des patients en échec thérapeutique présentait des mutations identiques sur des positions de résistance déjà connues (Hammond et al. 1998). Ils démontrèrent alors l'émergence de mutations convergentes en réponse au traitement antiviral.

Dans l'introduction, je me suis concentrée sur la résistance aux traitements chez le VIH-1 et le VHC, mais la résistance aux traitements concerne également d'autres virus. Le virus de la grippe est notamment traité dans certains cas avec des antiviraux (Jefferson et al. 2014), ce qui peut entrainer l'apparition de DRMs (Pizzorno, Abed, et Boivin 2011; Foll, Poh, et al. 2014).

2.3 QUELLES METHODES PERMETTENT D'ETUDIER LA CONVERGENCE AU NIVEAU MOLECULAIRE

Il est important d'étudier les mutations convergentes car si elles surviennent en réponse à des pressions de sélection fortes alors elles peuvent jouer un rôle important dans l'adaptation des espèces. Avant de les étudier, il faut pouvoir les détecter et distinguer les mutations dites adaptatives (ou de premier plan) des mutations d'arrière-plan. Il existe plusieurs méthodes pour détecter la convergence moléculaire, qui diffèrent entre autres par leur définition de la convergence moléculaire ainsi que l'échelle de détection (gène, position, mutation). Certaines méthodes présentées dans cette partie n'ont initialement pas été développées spécifiquement pour détecter la convergence, mais par leur principe, elles peuvent être appliquées à cette tâche.

2.3.1 Recherche manuelle des mutations

La plupart du temps, pour des petits jeux de données, il est possible d'observer des mutations convergentes sans avoir recourt à une méthode spécifique. En conditions expérimentales par exemple, on peut séquencer un échantillon du virus, laisser évoluer plusieurs réplicas, puis reséquencer les différents virus obtenus. Dans ce contexte, il est envisageable de comparer les séquences ainsi obtenues et d'identifier les mutations qui ont émergé indépendamment dans plusieurs réplicas. D'un point de vue phylogénétique cela revient à considérer une phylogénie en étoile dont toutes les feuilles sont issues d'une séquence racine connue : la séquence de l'échantillon initial. Dans l'exemple donné Figure 22, on voit par exemple que la substitution d'un P par un T à la position 4, ainsi que la substitution du K par un E à la position 12 sont probablement convergentes. En effet elles apparaissent de manière récurrente dans les séquences après évolution sous contraintes fortes. Cette approche a par exemple été utilisée dans ces différents travaux (J. J. Bull et al. 1997; Crandall et al. 1999). Cette approche ne permet cependant pas de donner de support statistique aux mutations détectées et ainsi de savoir si la convergence observée est due au hasard, aux contraintes évolutives ou à d'autres facteurs. On peut alors confirmer que les mutations ont un effet sur le phénotype expérimentalement, en regardant les structures 3D des protéines, etc.

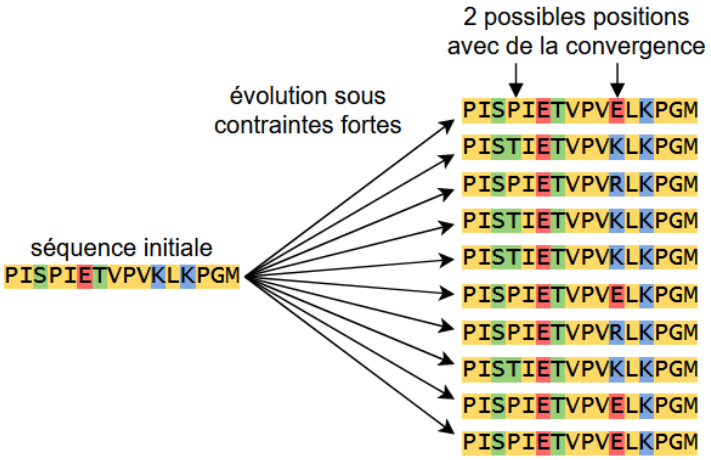


Figure 22: Représentation de convergence moléculaire pouvant être observée à partir de données expérimentales.

2.3.2 Association phénotype-génotype

A partir d'un phénotype connu, associé à chaque virus étudié, il est possible de calculer la fréquence des mutations présentes ou absentes dans chacune des classes phénotypiques. Des tests statistiques (test du chi-deux d'indépendance ou test exact de Fisher) peuvent ensuite être appliqués sur la table de contingence obtenue pour identifier les mutations significativement associées au phénotype convergent. Ce type d'approche est par exemple utilisé pour l'identification de DRMs chez le VIH (Villabona-Arenas et al. 2016). Toutefois cette approche ne prend pas en compte de possible corrélations phylogénétiques et pourrait conduire à l'identification de faux positifs. Par exemple, dans le cas d'une mutation présente dans l'ensemble des virus ayant le phénotype convergent (et seulement dans ceux-là), et si tous ces virus forment un clade monophylétique, on ne veut pas conclure à une mutation convergente. Il faut donc vérifier par des analyses phylogénétiques que les mutations corrélées au phénotype ne se sont pas propagées à partir d'un seul évènement de mutation ou évènement fondateur. Il existe aussi des méthodes de corrélations prenant en compte l'histoire évolutive des espèces étudiées (Pagel 1994; Pagel et Meade 2006).

2.3.3 Sélection positive directionnelle

Habituellement, la recherche de convergence moléculaire est faite dans un contexte d'adaptation à une contrainte, un environnement etc. On recherche en effet quelles mutations sous-tendent un phénotype convergent apparu en réponse à une certaine contrainte. L'adaptation au niveau moléculaire en réponse à des pressions de sélection se caractérise généralement par un rapport élevé entre le taux de substitutions non-synonymes et le taux de substitutions synonymes (dN/dS) (Goldman et Yang 1994). Cependant, la convergence moléculaire révèle une signature différente de la sélection positive globale, car on cherche un biais de mutation vers un type d'acide aminé particulier et dans certaines branches seulement, celles qui conduisent aux espèces possédant le phénotype convergent. Voir la partie introductive sur la sélection positive pour plus de précision sur les différentes méthodes de détection de la sélection positive.

Plusieurs travaux sur la convergence évolutive ont donc utilisé des approches de sélection positive (Christin et al. 2007; Besnard et al. 2009). Le principe est le suivant : après avoir identifié des sites sous sélection positive dans un gène impliqué dans un phénotype convergent, il est possible d'identifier les mutations qui se répètent dans des lignées indépendantes possédant le phénotype convergent. Les méthodes de sélection positive utilisées dans ce cas recherchent souvent la sélection qui s'opère vers un type d'acide aminé en particulier (sélection directionnelle), présente uniquement dans certaines lignées et à certains sites. Pour cela, il faut désigner a priori les branches menant aux taxons possédant le phénotype étudié (convergent). C'est un prérequis nécessaire car s'il y a alternance entre sélection positive et négative et le site sera détecté sous sélection neutre. Si la sélection positive est normalement recherchée au niveau des codons et donc des séquences nucléotidiques codantes, des méthodes ont également été adaptées aux séquences protéiques (Murrell, Oliveira, et al. 2012).

Une alternative à la détection de sélection positive vers un type d'acide aminé particulier est de rechercher un changement dans l'usage du type d'acide aminé à une position donnée. Cette approche suit le principe que des acides aminés suffisamment proches peuvent avoir les mêmes effets à une position donnée. Comme introduit dans l'introduction générale, les modèles de

mélange à profils (vecteurs de fréquences à l'équilibre des 20 acides aminés) permettent de modéliser des préférences dans l'utilisation des acides aminés notamment selon leurs propriétés physico-chimiques. Ainsi le modèle TDG09 (Tamuri et al. 2009) vise à détecter des changements dans les contraintes sélectives. Pour cela il permet d'appliquer des profils de fréquence différents selon les conditions et les sites. Ce modèle a été initialement construit pour l'étude de l'adaptation des virus de la grippe aviaire à l'hôte humain. A partir de deux modèles (contraintes sélectives indépendantes ou non de l'hôte) et en mesurant le rapport des vraisemblances entre ces modèles, ils ont déterminé les positions sous sélection différentielle selon les hôtes. Cette méthode a ensuite été généralisée par des modèles à codons (Tamuri, dos Reis, et Goldstein 2012; Tamuri, Goldman, et dos Reis 2014).

Dans la même logique, Parto et Lartillot (2017) ont mis au point un modèle de codon capable d'identifier les changements vers des acides aminés préférentiellement sélectionnés en fonction de prédicteurs connus (comme un phénotype convergent). Ce modèle estime par des méthodes d'inférence bayésienne, un facteur de sélection spécifique par site et par branche appelé profil de fitness d'acides aminés. Les auteurs ont d'abord appliqué ce modèle à l'étude des séquences du VIH-1 sous la pression de sélection du système immunitaire. Puis ils ont identifié les acides aminés différentiellement sélectionnés à des positions spécifiques le long de la séquence de la Rubisco (une enzyme servant à la photosynthèse chez les plantes), en fonction de la voie de photosynthèse utilisée par les plantes (Parto et Lartillot 2018).

Ces méthodes permettent de détecter de la convergence évolutive à l'échelle d'une position dans un alignement multiple, voire de trouver vers quel acide aminé il y un a biais de mutation. Pour cela, il est nécessaire de spécifier au préalable les lignées convergentes (branches avec un phénotype convergent). Cette étape peut néanmoins comporter des incertitudes, car on ne peut pas reconstruire avec certitude l'histoire évolutive des phénotypes ou des pressions environnementales. De plus, même si leur logique s'en rapproche, elles n'ont pas été spécifiquement conçues pour la détection de convergence et n'imposent pas forcément l'émergence indépendante et répétée des mêmes mutations. Par ailleurs, un même acide aminé peut émerger de manière répétée et indépendante sans pour autant entrainer un changement de profil et ne sera donc pas détecté par les méthodes de profils. Enfin, la convergence évolutive peut aussi survenir en réponse à des pressions de sélection négative ce qui ne serait pas détecté par ces approches.

2.3.4 Topologies et autres signaux phylogénétiques

La présence de convergence au niveau des séquences peut avoir des conséquences lors de la reconstruction phylogénétique. En effet, l'arbre issu de la reconstruction d'un gène portant des mutations convergentes peut grouper ensemble les espèces portant la convergence et ne plus concorder avec la phylogénie d'espèce (Doolittle Russell F. 1994). Un test de détection de convergence reposant sur cet effet, le test de Winning, avait d'ailleurs été proposé dans les années 1990 (Stewart, Schilling, et Wilson 1987; Swanson, Irwin, et Wilson 1991) avant d'être contesté (Zhang et Kumar 1997). Ce phénomène de discordance phylogénétique avait également été démontré par Bull et al. (1997) lors de leur expérience sur les bactériophages X174. La présence de plusieurs mutations de convergence ne permettant pas de reconstruire la véritable histoire évolutive des différentes lignées. Plus tard, Castoe et al. (2009) ont trouvé que la phylogénie

reconstruite à partir des génomes mitochondriaux de certains serpents et de lézards était discordante avec la phylogénie d'espèce. Ils ont pu démontrer que c'était la conséquence de la présence de nombreuses mutations convergentes dans les génomes mitochondriaux.

Parker et al. (2013) ont proposé une méthode « topologique » inspirée de ces observations pour détecter la convergence au niveau des génomes. En considérant un alignement de 2326 gènes orthologues chez un groupe d'espèces comprenant des animaux écholocateurs (plusieurs chauves-souris et un cétacé), ils ont reconstruit 3 phylogénies illustrées en Figure 23 : la phylogénie d'espèce communément admise (H0) et deux phylogénies alternatives groupant respectivement toutes les chauves-souris écholocatrices (H1) ainsi que toutes les chauves-souris écholocatrices avec le cétacé (H2). Ils ont ensuite mesuré la différence de vraisemblance (ΔL) pour chaque position de l'alignement entre les différentes phylogénies. Si la différence de vraisemblance est négative ($\Delta L_{H0\text{-}H1}$ ou $\Delta L_{H0\text{-}H2} < 0$) alors la phylogénie la plus vraisemblable est une de celle où les espèces convergentes sont regroupées. Ils ont considéré ce résultat comme un indicateur de convergence moléculaire. En couplant cette mesure à la détection de sélection positive, ils ont conclu à de la convergence généralisée sur les génomes d'animaux écholocateurs.

Lors de sa parution, cette méthode a été contestée (Thomas et Hahn 2015; Zou et Zhang 2015a). D'une part parce que si des phylogénies discordantes sont bien une des conséquences de convergence, d'autres raisons peuvent conduire à ce phénomène. D'autre part, parce que Parker et al, dans leur approche n'ont pas proposé de modèle nul, permettant d'estimer le nombre de mutations convergentes d'arrière-plan. Zou et Zhang (2015a) ont par exemple proposé comme modèle nul de mesurer la quantité de convergence trouvée avec une topologie alternative à H0 mais ne groupant pas les écholocateurs. Ils ont d'ailleurs pu observer plus de convergence qu'entre les animaux écholocateurs.

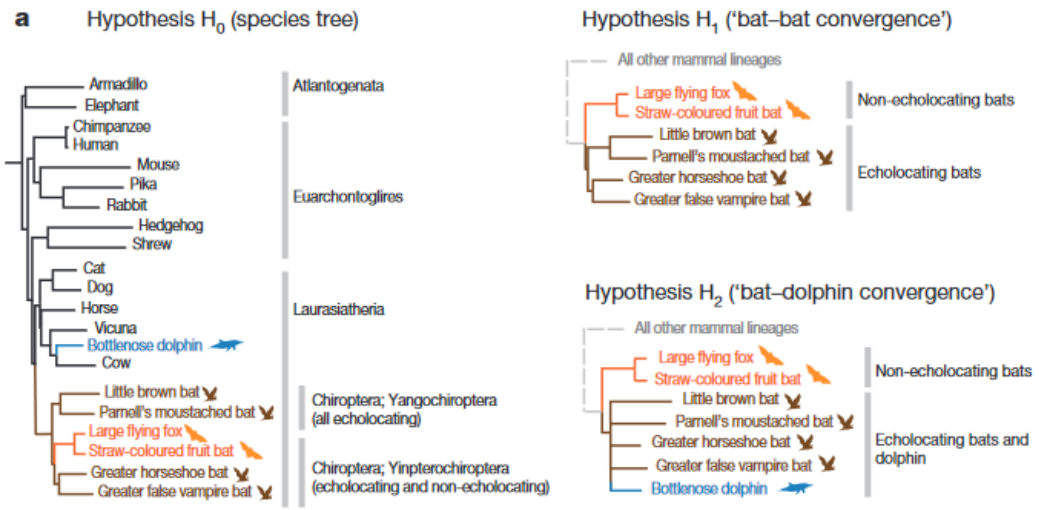


Figure 23: Phylogénie d'espèce et phylogénies alternatives proposées par Parker et al pour tester la présence de convergence dans les gènes d'animaux écholocateurs.
De (Parker et al. 2013). Les chauves-souris colorées en orange ne sont pas écholocatrices contrairement à celles en marron et au cétacé en bleu.

2.3.5 Détection de mutations spécifiques aux espèces cibles

2.3.5.1 Définition stricte

La méthode la plus instinctive pour détecter de la convergence moléculaire est sans doute celle introduite par Zhang et Kumar (1997). Selon cette approche, la convergence est identifiée par reconstruction de séquences ancestrales, en inférant les acides aminés aux nœuds internes. Cela permet de déterminer sur quelles branches de la phylogénie a eu lieu l'émergence des mutations et par extension de déterminer les sites portant des mutations parallèles ou convergentes dans les clades convergents. Ces sites sont ceux où l'on observe des mutations répétées et indépendantes vers le même acide aminé dans toutes les lignées avec le phénotype convergent et uniquement celles-ci. Seulement, cela ne permet pas de conclure si ces changements convergents sont dus à la chance ou sont un signe d'évolution adaptative en lien avec le phénotype d'intérêt, ce que l'on a appelé précédemment les mutations convergentes de premier plan. Pour tester la significativité de leurs détections, Zhang et Kumar ont proposé d'estimer (grâce à une approximation par la loi de Poisson selon un modèle d'évolution) le nombre de changements convergents attendus dans les lignées convergentes à l'échelle de la protéine étudiée. Si le nombre attendu est significativement inférieur au nombre observé alors on peut conclure à de la convergence évolutive adaptative à l'échelle du gène.

En appliquant cette méthode, ils ont identifié 2 positions dans les lysozymes (75 et 87) présentant des mutations parallèles dans les espèces convergentes. Ils ont ensuite déterminé que ces mutations n'ont pas pu être obtenues par chance, pour plusieurs raisons. D'une part ils ont observé des acides aminés différents chez les espèces non convergentes, et d'autre part, en retirant ces deux positions de l'alignement la différence entre le nombre de convergence observé et attendu n'était plus significative. Selon eux, seules les positions 75 et 87 des lysozymes stomachaux semblaient être soumises à adaptation évolutive.

Sans démonstration expérimentale de l'action de ces mutations sur le rôle de la protéine, on ne peut pas conclure avec certitude que ces mutations sont responsables du phénotype convergent observé. A cette fin, Zhang (2006) a énoncé quatre conditions requises pour la détection de convergence.

> *Premièrement, des changements similaires dans la fonction des protéines se produisent dans des lignées évolutives indépendantes. Deuxièmement, des substitutions parallèles d'acides aminés sont observées dans ces protéines. Troisièmement, les substitutions parallèles ne sont pas attribuables au seul hasard et doivent donc avoir été entraînées par une pression sélective commune. Quatrièmement, les substitutions parallèles sont responsables des changements fonctionnels parallèles.*

Il a ensuite appliqué ces principes à l'étude des RNases digestives chez des singes se nourrissant de feuilles et venant d'Asie ou d'Afrique. Il a ainsi pu montrer par des tests fonctionnels et des expériences de mutagénèse dirigée, que l'efficacité digestive chez ses singes a été accrue par de la convergence moléculaire adaptive.

La Figure 24 ci-dessous représente de manière concrète la définition employée dans cette approche. On voit par exemple qu'il y a eu 3 émergences vers le caractère rouge depuis le caractère ancestral jaune, ces émergences étant couplées à l'apparition d'une contrainte

environnementale représentée par la zone grisée. La recherche de convergence peut ainsi se faire en détectant toutes les mutations émergeant de manière répétée et indépendante puis en vérifiant qu'elles sont bien présentes uniquement chez les espèces avec le phénotype convergent. On retrouve par la suite l'utilisation de cette méthode dans divers travaux (Foote et al. 2015; Thomas et Hahn 2015; Zou et Zhang 2015a) sans nécessairement appliquer toutes les vérifications énoncées par Zhang (2006).

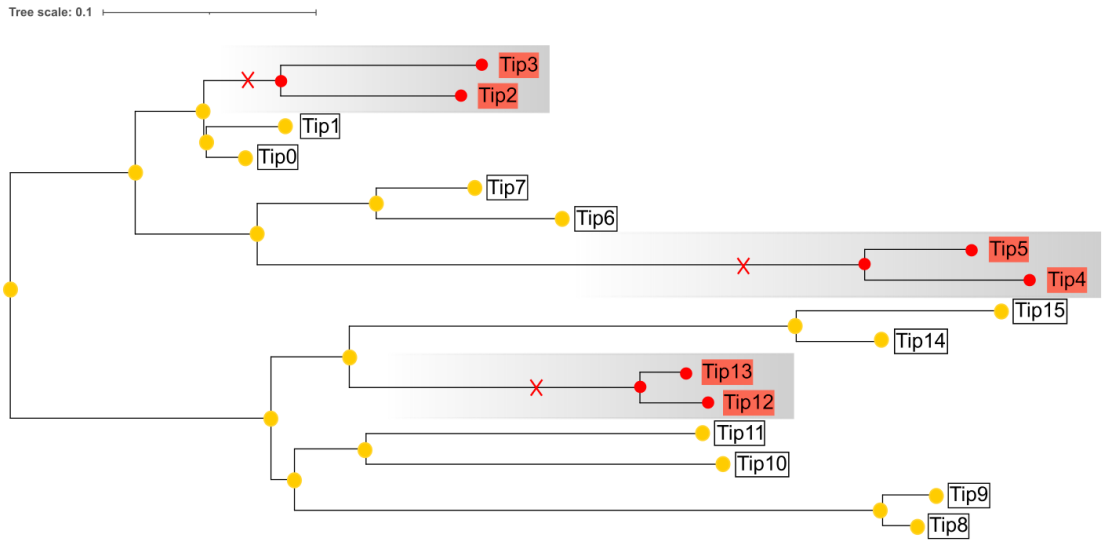


Figure 24: Principe de détection des mutations convergentes spécifiques aux espèces cibles.
Les ronds jaunes représentent le caractère ancestral et les ronds rouges le caractère convergent. Les étiquettes rouges représentent les taxons avec le phénotype convergent. Les croix rouges représentent l'évènement d'émergence de la mutation convergente vers le caractère rouge. Les zones grisées montrent un changement dans les forces de sélection. La barre d'échelle représente le nombre de substitution par sites.

Cette approche de détection de la convergence moléculaire est stricte car elle autorise le changement vers un seul type d'acides aminés et parce qu'elle impose que les mêmes mutations soient observées dans toutes les espèces avec le phénotype convergent et uniquement celles-ci. Elle est donc applicable à un nombre restreint d'espèces car en augmentant le nombre d'espèces on a plus de chance d'observer un phénotype convergent sans observer la mutation convergente et inversement (si le phénotype est obtenu par un autre acide aminé ou une mutation à un autre locus). Thomas, Hahn, et Hahn (2017) ont d'ailleurs montré qu'augmenter le nombre d'espèces dans l'analyse de la convergence, réduisait la probabilité d'observer des changements génomiques uniques dans les seuls taxons présentant les phénotypes d'intérêt. Leur analyse des génomes des mammifères marins en complétant le nombre de taxons a par exemple permis de révéler que certaines substitutions associées uniquement aux mammifères marins dans des petits jeux de données étaient en fait plus répandues, voire partagées par toutes les espèces marines.

2.3.5.2 Changements de profils

Des variations de cette dernière approche ont été proposés, notamment afin d'assouplir sa définition. Rey et al. (2018) ont par exemple cherché à détecter des changements dans les profils d'acides aminés utilisés dans les lignées convergentes plutôt que des mutations vers le même acide aminé. Ils ont donc repris la logique de travaux détaillés plus haut (Tamuri et al. 2009; Parto et Lartillot 2017) à la différence qu'ici les auteurs imposent qu'un changement d'acides aminés ait lieu sur les branches menant aux clades avec le phénotype convergent. Cette approche utilise deux sous modèles, l'un pour la détection de changement de profil (PC) et l'autre pour la détection

de mutations (OC) combinés pour donner PCOC, leur logiciel de détection. PCOC considère que les positions non convergentes évoluent selon un seul profil d'acides aminés, le profil ancestral.

Comparé aux autres approches à profil, PCOC a été développé spécifiquement pour la détection de convergence et donne en sortie la liste des positions sous convergence pour un phénotype convergent donné. Ce logiciel présente toutefois les mêmes inconvénients que les autres approches fondées sur les profils (définition a priori des lignées convergentes, toutes les mutations n'entrainent pas des changements de profils). Par ailleurs le modèle OC impose qu'une substitution ait lieu sur toutes les branches où le phénotype ancestral change vers le phénotype convergent ce qui peut se révéler être une condition assez stricte.

2.3.5.3 Index de convergence

Les approches précédentes reposent sur le fait que les taxons présentant le phénotype convergent possèdent le même acide aminé (ou proche dans le cas des profils) sur la position étudiée et que cet acide aminé a émergé indépendamment. Pour ce faire, les acides aminés aux feuilles sont comparés aux acides aminés ancestraux inférés par reconstruction de séquences ancestrales (en maximum de vraisemblance). Ces méthodes comportent deux principales incertitudes liées à la reconstruction ancestrale. Tout d'abord au niveau de l'inférence des acides aminés ancestraux aux nœuds interne qui peut comporter des erreurs, mais également pour désigner les lignées convergentes. En effet, on ne peut pas déterminer avec certitude la branche où est apparu le phénotype convergent. Pour pallier cela, Chabrol et al. (2018) proposent un indice de convergence par site, indépendant de la reconstruction de séquences ancestrales. Le but de leur approche est de sélectionner les gènes qui pourraient expliquer un phénotype convergent en fonction de l'indice de convergence de leurs sites. Ils définissent l'indice de convergence d'un site comme le nombre maximum attendu (inféré par simulations) de substitutions vers un acide aminé. Ils comparent ensuite cet indice de convergence par site à une distribution empirique d'indice de convergence obtenus par simulations sous modèle neutre et infèrent des p-valeurs pour chaque site. La p-valeur de convergence d'un gène est ensuite obtenu grâce à une combinaison des p-valeurs de ses sites.

Avec cette approche, il n'est pas nécessaire de spécifier les lignées convergentes mais seulement les phénotypes aux feuilles. Cette approche a été pensée pour la détection de gènes convergents à l'échelle des génomes complets de quelques espèces en classant les gènes selon l'indice de convergence de leurs sites. Elle n'est donc pas applicable en l'état à l'étude des mutations convergentes dans de grands alignements ce qui nous intéressent pour l'étude de la résistance chez les virus.

2.3.6 Nécessité d'une nouvelle méthode

Les approches décrites ici ont chacune leurs spécificités et sont plus ou moins adaptées à l'étude de grands jeux de données, la recherche au niveau des génomes entiers, d'un gène en particulier voire de quelques sites. Pour les virus, les méthodes privilégiées sont celles reposant sur la détection de sélection positive directionnelle en désignant à priori les branches convergentes (Tamuri et al. 2009; Murrell, Oliveira, et al. 2012; Parto et Lartillot 2017). Seulement, ces méthodes n'imposent pas forcément l'émergence indépendante et répétée des mêmes mutations. De plus, l'étape où l'on indique à priori les branches des lignées convergentes est complexe et repose sur des incertitudes. Dans le cas de la résistance aux traitements par exemple, il s'agit en majorité de

résistance acquise en réponse au traitement et l'inférence du phénotype aux nœuds internes ne suit pas la même logique que pour un phénotype transmis de manière héréditaire.

Les méthodes spécifiquement conçues pour la détection de convergence présentent également des inconvénients. Dans le cas des virus nous n'avons pas directement accès au phénotype mais plutôt à un prédicteur (ou proxy) de celui-ci. La prise d'un traitement antirétroviral est par exemple un prédicteur de résistance chez le VIH-1. En effet, certaines séquences peuvent naturellement présenter des mutations de résistance même en l'absence de traitement ou des mutations de résistance peuvent être transmises d'un individu traité à un individu non traité. Par ailleurs, comme nous l'avons vu dans l'introduction, il existe des mutations différentes selon le traitement administré et la protéine ciblée. Ainsi toutes les séquences annotées comme traitées ne vont pas toutes posséder les mêmes mutations. De plus, les méthodes présentées ici sont plutôt adaptées à l'étude d'alignements avec quelques dizaines de séquences.

Le chapitre suivant décrit la méthode que j'ai développé, qui est adaptée à de grands alignements de séquences (plusieurs centaines) et lorsque l'information sur le phénotype est approximative : traitement antiviral sans spécifier la molécule antivirale utilisée, adaptation partielle à un nouvel environnement…

3 DEVELOPPEMENT D'UNE METHODE POUR DETECTER LA CONVERGENCE MOLECULAIRE

Nous avons vu dans le chapitre précédent un cas particulier d'adaptation des génomes à des pressions de sélection via la convergence moléculaire. Nous avons vu que chez les virus ce phénomène n'était pas rare et que son étude était fondamentale afin de prévoir et comprendre les différents chemins évolutifs que les virus pouvaient emprunter. Nous avons également décrit les méthodes actuelles et leurs limites, notamment qu'elles n'étaient pas toujours adaptées à la détection de mutations de résistance chez les virus. Je présente dans ce chapitre la méthode de détection de convergence moléculaire que j'ai développé avec l'encadrement d'Olivier Gascuel et Frédéric Lemoine et qui permet de répondre à certains points soulevés précédemment.

En particulier, cette méthode est applicable à la détection de mutations de convergence dans de grands alignements lorsqu'un gène a préalablement été identifié comme convergent. Il n'est pas nécessaire d'annoter les lignées convergentes, mais seulement d'associer un phénotype (ou un prédicteur du phénotype) convergent avec les taxons existants.

Ce chapitre se présente sous la forme d'une brève introduction des concepts utilisés pour construire la méthode puis d'un article décrivant la méthode. Nous avions soumis une première version de cet article à la revue MBE le 01/07/2021 (version disponible en prépublication sur le site www.biorxiv.org (Morel, Lemoine, et Gascuel 2021)). La première version ayant été rejetée avec encouragement à resoumettre, nous présentons ici les corrections en vue d'une re-soumission.

3.1 UNE APPROCHE PAR SIMULATION ET CORRELATION

Au-delà de la détection de l'émergence répétée et indépendante d'une même mutation, il est essentiel pour une méthode de détection de convergence d'associer ces mutations à une pression de sélection. En d'autres termes il faut que la méthode soit capable de discriminer entre les mutations de convergence de premier plan et d'arrière-plan.

Cette tâche passe la plupart du temps par l'estimation du nombre attendu de mutations à l'échelle d'un gène par des modèles d'évolution. Toutefois, selon le modèle utilisé ou la méthode appliquée, les résultats obtenus peuvent être contradictoires. Ainsi en réanalysant des données sur l'écholocalisation (Parker et al. 2013), Chabrol et al. (2018) sont parvenus à une conclusion différente de celle de travaux précédents (Thomas et Hahn 2015; Zou et Zhang 2015a).

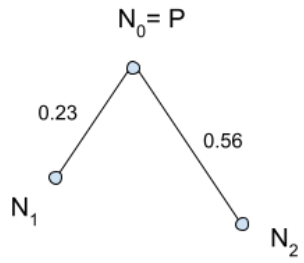
Ici, nous avons fait le choix d'utiliser une approche fondée sur des simulations afin de déterminer la distribution attendue sous un modèle nul, c'est-à-dire ici sans évolution convergente. Lorsque l'on étudie l'évolution, on est obligé de faire des hypothèses car on ne peut pas connaître le passé avec certitude. Les simulations permettent ainsi à partir d'une séquence ancestrale initiale de simuler son évolution sous plusieurs hypothèses. En répétant un grand nombre de fois l'évolution d'une séquence à partir d'une phylogénie on espère ainsi approcher la distribution des mutations

possibles (chemin évolutif), du moins du champ des possibles, à laquelle nous pouvons comparer ce que l'on observe dans les données afin de tester leur écart aux simulations.

Si l'on considère que les simulations représentent la distribution des hypothèses selon le modèle nul alors on peut facilement extraire une p-valeur en comparant les observations aux simulations. En effet la p-valeur correspond à la probabilité d'obtenir sous un modèle nul la même valeur ou une valeur encore plus extrême que celle observée.

3.1.1 Des simulations pour estimer le nombre d'émergence d'une mutation.

A partir d'une séquence racine, d'un arbre et d'un modèle d'évolution, il est possible de simuler l'évolution de cette séquence racine dans les nœuds enfants de l'arbre jusqu'au feuille. A partir de l'état ancestral (un acide aminé dans notre cas) nous pouvons calculer la probabilité d'obtenir chacun des états enfants en fonction de la longueur de branche séparant le nœud ancestral d'un nœud enfant. La figure 25 illustre le résultat des probabilités obtenues aux nœuds enfants lors d'une simulation de l'évolution de l'état au nœud parent N_0. Au nœud ancestral, on voit que l'acide aminé est un P. Selon le modèle d'évolution HIVb (Nickle et al. 2007), et étant donné les longueurs de branches, on constate qu'on a une grande probabilité de conserver P aux nœuds enfants. Pour déterminer l'acide aminé aux nœuds enfants on peut faire un tirage dans la liste des probabilités associées à chaque acide aminé. A partir des acides aminés tirés aux nœuds enfants, on peut ensuite déterminer les acides aminés des nœuds enfants suivants et ainsi de suite jusqu'aux feuilles.



	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
N_0	0.014	0.01	0.001	~0	~0	0.026	0.001	0.001	0.006	0.001	0.022	0.002	~0	~0	0.875	0.028	0.012	~0	~0	0.001
N_1	0.029	0.021	0.006	0.001	0.001	0.052	0.003	0.004	0.012	0.006	0.046	0.007	0.001	0.002	0.726	0.052	0.025	0.001	0.001	0.003

Figure 25: Probabilités associées aux états des nœuds enfants N_1 et N_2, lors de l'évolution de l'état au nœud parent N_0.

Dans notre approche nous avons suivi ce procédé en simulant l'évolution de chaque position de l'alignement un grand nombre de fois (10'000 fois dans nos expériences). Etant donné que nous connaissons l'état de chaque nœud il est ensuite possible de compter les émergences de chaque acide aminé à chaque position. Nous définissons une émergence comme un changement d'acide aminé entre un nœud parent et un nœud enfant à la condition que cet acide aminé nouvellement acquis soit conservé au moins jusqu'à une feuille (suivant un chemin non interrompu de son émergence à une feuille). Cette approche est décrite plus en détail dans la section 3.2 consacrée à l'article à ce sujet.

Les 10'000 simulations permettent d'obtenir une distribution du nombre d'émergence de chaque acide aminé à chaque position. Il est alors possible de comparer le nombre observé d'émergences à chaque position dans l'alignement de référence avec le nombre attendu estimé par les

simulations, tel qu'illustré Figure 26: Comparaison entre la distribution du nombre attendu d'émergences sur 10'000 simulations et le nombre observé à la position 123 d'un alignement VIH-1 de la reverse transcriptase.Figure 26.

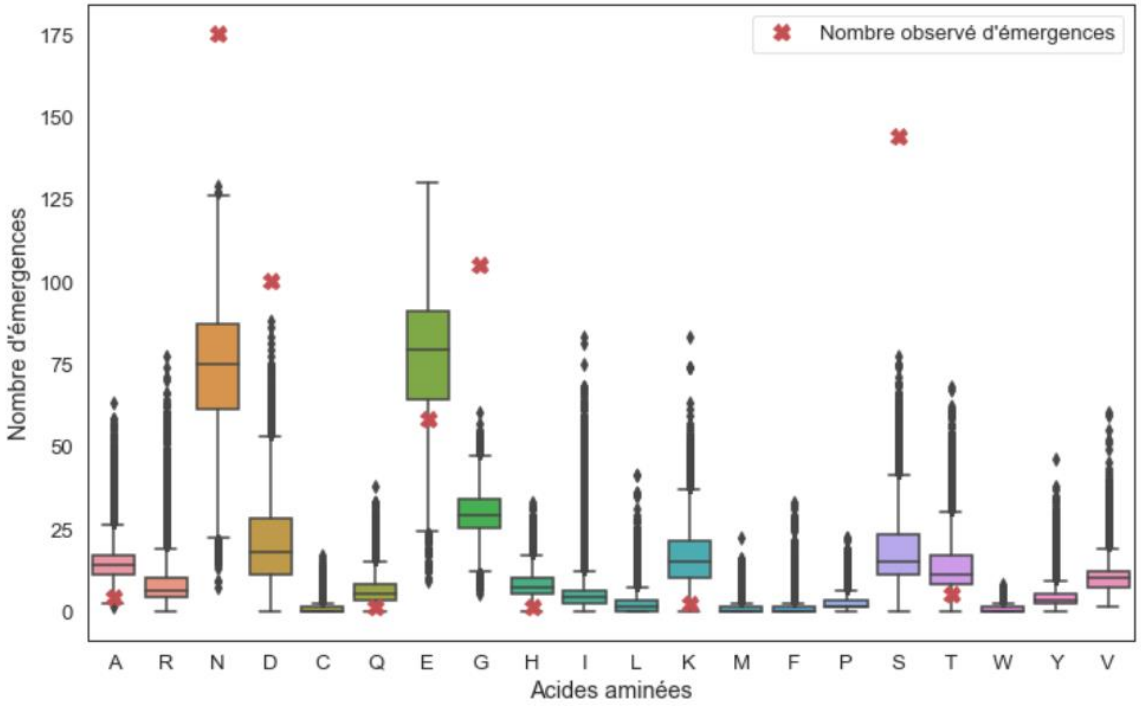


Figure 26: Comparaison entre la distribution du nombre attendu d'émergences sur 10'000 simulations et le nombre observé à la position 123 d'un alignement VIH-1 de la reverse transcriptase.
Les boxplot (boites à moustaches en français) représentent la distribution du nombre d'émergences de chaque acide aminé calculé pour 10'000 simulations.

Sur cette Figure 26, nous avons représenté par des boxplots le nombre d'émergence de chaque acide aminé sur 10'000 simulations sous le modèle HIVb (Nickle et al. 2007) notre modèle nul (supposé sans convergence). Nous comparons cette distribution au nombre observé d'émergence de chaque acide aminé. Nous voyons ainsi que 4 acides aminés (N, D, G et S) émergent plus souvent que ce qui est attendu dans le modèle nul. A chaque observation, nous associons ainsi une p-valeur qui correspond au nombre de simulations avec un nombre d'émergence supérieur à celui observé que nous divisons par le nombre total de simulations. Si aucune simulation n'atteint le nombre d'émergence observé nous associons à la mutation une p-valeur inférieure à 1/10'000.

Le résultat des simulations permet de définir les mutations qui émergent significativement plus souvent qu'attendu sous un modèle nul. Cela ne permet toutefois pas de conclure quant à la raison de ces émergences répétées (pression de sélection, biais de mutation, biais de fixation…). Afin de se concentrer sur la convergence liée au phénotype d'intérêt (convergence de premier plan), nous appliquons des mesures de corrélation aux résultats de la composante émergence de ConDor. Dans l'exemple précédent, l'émergence de ces mutations s'est révélée être indépendante de la résistance et nous n'avons pas retenu ces mutations comme convergentes.

3.1.2 Une mesure de corrélation pour estimer si une mutation émerge indépendamment du phénotype convergent ou non.

Pour déterminer si une mutation est corrélée à un phénotype convergent nous avons utilisé le logiciel BayesTraits (Pagel 1994; Pagel et Meade 2006) qui permet de déterminer si deux variables sont indépendantes ou non tout en tenant compte de la corrélation phénotypique. Cette approche présente l'avantage de ne pas avoir à annoter le phénotype aux nœuds ancestraux tout en tenant compte de la phylogénie. Comme précédemment plus de détails sont apportés dans la section 3.2.

3.1.3 Implémentation dans le logiciel ConDor

J'ai implémenté cette méthode dans un logiciel appelé ConDor pour « *convergence detector* ». J'ai essayé de respecter au maximum les règles de reproductibilité en utilisant le gestionnaire de workflow Nextflow, le logiciel de suivi de version Github ainsi que la mise sous container des différentes dépendances de la méthode grâce à Docker. Cela permet que le logiciel soit facilement téléchargeable et utilisable par les utilisateurs mais aussi d'assurer la répétabilité des expériences. La traçabilité des résultats et des expériences a également été permise grâce à l'utilisation de Notebooks Jupyter.

ConDor est composé d'un certain nombre d'étapes qui font appel à des outils déjà existant ou nouvellement implémenté en python. En particulier ConDor fait appel aux logiciels ModelFinder (Kalyaanamoorthy et al. 2017), IQtree (Nguyen et al. 2015), PastML (Ishikawa et al. 2019), Gotree, Goalign (Lemoine and Gascuel 2021) et BayesTraits (Pagel 1994; Pagel et Meade 2006). J'ai implémenté en python la composante émergence qui est nouvelle, et notamment les étapes de simulations, comptages et calcul des p-valeurs à l'aide de scripts python.

Pour utiliser ConDor, l'utilisateur doit fournir un alignement multiple de séquences protéiques, la phylogénie correspondante, une liste des noms des séquences avec le phénotype convergent, ainsi que les noms des séquences constituant l'*outgroup*. L'utilisateur peut ensuite définir deux limites pour la recherche de convergence qui sont le nombre minimum de séquences où une mutation doit se trouver pour être analysée, ainsi que son nombre minimum d'émergence. Cela permet, lors de l'étude de grands alignements, de se concentrer sur les mutations les plus fréquentes et réduire ainsi le temps d'exécution.

L'exécution de ConDor peut également se faire grâce à un site web hébergé à l'adresse condor.pasteur.fr, ou en téléchargeant un containeur pour une utilisation en ligne de commande. En sortie d'exécution, ConDor renvoie un fichier contenant toutes les mutations testées ainsi que diverses statistiques associées ainsi qu'un deuxième fichier contenant uniquement les mutations qui passent le seuil de détection. Il est également possible de connaitre les résultats de chacune des composantes individuellement (émergence, corrélation).

3.2 Article : Accurate detection of Convergent Mutations in Protein Alignments with ConDor

Accurate Detection of Convergent Mutations in Large Protein Alignments with ConDor

Marie MOREL[1,2], Frédéric LEMOINE[1,3], Anna ZHUKOVA[1,3,4] and Olivier GASCUEL[1,5]

1- Institut Pasteur, Université Paris Cité, Unité Bioinformatique Evolutive, Paris, FRANCE

2- Université de Paris, 5 rue Thomas Mann, 75013 - Paris, FRANCE

3- Institut Pasteur, Université Paris Cité, Bioinformatics and Biostatistics Hub, Paris, FRANCE

4- Institut Pasteur, Université Paris Cité, Epidemiology and Modelling of Antibiotic Evasion, Paris, FRANCE

5- Institut de Systématique, Evolution, Biodiversité (UMR 7205 - CNRS, Muséum National d'Histoire Naturelle, SU, EPHE, UA), Paris, FRANCE

Corresponding Authors: marie.morel@pasteur.fr, olivier.gascuel@mnhn.fr

Abstract:

Evolutionary convergences are observed at all levels, from phenotype to DNA and protein sequences, and changes at these different levels tend to be highly correlated. Notably, convergent and parallel mutations can lead to convergent changes in phenotype, such as changes in metabolism, drug resistance, and other adaptations to changing environments.

We propose a two-step approach to detect mutations under convergent evolution in protein alignments. We first select mutations that emerge more often than expected under neutral evolution and then test whether their emergences correlate with the convergent phenotype under study. When no phenotype is available, as is often the case with microorganisms, the first step can be used alone. To do this, a phylogeny is inferred from the data and used to simulate the evolution of each alignment position. These simulations are used to estimate the expected number of mutations under neutral conditions, which is compared to what is observed in the data. Next, we measure using a comparative phylogenetic approach, whether the presence of mutations occurring more often than expected correlates with the convergent phenotype.

Our method is implemented in ConDor which is available in a webserver and as a standalone. ConDor is applied to three real datasets: sedges PEPC proteins, HIV reverse transcriptase, and fish rhodopsin. ConDor compares favorably to other available tools, especially on large datasets.

Key Words: molecular evolution, phylogenetics, selection, adaptation, convergence, C4 metabolism, HIV, resistance to drugs, rhodopsin.

Introduction

Convergent evolution is often defined as the independent acquisition of similar traits in distinct lineages over the course of evolution (Arendt et Reznick 2008; Losos 2011; Stern 2013). The studied traits can be behavioral, morphological, molecular, etc. In each category, traits can be quantitative (size, length, weight…), binary (presence or absence of a given phenotype), or categorical (a trait is subdivided into several classes). The presence of convergence, especially at the phenotypic level, is often seen as evidence of adaptation in the sense that similar solutions are found in response to the same evolutionary constraints (Castoe et al. 2009; Losos 2011). Many studies focus on the molecular level, assuming that convergent phenotypes may result from the same genetic changes (Stern 2013; Rosenblum, Parent, et Brandt 2014; Storz 2016). At the protein level, it is common to distinguish (J. Zhang and Kumar 1997) between parallel mutations (a change toward the same amino acid is observed from the same ancestral amino acid), convergent mutations (change toward the same amino acid, from different ancestral amino acids), and reversions (mutations that restore an amino acid previously lost during evolution). For the sake of simplicity, in what follows we will refer to these three types of mutations as "convergent mutations", unless explicitly stated.

Examples of evolutionary convergence at the molecular level have been demonstrated in higher eukaryotes, related to adaptation to certain environments (Muschick, Indermaur, et Salzburger 2012; Foll, Gaggiotti, et al. 2014; Foote et al. 2015; Hill et al. 2019; Lu, Jin, et Fu 2020; Shaohua Xu et al. 2020), diet (J. Zhang 2006; Zhen et al. 2012; Ujvari et al. 2015; Hu et al. 2017), changes in metabolism (Besnard et al. 2009; Parto and Lartillot 2018), morphological transformations (Larter et al. 2018) and acquisition of new abilities (Davies et al. 2012; Parker et al. 2013; Lee et al. 2018; Marcovitz et al. 2019; Chai et al. 2020). Similarly, when submitted to constraints such as experimental conditions and drug treatments, microorganisms and viruses adapt and are likely to exhibit similar escapes. This has been demonstrated in HIV after exposure to antiviral drug treatments in several patients (Crandall et al. 1999) and within a single treated patient (Holmes et al. 1992). Similarly, Cuevas et al. (2002) found adaptive convergence in experimental populations of RNA viruses, and van Ditmarsch et al. (2013) in pathogenic bacteria. In natural conditions, evolutionary convergence was found in viruses having experienced host shifts (Longdon et al. 2018; Escalera-Zamudio et al. 2020; Martin et al. 2021) and changes in vector specificity (Tsetsarkin et al. 2007).

Several methods have been developed to detect convergent evolution at the molecular level (J. Zhang and Kumar 1997; J. Zhang 2006; Tamuri et al. 2009; Parker et al. 2013; Thomas et Hahn 2015; Zou and Zhang 2015b; Parto and Lartillot 2017; Chabrol et al. 2018; Rey et al. 2018). They are all based on prior knowledge of a convergent phenotype and aim to identify the protein mutations underlying the phenotypic trait in question. However, they differ in the scale at which molecular convergence is sought and the exact definition of what a convergent mutation is.

Some approaches aim to identify which coding genes show mutations supporting a convergent phenotype, while others study which amino-acid changes can explain convergent changes at the scale of a single protein. Methods of the first category are commonly applied to eukaryotic and prokaryotic genomes and perform genome-wide analyses to detect convergent genes by considering simultaneously all positions of the corresponding protein sequences; for

example, the methods developed by Parker et al. (2013), Zou and Zhang (2015b), Thomas and Hahn (2015) and Chabrol et al. (2018) were applied to the search of genes responsible for echolocation in mammals. In the second configuration, the coding genes responsible for the convergent phenotype have already been identified and the methods focus on the detection of convergent evolution at the position level; for example, Zhang and Kumar (1997) identified convergent and parallel mutations in stomach lysozyme sequences of foregut fermenters. Similarly, Zhang (2006) found parallel substitutions in colobine pancreatic ribonucleases, and Rey et al. (2018) found positions with convergent substitutions in the PEPC protein occurring jointly with the transition toward C4 metabolism in sedges. In fact, testing the significance of convergent changes at individual protein positions has many potential applications. In the case of complex eukaryotic or bacterial organisms, there are few examples of a single amino-acid change that could explain a convergent phenotype (Storz 2016). However, in the case of viruses with rapid evolution, and whose (small) genomes are strongly constrained, only a few amino-acid changes are generally possible at a given position (Pond et al. 2012) and position-wise convergent evolution is expected to be relatively frequent (Gutierrez et al. 2019). Determining molecular changes that deviate from what is expected by chance can thus be indicative of adaptive phenomena. This was the case for SARS-CoV-2, where one first identified mutations in the Spike protein, which were spreading within the viral population and appeared multiple times independently, before being demonstrated to be evolutionarily advantageous for the virus (van Dorp et al. 2020; Korber et al. 2020; Martin et al. 2021). Note, however, that mutations that were initially thought to be adaptive were eventually shown to be simply the result of founder events (Hodcroft et al. 2021), demonstrating the difficulty of detecting convergent mutations without access to the phenotype.

Most importantly, different methods have different ways of selecting which mutations underly the studied convergent phenotype. In the most intuitive definition, one aims to detect mutations toward the same amino acid, which occurred in all clades with the convergent phenotype. This is the definition used first in (J. Zhang and Kumar 1997) and then in (Zhang 2006; Foote et al. 2015; Thomas and Hahn 2015; Zou and Zhang 2015b). An extension was proposed by Chabrol et al. (2018), where the convergent amino acid may only be found in a subset of the convergent species, as well as in some non-convergent species. Considering that a change toward the same amino acid may be too strict since several amino acids have similar physicochemical properties, Rey et al. (2018) relaxed this constraint by considering changes in amino-acid profiles (Le et al. 2008). Their work on amino-acid profiles follows previous works aimed at detecting positions under condition-dependent selection, but which did not focus solely on convergent evolution (Tamuri et al. 2009; Parto and Lartillot 2017; 2018). A radically different approach, proposed by Parker et al. (2013) and inspired by Castoe et al. (2009) work, relies on the fact that convergence can lead to errors in phylogenetic reconstruction by artificially bringing convergent species together. These authors proposed selecting positions that best support the phylogeny that groups species with the convergent phenotype together, rather than the species tree (but see the critiques of this method by Thomas and Hahn (2015) and Zou and Zhang (2015b)).

One of the main challenges in detecting molecular convergence is to identify only the convergent mutations that are linked to the studied convergent phenotype. In their review of methods for detecting molecular convergence, Rey et al. (2019) referred to this type of mutation as foreground convergence (or foreground convergent mutations) in opposition to background

convergence which is unrelated to the convergent phenotype. Indeed, at the molecular level, one can find patterns of convergent mutations linked to another convergent phenotype, or occurring because of mutational biases, protein conformation limitations, constraints at the molecular level, or epistatic forces (Zhang and Kumar 1997; Rokas and Carroll 2008; Storz 2016; Stoltzfus and McCandlish 2017). It has been shown that most (if not all) substitution models may fail at distinguishing between foreground convergent mutations and background ones (also called non-adaptative convergent mutations), especially in close taxa between highly exchangeable amino acids, and on fast-evolving sites (Goldstein et al. 2015; Zou and Zhang 2015b). In other words, the finding of several independent mutations to the same (or a similar) amino acid should be tested carefully, even when the number of such mutations appears to be high. We shall see that our findings confirm this.

Another difficulty is the definition of the convergent phenotype and the annotation of taxa that do or do not have this phenotype. For example, in the case of viruses, we usually do not know the exact phenotype, but use a proxy. In the case of drug resistance mutations (DRMs) that occur repeatedly in different patients treated with antiviral drugs, being on treatment is a proxy for resistance. Although we expect that most (but not all, e.g., due to poor adherence) sequences from patients who fail drug treatment will contain resistance mutations, we also expect that some drug resistance mutations will be found in untreated (naive) patients in the case of resistance transmission (Blassel et al. 2021). Similarly, environmental constraints are not strictly speaking phenotypes, but act as selective forces that can lead to molecular convergence. However, we do not expect all organisms living under the same environmental conditions to exhibit the same recurrent mutations.

In some respects, the identification of convergent mutations has similarities to the detection of positions under positive selection (Goldman et Yang 1994). The idea is indeed to identify mutations that might be advantageous, as they are found more often than expected in a neutral (or purifying) model of evolution. In the positive selection framework, these mutations can be directed to a specific amino acid, or correspond to any change that differs from the original amino acid. This is the case, for example, with immune avoidance where mutations towards any new amino acid at antigenic sites are generally favorable and positively selected. Conversely, in the case of convergent evolution, we are interested in substitutions towards one or a few similar amino acids, in the branches leading to the convergent taxa. Thus, a large number of non-synonymous substitutions on convergent positions are expected, but the simple criterion of positive (or neutral) selection is not sufficient to assert convergence. The FADE software (Murrell et al. 2012) in the HyPhy suite (Pond et al. 2005) tests whether positions in a protein alignment are subject to directional selection (or mutational bias) within a specified set of "foreground" branches that typically correspond to convergent taxa. This tool thus has its roots in positive selection approaches, but is closely related to convergence detection.

Here we propose a method for detecting convergent evolution at the position (or site) scale in large amino-acid alignments while relaxing the constraint that convergent mutations must be found only in organisms with the convergent phenotype and in all such organisms. Our method does not require specifying the branches where molecular convergence occurred (as with PCOC or FADE, for example), which is a complex step, especially with large data sets and when using a proxy for the phenotypic convergence. The taxa are simply annotated as convergent or non-

convergent, and the method searches for mutations correlated with this status. We are interested in changes to a target amino acid, independent of the ancestral amino acids that lead to that mutation in contemporary sequences. In other words, parallel, convergent and revertant mutations are considered indifferently, and we consider mutations to different target amino acids as different events. With this definition, our method is in line with methods aiming at detecting changes towards the same amino acid, as opposed to detecting changes in profiles (Rey et al. 2018). Indeed, there are many examples of known convergent mutations, where the changes involve highly exchangeable amino acids that have very similar biochemical profiles. For example in HIV drug resistance, there are convergent mutations from Isoleucine to Valine and Tyrosine to Phenylalanine (the two most exchangeable amino acid pairs, cf. BLOSUM62) that confer resistance to certain drugs (Wensing et al. 2019).

In the following sections, we describe this approach, which is implemented in a workflow called ConDor (for Convergence Detector), available via a web service (condor.pasteur.cloud) and as standalone software. Its performance is evaluated using three real-world datasets involving sedge PEPC proteins, HIV reverse transcriptase, and fish rhodopsin. The results are compared to those of PCOC and FADE, which are based on different assumptions.

New Approaches

Method overview

Our method aims to identify amino-acid mutations that emerged multiple times in independent lineages, occurred more frequently than expected under a neutral (or null) substitution model, and are correlated with a known convergent phenotype. The method is subdivided into two independent components: (1) the “Emergence” component that detects mutations emerging more often than expected under neutral evolution, and (2) the “Correlation” component that identifies mutations that are positively correlated with the convergent phenotype. The combination of the two components accurately identifies amino-acid mutations resulting from convergent evolution associated with the convergent phenotype (foreground mutations), although it is also possible to interpret the results of the two components independently.

A representation of the ConDor workflow is shown in Figure 1. Inputs are: a multiple protein sequence alignment (MSA), a phylogeny, an outgroup, the phenotype of each of the taxa, and user-supplied thresholds to select convergent mutations. The two first steps are common to the Emergence and Correlation components. In step (1), we estimate the parameters of the null model from the MSA (parameters of the substitution model, ML-based branch lengths of the phylogeny, evolutionary rate per position, etc.). In step (2), we reconstruct the substitution history and count the number of emergence events of mutation (EEMs) observed for every position and amino acid of interest in the MSA (i.e., those present in a sufficient number of sequences). In the Emergence component, for each position and amino acid of interest, the two main steps are: (3) simulation of new datasets under the null model and counting of simulated EEMs; (4) comparison of observed and simulated EEM numbers to identify the mutations that occurred significantly more often than expected assuming the null model. The Correlation component is applied to all mutations with more than *n* EEMs (*n* is user-defined). The two mains steps are: (3’) computation of the log Bayes factor of the model assuming a dependence between the presence/absence of

the phenotype and the given mutation, versus the model assuming their independence, using BayesTraits (Pagel 1994; Pagel and Meade 2006); (4') determination for mutations associated with the phenotype if the dependence is positive or negative. The final step (5) combines the results of steps (4) and (4') and provides a list of potential convergent mutations. The results of steps (4) and (4') are also provided to the user and can be interpreted independently. Note that step (4) does not require knowledge of the phenotype (or a proxy for it, as with DRMs in HIV). For all selected mutations, Condor provides the evolutionary rate of the corresponding position, the nature of the mutation (convergent, parallel, revertant), the number of EEMs, the genetic barrier, the BLOSUM score, etc. All these statistics can be used to further analyze the results and select the most relevant mutations.

The null model and all its components are inferred from the input alignment and phylogeny using ModelFinder (Kalyaanamoorthy et al. 2017) and IQtree (Nguyen et al. 2015). The selected substitution model, along with amino-acid frequencies, rates-across-sites distribution parameters, branch lengths, and evolutionary rate per site are assumed to represent the data without convergence. We make this assumption because using large alignments (>1000 sequences), we consider that mutations resulting from convergent evolution are rare enough to have a negligible influence on parameter inference. The phylogeny with optimized branch lengths is then rooted using the user-supplied outgroup. This is necessary to infer the ancestral sequence at the root of the tree, run simulations starting from this sequence, and count simulated EEMs. For the rest of the analyses, we restrict ourselves to mutations present in at least 0.5% of the sequences (default value, this threshold is adjustable by the user). Ancestral character reconstruction (ACR) for positions with mutations of interest is performed using a maximum likelihood approach, implemented in PastML (Ishikawa et al. 2019). We use the "maximum a posteriori" (MAP) method in which the state with the highest marginal posterior is selected at each tree node. Once all ancestral amino acids are reconstructed and associated with all internal nodes in the phylogeny, we identify where independent amino-acid changes occurred in the tree and count them as explained in the subsection "Counting emergence events". Using this count of the observed number of EEMs, we restrict the Emergence and Correlation analyses to mutations with more than n EEMs (i.e., found in more than n independent clades, n=2 in all analyses below).

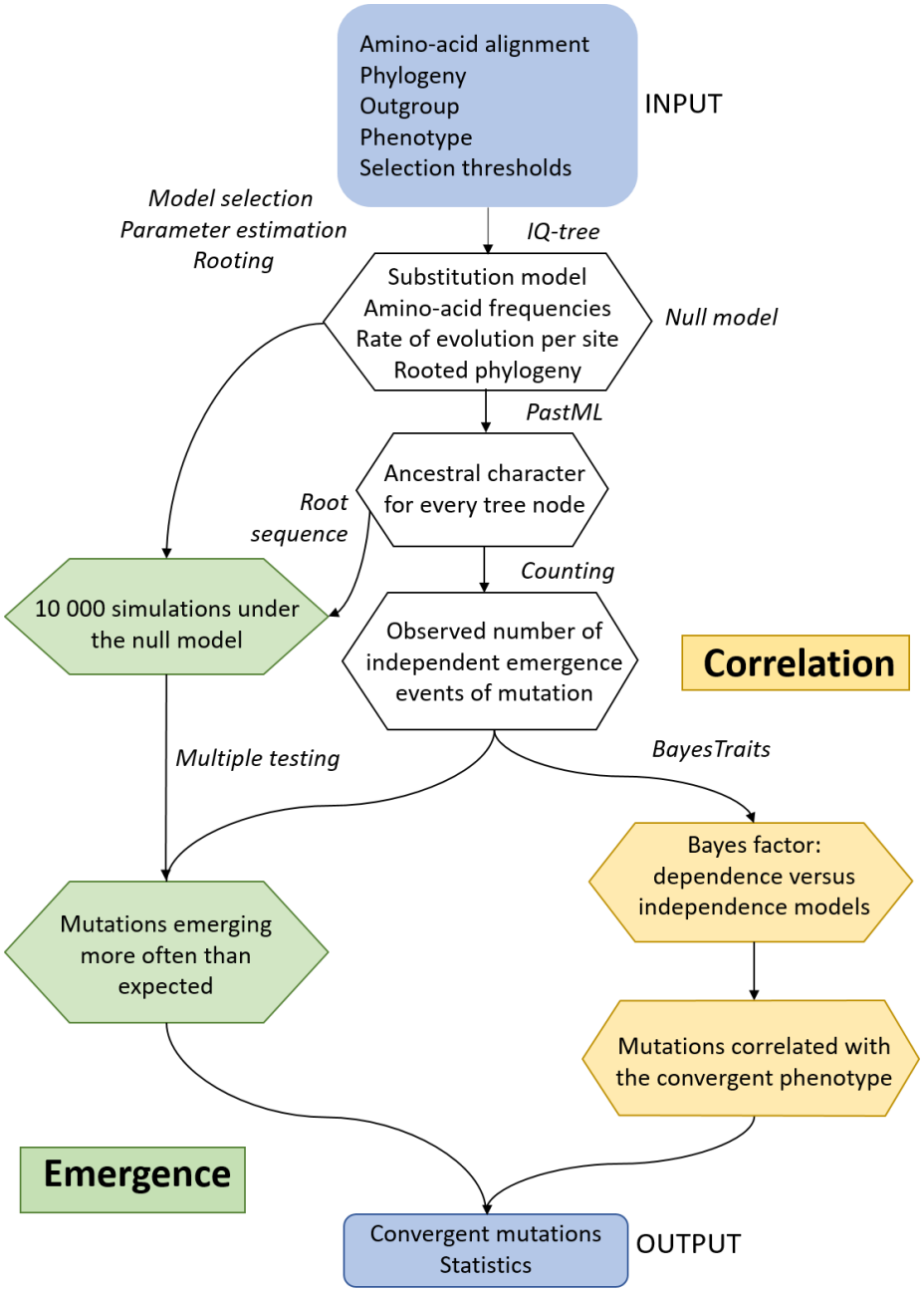


Figure 1: Flowchart of the method. *The method takes as input an amino-acid alignment as well as the corresponding phylogeny and phenotype metadata. The MSA and phylogeny are used for inference of the null model (branch lengths, substitution model and its parameters, evolutionary rate per site, etc.) and ancestral character reconstruction. In the emergence component, the tree and root sequence are used to simulate 10,000 alignments under the null model; the output is the list of mutations that emerged more often in the input alignment than in the simulations. In the correlation component, we first select mutations whose number of emergences is greater than a user-specified threshold, and then those that are correlated with the phenotype. The combination of the two components gives the list of mutations proposed as convergent.*

Estimating with simulation the expected number of emergences

The emergence component consists of simulating the convergence-free evolution that is expected for each tested position of the alignment. In our experiments, we performed many (10,000) simulations per position. Our implementation does not use the exact root amino acid

reconstructed by ACR as a starting point, but draws amino acids based on their marginal posterior probabilities to account for reconstruction uncertainty (e.g., two amino acids with posteriors of 0.55 and 0.45). After the simulations of sequence evolution along the tree, we count the simulated numbers of EEMs (10 000 values per position and AA studied) using the algorithm detailed below. For example, let us consider the mutation M41L from our real HIV dataset, where at position 41, a Methionine (M) is substituted into a Leucine (L) in 211 sequences. The observed number of EEMs toward L is 47, which is smaller than 211 as in some subtrees all tips have L, corresponding to only 1 EEM. This number is compared to the distribution of the number of EEMs toward L, starting from an M at the tree root every time (no ambiguity in ACR), among 10,000 simulations in the null model; this number of simulated EEMs ranges from 0 to 31 with an average of 12. From the observed number of EEMs and the distribution of simulated EEMs, we estimate a p-value for each observed mutation, which is equal to ~0 (<1/10,000) in our M41L example. Since we test many positions and mutations, we use the Holm–Bonferroni method (Holm 1979) to counteract the problem of multiple comparisons, with a default rejection threshold of 10% (adjustable by the user). After correcting for multiple testing, we consider that mutations with p-values lower than this rejection criterion did not occur by chance. Thus, the Emergence component provides a list of mutations occurring more frequently than expected under the null model inferred from the data, some of which are the result of convergent evolution (including parallel and reverting mutations). Although these mutations can be studied on their own in the absence of an identified phenotype, we know from previous studies that background convergent mutations in real data tend to be more frequent than expected under any available substitution model, due to model approximations, epistatic constraints, etc. (Rokas and Carroll 2008; Castoe et al. 2009; Goldstein et al. 2015; Zou and Zhang 2015b). Moreover, some of these very frequent mutations may be truly adaptive and convergent, but for other phenotypic traits than the one studied. Thus, we expect that a significant fraction of these frequent mutations are false positives for the studied phenotype. The Correlation component complements the Emergence component to focus on mutations that correlate positively with the phenotype, that is, foreground convergent mutations.

Counting independent emergence events of mutation (EEMs)

The observed number of EEMs is inferred by ACR based on the input sequences, while the expected number of EEMs and its distribution are estimated from many simulations evolving the probabilistic root sequence along the tree using the inferred null model. In simulations, changes may appear which cannot be inferred by ACR, in particular when they are not transmitted to any tree leaf. In this case, the expected number of changes artificially deviates from the ACR-based “observed” number of EEMs. This effect is even more pronounced on fast-evolving positions since more changes are expected. Thus, only the changes transmitted to at least one leaf are counted by our method, since they are the only ones that could be found by ACR.

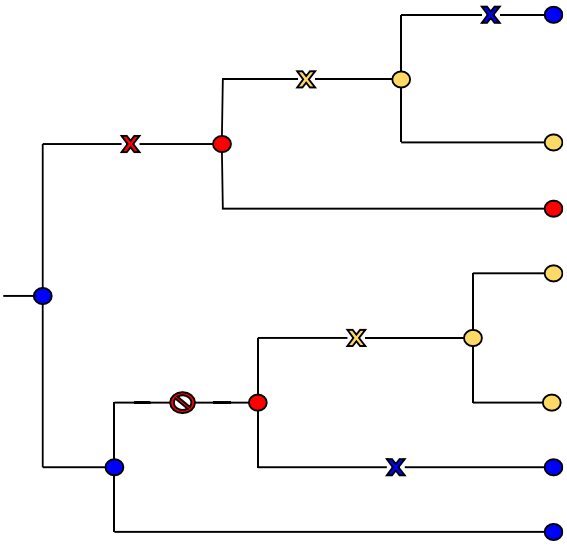


Figure 2: *Counting the emergence events of mutations (EEMs). In this tree, we count two (parallel) EEMs toward the yellow state, two reversions toward the blue state, and only one EEM toward the red state since the mutation at the bottom of the tree is not transmitted to any leaf and thus not counted.*

In the tree illustrated in Figure 2, we have 6 changes along the branches, which are represented either by a cross or a NO symbol. The NO symbol stands for a change of the blue state toward the red one, but the red state is then lost and not transmitted to any tree tip. With ACR, this node would have been either blue or yellow, but never red, while this might occur in simulations. Thus, in our counting, we do not consider this EEM toward the red state, and only the one in the upper subtree is counted. The two yellow crosses mark changes transmitted to one leaf in the upper subtree and two tips in the bottom subtree; thus, both are counted. The two blue crosses mark a return to the ancestral state present at the tree root and are reversions. Even though we count EEMs without making the difference between convergent, parallel, and revertant events, we retain the information during the counting process for interpretation afterward. The algorithm is further described in Material and Methods. It has linear time complexity in the number of tips, just as PastML (Ishikawa et al. 2019) and the simulation algorithm, which explains the relatively fast computing times of the Emergence component (12 minutes on average per mutation on the rhodopsin dataset with 1,500 sequences, see below), though it is based on many (10,000) simulations.

Correlation with the convergent phenotype

The correlation component of ConDor is based on the 'Discrete' method from BayesTraits (Pagel 1994; Pagel and Meade 2006), which combines Markovian modeling of trait evolution and Bayesian model comparison, to distinguish between the two hypotheses of independent (H_0) versus dependent (H_1) evolution of two traits along a phylogeny. Here, we apply BayesTraits to the analysis of two binary traits: presence/absence (1,0) of the mutation and convergent/not convergent (1,0) phenotype. For each of the hypotheses (corresponding to different evolutionary models), the marginal log-likelihood (approximated by the harmonic mean of the likelihoods after several millions of iterations) is calculated using a stepping stone sampler. BayesTraits then calculates the log Bayes factor (logBF) to decide if there is some support for the (H_1) dependence hypothesis:

$$\text{logBF} = 2\log\left[\frac{\text{MarginalLikelihood}(H_1)}{\text{MarginalLikelihood}(H_0)}\right].$$

As described in the BayesTraits manual (www.evolution.reading.ac.uk/Files/BayesTraits-V1.0-Manual.pdf), a "logBF greater than 2 is taken as 'positive' evidence, greater than 5 is 'strong' and greater than 10 is 'very strong' evidence". To account for different sizes of datasets, we chose different thresholds for log BF (2 for the Sedges PEPC dataset, with 50 sequences, and 20 for the other datasets, with >1,000 sequences). Moreover, using the two components of ConDor, BayesTraits is only applied to mutations that are already selected by the Emergence component and corrected for multiple testing (regarding EEMs).

Once we know that the evolution of the two traits is correlated, we need to determine the direction of the correlation: Is the presence of the mutation favored in the convergent phenotype (positive correlation) or the non-convergent phenotype (negative correlation)? To do this, we check whether the proportion of sequences with the mutation is greater among the sequences with the convergent phenotype (positive correlation) or with the non-convergent phenotype (negative correlation). With the rhodopsin dataset, we were interested in adaptation toward both "convergent" (fresh/brackish) and "non-convergent" (marine) environments and considered both positive and negative correlations. With the HIV sequences, we were only interested in (resistance) mutations that are positively correlated with treatment.

This method has been widely used in evolutionary biology and ecology to test correlations among behavioral, morphological, genetic, and cultural characters, and for predicting functional gene linkages (Barker and Pagel 2005). To our knowledge, it has not been used to detect evolutionary convergence. One of the main advantages of this method is that it takes into account the phylogenetic correlation between taxa (as opposed to simple association tests, such as Fisher's exact test that is commonly used for the detection of DRMs in HIV (Blassel et al. 2021)). Furthermore, it does not force the emergence of molecular convergence in all species with the convergent phenotype, as does the 'One Change' (OC) model of PCOC, for example (Rey et al. 2018). This characteristic is especially important as in most analyses we do not know the exact phenotype, but use a proxy. However, it should be kept in mind that the Correlation component in isolation can identify mutation events that fall outside the scope of convergent evolution. For example, a perfect correlation between a mutation and phenotype can arise from a single mutation event which is then propagated until the tips of the corresponding subtree (a so-called "founder" event; Bhattacharya et al. 2007; Gutierrez et al. 2019). As we used a Bayesian framework through BayesTraits, the analysis of large trees can require a relatively large amount of computing resources (~30 min per mutation on the rhodopsin dataset, parallelized in ConDor). However, this computing time could be reduced using a similar maximum-likelihood approach (e.g., based on the 'ace' routine from APE; Paradis et al. 2004).

Results

Overview: data, methods, and comparison criteria

We applied ConDor to three real datasets with known convergent mutations: (1) a sedge phosphoenolpyruvate carboxylase (PEPC) protein dataset with mutations associated with the acquisition of C4 metabolism; (2) a real HIV dataset of reverse transcriptase with 26% sequences

with DRMs; and (3) a dataset of fish rhodopsin, a light-sensitive receptor protein that is highly conserved but known to vary at certain positions among species depending on their environment. On simulated data, we know exactly which mutations are truly convergent or not. With real data, even if we know certain convergent mutations, some other mutations are likely convergent but correspond to other phenotypic traits (background convergence). In other words, the question of assessing the rate of false is questionable and method sensitivity in detecting the known convergent mutations will be the main criterion.

For HIV and rhodopsin data sets, we reconstructed the phylogeny from the sequences (nucleotide data and protein data respectively), using ModelFinder (Kalyaanamoorthy et al. 2017) and IQtree (Nguyen et al. 2015) with standard options (see Material and Methods). For Sedge data we used the provided phylogeny. Each phylogeny (with branch lengths reoptimized with amino-acid sequences for HIV and Sedge) was used as input of ConDor, PCOC (Rey et al. 2018), and FADE (Murrell et al. 2012). For all tested methods, we evaluated the same mutations and positions, corresponding to the mutations present in at least 0.5% of the sequences and in more than 2 independent clades.

Given a rooted phylogeny, aligned amino acid sequences, and a list of convergent clades, PCOC performs a detection analysis for its three models (Profile Change, One Change, or both) for which we can set independent significance thresholds. Instead of detecting a change toward the same amino acid, the Profile Change (PC) component aims at detecting positions for which the general use in amino acid preference has changed in the convergent lineages. This preference is modeled by a vector of amino acid frequencies or 'profile' and, at a convergent position, the profile used in all clades with the convergent phenotype must be different from the ancestral profile. Thus, for a non-convergent position, the same profile is used all along the tree. In addition, the One Change (OC) model forces that the switch of profile occurs along with at least one substitution. Positions that verify the two sub-models are retained as convergent by PCOC, using a specific approach to combine the p-values from both sub-models. For the profiles, we used the C10 model that combines 10 profiles to represent the diversity of biochemical and mutational properties among amino acids (Le et al. 2008; default option in PCOC). Before running PCOC, users have to annotate the clades for their convergent status, using the list of species having the convergent phenotype. According to (Rey et al. 2018), a clade is said to be convergent if all its tips possess the convergent phenotype, and the branches yielding convergence (where PC and OC are expected) are those rooting the maximal convergent clades. PCOC aims to detect positions with molecular convergence, but does not return a list of mutations (it returns a list of positions). We considered in our experiments that a mutation was detected by PCOC, when it was present at a position where at least one of the sub-models (PC and/or OC) was verified (threshold above 0.8), and the mutation was more frequent in species with the convergent phenotype than in the non-convergent ones.

FADE is one of the methods to detect selection available in the HyPhy package (Pond et al. 2005). FADE replaces previous approaches to test for episodic directional selection in protein alignments, which showed high detection power with DRMs in HIV (Murrell et al. 2012). To launch FADE, the users have to specify the branches that are expected to have undergone directional selection, called "foreground" branches. These typically correspond to all branches in convergent clades (and non-solely to the clade rooting branches, as with PCOC; see Material and Methods for

details). FADE tests for each position in the alignment if there is a “substitution bias toward a particular amino acid in the foreground branches, compared to the background branches”. The method relies on a Bayesian framework and a Bayes Factor > 100 provides strong evidence that the site is evolving under directional selection.

ConDor aims at detecting mutations emerging more often than expected under a null model and which are correlated with the convergent phenotype (or its proxy). In our experiments, the null model corresponded to the best substitution model according to BIC, as inferred by ModelFinder (Kalyaanamoorthy et al. 2017). However, we also tested alternative models to check the robustness of the method to model violation (inevitable with real data). Both ConDor components use selection thresholds, which were fixed to p-value <10% and Bayes factor >20 for Emergence and Correlation, respectively, unless otherwise specified. A mutation that verified both conditions was retained as a sign of foreground convergence.

We compared the three methods (PCOC, FADE, and ConDor) using common statistics such as the number of true positives (TP: number of detected convergent mutations), true negatives (TN: number of non-detected non-convergent mutations), false positives (FP: number of detected non-convergent mutations) and false negatives (FN: number of non-detected convergent mutations). We also computed the recall (TP/(TP+FN)), precision (TP/(TP+FP)) and F1 score for each method. The F1 score is the harmonic mean of recall and precision:

$$\text{F1 score} = \frac{2\times TP}{2\times TP + FP + FN}.$$

The F1 score provides a balanced view between recall and precision, which are generally in tension (improving precision typically reduces recall and vice versa). The F1 score is robust to class imbalance, as is usually the case with convergent mutations that are much less frequent than non-convergent mutations.

Sedges PEPC protein dataset

We selected this dataset on C4 metabolism because it was the one used as a reference to evaluate PCOC in (Rey et al. 2018) and thus allows for a fair comparison of convergence detection methods with a small dataset (79 sequences). C4 metabolism is a recognized case of convergence in plants, with multiple independent appearances, especially in cases of adaptation to arid environments (Ehleringer et al. 1997). Among the many proteins involved in the C4 photosynthetic pathway, phosphoenolpyruvate carboxylase (PEPC) has been studied to find the molecular basis for phenotypic convergence. Within this enzyme, several convergent amino acid mutations were found in both grasses (Christin et al. 2007) and sedges (Besnard et al. 2009) with an intersection of 5 convergent positions that are identical between these two distant clades. Here, we focused on sedges, based on the dataset used in (Besnard et al. 2009) and (Rey et al. 2018). This dataset consists of a multiple sequence alignment of 79 sequences and 458 positions. Besnard et al. (2009) found 16 positions under positive selection that carry parallel amino-acid mutations in species with C4 metabolism. Although not fully confirmed experimentally, Rey et al. (2018) used these potentially convergent positions to evaluate the application of PCOC (and other approaches) to this dataset. We used the same reference positions in our analyses and comparisons. We tested 66 mutations for convergent evolution, those present in at least 3 sequences and in more than 2 independent clades. These 66 mutations are spread over 56 positions. However, 4 positions

proposed by Besnard et al (2009) did not have mutations meeting the above criteria and were not analyzed here, which means that 12 positions are (most likely) convergent, among the 56 tested. The results of the method comparisons are presented in Table 1.

Using PCOC on this dataset with a posterior probability threshold ≥0.8 (as used in (Rey et al. 2018)), 10 positions are detected among which 7 are true positives (TP). We thus find a large intersection between Besnard et al. and PCOC results, as previously described in (Rey et al. 2018). PCOC results are mostly driven by the OC component, which detects 14 positions including 8 true positives. The PC component, on the other hand, leads to lower accuracy and finds only 1 true convergent position and 4 false positives with this threshold of 0.8.

FADE detects 15 positions, including 11 TP. Even though this tool and model were designed in a different context (typically the detection of DRMs in viruses; Murrell et al. 2012), it performs very well on this dataset and outperforms PCOC with a F1 score of 0.81 (against 0.63 for PCOC).

On this dataset, ConDor selected as null model the JTT substitution matrix associated with 'freerate' rates across-sites model (Susko et al. 2003; Soubrier et al. 2012) with 3 categories (R3). The Emergence component of ConDor detects 21 positions with a higher number of EEMs than expected (Holm-Bonferroni adjusted p-value <10%), 7 of which are true positives. Emergence does not use any phenotype information and likely detects convergent mutations linked to factors other than C4 metabolism, hence the high number of detected positions that do not belong to Besnard et al. list. The Correlation component refines these results, as expected since it accounts for the phenotype and focuses on foreground convergent mutations: 7 of these 22 positions carry mutations that are positively correlated with C4 metabolism (BF>2; A780S, I588L, P540T, E572Q, S620C, H665N, F611L), among which 4 are TP. They correspond to the mutations present in the 'ConDor detection zone' in Figure 3. We notice that Correlation alone works fairly well (Table 1), without using any information on amino-acid exchangeability and biochemistry, as constitutive of the Emergence component. The combination of the two components in ConDor increases the precision and F1 score of the two components individually without however reaching those of FADE and PCOC, resulting in mild F1 (0.47). The 4 TP found by ConDor are also found by the two other methods. Among the convergent candidates, two are found either by FADE or PCOC and one only by ConDor. Similarly, most of the convergent candidates of PCOC and FADE are different, indicating that the three methods do not rely on the same mechanisms to detect mutations. Even if positions 620 and 611 were not retained by Besnard et al. as under positive selection, the results of ConDor and either PCOC or FADE support that S620C and F611L respectively, could be convergent mutations. Mutations on the two positions 780 and 665 were confirmed experimentally to have an impact on the catalytic activity and folding of the PEPC (Svensson, Bläsing, and Westhoff 2003; Christin et al. 2007). They are both found by PCOC, FADE, and ConDor demonstrating the suitability of these three approaches to detect foreground convergent mutations. Moreover, the experimental evidence on this dataset is still partial and additional convergent mutations could likely be discovered.

It is important to note that, in this dataset, the annotation of the convergent clades (C4 metabolism) was made according to the presence or absence of mutation A780S (Besnard et al. 2009) in various paralogues of the pepc genes. This annotation follows the results from this study (Bläsing, Westhoff, et Svensson 2000) where they demonstrated experimentally that mutation

A780S was a major determinant for C4-specific characteristics. We thus expect methods to find this mutation since it defines the convergent phenotype. If we were completely agnostic of convergence mutations, we would not know for sure which genes encode for a PEPC protein responsible for a C3 or C4 metabolism. To account for phenotype uncertainty, we annotated all paralogues with the phenotype of the corresponding sedge (annotated according to (Bruhl et Wilson 2007)). Hence, in this new dataset, all paralogues of a same species have the same annotation). This resulted in a change of annotation for 4 genes, from C3 to C4. We also removed from our analysis a clade of 7 sedges which were both C3 and C4. This new dataset is thus composed of 72 protein sequences, 22 being annotated as C4 (vs 23 before for 79 sequences). We tested 51 positions and 11 positions were considered as TP. By doing so, we introduced uncertainty in the annotation of phenotype, making it imperfect.

This procedure mostly affected results of PCOC and its components individually, which did not retrieve TP anymore. FADE accuracy also decreased but still lead to the best performance. Finally, ConDor was the less affected by this distortion of the phenotype and maintained a mild F1-score. ConDor is thus a robust method when phenotype is not perfectly characterized.

56 positions tested	TP	FP	FN	TN	Recall	Precision	F1 score
PC	1	4	11	40	0.08	0.20	0.12
OC	8	6	4	38	0.67	0.57	0.62
PCOC	7	3	5	41	0.58	0.70	0.64
FADE	**11**	4	**1**	40	**0.92**	0.73	**0.81**
Emergence	7	14	5	30	0.58	0.33	0.42
Correlation	7	6	5	38	0.58	0.54	0.56
ConDor	4	**3**	8	**41**	0.33	**0.57**	0.42
Phenotype uncertainty 51 positions tested							
PC	0	3	11	37	0	0	0
OC	0	1	11	39	0	0	0
PCOC	0	0	0	0	0	0	0
FADE	**9**	7	**2**	33	**0.82**	0.56	**0.69**
Emergence	6	18	5	22	0.55	0.25	0.34
Correlation	5	7	6	33	0.45	0.42	0.43
ConDor	4	**3**	7	**37**	0.36	**0.57**	0.44

Table 1: Method comparison with sedge PEPC protein dataset. *We display for PCOC, FADE, ConDor, and their sub-components, several performance indicators on the detection of convergent positions. TP: true positives. FN: false negatives. FP: false positives. TN: true negatives. Recall (TP/(TP+FN)): proportion of TP among all positions retained by Besnard et al. (2009; 12 positions tested). Precision (TP/(TP+FP)): proportion of TP among all positions retained by the given method. F1 score: harmonic mean between recall and precision. PC: positions detected with a profile change in convergent clades, with posterior probability >0.8 (as used in (Rey et al. 2018)). OC: positions with one mutation on the branches leading to the convergent clades, with posterior probability >0.8. PCOC: combination of PC and OC components with posterior probability >0.8. FADE: positions with mutations showing evolution under directional selection in convergent clades, with a Bayes Factor >100. Emergence: positions with mutations showing a number of EEMs statistically higher than expected at a p-value <0.1 (after Holm-Bonferroni correction for multiple tests. Correlation: positions with mutations positively correlated with the C4 metabolism, with a log Bayes factor >2. ConDor: combination of Emergence and Correlation. In bold: best result for each indicator.*

Based on these results, we see that the Emergence component of ConDor lacks precision and may be "too" sensitive in the absence of phenotype knowledge. This result is expected as Emergence alone cannot distinguish between foreground and background convergence. We shall see on larger datasets that both ConDor components are needed and complementary. With this dataset, PCOC results are derived primarily from the OC component that "assumes that convergent positions must have undergone a substitution on the branches where the adaptation took place" (Rey et al. 2019). With small datasets like this one, one can reasonably use PCOC recommended method, which infers "the branches where the adaptation took place" as the ones rooting the convergent clades, where all tips have the convergent phenotype. This works very well here (see Fig. 4 in Rey et al. 2018), hence the performance of PCOC. However, it is difficult (if not impossible) to define the position of these branches in larger and more complex phylogenies, due to phylogenetic uncertainty, reconstruction errors, and the use of a proxy for the phenotype. In addition, there may be confounding factors that make the OC rule too strict. In particular, we will see that the OC component does not work well on datasets with reversions or if only a subset of the convergent clades harbors a convergent mutation. Even though FADE also needs the user to define the foreground branches where adaptation took place, the hypotheses behind directional selection are less strict than with OC, as one simply assumes a mutational bias towards a certain amino acid in all branches of the convergent clades. Our approach does not require defining convergent clades (and thus branches where adaptation took place), but only the extant taxa showing the convergent phenotype, or a proxy for this phenotype, as the treatment status in HIV that we analyze in the next subsections.

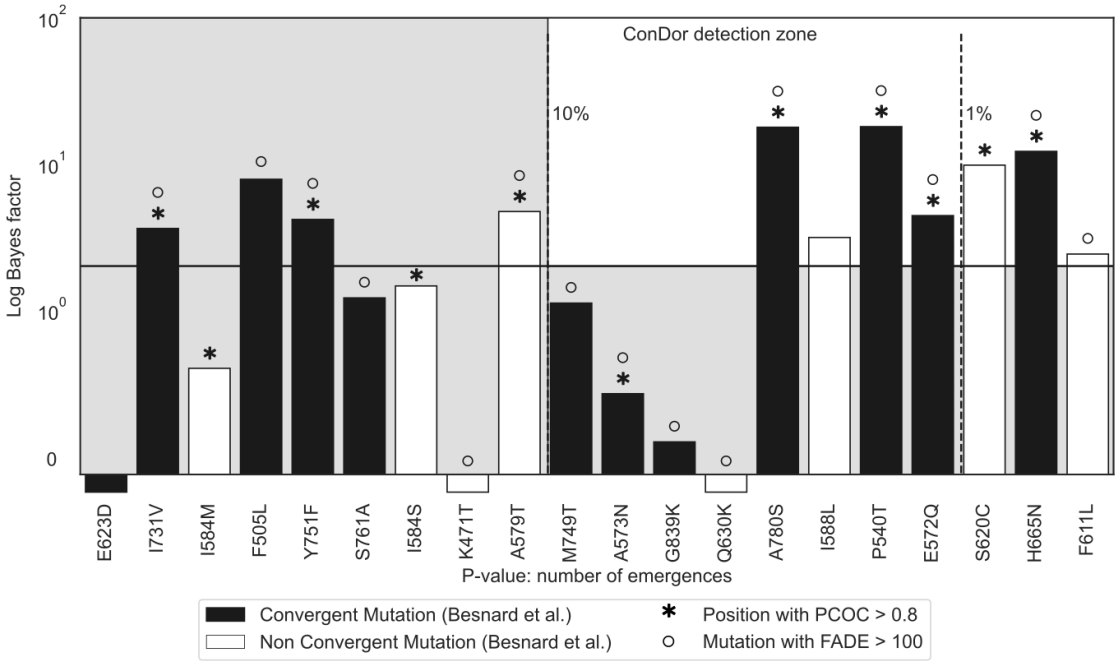


Figure 3: ConDor, PCOC, and FADE detections at the mutation level on sedges PEPC protein dataset. *We display mutations associated with a change of metabolism C3 – C4, using a data set (Besnard et al. 2009) analyzed in PCOC publication (Rey et al. 2018). Convergent mutations as retained by Besnard et al. are in black, and non-convergent mutations proposed by PCOC, FADE and/or ConDor in white. Mutations proposed by PCOC and FADE are surmounted with an asterisk and a circle, respectively. Mutations proposed by ConDor are present in the 'ConDor detection zone', corresponding to the upper-right white rectangle. Mutations are sorted on the x-axis by the p-value associated with the number of emergences (EEMs). The dashed lines represent various thresholds of Holm-Bonferroni adjusted p-values. We report on the y-axis the log Bayes Factor as obtained with BayesTraits. The*

plain horizontal line represents the threshold for strong evidence of dependence between mutations and convergent metabolism (logBF > 2).

HIV-1 M Real data

Drug resistance mutations (DRMs) occur independently in patients undergoing drug therapy and are therefore a perfect example of molecular convergence. In the case of HIV, they are well characterized and extensively studied, as their emergence can lead to treatment failure and transmission of resistant virus strains. In particular, to be identified as DRMs, mutations must meet certain criteria, including experimental validation (Wensing et al. 2019). DRMs are primarily found in the proteins targeted by antiretroviral therapy: protease, reverse transcriptase and integrase. The list of known DRMs affecting these proteins is publicly available at https://hivdb.stanford.edu/ and updated regularly. DRMs are written as "XposY", with X the ancestral (or wild-type) amino acid, "pos" the position of the substitution, with numbering based on the HXB2 reference sequence, and Y the mutated amino acid, that is, the amino acid conferring resistance. We will use this notation for all our analyses.

In our case, we are interested in mutations on the reverse transcriptase, where DRMs are numerous, varied, and have been confirmed experimentally. Furthermore, not all mutations that occur at a resistance-associated position necessarily render the virus resistant. For some positions, only a subset, or even a change toward a specific amino acid, makes the virus resistant. This case is thus particularly adapted to our method which aims at detecting convergence at the mutation level and not only at the position level. We analyze here a dataset of reverse transcriptase from HIV-1 group M sampled from 10 West and Central African countries and associated with metadata such as the treatment status of patients. We use the treatment status of the host patient as a proxy for phenotype, on the basis that most patients with detectable viral load (the virus is still circulating in the host organism) are either treated patients with treatment failure because of the development of DRMs or untreated (naïve) patients without DRMs. However, some treated patients may have an unsuppressed viral load for other reasons (e.g., poor adherence to treatment), and some naïve patients may have been infected by resistant strains with DRMs. It was first studied in (Villabona-Arenas et al. 2016) and then in (Blassel et al. 2021) from which we retrieved the data. After the removal of recombinant sequences (for which the recombination occurs within the reverse transcriptase), it contains 1858 sequences of 747 nucleotide positions that were translated into 249 amino acids. Ten subtypes are represented in this data, the major one being subtype C (37%). This dataset has several advantages for convergence detection. First, a large percentage of the sequences are from treated patients (31%). Second, the DRMs are frequent and ~26% of the sequences harbor at least 1 DRM present in at least 10 sequences. Finally, there is little transmitted resistance (12% of naïve sequences have one or more DRMs) (Villabona-Arenas et al. 2016), which means that the correlation between treated status and the presence of resistance mutations is strong (providing that DRMs are frequent enough) but not perfect.

We tested 240 mutations, corresponding to 95 positions, in total: those present in at least 10 sequences and showing more than 2 EEMS. Overall, there are 29 DRMs present in at least 10 sequences. The most common one, M184V, is found in 383 sequences. They are distributed in 24 positions. We focused on these 29 DRMs to assess the performance of our approach.

The PC component of PCOC works with a mild accuracy (F1 = 0,41) and retrieves 12 positions (corresponding to 20 mutations positively correlated with phenotype) associated with a shift in profile, of which 8 harbor DRMs. However, no position is significant for the OC component (nor PCOC).

Mutation level – Real HIV dataset	TP	FP	FN	TN	Recall	Precision	F1 score
PC	10	10	19	201	0.35	0.50	0.41
FADE HIVb > 100	**22**	66	**7.0**	145	**0.76**	0.25	0.38
FADE HIVb ~inf	13	6	16	205	0.45	0.68	0.54
FADE JTT > 100	24	69	5.0	142	0.83	0.26	0.39
Emergence HIVb	20	67	9.0	144	0.69	0.23	0.34
Emergence JTT	21	76	8.0	135	0.72	0.22	0.33
Correlation	16	3	13	208	0.55	0.84	**0.67**
ConDor HIVb	15	**2.0**	14	**209**	0.52	**0.88**	0.65
ConDor JTT	16	2	13	207	0.55	0.89	0.68

Table 2: ConDor, PCOC, and FADE performance on convergent mutation detection applied to HIV dataset. *We display for PC, FADE, ConDor, and ConDor sub-components, several performance indicators on the detection of convergent positions. TP: DRMs found by the given method. FN: DRMs not found by the given method. FP: mutations found by the given method and not DRM. TN: non-DRM mutation and not found by the given method. Recall: proportion of TP among all DRMs (29 DRMs tested). Precision: proportion of TP among all positions found by the given method. Balanced accuracy: classification accuracy while accounting for the imbalance of both classes (convergent: 29, non-convergent: 211). F1 score: harmonic mean between recall and precision. PC (profile change): mutations detected on positions with a profile change in convergent clades with a posterior probability > 0.8* (Rey et al. 2018)*. FADE (inf): mutations showing evolution under directional selection, with a Bayes Factor >100 (or inf). Emergence: mutations showing a number of EEMs statistically higher than expected at a p-value ≤ 0.1 after Holm-Bonferroni correction. Correlation: mutations positively correlated with the treatment status, with a log Bayes factor > 20. ConDor: a combination of emergence and correlation. We represent in bold the best result for each indicator.*

On HIV data, FADE has an excellent recall but also detects many non-DRM mutations (66, Table 2) results, leading to a mild F1 score. This is about the same level that the emergence component of ConDor whereas this component does not have any information on phenotype. Focusing on detections with the highest BF (~inf) FADE detects fewer DRMs than ConDor and more non-DRM mutations. Overall the performance of FADE on this dataset is good which is coherent as the models EDEPS and MEDS (now replaced by FADE) were designed for drug resistance detection in HIV (Murrell et al. 2012)

The null model inferred in this dataset with ModelFinder (Kalyaanamoorthy et al. 2017) is HIVb (Nickle et al. 2007), with 'freerates' (Soubrier et al. 2012) rates-across-site model and 4 rate categories. Using the emergence component of ConDor, we detect 87 mutations presenting more EEMs than expected, after applying a Holm-Bonferroni correction for multiple testing (adjusted p-value < 10%). Among these detections, 20 are DRMs, which represents 69% of true positives and is a higher proportion than what is expected by chance (Fisher's exact test p-value = 2e-4). However, 67 mutations are non-DRM events and we do not know if they are false positives or possible convergent candidates associated with a phenotype different from drug resistance. The mutation-phenotype correlation, at a log Bayes factor of 20, detects 19 mutations including 16 DRMs which are correlated with the treated status of patients. With this data, the phenotype-correlation component of ConDor is thus basically sufficient, with results similar (slightly more sensitive, but slightly less precise) to those obtained using both components (F1 score = 0.67 vs

0.65 with ConDor; Table 2). As expected, the correlation between DRM and treatment status is strong and phenotype is thus a good proxy for resistance. We shall see that this configuration does not happen on the rhodopsin dataset, where both components are needed. With both ConDor components, 17 mutations are detected convergent in the context of treatment resistance. Of these 17 detected mutations, 15 are true DRMs and 2 are convergent candidates (T48S and L228R).

In the case of model misspecification, illustrated by JTT line Table 2, we expect to increase the number of background convergent substitutions detected by the emergence component, represented by the FP. Indeed, we slightly increase their number, but it does not significantly lower our precision (from 0.23 to 0.22) The correlation component however allows us to smooth out this effect which shows that our method as a whole is robust to model misspecification. We note that FADE is also robust to the change of model.

As illustrated in Figure 4, the DRMs V90I, E138A, K219E, P225H, and M230L are not detected because they do not pass the log Bayes factor threshold limit even though they have a significant p-value in terms of EEMs. We read in the literature (Sluis-Cremer et al. 2014), that mutation E138A is a polymorphic mutation found naturally in naïve patients and especially in subtype C which happens to be mostly sampled from naïve patients in this dataset and the major subtype sampled (Villabona-Arenas et al. 2016). DRM E138A is thus spread with no significant difference between treated and naïve patients which explains its negative log Bayes factor. Mutations V90I, K219E, and P225H almost reach the log Bayes factor limit and are significant in terms of EEMs. Mutation M230L is present in a small number of sequences (n=14) and the correlation with phenotype is difficult to estimate with such a small number. However, this DRM is significant for the emergence component. 9 DRMs were not detected because their p-values did not pass the Holm Bonferroni correction; 8 of them have a log Bayes factor below 20, indicating a strong agreement between both ConDor components, mostly when mutations have a small number of EEMs.

A brief literature review confirmed that 1 of the 2 convergent candidates was already described as related to treatment intake, as L228R (Rhee 2003; Blassel et al. 2021) has been described as accessory mutations occurring in response to certain HIV treatments.

The emergence component of ConDor is mostly driven by the number of EEMs and the exchangeability between amino acids. A mutation with a high number of EEMs and between amino acids with a low probability of exchange will not emerge often enough under neutral evolution and will thus be detected by the emergence component. On the other hand, mutations between amino acids V and I have a high probability which explains why DRM V108Iis detected by the mutation-phenotype correlation only. This dataset contains several examples of DRMs between highly exchangeable amino acids (some of which are detected by ConDor: K70R, M41L, or almost: V90I) demonstrating that convergent mutations with effects on phenotype can arise even between amino acids sharing similar physico-chemical properties. In this case, the PC component of PCOC may not always be appropriate.

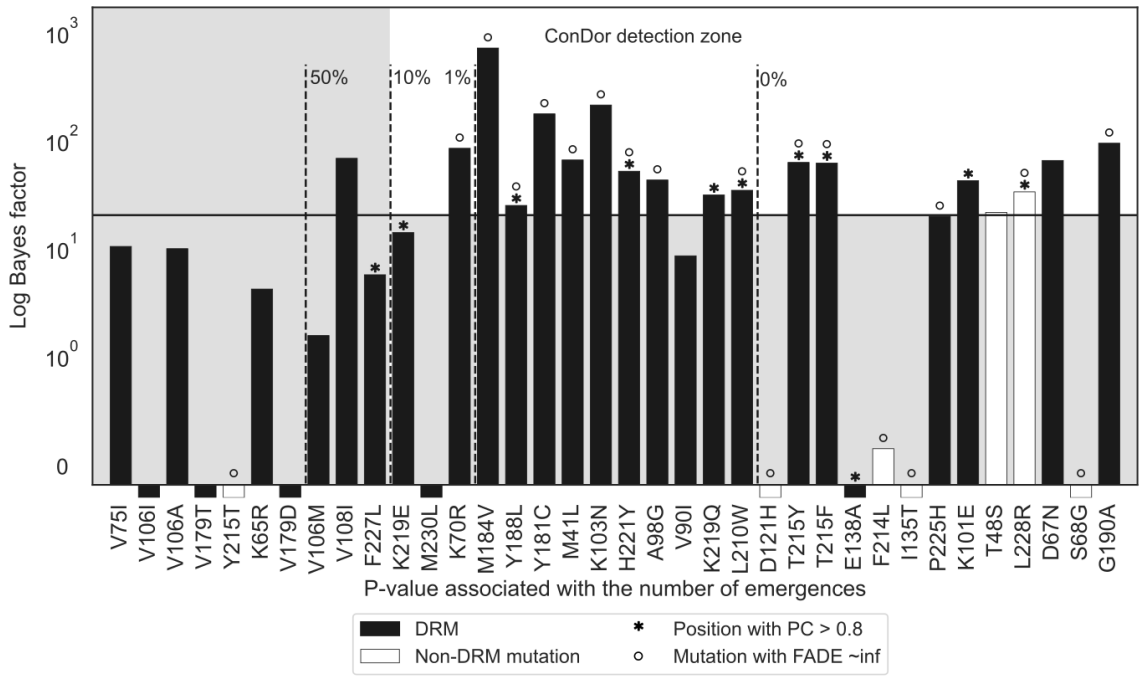


Figure 4: DRMs detection and convergent candidates on HIV real data. *We display DRMs (black) and non-DRM mutation (white), as obtained using PCOC, FADE, and/or ConDor on the HIV-1 group M MSA. If those mutations were found on positions associated with a shift in profile using PCOC, the bar is surmounted with an asterisk. If they were found associated with an infinite BF using FADE, the bar is surmounted with a circle. Mutations found by ConDor are present in the 'ConDor detection zone', corresponding to the upper-right white rectangle. Mutations are sorted by their p-value (emergence component) on the x-axis. The dashed lines represent various thresholds of adjusted p-values using a Holm-Bonferroni correction. We report on the y-axis the log Bayes Factor as obtained with BayesTraits. The plain horizontal line represents the threshold for strong evidence of dependence between a mutation and treated patients (log Bayes factor >20). Mutations that display a bar below the x-axis, are DRMs that were found to be independent or negatively correlated with patient status (treated or not).*

Rhodopsin data

Rhodopsin is a photosensitive protein pigment responsible for the eye's sensitivity to light. It is found in many vertebrates and has been shown to be under positive selection among species that evolve in different environments (Spady et al. 2005). Depending on the habitat and the amount of light available, different amino acids are observed at the same positions, which results in variations in rhodopsin structure and different maximum wavelength absorption (λmaxs). Mutagenesis experiments of engineered pigments revealed that the difference of λmaxs between most rhodopsins could be explained by 9 amino-acid mutations. In particular D83N, E122Q, F261Y, and A292S (using similar substitution encoding as with HIV) occurred several times independently (Yokoyama 2008) and resulted in functional changes. The other changes which were experimentally demonstrated to change wavelength absorption did not emerge multiple times across evolution and are thus not investigated in this analysis.

The dataset we used comes from a study in which the authors characterized substitution F261Y as convergent in fish rhodopsin, as a possible result of a transition from marine to brackish or freshwater environments (Hill et al. 2019). It contains an alignment of 2,047 sequences with 308 amino-acid positions. The sequences have been classified by the authors into two groups: species found only in marine water and species that can live (exclusively or not) in brackish or freshwater. Species annotated within the habitats brackish or freshwater can therefore also be found in

marine water. The proxy for the λmax is thus given by the environmental condition, depending on whether the fish species are found exclusively in marine water or not. This approximation of the phenotype is rather imprecise, and we expect the correlation component to work less well on this dataset than for the HIV dataset.

The reconstructed tree is well supported with three-quarters of the bootstrap supports above 70%. We tested 358 substitutions, the ones present in at least 11 sequences and more than 2 EEMs. We applied PC (profile change) and OC (one change) components of PCOC on this dataset for both environmental annotations. In total, 12 positions were found to be associated with PC, 1 of which present a mutation involved in a change of wavelength absorption of rhodopsin (E122Q, not detected by ConDor, with an adjusted p-value of ~1 and a log Bayes factor of 5.3 associated with the marine environment). However, as for the other large datasets, no position was significant for OC and by extension, PCOC.

FADE on the other hand shows a large number of detections with 74 mutations under directional selection when the foreground branches lead to the taxa in fresh/brackish water and 55 mutations with marine water. This is hardly surprising given that the environment used as a proxy for the phenotype is very vague. A large proportion of the branches are therefore considered as foreground which reduces the sensitivity of the method. Given this low sensitivity, mutations A292S, N83D, D83N, and F261Y are detected by FADE.

	PC Probabilité postérieure (> 0.8)	FADE BF (> 100)	Correlation Log-BF (> 20)	ConDor p-value (< 0.1)
N83D	150 (1.9e-11)	**7** (1.2e10)	**1** (134)	**1** (<1/10,000)
F261Y	140 (3.5e-7)	**39** (7,094)	**19** (43)	**5** (<1/10,000)
A166S*	**12** (0.95)	**34** (1.3e4)	**19** (56)	**3** (<1/10,000)
D83N	89 (1.3e-17)	**37** (1730)	**3** (102)	**1** (<1/10,000)
E122Q	**5** (0.98)	214 (1.8)	55 (6)	X
A292S	50 (6.3e-3)	**6** (6.5e11)	**12** (59)	**5** (<1/10,000)

Table 3: Detection ranking of the wavelength-associated mutations for each method. *We indicate the detection ranking of the convergent mutations associated with a change in wavelength for each method. We also precise the corresponding threshold in parenthesis. The ranks in bald indicate that the mutation is detected. A mutation is considered as detected if it verifies the following thresholds: >0.8 for PC, >100 for FADE, >20 for correlation and <0.1 for the emergence component of ConDor. The mutation marked with * is not wavelength-associated but was detected by the three methods.*

The neutral model inferred by ModelFinder on this dataset was 'MtZoa' and 'freerates' with 8 rate categories. However, we analyzed the data using LG (S. Q. Le et Gascuel 2008) to ensure a fair comparison with FADE (MtZOA was not among the possible models) which resulted in a slightly less good likelihood of the tree (-LnL = 53419.64 vs 54215.26 using LG). On this dataset, 60 mutations exhibit a number of EEMs significantly higher than expected. Combining now both

ConDor indicators, we find 19 convergent mutations which are correlated with the environment (9 with the fresh brackish water and 10 with marine water). For this dataset, we are interested in adaptations toward the marine environment as well as in fresh or brackish water. Thus, we look at all the correlations (positively correlated or negatively correlated). This is not the case for the other datasets. As shown in Figure 6, we retrieve substitutions F261Y, D83N, and A292S, and reversion N83D is also found convergent. Substitution E122Q is not found as convergent since glutamine (Q) independently emerged only 3 times according to ACR, but emerged up to 11 times in simulations.

Focusing on the five mutations experimentally demonstrated to change wavelength absorption in rhodopsin, ConDor gives the highest ranking among all methods for most of the mutations (Table 3). The correlation component alone gives high rankings for these mutations but when adding the emergence component, helps refine the results. Interestingly, mutation A166S was found by the three methods. A brief literature review indicates that it might be associated with a blue-shifting absorption (Malinsky et al. 2015; O'Reilly et al. 2016).

Using the mutation-phenotype correlation alone at a log Bayes factor of 20, we would have detected 73 mutations (40 correlated with fresh/brackish water and 33 with marine water). With this dataset, both components are needed to focus on a reasonable number of convergent candidates. This behavior was expected and highlights the benefit of ConDor to detect foreground convergent mutations without clear/precise knowledge of phenotype.

Vertebrate rhodopsin has been widely studied as an example of phenotypic adaptations resulting from specific molecular changes which are demonstrated experimentally. Notably, it has been used in the development of methods for detecting adaptive sites to compare predicted sites with real observations (Nozawa, Suzuki, et Nei 2009; Murrell, Wertheim, et al. 2012). We thus investigated whether any of the sites detected by ConDor had been previously found in positive selection analyses. While developing MEME (Murrell, Wertheim, et al. 2012) to detect positions subject to episodic diversifying selection, Murrell et al. found 19 adaptive positions in fish rhodopsin. Among them, 3 had been reported to affect the wavelength of maximum absorption in (Yokoyama et al. 2008). However, these positions did not present mutations occurring multiple times during rhodopsin evolution and are thus not convergent. Looking at their other detections, 4 positions (165, 205, 210, and 236) harbor mutations detected with ConDor as potential convergent candidates.

Discussion

In this work, we have developed the ConDor approach, which detects evolutionary convergence at mutation resolution using two components: emergence and correlation. ConDor uses only the knowledge of the phenotype (or constraint) at the level of the existing taxa without the need to infer the phenotype state of the internal nodes as in other methods. As we developed this method for the study of viruses where the phenotype is difficult to access, ConDor allows the use of environmental constraints (or other selection pressures) as a proxy for the phenotype. Thus, convergent mutations can be found even if they are present in a subset of convergent taxa or even if they occur in some taxa that are not convergent. This is particularly suitable for the study of large datasets with several thousand sequences. For example, we were able to find more than half of the DRMs on a large real HIV dataset, where the application of PCOC was not

appropriate since the underlying assumptions (OC and PC) were poorly verified. We also detected more DRMs with ConDor than using FADE whereas the assumptions of this software were made for the detection of DRMs in HIV. Although ConDor was primarily developed for the analysis of large viral datasets, it was able to detect several convergent mutations involved in the change in metabolism in a small sedge PEPC protein dataset or the absorption wavelength in a fish rhodopsin dataset. These results confirm that our method detects a realistic convergent evolution signal and could be applied to a wide range of organisms and dataset sizes. On a smaller dataset, ConDor was outperformed by PCOC and FADE but was the least affected when we used an approximate phenotype.

We tested the robustness of the emergence component of ConDor to model violation by using JTT (Jones, Taylor, and Thornton 1992) instead of HIVb (Nickle et al. 2007) as a neutral model of evolution for the study of the HIV dataset. In doing so, the sensitivity remained high, and we still detected the most frequent DRMs with the emergence component. However, the number of false positives slightly increased. With the addition of the correlation component, ConDor is robust to model violation. In real data, as we do not know the true evolutionary model, a high number of false positives could be observed with the emergence component even using HIVb substitution matrix. This confirms previous observations on the difficulty to account for background convergent mutations using standard null models (Goldstein et al. 2015; Zou et Zhang 2015b). Indeed, patterns of convergent mutations can arise by chance (not driven by adaptive forces) at positions with very high evolutionary rates and between highly exchangeable amino acids. The emergence component of ConDor detects both those types of mutations. This effect is especially strong in HIV data where sequences are closely related as it has been demonstrated that non-adaptive amino acid convergence rates decrease over time (Goldstein et al. 2015). More advanced substitution models based, for example, on mixtures or some ideas derived from the CAT model (Le, Gascuel, et al., 2008) are also used in (Rey et al. 2018) but in a different setting, or a mixture of matrix models which accounts for structural features of the positions (Le, Lartillot, and Gascuel 2008) or their evolutionary rate (Le, Dang, and Gascuel 2012), should likely enhance the emergence component, make the simulations more realistic and lower the number of background convergent detections in this component. However, those approaches are resource-intensive, and the correlation component already completes well the emergence component.

ConDor was developed to detect convergent amino-acid changes and not convergent positions, which complicates the comparison with existing approaches based on convergent site detection (e.g. PCOC (Rey et al. 2018), and to some extent positive selection methods (MEME; Murrell et al. 2012)). An adaptation of ConDor to work at the position level could be an interesting feature to add to the program. Our approach is made possible since we are working at the scale of a single protein with thousands of sequences, which provides sufficient signal and detection power. By working on thousands or even millions of positions (e.g., with bacterial genomes), ConDor would likely lack the statistical power to work at the scale of a single mutation due to multiple testing. An extension of ConDor to work at the gene level (similarly to (Chabrol et al. 2018)), or to detect convergence within a sliding window, would certainly be a useful development.

Other improvements could concern firstly, the correlation component which for now, uses discrete trait evolution models in a Bayesian framework. An extension using a more complex

evolutionary model such as amino acids substitution matrix in a maximum likelihood framework could be a useful improvement such as the model proposed in (Escalera-Zamudio et al. 2020). Secondly, the simulations in the emergence component are computationally intensive, and analytic approaches, similar to those used in (Chabrol et al. 2018) would help reduce the computing time of the approach.

Materials and Methods

Sedge PEPC protein dataset

Protein data of sedge PEPC, associated phylogeny, and data on C3/C4 metabolism were retrieved from https://github.com/CarineRey/pcoc/tree/master/data/det. We used the sequence 'Chrysithr' as outgroup following the work from (Rey et al. 2018). We used the provided phylogeny and reestimated branch lengths using protein sequences and JTT+R3 substitution model. In the experiment with phenotype uncertainty, we annotated each specie according to (Bruhl and Wilson 2007). We removed the clade consisting of Eleocharis baldwinii and Eleocharis vivipara as they are annotated as both C3 and C4.

Real HIV dataset

The HIV reverse transcriptase dataset we analyzed is based on the nucleotide alignment provided in (Blassel et al. 2021) which we downloaded from https://github.com/lucblassel/HIV-DRM-machine-learning/tree/main/data/African_dataset. This is an alignment of 3,990 HIV-1 group M polymerase sequences subdivided into naïve and treated sequences along with a metadata file indicating the treatment status if the sequence has DRM(s) and the subtype. The subtype annotation indicates that many sequences are recombinant forms which we removed if the recombinant breakpoints were found within the reverse transcriptase (if so, we cannot reconstruct the phylogeny). We then ran JPHMM (Schultz et al. 2012) to identify other possible recombinant forms. We added to this MSA, 3 subtype N reference sequences downloaded from https://www.hiv.lanl.gov/content/ sequence/NEWALIGN/align.html of reverse transcriptase (user-defined range 2550 - 3297) as outgroup, following the phylogenetic tree of HIV (https://www.hiv.lanl.gov/content/sequence/HIV/COMPENDIUM/1999/1/intro.pdf). The resulting alignment contains 1858 sequences which we translated into amino acids. DRMs were identified in the translated MSA using the 2019 list of DRMs in HIV-1 (Wensing et al. 2019)

Rhodopsin dataset

Protein data of rhodopsin and fish habitat were retrieved from https://github.com/Clupeaharengus/rhodopsin/tree/master/phylogeny_habitat. We extracted the 2,056 sequences from "spp_to_keep.txt" from the file "final_alignment.translated.fullrhodopsin.fasta" and removed 7 badly aligned sequences. We checked the quality of the alignment with TCS (Chang, Di Tommaso, and Notredame 2014) and obtained a score of 997/1000 demonstrating good reliability. We used the same sequences used for rooting as in (Hill et al. 2019) (*Huso huso* and *Polyodon spathula*). The habitat was provided in the file "rabo_allele_hab.tsv" from the repository provided in (Hill et al. 2019).

Tree reconstruction by maximum likelihood

Rhodospin phylogeny was reconstructed from the corresponding protein MSAs, using the following procedure and options. We used Model Finder (Kalyaanamoorthy et al. 2017), corresponding to parameter `-m MFP` in IQtree version 1.6.8 (Nguyen et al. 2015), to select the model of sequence evolution (substitution matrix, gamma categories or freerates model, presence of invariant positions, etc.). We used amino acid equilibrium frequencies from the model and site-specific evolutionary rates were estimated using option `-wsr`.

Since the amino acid sequences of the real HIV data were not divergent enough, we used nucleotide MSA to infer the tree topology while selecting the model with Model Finder (GTR+R10 in this case). We then reoptimized branch lengths and evolutionary rates using option -m MFP while fixing the topology using option `-te` in IQtree. We followed the same procedure to re-estimate branch lengths and rates of evolution in the sedge dataset and for the model misspecification test in HIV, although we optimized the frequencies for JTT substitution matrix using option `-FO` in the last case .

Ancestral character construction by maximum likelihood

Ancestral character reconstruction was achieved using PastML version 1.9.33 (Ishikawa et al. 2019) with option `--prediction_method MAP`. We provided one parameter file (option `--parameter`) per position, in which are written (1) the amino acid frequencies for the whole alignment and (2) the scaling factor for the studied position, corresponding to the rate of evolution of the site as estimated by IQtree. The selected substitution matrix (HIVb, JTT, resp. MtZoa) was given as input (`--rate_matrix`) using PastML option -m `CUSTOM_RATE`.

Correlation between discrete traits using BayesTraits

Correlations between convergent phenotype and mutations emerging more often than expected were measured with BayesTraits 'discrete dependent' model (Pagel 1994; Pagel and Meade 2006). Before running the software, we transformed our data into discrete binary traits. This way, the convergent phenotype was annotated as 1 and not convergent phenotype as 0. Similarly, for a given position, the amino acid change of interest had the value 1, and the other amino acids at that position had the value 0. We followed the procedure detailed in http://www.evolution.rdg.ac.uk/BayesTraitsV3/Files/-BayesTraitsV3.Manual.pdf to assess whether the dependence between both traits was more likely than their independence. The dependence hypothesis was retained if the log Bayes factor was greater than 2 for the sedge PEPC dataset and 20 for the others, these thresholds given as strong and very strong evidence against H_0 in (Kass et Raftery 1995). Priors for the transition rates were set to uniform with a range between 0-100 as described in the user guide. Mutations for which the distribution was found dependent with the phenotype by BayesTraits were retained as convergent if the correlation was positive, i.e., correlated with the convergent phenotype (a negative correlation would be with the non-convergent/ancestral phenotype). To do so, we checked that the proportion of the mutation emerging more often than expected was greater in the sequences with the convergent phenotype than in the sequences with the non-convergent phenotype. More formally, let us denote M_c the number of sequences that have the mutation M and are annotated with the convergent phenotype, M_{NC} the number of sequences that also have the mutation M but are annotated with the non-convergent phenotype NC, C the total number of sequences annotated with the

convergent phenotype, and NC the total number of sequences annotated with the non-convergent phenotype. If $\frac{M_C}{C} > \frac{M_{NC}}{NC}$, then the correlation is positive, and M is considered as a convergent mutation by ConDor.

Technical details

The whole method is implemented in a Nextflow pipeline (Tommaso et al. 2017) taking as input an amino-acid alignment, a rooted/unrooted tree, and a file containing outgroup sequences identifiers. The python libraries `numpy` (Harris et al. 2020), `pandas` (McKinney 2010), and `scipy` (Virtanen et al. 2020) were used for data frames and matrices manipulations and for the statistic tools they provide. We used `biopython` (Cock et al. 2009) for sequences and alignments manipulations. Tree traversals and analyses were achieved with `ETE 3` (Huerta-Cepas et al. 2016). Graphics were obtained using `matplotlib` (Hunter 2007) and `seaborn` libraries. All MSA (translation to amino acids, subalignments, etc.) and tree manipulations (pruning, rooting, etc.) were achieved using `goalign` and `gotree` (Lemoine and Gascuel 2021). Simulations and counting of EEMs were computed using homemade python scripts. Mutations emerging more often than expected were selected based on their p-value after correcting for multiple testing with a Holm – Bonferroni correction, with a risk alpha of 0.1.

JPHMM

We ran jpHMM to identify recombinants form on HIV-1 group M MSA and remove them if their breakpoints were found inside the reverse transcriptase gene. We used the default settings for HIV `-v HIV` and the priors provided in the jpHMM folder `-a priors/emissionPriors_HIV.txt -b priors/transition_priors.txt`.

PCOC

We used PCOC (Rey et al. 2018) to detect convergent positions based on the knowledge of the metabolism (C3 vs C4), treatment status (treated vs naive), and habitat (marine vs fresh/brackish water). We used the profile C10 (`-CATX_est 10`) with 4 gamma categories (`--gamma`) and fixed the posterior probability threshold of a position to be above 0.8 for all datasets (`-f 0.8`). For the convergent scenario (`-m`) corresponding to the different clades of nodes that exhibit the convergent transition, we first retrieved the tips with the convergent phenotype (C4 metabolism, treated, fresh/brackish water, and in a second time marine water). Then, we retrieved all internal nodes the tips of which were in the convergent environment, and completed the scenario, as described in the user guide (https://github.com/CarineRey/pcoc). The selection of the internal nodes corresponds to the conjunction mode used for FADE.

FADE

We used FADE (unpublished to date) from the HyPhy package (Pond et al. 2005) to detect mutations under directional selection. FADE requires as input a rooted tree with annotations for the set of foreground branches suspected to have undergone directional selection. We annotated the foreground branches using http://phylotree.hyphy.org/. We first selected all the branches leading to the tips with convergent phenotype and then added the internal nodes using the conjunction mode, corresponding to the annotation used in PCOC. FADE was then run using the same substitution matrix as ConDor (JTT, HIVb, and LG) and providing the same amino acid alignments as for PCOC or ConDor.

3.3 ANALYSES COMPLEMENTAIRES

3.3.1 Performances sur simulations

Dans l'article précédent, nous avons évalué et comparé les performances de ConDor sur des jeux de données réelles. Cependant avec les données réelles, nous ne pouvons pas connaitre avec certitude quelles sont les mutations convergentes vraies et les faux positifs. Il se peut que certaines mutations classifiées comme faux positifs soient de vraies convergences de premier plan mais n'ayant pas été confirmées expérimentalement.

3.3.1.1 Génération des données simulées

Afin d'évaluer le nombre de vrais et faux positifs dans les résultats de ConDor, nous avons créé un jeu de données simulé inspiré d'un cas réel, et contenant des mutations convergentes : les DRMs de la transcriptase inverse du VIH-1.

Pour cela, nous avons repris le jeu de séquences nucléotidiques du VIH-1 décrit et analysé dans l'article ci-dessus (sections Résultats et Matériel et méthodes). Nous avons extrait les positions et les séquences avec des DRMs dans l'alignement, et remplacé les résidus correspondants par des gaps. Les DRMs extraites sont celles de la liste des mutations de résistance aux traitements mise à jour en 2019 (Wensing et al. 2019). Nous avons ensuite reconstruit un arbre à partir de cet alignement sans DRM, estimé les paramètres du modèle de substitution et inféré les taux d'évolution par position. Cet arbre représente les relations de parenté entre séquences dans les données réelles et sa topologie n'est pas affectée par les DRMs (la suppression des DRMs est un procédé standard lors de l'inférence des arbres à partir des séquences du VIH). Nous avons ensuite simulé le long de cet arbre l'évolution de la séquence ancestrale déduite par PastML (Ishikawa et al. 2019) avec l'option MAP et la matrice de substitution HIVb (Nickle et al. 2007). Pour assurer la robustesse des résultats, nous avons effectué 5 simulations, donnant lieu à 5 alignements multiples sans convergence. Les DRMs ont ensuite été ajoutées aux alignements simulés aux mêmes séquences et positions que dans l'alignement original (réel). Par exemple, la mutation M184V a été trouvée dans les données réelles dans 383 séquences. Comme nous avons utilisé la topologie de l'arbre correspondant aux données réelles pour les simulations, nous avons remplacé dans les 383 séquences simulées correspondantes, tout acide aminé trouvé en position 184 par une Valine (V). Sans utiliser l'arbre et les données réelles pour créer notre jeu de données synthétiques, nous aurions placé les DRMs de façon aléatoire, ce qui aurait rendu la tâche de détection beaucoup plus facile. Nous avons appliqué cette procédure d'ajout des DRMs pour l'ensemble des DRMs connues présentes dans le jeu de données réelles. Les cinq jeux de données simulées ne présentent donc aucun événement convergent, à l'exception des DRMs ajoutées de manière « réaliste ». Afin d'éviter toute confusion entre les jeux de données simulées et les simulations nécessaires à l'estimation du nombre de convergence attendu (10'000 simulations dans ConDor), je référerai désormais aux cinq jeux de données simulés sous le terme « données synthétiques ».

La répartition des séquences traitées/non traitées dans les données synthétiques suit le même principe et correspond à la répartition dans les données réelles. Comme pour les données réelles, cela implique que le phénotype dans les données synthétiques n'est pas parfaitement connu mais

correspond à un proxy, avec de nombreuses exceptions (par exemple, la mutation M184V est présente respectivement dans 66% et 0.3% des séquences traitées et naïves).

3.3.1.2 Résultats

Les données synthétiques sont constituées de cinq alignements de séquences multiples de 1858 séquences et 249 acides aminés chacun, imitant la réverse transcriptase du VIH-1 groupe M. Au total, 53 DRMs ont été placées dans chacun des alignements sur 32 positions. Comme précédemment, nous nous sommes concentrés uniquement sur les DRMs trouvées dans au moins 10 séquences et avec plus de 2 émergences, soit, selon les simulations, en moyenne ~32 DRMs par jeu de données sur ~25 positions. La mutation la plus commune, M184V, est présente dans ~384 (21%) séquences. Cependant, ~7 DRMs sont présentes dans moins de 1% des séquences (c'est-à-dire dans 10 à 18 séquences) et devraient être difficiles à détecter. Nous avons testé en moyenne 387 mutations par jeu de données synthétiques.

Mutations – Données synthétiques	TP	FP	FN	TN	Rappel	Précision	Balanced Accuracy	Score F1
PC	11.6	10.8	20.0	344.6	0.37	0.53	0.67	0.43
FADE	**22.2**	4.40	**9.40**	351.0	**0.70**	0.84	**0.85**	0.76
FADE ~inf	15.4	1.00	16.2	354.4	0.49	0.94	0.74	0.64
FADE JTT	24.2	10.4	7.40	344.8	0.77	0.70	0.87	0.73
FADE JTT ~inf	14.8	0.8	16.8	354.4	0.47	0.95	0.73	0.63
Emergence HIVb	21.2	0.40	10.4	355.0	0.67	0.98	0.84	**0.80**
Emergence JTT	21.4	11.8	10.2	343.4	0.68	0.65	0.82	0.66
Corrélation	16.8	1.80	14.8	353.6	0.53	0.91	0.76	0.67
ConDor HIVb	16.8	**0.00**	14.8	**355.4**	0.53	**1.00**	0.77	0.69
ConDor JTT	16.8	0.00	14.8	355.4	0.53	1.00	0.77	0.69

Tableau 4: Performances de PC, FADE, ConDor et ses sous composantes sur données synthétiques.
Nous présentons pour PC, FADE, ConDor, et les sous-composantes de ConDor, plusieurs indicateurs de performance sur la détection des mutations convergentes. TP : DRMs trouvées. FN : DRMs non trouvées. FP : mutations non DRM trouvées. TN : mutations non DRM et non trouvées. Rappel : proportion de vrais positifs (TP) parmi toutes les DRMs (~32 DRMs testées). Précision : proportion de TP parmi toutes les mutations détectées. Balanced Accuracy : précision de la classification en tenant compte du déséquilibre des deux classes (convergente : ~32, non convergente : ~355). Score F1 : moyenne harmonique entre le rappel et la précision. PC (changement de profil) : mutations détectées sur les positions avec un changement de profil dans les clades convergents avec une probabilité postérieure > 0,8. FADE (inf) : mutations montrant une évolution sous sélection directionnelle, avec un facteur de Bayes >100 (ou infini). Emergence : mutations montrant un nombre d'émergences statistiquement plus élevé que prévu à une valeur $p \leq 0,1$ après correction de Holm-Bonferroni. Corrélation : mutations positivement corrélées avec le statut du traitement, avec un facteur log Bayes > 20. ConDor : une combinaison d'émergence et de corrélation. Le meilleur résultat pour chaque indicateur est représenté en gras.

Sur ces données synthétiques, PCOC a détecté ~22 mutations associées à un changement de profil (sous-modèle PC avec un seuil de probabilité postérieure > 0.8), dont ~12 avec des DRMs. Cependant, aucune position n'était significative pour OC (même à un seuil inférieur) et par extension, aucune n'était significative pour PCOC non plus, comme précédemment observé sur de grands jeux de données (données VIH ou de rhodopsine dans l'article). Les performances de PCOC sur données synthétiques sont d'ailleurs très proches des performances observées sur les vraies données VIH. D'après le Tableau 4, nous constatons que les indicateurs de performance « Balanced Accuracy » et score F1 sont les plus faibles pour PC (moyenne sur les cinq jeux de données de 0.67 et 0.43). L'absence de résultats pour la composante OC (et donc PCOC), peut s'expliquer à la fois par la taille des jeux de données et par la nature des mutations que nous

cherchons à détecter. En effet, toutes les DRMs ne sont pas trouvées chez tous les patients traités et des DRMs peuvent également être trouvées chez des patients non traités. De plus, de nombreuses DRMs se produisent entre des acides aminés partageant des propriétés biochimiques similaires (par exemple, les mutations entre V et I, V et M, ou K et R qui ont toutes des scores BLOSUM positifs) et ne vérifient donc pas nécessairement un changement de profil. Par conséquent, nous nous attendions à ce que PCOC fonctionne avec une précision modérée.

FADE au seuil de 100 et avec le modèle HIVb, obtient de bonnes performances sur les données synthétiques avec le plus grand nombre de vrais positifs (TP) parmi toutes les méthodes ainsi qu'une « balanced accuracy » et un score F1 élevés (0.85 et 0.76 Tableau 4). FADE double d'ailleurs son score F1 par rapport aux vraies données (0.76 vs 0.38 dans les vraies données) ce qui s'explique notamment par une très forte diminution du nombre de faux positifs (~4 vs 66 dans les vraies données à un seuil de 100).

En ce qui concerne ConDor, le modèle inféré des données par ModelFinder (Kalyaanamoorthy et al. 2017) était HIVb+R4 (Nickle et al. 2007), qui est la matrice de substitution que nous avons utilisée pour les générer. Ainsi, l'ensemble de l'analyse (optimisation des longueurs de branche, reconstruction ancestrale et simulations) a été réalisée avec le vrai modèle d'évolution. Dans une deuxième série d'expériences, nous avons utilisé JTT+R4 (Jones, Taylor, et Thornton 1992), pour tester l'impact d'une violation du modèle sur la composante émergence de ConDor ainsi que sur FADE. Nous voyons dans le Tableau 4 qu'en utilisant le vrai modèle, on trouve en moyenne ~22 mutations qui ont un nombre d'émergences plus élevé qu'attendu, dont ~21 sont de vraies DRMs (sur ~32 DRMs ciblées). En corrélant ces mutations avec le statut du traitement (c'est-à-dire en utilisant les deux composantes de ConDor), ~17 sont positivement associées au phénotype convergent, toutes étant de véritables DRMs. En changeant le modèle (ligne JTT dans le Tableau 4), nous augmentons le nombre de faux positifs de la composante émergence (de <1 à ~12) et diminuons sa précision. Cependant, en ajoutant les informations sur le phénotype avec la composante corrélation, nous trouvons des niveaux de précision similaires aux précédents. Nous avons également appliqué le changement de modèle à FADE qui s'est avéré moins sensible à ce changement. Toutefois, ConDor obtient la meilleure précision dans les deux séries d'expériences pour des indicateurs de performances corrects (0.77 et 0.69).

Avec le vrai modèle d'évolution, la composante émergence seule conduit au meilleur score F1 sur ce jeu de données. En effet, dans le cas des données synthétiques, nous connaissons le modèle d'évolution exact et les mutations convergentes sont les DRMs que nous avons ajoutées après avoir simulé les jeux de données. Dans ce cas spécifique, le nombre d'émergences est suffisant pour retrouver les mutations convergentes. Toutefois, dans les vraies données nous ne connaissons jamais avec exactitude le modèle d'évolution. Ici le changement de modèle (HIVb/JTT) entraîne une diminution des performances de la composante émergence seule qui n'a aucune information sur le phénotype convergent. L'analyse des données synthétiques démontre ainsi l'importance de coupler la composante émergence à la composante corrélation.

Aux seuils par défaut des méthodes, FADE se révèle plus performant que ConDor sur les données synthétiques. Toutefois, à précision équivalente (lignes FADE ~inf Tableau 4), ConDor obtient des meilleures performances que ce soit en termes de « balanced accuracy » ou de score F1 et ce pour les deux modèles testés.

Dans cette analyse, toutes les DRMs ne sont pas détectées par les différentes méthodes. Pour la composante émergence, il s'agit notamment des DRMs avec un faible nombre d'émergences ou de DRMs entre des acides aminés avec une forte échangeabilité (V↔I ; score BLOSUM 3). Je décrirai plus en détail ce comportement dans la section suivante. On remarque d'ailleurs avec les données synthétiques comme avec les vraies données, que le rappel de chaque méthode reste à peu près équivalent, démontrant la difficulté à détecter les DRMs les plus rares.

La composante corrélation seule donne des résultats raisonnables, mais nous perdons quelques DRMs et la « balanced accuracy » diminue légèrement par rapport à la composante émergence. Dans la plupart des cas, cela se produit lorsqu'une mutation a un faible nombre d'émergences. Comme la mutation est rare, même si elle est principalement trouvée chez les patients traités, le signal pour la corrélation sera faible et associé à un facteur logarithmique de Bayes non significatif. Une autre raison à cela est la façon dont nous avons généré les données. En effet, lors de la génération de ce jeu de données, nous avons ajouté les DRMs majoritairement dans les séquences provenant de patients traités, reproduisant les données réelles (voir section précédente sur la génération des données). Cependant, lors de la simulation du jeu de données, certaines mutations correspondant aux DRMs peuvent émerger par hasard dans des séquences provenant de patients non traités. Dans ce cas, la corrélation DRM-phénotype est faussée. C'est d'ailleurs ce que nous observons à la ligne corrélation du Tableau 4), où seulement 53% des DRMs ciblées sont positivement associées au phénotype convergent. Cela explique également que nous observons quelques faux positifs. Dans les données réelles, la corrélation DRM-phénotype n'est pas non plus parfaite puisque toutes les DRM ne se trouvent pas nécessairement chez les patients traités et que les patients non traités peuvent avoir hérité de souches résistantes.

3.3.1.3 Conclusions

Dans cette étude, nous avons cherché à simuler des données de convergence de la manière la plus réaliste possible. Pour cela nous avons repris la topologie correspondant aux vraies données et simulé l'évolution d'une séquence ancestrale inférée depuis les vraies données. Les mutations convergentes correspondent aux DRMs des vraies données et sont placées dans les mêmes séquences. Pour autant il est difficile de déterminer à quel point ces simulations sont réalistes. Sur les vraies données nous avions testé 240 mutations et ici ~387 pour chacun des cinq alignements synthétiques. En effet, dans les simulations, étant donné que l'on utilise des modèles de Markov, on ne tient pas compte des états des acides aminés précédemment simulés et on simule une plus grande variété d'acides aminés que dans les vraies données. Sur ce point, nos données synthétiques sont à améliorer et souffrent du même biais que la composante émergence de ConDor (voir discussion dans l'article section 3.2).

Analyser les performances des méthodes avec la véritable matrice de substitution (ici HIVb) est également peu réaliste. En changeant la matrice de substitution, le nombre de faux positifs augmente que ce soit pour FADE ou pour la composante émergence de ConDor. Ce résultat laisse penser que sur les vraies données, où l'on ne connait pas le vrai modèle d'évolution, une part non négligeable des détections non-DRMs sont des faux positifs. Au taux le plus strict de FADE (~inf) le nombre de faux positifs est toutefois réduit même avec le modèle JTT. Dans ce cas, ConDor présente les meilleures performances grâce à la composante corrélation qui n'est pas affectée par le changement de modèle. PCOC, qui repose sur les matrices à profils (C10), est la méthode la plus éloignée du véritable modèle d'évolution sur les données synthétiques ce qui explique également ses performances moyennes par rapport à FADE ou ConDor.

3.3.2 Sensibilité de la composante émergence

Dans cette partie, j'approfondis les facteurs qui influencent les performances de la composante émergence de ConDor. En l'absence de phénotype connu, il est possible d'utiliser les résultats de cette composante individuellement.

Sur les vraies données VIH-1 (voir article section 3.2), la composante émergence seule détecte 87 mutations dont 20 sont des DRMs. Parmi les 67 mutations non DRMs, nous nous attendons à trouver des vraies mutations convergentes liées à d'autres phénotypes que la résistance, mais également des faux positifs ou des mutations d'arrière-plan (voir section 2.1.4). En effet, les modèles utilisés par ConDor sont des modèles markoviens dont il a été montré qu'ils pouvaient conduire à la détection de convergence non-adaptative ou d'arrière-plan (Goldstein et al. 2015; Zou et Zhang 2015b). Dans un usage normal, ConDor discrimine les faux positifs grâce à la connaissance du phénotype. En l'absence de phénotype, nous nous sommes demandés si certaines statistiques pourraient aider à cette tâche.

Nous avons ainsi représenté sur la Figure 27 plusieurs statistiques sur les détections de la composante émergence en distinguant les 20 DRMs détectées des autres détections. Le principe est d'identifier des seuils où les distributions des statistiques correspondantes aux DRMs et aux autres détections ne se chevauchent pas. Nous avons choisi de représenter le nombre d'émergences des mutations, leur p-valeur, leur taux de substitution, le taux d'évolution du site ainsi que le z-score (écart entre la distribution des simulations et la valeur observée).

La distribution des taux d'évolution par site se révèle être différente pour les DRMs et les autres détections. A des taux d'évolution très élevés, on ne trouve pas de DRMs mais seulement des mutations candidates à la convergence. La composante émergence serait donc biaisée sur les sites rapides et sous-estimerait l'émergence de mutations. Cette observation corrobore les résultats de la littérature (Goldstein et al. 2015).

De la même manière nous avons représenté sur la Figure 28 les mêmes statistiques pour les DRMs détectées et non détectées afin de comprendre pourquoi certaines DRMs n'étaient pas trouvées par la composante émergence. On remarque que les DRMs détectées sont celles ayant le plus grand nombre d'émergences. Toutefois, certaines DRMs détectées émergent moins souvent que d'autres DRMs non détectées. Il semblerait que ce soit les DRMs entre acides aminés peu échangeables.

A partir de ces analyses, en l'absence de phénotype connu, je recommanderai aux utilisateurs de ConDor d'être vigilants quant aux mutations de convergence trouvées sur des sites avec des taux d'évolution très élevés. Une représentation telle que celle proposée Figure 29 permet facilement d'identifier les mutations aux sites les plus rapides ou entre acides aminés peu échangeables afin de se concentrer sur les mutations les plus intéressantes. Sur cet exemple, en filtrant les sites les plus rapides (top 5% : positions 40, 122, 123, 126, 135, 173, 174, 177, 178, 207, 245 et 248), on retire ainsi 17 des 67 mutations non-DRMs. Cette manipulation ne permet pas d'améliorer de beaucoup la précision de la composante émergence. Cela démontre d'une part le défaut des modèles markoviens pour l'analyse de convergence et d'autre part l'importance de la composante corrélation pour filtrer les détections.

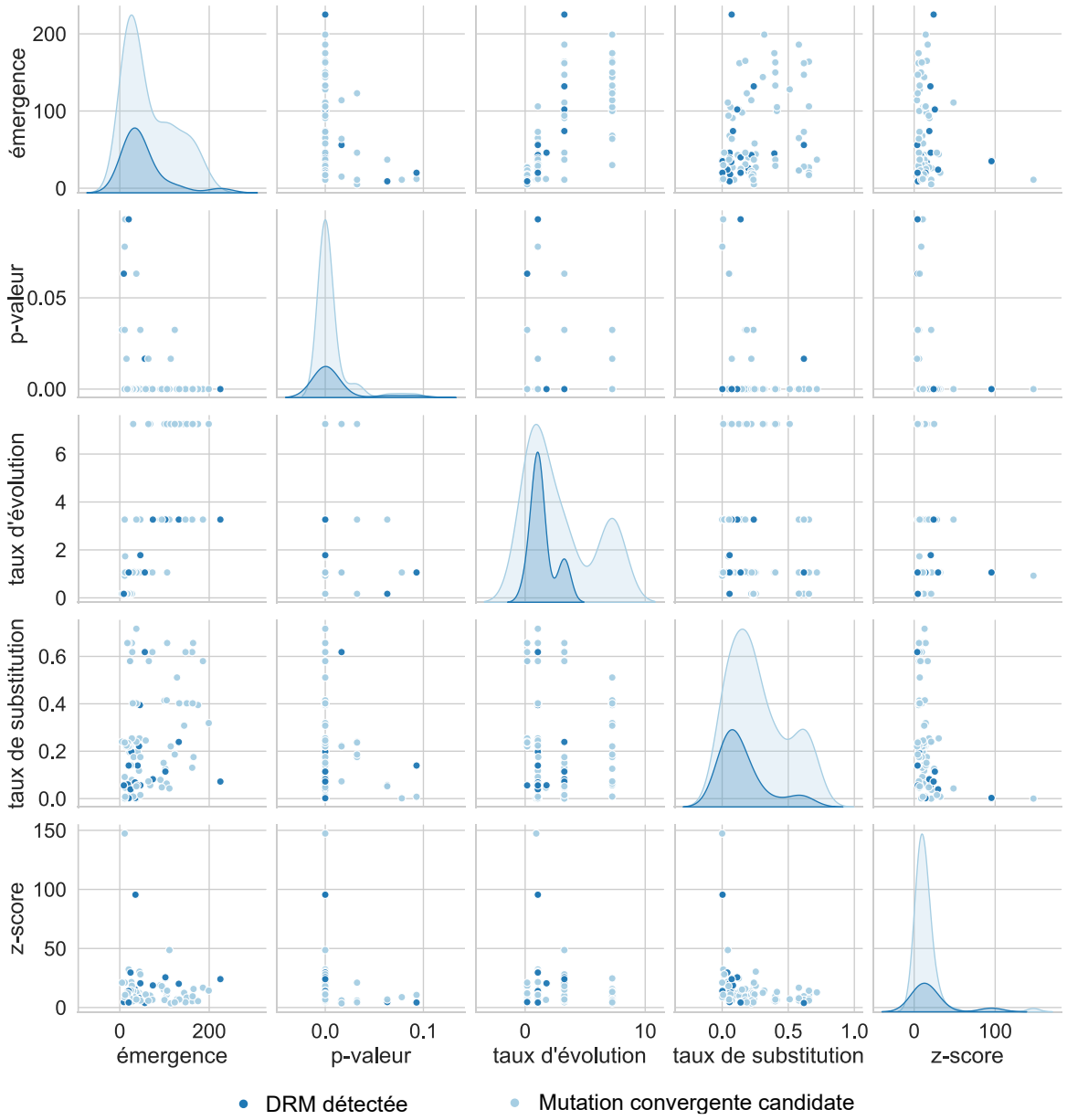


Figure 27: distributions bivariées par paires de plusieurs statistiques pour les DRMs détectées et les mutations convergentes candidates sur le jeu de données vraies du VIH-1.

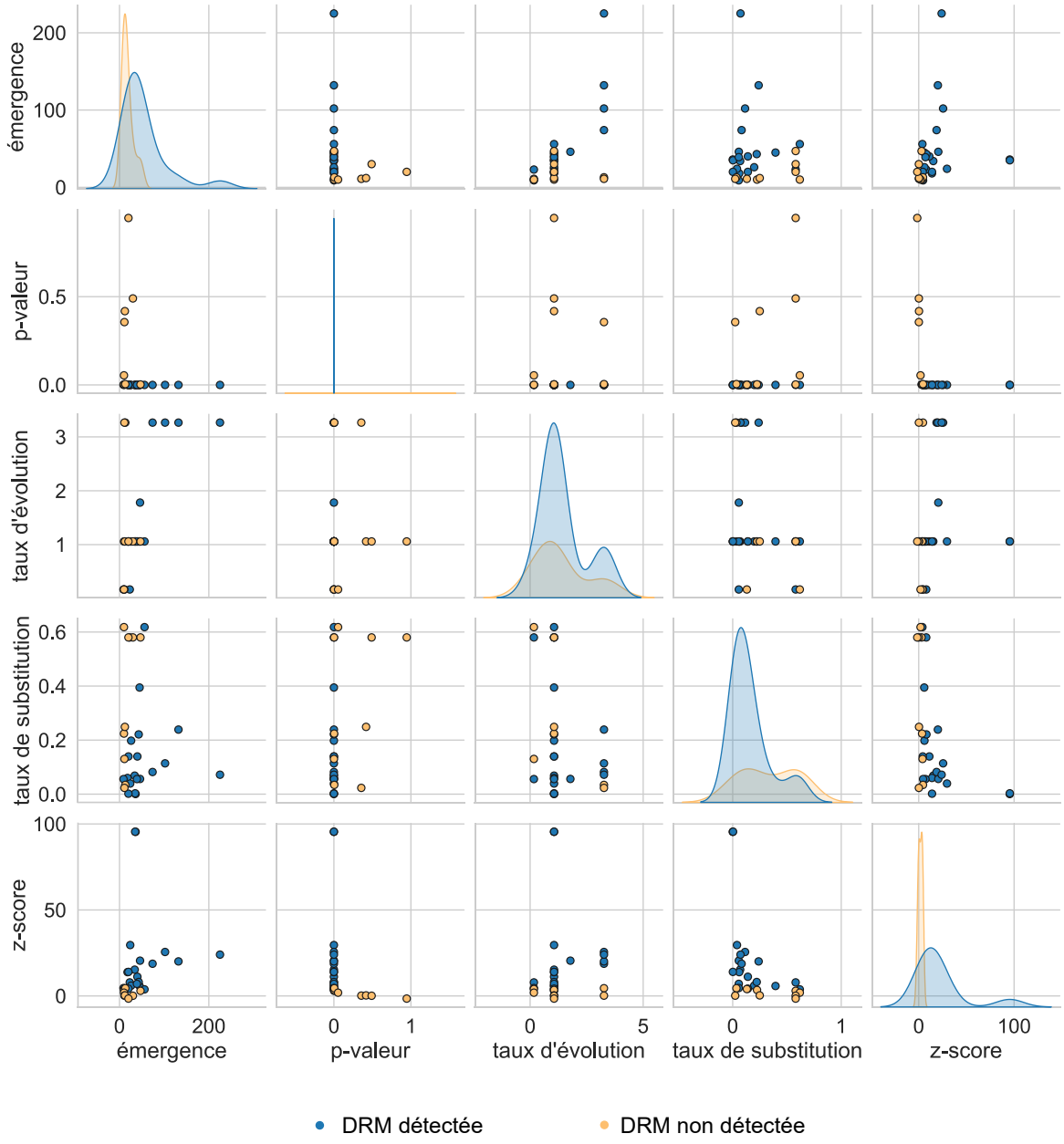


Figure 28 : distributions bivariées par paires de plusieurs statistiques pour les DRMs détectées et non détectées sur le jeu de données vraies du VIH-1.

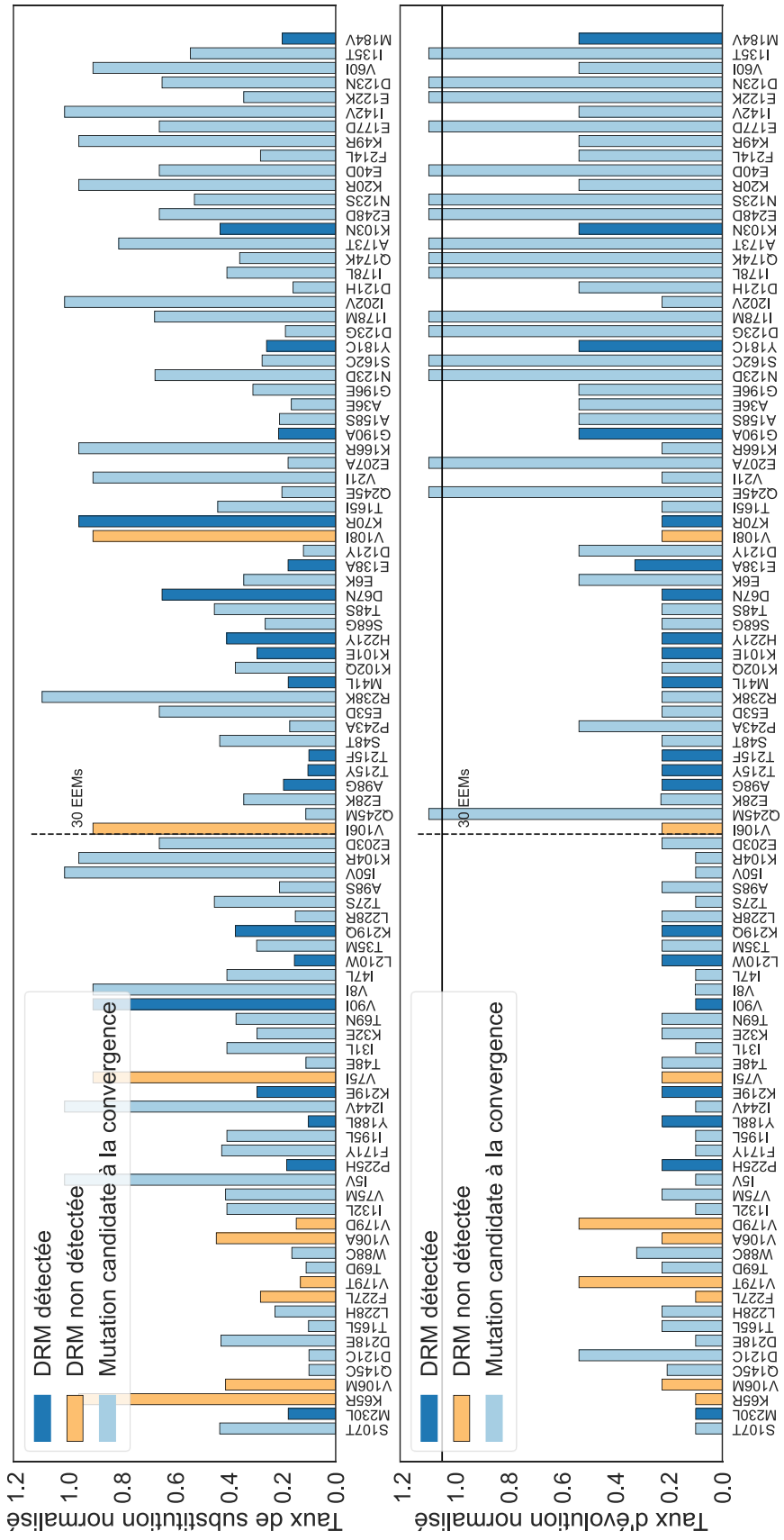


Figure 29: Mutations détectées par la composante émergence, triées par nombre d'émergence.

3.3.3 Sélection positive

J'ai appliqué aux données réelles du VIH-1 plusieurs méthodes de sélection positive de la suite logicielle Hyphy (Pond, Frost, et Muse 2005) afin d'évaluer si les codons correspondant aux DRMs apparaissaient sous sélection positive. En particulier, j'ai appliqué FEL (***F**ixed **E**ffects **L**ikelihood*; (Pond, et Frost 2005)), MEME (***M**ixed **E**ffect of **M**olecular **E**volution ;* (Murrell, Wertheim, et al. 2012)) et FUBAR (***F**ast, **U**nconstrained **B**ayesian **A**pp**R**oximation ;* (Murrell et al. 2013)).

FEL permet de calculer par maximum de vraisemblance, site par site, les taux de substitutions non synonymes (dN) et synonymes (dS). Les pressions de sélection pour chaque site sont supposées constantes tout au long de la phylogénie. FUBAR est une approche similaire à FEL à la différence qu'il s'agit d'une approche bayésienne qui renvoie ainsi une probabilité postérieure et non une p-valeur. Par ailleurs, cette approche est plus rapide et plus adaptée à de grands alignements de données. MEME se différencie de ces approches et vise à détecter si la sélection positive à certains sites s'applique uniquement sur certaines branches. Les pressions de sélection pour chaque site ne sont donc plus supposées constantes tout au long de la phylogénie mais peuvent être épisodiques. A priori, les DRMs ne devraient pas correspondre à des pressions constantes tout au long de la phylogénie. En effet les DRMs apparaissent en réponse à la prise d'un traitement ce qui devrait se traduire par des pressions de sélection épisodiques.

Sur les vraies données VIH-1, avec les options par défaut, FEL et FUBAR détectent respectivement 12 et 9 sites sous sélection positive, dont 8 sont communs (positions : 48, 49, 102, 135, 165, 173, 200 et 245). Parmi ces positions, aucune n'arbore de mutation de résistance probablement parce que les mutations de résistance correspondent à des pressions épisodiques. Comme autre hypothèse explicative, il se peut que les positions avec DRMs correspondent à une alternance entre sélection positive et négative. MEME ne parvient pas non plus à détecter les DRMs. On retrouve 4 positions avec des DRMs parmi 57 sous sélection positive ce qui n'est pas significatif (p-valeur test exact de Fisher = 0.6). MEME est particulièrement adapté à la détection de sélection de diversification et dans le cas des DRMs il s'agit plutôt de sélection directionnelle.

249 positions testées	Seuils de détection	Positions sous sélection positive	Positions avec DRMs sous sélection positive
FEL	0.1 (p-valeur)	12	0
FUBAR	0.9 (proba postérieure)	9	0
MEME	0.1 (p-valeur)	57	4

Tableau 5: résultats de FEL, FUBAR et MEME sur la détection des DRMs dans le jeu de données réelles du VIH-1. FEL et FUBAR détectent respectivement 219 et 217 positions sous sélection négative/purificatrice.

Il est important de rappeler que ces 3 méthodes ne permettent pas à priori de désigner les branches supposément sous sélection (parfois appelées branches de premier plan ou *foreground branches*). Elles ne permettent donc pas de donner une information sur le phénotype. Etant donné que les DRMs apparaissent sur une très faible fraction de séquences, il est difficile de les détecter sans cibler les branches correspondant à une prise de traitement. Cela explique sans doute qu'on ne parvienne pas à trouver les positions avec DRMs avec ces 3 méthodes. Il serait d'ailleurs intéressant de tester ces méthodes sur un jeu de données contenant uniquement des séquences provenant de patients traités.

Rappelons que pour identifier des pressions de sélection s'exerçant uniquement sur certaines branches, une des premières approches proposée (Z. Yang 1998) reposait sur la spécification a priori des branches d'intérêt. Cette approche reposait sur l'hypothèse que le reste des branches était soumis à une même pression de sélection uniforme. Si cette hypothèse n'était pas vérifiée (notamment avec un nombre de taxons important), un biais pouvait apparaitre et augmenter le nombre de faux positifs : une branche de premier plan est faussement détectée sous sélection positive.

3.4 LIMITES : ETUDE DU SARS-COV-2

Je présente ici des analyses préliminaires réalisées dans le cadre de l'étude de la convergence chez le SARS-CoV-2. En mai 2020, nous nous sommes posé la question d'appliquer la composante émergence de ConDor aux premières séquences de SARS-CoV-2 disponibles à cette époque telles que récupérées depuis la base de données du GISAID[15]. Plusieurs limites à l'analyse de ces données se sont alors imposées à nous.

Tout d'abord ConDor a été développé pour l'étude de séquences en acides aminés et malgré l'évolution rapide de ce virus, les séquences protéiques n'étaient pas suffisamment divergentes pour permettre leur analyse directe avec notre méthode. Nous nous sommes alors penchés sur la possibilité d'établir une version nucléotidique de ConDor.

Comme nous l'avons montré, les performances de la composante émergence de ConDor dépendent en grande partie de la bonne représentation des données par le modèle nul. Pour vérifier ce critère, nous avons comparé les fréquences des nucléotides obtenues par simulation avec celles des données réelles. Comme représenté Figure 30, les fréquences des nucléotides C et T (U) n'étaient pas bien simulées avec le modèle nul inféré sur les données. Il y avait en effet une forte dérive des séquences du SARS-CoV-2 vers une augmentation de l'usage de la thymine et une réduction de l'usage des cytosines. Cela a d'ailleurs été confirmé dans le travail de (Rice et al. 2021). Dans notre expérience ,nous avons utilisé les fréquences d'équilibre estimées par maximum de vraisemblance. En essayant de définir manuellement des fréquences avantageant l'émergence de T et la disparition de C, nous n'avons pas non plus réussi à simuler correctement l'évolution des séquences de SARS-CoV-2.

Notre approche repose en grande partie sur la reconstruction phylogénétique des séquences, ce qui s'avère être une tâche difficile pour les données SARS-CoV-2 (B. Morel et al. 2021). Une mauvaise reconstruction de la racine ainsi que la présence de nombreuses multifurcations (un nœud interne ayant plus de 2 enfants) pourraient impacter l'étape de reconstruction de séquences ancestrales essentielle à l'inférence du nombre d'émergences des mutations.

[15] https://www.gisaid.org/

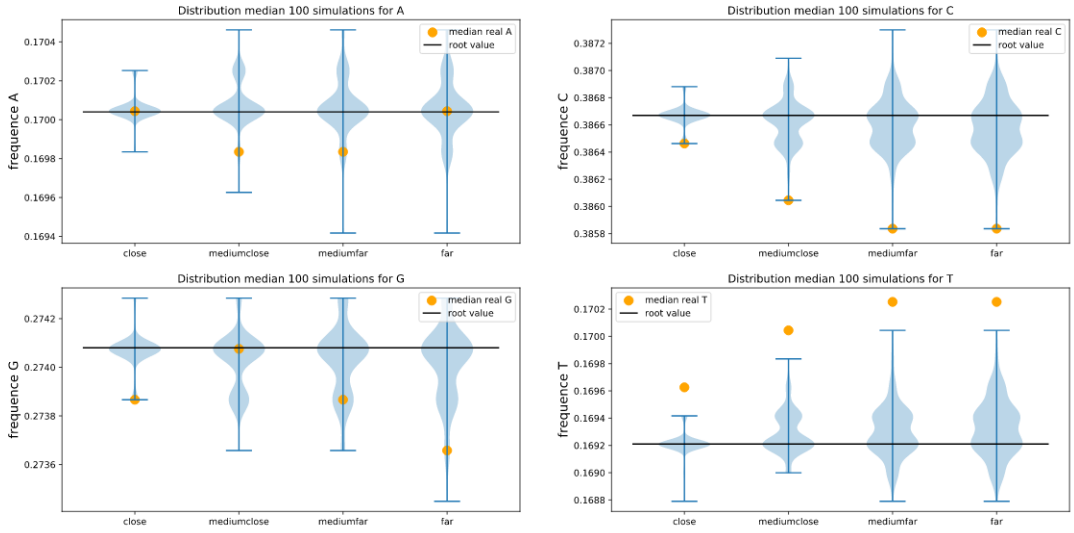


Figure 30: Comparaison de la fréquence des nucléotides dans les vraies données et dans les simulations après différents temps d'évolution.
Close, Medium close, Medium far et Far correspondent à l'éloignement (somme des longueurs de branches) des séquences par rapport à la racine.

Ces résultats préliminaires démontrent que même si notre méthode a été conçue pour l'analyse des virus, il est nécessaire que les séquences soient suffisamment divergentes et que des modèles d'évolution permettent correctement de modéliser l'évolution des séquences.

Nous n'avons pas réitéré l'analyse des données SARS-CoV-2 depuis mai 2020 et sans doute qu'en tenant compte de l'évolution des séquences et des améliorations faites dans les reconstructions phylogénétiques de ce virus, nous aurions aujourd'hui des résultats différents. Toutefois en l'absence de phénotype associé aux séquences de SARS-CoV-2 nous nous attendons à trouver de nombreuses émergences de mutations dont on ne pourra pas conclure qu'elles présentent un avantage évolutif pour le virus ou pas.

J'ai collaboré à une revue de la littérature sur les origines du SARS-CoV-2 publiée en 2021 dans le journal Comptes Rendus Biologie (Zhukova et al. 2021).

3.5 CONCLUSIONS

Dans ce chapitre, j'ai présenté ConDor, une méthode de détection de mutations de convergence pour de grands alignements protéiques. Son fonctionnement repose sur deux composantes. La composante émergence s'appuie sur des simulations et la composante corrélation utilise BayesTraits (Pagel 1994; Pagel et Meade 2006). ConDor a été conçu en première intention pour l'étude de la convergence moléculaire chez les virus. Appliqué à un jeu de données de la reverse transcriptase du VIH-1 constitué à 20% de séquences annotées convergentes, ConDor a permis de détecter plus de la moitié des DRMs avec une précision autour de 80%. L'usage de ConDor chez les virus n'est pas limité à la recherche de résistances et pourrait sans doute être employé dans le cas de changement de contraintes sélectives comme un changement d'hôte par exemple.

Cette méthode est présentée en détail dans l'article joint (partie 3.2) dans lequel nous avons montré que l'utilisation de ConDor n'était pas limitée à l'étude des virus. En particulier, ConDor a permis de détecter des mutations liées à des changements de la longueur d'onde d'absorption d'une protéine photosynthétique chez les poissons et des changements de métabolisme chez les plantes. En comparaison avec d'autres méthodes (FADE et PCOC), ConDor s'est révélé particulièrement adapté à l'étude de la convergence moléculaire lorsque la connaissance du phénotype est limitée ou quand le prédicteur du phénotype est imparfait.

Sur données synthétiques reproduisant l'évolution de séquences du VIH-1, nous avons obtenu des résultats proches des vraies données. Cette analyse a permis de montrer que la composante émergence de ConDor était sensible au changement de modèle et que cela se traduisait par une baisse de précision sur les vraies données. Nous avons montré sur données simulées que ConDor témoignait d'une excellente précision indépendamment du modèle utilisé dans l'analyse, grâce à la composante corrélation. Même si les simulations ne sont pas tout à fait réalistes, elles confirment la capacité de ConDor à détecter de manière précise les mutations de convergence.

FADE, encore en développement mais disponible dans la suite logicielle Hyphy (Pond, Frost, et Muse 2005) s'est avéré être une méthode concurrente très intéressante. Les motivations sont d'ailleurs très semblables à celles de ConDor puisque FADE cherche s'il existe un biais de substitution vers certains acides aminés à chaque position et dans certaines branches sélectionnées a priori. Sur un petit jeu de données (métabolisme des sedges), FADE s'est révélé plus performant que ConDor. En revanche, sur les jeux de données réelles VIH-1 et rhodopsine, ConDor est apparu plus performant. De même, sur les données synthétiques, à précision équivalente, ConDor a obtenu un meilleur rappel.

ConDor présente l'intérêt de travailler à l'échelle des mutations et non pas des positions. Il permet donc de déterminer quel acide aminé émerge plus souvent qu'attendu dans la condition d'intérêt. Par ces aspects il se rapproche des méthodes de détection de sélection directionnelle comme FADE auquel nous nous sommes comparés. D'autre part, l'annotation du phénotype est seulement nécessaire au niveau des taxons existants ce qui limite les incertitudes sur la reconstruction ancestrale du phénotype. Finalement, la composante émergence peut être utilisée seule, lorsqu'aucune donnée phénotypique n'est disponible par exemple. Les résultats de cette composante peuvent alors être utilisés comme première approximation de convergence moléculaire.

3.6 PERSPECTIVES

Dans la partie discussion de l'article, nous avons évoqué la possibilité d'améliorer la composante émergence de ConDor par une approche analytique. Cela permettrait à la fois de réduire le temps de calcul de cette composante et de tester des modèles mixtes plus complexes (et donc plus gourmands en calcul) que les modèles markoviens qui sont le principal défaut de notre approche. Lors des travaux préliminaires de cette thèse, nous avions par exemple testé la théorie des valeurs extrêmes pour estimer la probabilité des mutations observées dans les données. Comme certaines des hypothèses nécessaires n'étaient pas respectées par nos données (données discrètes et non continues, distribution du nombre d'émergence ne pouvant pas être approximée par des lois de probabilités usuelles) cet axe de recherche n'a pas pu aboutir. Nous restons convaincus que des travaux dans cette direction pourraient être utiles.

Dans le chapitre précédent nous avons vu qu'il existait plusieurs méthodes de détection de convergence ainsi que des méthodes de sélection pouvant être assimilées à de la détection de convergence. Ces méthodes reposent sur des définitions parfois différentes de la convergence ou ont des buts différents : recherche de mutations, positions, gènes… Selon leur définition elles sont également conçues pour l'étude de quelques séquences ou de grands alignements. Une comparaison équitable de ces différentes méthodes se révèle donc difficile. Dans ce chapitre, nous avons par exemple comparé PCOC, adapté à la détection de positions avec FADE et ConDor, adaptés la détection de mutations. Pour comparer ces deux types de résultats, nous avons ajusté les détections de PCOC aux mutations ce qui introduit sans doute un biais. Par ailleurs, les différentes méthodes reposent sur des statistiques et des seuils de significativité différents (p-valeur, facteur de bayes, probabilités postérieures) à l'origine de biais possible en faveur ou en défaveur de la méthode. Une vigilance doit donc être apportée quant aux choix des seuils de significativité pour inférer les performances d'une méthode vis-à-vis d'une autre. Dans nos travaux, nous avons fait le choix de privilégier la précision au détriment du rappel car nous ne cherchions pas à caractériser de nouvelles mutations convergentes mais à retrouver celles connues dans la littérature. Les seuils que nous avons défini pour les deux composantes de ConDor sont donc plutôt stringents et une véritable analyse de l'impact des seuils de significativité en fonction de la taille des jeux de données, du nombre de taxons convergents etc., sera à mettre en avant pour valoriser cette méthode.

De manière générale, la détection de convergence est un domaine de recherche qui pourrait être amélioré par des définitions plus claires des objectifs des différentes méthodes et de leur champ d'application. Combiné à des jeux de données simulées de convergence réalistes, cela permettrait une comparaison juste des différentes méthodes afin d'aider d'éventuels utilisateurs dans leur choix. De nombreuses méthodes présentées dans le chapitre 2 ne sont d'ailleurs pas facilement téléchargeables ou utilisables pour un utilisateur non averti. C'est pourquoi nous proposons ConDor à la fois sous la forme d'un site web et d'un standalone facilement utilisables.

4 ETUDE DE LA RESISTANCE AUX TRAITEMENTS CHEZ LE VHC

Dans ce chapitre je me suis intéressée à l'étude de la résistance chez le virus de l'hépatite C (VHC). Pour ce projet j'étais sous la supervision d'Etienne Simon-Lorière. Grâce à une collaboration avec Médecins sans Frontières (MSF) et l'Institut Pasteur du Cambodge, nous avons eu accès à des échantillons de patients traités contre le VHC en échec thérapeutique. La question générale abordée dans ce chapitre est relativement similaire au chapitre précédent puisque j'ai cherché à identifier des mutations de résistance dans des génomes viraux. La structure des données est toutefois différente puisque j'ai analysé ici des données de séquençage de nouvelle génération sur un nombre limité d'échantillons.

Je présente ici les résultats de l'assemblage des génomes viraux du VHC et de leur analyse sous la forme d'une publication en préparation. En particulier, j'ai exploré la présence de résistances avant et après traitement dans des séquences consensus ainsi qu'au niveau de la population virale par l'étude des variants mineurs.

4.1 PRESENTATION DES DONNEES ET DU PROBLEME

Des échantillons de plasma ont été prélevés chez 21 patients infectés par le génotype 6 du VHC en échec thérapeutique après un traitement antiviral. Les échantillons plasmatiques de ces patients avant administration du traitement ont également été récupérés. Nous avons également demandé à avoir accès à des échantillons de patients traités avec succès à des fins de comparaison. L'origine des différents échantillons auxquels nous avons eu accès est résumée Figure 31.

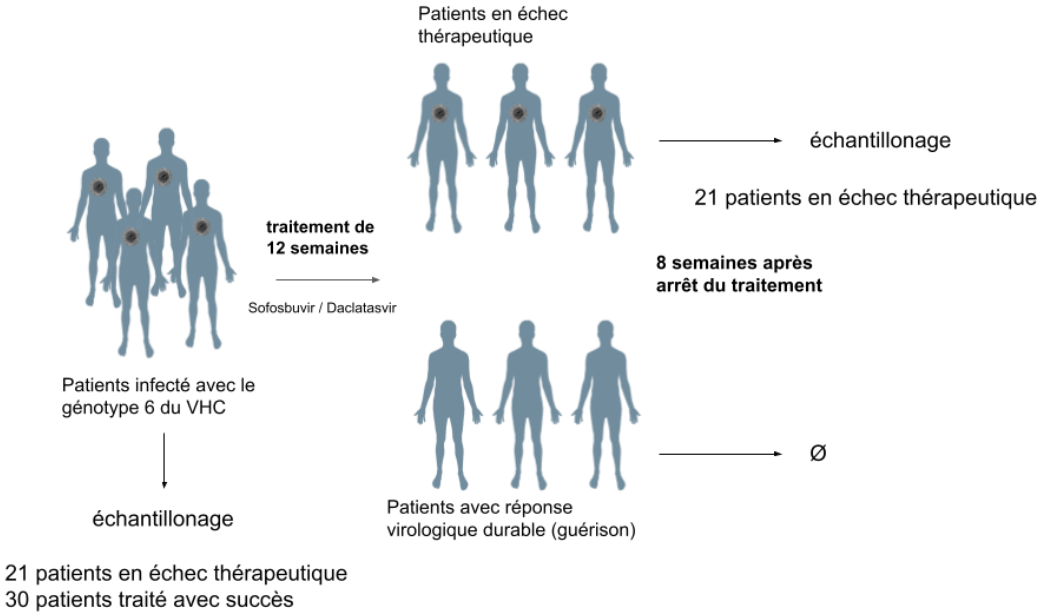


Figure 31 : Provenance des différents échantillons plasmatiques des patients infectées avec le génotype 6 du VHC.

Les objectifs de ce projet étaient multiples. Tout d'abord nous devions séquencer et assembler les génomes complets du VHC de ces échantillons afin de venir augmenter le nombre de génomes viraux du génotype 6. Ensuite, nous avons cherché à comprendre les causes de l'échec du

traitement chez ces patients en étudiant les variations génétiques au niveau des séquences consensus ou dans la diversité intra-hôte virale avant et après le traitement. Nous avons pour cela utilisé comme référence les patients traités avec succès. Ce deuxième aspect du projet avait pour but d'indiquer à MSF si un dépistage systématique des mutations de résistance était nécessaire avant administration du traitement.

4.2 ARTICLE : GENOMIC VARIATIONS ASSOCIATED WITH DRUG RESISTANCE IN HCV GENOTYPE 6 INFECTED PATIENTS FAILING DAA-BASED THERAPY.

Genomic variations associated with drug resistance in HCV genotype 6 infected patients failing DAA-based therapy.

Marie Morel, Matthieu Prot, Janin Nouhin, Jean-Philippe Dousset, Etienne Simon-Loriere* &

Philippe Dussart*

Médecins Sans Frontières – France, Phnom Penh, Cambodia

Abstract

Background: In 2016, Médecins Sans Frontières (MSF) conducted in Cambodia a project aiming at screening and treating patients infected with the Hepatitis C virus (HCV) with a regimen of Sofosbuvir and Daclatasvir (SOF+DCV). Despite a high rate (95%) of sustainable viral response 12 weeks after the end of treatment (SVR12), a few treated patients had experienced treatment failure. In the present analysis, we explored if genomic variations at the consensus level or in the intra-host diversity could cause treatment failure for selected patients infected with HCV genotype 6 (GT6), as drug resistance for this genotype is still poorly characterized. In addition, we aimed to investigate the interest in performing drug resistance screening before direct-acting antiviral (DAA) medication.

Methods: The analysis was conducted on samples obtained from 16 selected HCV-infected patients presenting treatment failure and 29 successfully treated patients. Focusing on the NS5A and NS5B proteins targeted by DCV and SOF, respectively, we studied variations at the consensus level at baseline (pre-treatment) and after failure, and in comparison to mutations observed in the pre-treatment samples of the successfully treated patients. We also explored minor variants to understand the effect of DAA treatment on intra-host diversity.

Results: We identified several mutations linked to DAA failure at the consensus level, especially with the acquisition, in post-treatment sequences, of either known resistance-associated substitutions (RASs) or mutations at resistance-associated positions. For 94% (15/16) of the samples, we also observed a significant decrease in nucleotide diversity after treatment. At baseline, only four samples presented mutations at resistance-associated positions, and no known RAS.

Conclusions: Overall, we cannot advise for resistance screening in GT6 prior to treatment and most mutations identified in this study should be further validated with in vitro phenotypic experiments. A study of genomic variations was done thanks to the sequencing of new complete genome sequences of HCV GT6, some of which are of undefined subtypes.

Introduction

Hepatitis C virus (HCV) can cause severe chronic liver disease and constitutes a major threat to human health, infecting approximately 71 million people worldwide. The development of direct-acting antivirals (DAAs) has been a breakthrough in the treatment of HCV, allowing sustainable viral response (SVR) rates of 90-95% to be achieved for most genotypes (GTs) (Falade-Nwulia et al. 2017).

Nevertheless, HCV remains a major concern for public health in absence of a preventive vaccine and with the risk of emergence and spread of DAA resistance-associated substitutions (RASs), which can compromise the effectiveness of DAA-based therapies (Lontok et al. 2015; Sarrazin et al. 2016; Kai et al. 2017; Pawlotsky et al. 2018; Dietz et al. 2018; Perales et al. 2018). It has been shown that RASs can have an impact on treatment outcome even at low frequency as they can be selected and become major variants. The clinical relevance of minor variants present at 1-15% is however still being discussed (Pawlotsky 2016; Paolucci et al. 2017; Perales et al. 2018; Nguyen et al. 2019).

HCV is genetically highly diverse, and its classification has recently expanded with currently 8 distinct GTs characterized (1–8), further subdivided into 90 subtypes (https://talk.ictvonline.org/ictv_wikis/flaviviridae/w/sg_flavi/56/hcv-classification, 2019). HCV GT1 is globally distributed, while other HCV genotypes are geographically restricted. For instance, HCV GT2 and GT5 are highly prevalent in West Africa (Markov et al. 2009; Shenge, Odaibo, and Olaleye 2019), GT3 in the Indian subcontinent (Narahari et al. 2009), GT4 in North Africa and the Middle East (Gower et al. 2014), and GT6 in China and Southeast Asia (Pybus et al. 2009; Gower et al. 2014; Blach et al. 2017).

If treatment effectiveness has been largely studied for genotypes 1 and 3, the two most common GTs, it is only until recently that several studies have evaluated the efficacy and safety of DAAs for patients infected with GT6 and living in China (Wu et al. 2019), Myanmar (Hlaing et al. 2019) Cambodia (M. Zhang et al. 2020) or Vietnam (Due et al. 2020). RASs in GT6 are thus still poorly characterized, with data available only for a few subtypes and treatments (Gottwein et al. 2018; Pham et al. 2018; Han et al. 2019; Sarrazin 2021).

In 2016, in collaboration with the Ministry of Health of Cambodia, Médecins Sans Frontières (MSF – Doctors without borders) started a pilot program providing free of charge HCV screening and treatment to any patients seeking HCV care in Cambodia (M. Zhang et al. 2020). The phylogenetic analysis of 3 000 HCV sequences available from this program revealed that GTs 1 and 6 were the most predominant (46% of each) in the country. Within GT6, the most common viral subtypes were 6e (44%) and 6r (23%) (Nouhin et al. 2019). Considering the high proportion of GT6 circulating in Cambodia, we further investigated genomic variations of HCV GT6 that might be associated with DAA treatment failure (TF) using a next-generation sequencing approach.

Results

Genome Reconstruction

Among 21 patients experiencing treatment failure, HCV full-length genome sequences at pre-treatment and post-treatment time points were obtained for 16 patients. Four patients had HCV genome sequences at only one timepoint, either pre-treatment (n=2) or post-treatment (n=2). We could not reconstruct any genome sequence in one patient. Additionally, HCV genome sequences at pre-treatment timepoint were reconstructed for 29 out of 30 patients achieving SVR12 (referred to as control sequences afterward). In total, we reconstructed 65 full-length HCV sequences from 49 different patients. Samples from patients who had experienced TF were named from HCV01 to HCV33 and control samples from HCV101 to HCV130.

Reconstructed genomes were 9339-9535 nucleotides (nt) long (5' et 3' UTR included) and each contained a single ORF of 9033-9069 nt which encode for a 3011-3023 amino-acid long polyprotein. The length of the polyprotein was constant for a given patient between pre- and post-treatment sequences.

The average sequencing depth for the 65 full-length sequences was 40,000x. However, 7 samples (pre-treatment: n=3 and post-treatment: n=4) did not reach a coverage higher than 10,000x (Supplementary Table S1 and S2). This threshold of 10,000x has previously been defined as a limit to identify minor variants present in at least 1% of the viral population with 99.9% confidence considering a noise level of 0.5% (Soria et al. 2018). In particular, two samples (HCV05 and HCV11) presented post-treatment coverage of 417x and 351x which affected the reliability of their minor variants frequencies estimations. On average post-treatment samples lead to lower sequencing depth than pre-treatment samples (p-value t.test = 0.04) without being associated with lower viral loads (Supplementary Table S1).

Phylogeny and genotyping

We reconstructed a maximum likelihood phylogeny using the 65 novel complete sequences alongside reference sequences (one sequence per each known subtype of HCV GT6 and sequences of undefined HCV GT6 subtypes) retrieved from ICTV (Lefkowitz et al. 2018) and ViPR (Pickett et al. 2012) databases (Figure 1). Each pair of sequences obtained from a given patient clustered together. Additionally, there was no specific clustering of sequences obtained from patients with TF compared to samples obtained from successfully treated patients.

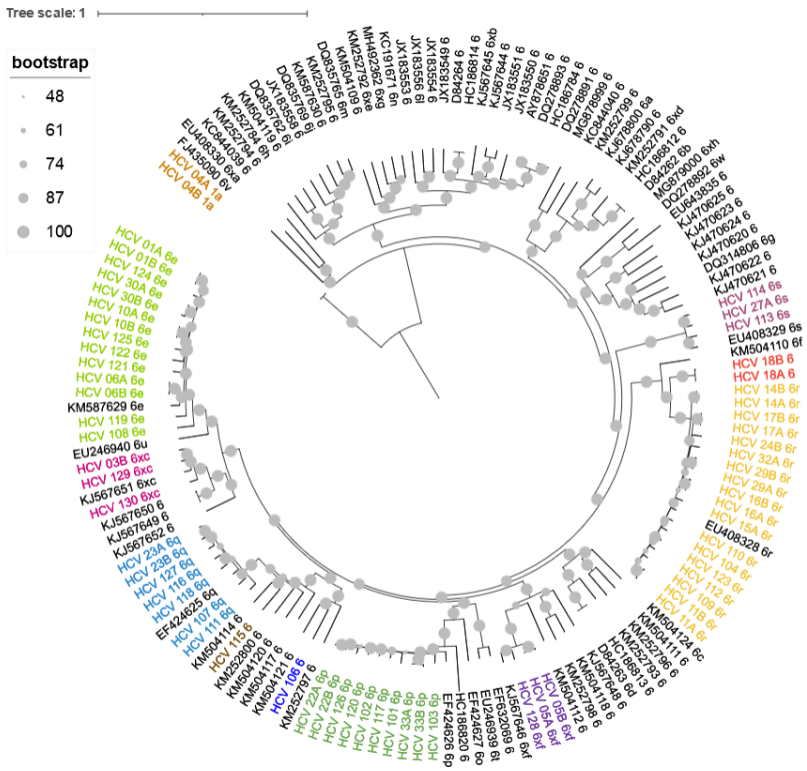


Figure 1: Phylogeny of HCV GT6 complete genomes. *The names of the newly sequenced genomes are highlighted with colors. Different colors stand for different subtypes which are indicated at the end of each name. The grey disks on branch nodes represent bootstrap supports. They ranged from 48 to 100. The tree scale indicates branch lengths of 1 substitution per site.*

Most of our samples clustered within the previously known subtypes 6a-6xh which was confirmed by the pairwise distance (>85% similarity with other sequences of the subtype) (Supplementary Figure S1). Samples HCV115, HCV106, and HCV18 did not cluster with known lineages and we could not find three isolates (partial CDS of core/E1 and NS5B or complete sequence) that shared >85% similarity with them (D. B. Smith et al. 2014; C. Li et al. 2015; Charlotte Hedskog et al. 2019). Surprisingly, sample HCV04 was not from GT6 but subtype 1a. We used it as an outgroup to root the tree. As subtype 1a is one of the most studied subtypes we kept this pair as a "positive" control to evaluate our approach on RASs. The summary of sample genotyping is presented in Table 1.

Genome lengths were variable between and within subtypes, demonstrating the high genetic diversity of GT6 (Supplementary Table S1 and S2). NS5B, which is known to be a highly conserved protein, was always the same length between different subtypes.

Table 1: Samples from HCV-infected patients for which complete genomes were reconstructed. *Samples noted with* ^{low} *had post-treatment sequences with sequencing depth lower than 500x. Samples annotated with (A) or (B) correspond to uncomplete pairs for which we could only reconstruct the genome before (A) or after (B) treatment. Samples annotated with ‡ could not be sequenced.*

Subtype	Sample ID « Control »	Sample ID « Treatment failure»
6e	HCV108, HCV119, HCV121, HCV122, HCV124, HCV125	HCV01, HCV06, HCV10, HCV30
6p	HCV101, HCV102, HCV103, HCV117, HCV120, HCV126	HCV22, HCV33
6q	HCV107, HCV111, HCV116, HCV118, HCV127	HCV23
6r	HCV104, HCV105‡, HCV109, HCV110, HCV112, HCV123	HCV11^{low}, HCV14, HCV15, HCV16, HCV17, HCV24(B), HCV29, HCV32(A)
6s	HCV113, HCV114	HCV27(A)
6xc	HCV129, HCV130	HCV03(B)
6xf	HCV128	HCV05 ^{low}
6		HCV18, HCV28‡
6	HCV115	
6	HCV106	
1a		HCV04

Mutations on NS5A and NS5B consensus sequences

We first explored the presence of RASs at the consensus level either at baseline in pre-treatment sequences or post-treatment sequences. Mutations found at resistance-associated positions that were not known as RASs for GT6 were defined in this manuscript as resistance-associated polymorphisms (RAPs). RAPs are therefore mutations found at positions associated with resistance, but which have never been shown experimentally to induce DAA resistance in GT6. Some RAPs can however induce resistance (with experimental validation) in other HCV GTs. This analysis was conducted at the amino acid level to focus on non-synonymous mutations which could have an impact on the protein structure.

Post-treatment mutations

Here we investigated breakthrough mutations in NS5A and NS5B regions that could have emerged in response to treatment. We thus examined amino-acid changes between pre- and post-treatment consensus sequences in the 16 complete sample pairs (4 pairs were uncomplete). To avoid finding emergent mutations that are not directly related to treatment, we removed from

the analysis mutations also found in our control sequences, i.e., sequences from patients who achieved SVR12.

For eight sample pairs (50%), we found mutations in the post-treatment consensus sequence that were not found in the corresponding pre-treatment sequence (Table 2) or the control sequences. Four of these mutations were found on resistance-associated positions (positions 30 and 31 for NS5A and 282 for NS5B) out of which two were known RASs. Indeed, S282T is known to reduce susceptibility to SOF (Lam et al. 2012; Han et al. 2019) in GT1 to GT6 and Q30R confers resistance to DCV for subtypes 1a and 4a. RAP S30G has been shown to confer resistance to DCV for subtype 1a and GT4 (Sorbo et al. 2018) and RAP L31I for subtypes 1a and 3a (Sorbo et al. 2018). In GT6, RAP L31I was shown to have a slight impact on the effect of DCV in vitro (McPhee et al. 2019) but not to the same extent as the resistance conferred by the RAS L31M (not found here).

We checked if these breakthrough mutations were already present in minor variants before treatment and if their frequency increased in response to treatment. None of the breakthrough RAS nor RAP were found as minor variants at baseline. However, in NS5B, R81K was found at 34% in the pre-treatment sequences from sample HCV30, S206N was found at 13% in sample HCV29, and I323V at 31% in sample HCV18. Regarding NS5A, K107E was found at 37% in sample HCV04.

Table 2: Breakthrough mutations found only in HCV-infected patients experiencing TF in response to treatment at the consensus level.
In bold are the known RAS. Mutations noted with [1] were found on resistance-associated positions.: stop codon*

Subtype	Sample ID	Mutations on NS5A	Mutations on NS5B
6e	HCV30	L31I[1]	R81K, C262R
6p	HCV33		K212R, V564L
6q	HCV23		N125K
6r	HCV11	V52I	R591Y, *592R
	HCV29		S206N, **S282T**[1], L285F
6xf	HCV05	S30G[1]	K151R, A335S, I363V
6	HCV18		I323V
1a	HCV04	**Q30R**[1], K107E	

Mutations identified by Geno2Pheno for HCV03(B) and HCV24(B) and not found in control sequences from the same subtype are presented in supplementary table S3. Without the corresponding baseline sequences, we cannot know If the mutations observed are breakthrough or not.

Baseline mutations

Resistance may not only be developed in response to treatment but could already be present at baseline in patients associated with TF (Lontok et al. 2015; Paolucci et al. 2017; McPhee et al. 2019; Sarrazin 2021). Mutations at baseline were analyzed using the Geno2Pheno algorithm (Kalaghatgi et al. 2016) by comparing with the closest reference viral sequence. Of all the mutations found at baseline, we only kept those that were found exclusively in TF-associated sequences. Indeed, if for a given subtype a mutation was found in the control sequences, it is probably not associated with resistance. In total, we analyzed 18 samples having pre-treatment sequences, including sample HCV04 from GT1.

Among 18 samples, four samples had baseline mutations at positions associated with resistance in NS5A (noted with [1] in Table 3). NS5A RAPs C54H, M62E and G28K have not been shown to confer

resistance to DCV in GT6 and only RAP Q62L is known to reduce susceptibility to DCV in other GTs (Sorbo et al. 2018).

The analysis of NS5B showed several mutations that have never been described as RAS, except mutations at position 237 (H237N of HCV subtype 6p and E237A of HCV subtype 6r) which were found in two samples (Table 3). Mutation at position 237 in NS5B is not usually described as associated with resistance but Xu et al. (2017) described a slight decrease in SOF susceptibility for mutation N237S in GT6a. Similarly, E237G was shown to reduce susceptibility to SOF (1.3-fold change) in a GT1a replicon assay (D. Wyles, Mangia, et al. 2017). We also noted that mutation Y203H was found in five samples of subtype 6r (3 times associated with C575S), which could be an indicator of convergent evolution. However, in HCV subtype 6e histidine (H) at position 203 is the wild-type amino acid.

Overall, there was no clear indication of resistance at baseline which could be observed at the consensus level for NS5B protein and only hypothetically for NS5A. All polymorphisms found at baseline on NS5A and NS5B resistance-associated positions are shown in Supplementary Tables S4 and S5.

Table 3: Baseline mutations found exclusively in sequences from HCV-infected patients experiencing TF*. Mutations noted with [1] were found on resistance-associated positions.*

Subtype	Sample ID	NS5A mutations	NS5B mutations
6e	HCV01	I63L, A135N	D202E, R206K, A273T
	HCV06	C54H[1], V83T, H159Q, S174T, S181G	D66H, H203Y, R566Q
	HCV10	R41K, Y161F	L47I, R206Q
	HCV30	Q62L[1], V153I, T213M	S29A
6p	HCV22	I52L, M62E[1], T64E, D103G	I131S
	HCV33	R6K, R123K, T213Q	I131M, K181R, S227T, H237N, V338A, T364S, L392M, V442I, V552T, V564M, L587W
6q	HCV23		D352E, L362M
6r	HCV11	Q44R, T204N	V59M, Y203H, R234C, E237A, S334G, V454I, I520V, C575S
	HCV14	I34V	V57L, T149I, Y203H, T235V, N257S, C575S
	HCV15	G28K[1], R41Q, Q44R, T64S, A146T, V156I	
	HCV16	A114G, E116D, N127S	A94T, A185E, R270Q, A324S, A327T, D353E, V450I, M469T
	HCV17	V99I, Q125H, A197V, V198A, A207T	A9T, I116V, Y203H, V450I, C575S
	HCV29	R176T	Y203H, D353H, V389T, A488V
	HCV32(A)	N127H, V131T, A146T, T213S	K77N, Y203H, S206G
6s	HCV27(A)	K73R, T116N, V119A	V59I, N310D, A487G, V585I
6xf	HCV05	T21S, T64A, K73R, H75V, T79M, V124R, Y129F, L158I, V173I	I11V, D73A, Q461A
6	HCV18	C53S, G125K, S176T, S180H	T85V, S131T, K181R, T248V, I520V
1a	HCV04	E181D, I144V	S5T, D135Q, T213N, N444D, S506N

Low-frequency variants at resistance-associated positions

As patients were sampled 12 weeks after the end of treatment, RASs could have disappeared from the consensus but still be present as low-frequency variants. It has also been shown that RASs present at baseline in low-frequency variants could be selected during treatment (Hedskog et al. 2015; Kai et al. 2017; Perales et al. 2018; Sarrazin 2021).

The analysis of the NS5A region showed that low-frequency variants at baseline (pre-treatment) did not appear to be associated with DAA resistance (Table 4). The variants found were not known as RAS and most of them were also found at consensus in other samples for the same subtype or in the sensitive samples which are detailed in supplementary table S5. Low-frequency variants M28R, C54S, M62T, and A92S were the only possible RAPs that could explain TF. Moreover, C54S and A92S were found at a frequency equal to or higher than 10%, a frequency at which baseline RAS have already been suspected to be selected upon treatment (Perales et al. 2018). None of these RAPs were however found in post-treatment samples either at consensus or in the low-frequency variants.

Table 4: Low-frequency variants found at resistance-associated positions. *Mutations in bold are RAS. Underlined mutations were not found at consensus in patients successfully treated, they are likely RAPs. Between brackets is noted the frequency associated with each minor variant. *: stop codon. S282T?: possible footprint of an ancient T at position 282. low: lower confidence for post-treatment low-frequency variants.*

		Pre-treatment samples		**Post-treatment samples**	
Subtype	**Sample ID**	**NS5A**	**NS5B**	**NS5A**	**NS5B**
6e	HCV01				
	HCV10	V28M (1.7%)		**P58S** (4%)	
	HCV30	C54Y (37%), C54S (10%), L62Q (34%)		S30A (7%)	
	HCV06	H54Y (2.5%)			
6p	HCV22	M28R (1%)			
	HCV33	M62T (1.2%)			L320* (1.1)
6q	HCV23	V28M (1.8)			
6r	HCV14				**S282T?** (8%)
	HCV15	V62A (6%), A92S (15%)			
	HCV16	V62A (2%)			
	HCV17	V62A (2%)	L320* (1%)		
	HCV11^{low}			T54A (5%)	
	HCV29	T28A (7%), V62A (1.5%)			
6xf	HCV05^{low}				
6	HCV18		V321I (1%)		
1a	HCV04			**R30H** (3.9%), **Y93H** (1.4%)	

Interestingly, in one sample of subtype 6e (HCV30), the RAP Q62L found previously at baseline (Table 3) was associated in pre-treatment sequences with L62Q at a frequency of 34% (Table 4). However, in post-treatment sequences, we could only find the leucine (L) at position 62. Thus, the wild-type amino acid Q at position 62 which were present at 34% in pre-treatment samples disappeared after treatment in sample HCV30.

After treatment, we found RASs in the low-frequency variants of NS5A in samples HCV10 and HCV04 (highlighted in bold in Table 4). In sample HCV04 we observed a RAS at position 30 at consensus level after treatment, which means that both major and minor variants at this position are RAS. Interestingly, minor variant R30H was obtained through two substitutions in the same codon that always occurred concomitantly.

In sample HCV14, we found two minor variants at codon 282 in NS5B that occurred on positions 1 and 2 of the codon. The one at the second position results in a resistance mutation (S282T), but as both variants occurred together, they resulted in a silent mutation in post-treatment sequences. In the pre-treatment sequence, the codon usage was almost exclusively AGC (99.9%) and in the post-treatment sequence, we observe both codons AGC (~92%) and TCC (~8%). Even though they both encode for a serine, during treatment there must have been an intermediate state being either TGC (encoding for a cysteine) or ACC (encoding for a threonine, associated with resistance).

Treatment bottleneck effect on the nucleotide diversity in response to treatment

Finally, we investigated the effect of treatment on viral nucleotide diversity. Intra-host viral diversity was compared between samples using an unbiased metric (Zhao and Illingworth, 2019) which takes into account the coverage and length of the samples.

We observed a strong decrease in the nucleotide diversity after treatment for all samples except sample HCV01 (Supplementary Figure S2). The difference in distribution was highly significant between pre- and post-treatment samples, as illustrated in Figure 2. Interestingly, the viral population from pre-treatment TF samples presented nucleotide diversity comparable to control samples that did achieve SVR12 (t-test p-value=0.11).

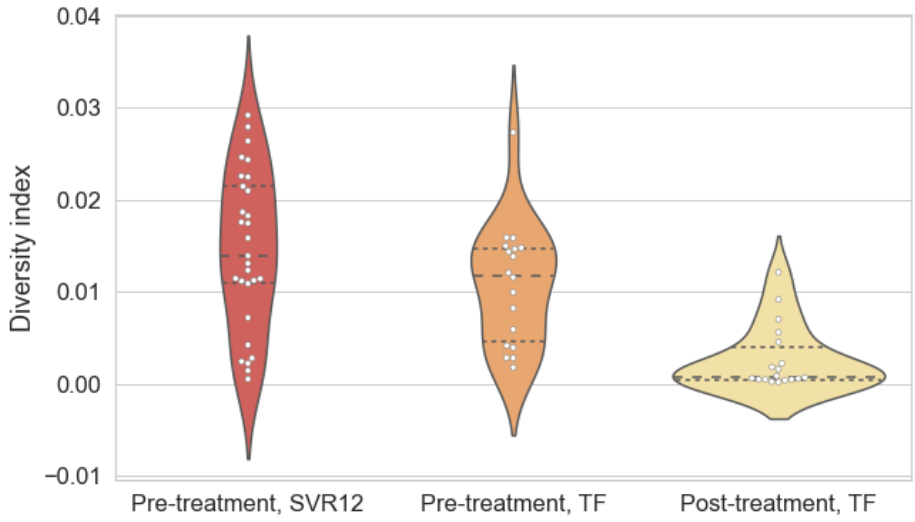


Figure 2: Distribution of the nucleotide diversity in pre- and post-treatment samples and samples achieving SVR12*. Paired t-test p-value=1.46e-4 between pre- and post-treatment samples. Independent t-test p-value=0.113 between pre-treatment samples with TF and achieving SVR12. Independent t-test p-value=1.1e-6 between post-treatment samples with TF and pre-treatment samples achieving SVR12. The dotted lines inside the violin plots represent the first, second, and third quartiles. The white circles represent the diversity index for one sample. We only plotted the diversity index for the 16 complete sequence pairs.*

Discussion

In this work, we explored if genomic variations at the consensus level or in the intra-host diversity could cause treatment failure for patients infected with HCV genotype 6 (GT6), as drug resistance mutations for this genotype are not exhaustively characterized.

To this aim, we selected 20 GT6 HCV-infected patients (16 with complete and 4 non-complete sequence pairs pre- and post-treatment) who failed to achieve SVR12 when treated with a

SOF/DCV regimen. We explored genetic variation in NS5A and NS5B which are the target of DAA at the level of consensus or intra-host viral diversity in pre- and post-treatment sequences, compared with GT6 sequences obtained from 29 successfully treated patients as controls.

At the consensus level, we identified the fixation of RASs (S282T in HCV29 NS5B and Q30R in HCV04 NS5A) and RAPs on NS5A (L31I in HCV30 and S30G in HCV05) in 20% of TF patients 12 weeks after the end of treatment. Some sequences may have been resistant prior to treatment, and we found baseline RAPs (C54H, Q62L, M62E, G28K on NS5A and possibly H237N and E237A on NS5B) in 6 samples. We also found recurrent baseline mutations in NS5B, present only in patients associated with TF, which could be interesting candidates for resistance (Y203H and C575S).

Within their host, RNA viruses circulate as a swarm of variants and RASs or RAPs may exist at low frequencies prior to treatment. We found evidence of resistance in the low-frequency variants with 3 RASs (P58S, R30H, and Y93H) and 2 RAPs (S30A and T54A) in NS5A after treatment, while we found a possible genetic footprint of RAS S282T in HCV14 and RAP L320* in HCV33 in NS5B. Indeed, the change in codon usage a position 282 is coherent with observations made in (Andreas Walker et al. 2017) where they report a viral breakthrough of a patient treated with a regimen of SOF/DCV. At the time of the breakthrough, they found that S282T was predominant and then disappeared during follow-up. After treatment, they still noticed a change in the codon usage at position 282 after the reversion toward the wild-type amino acid. At the pre-treatment timepoint, indications for resistance at low frequency were weaker with the identification of 4 RAPs in NS5A (M28R, C54S, M62T, and A92S) and 2 RAPs in NS5B (L320* and V321I) but no RAS.

We did not observe the emergence of mutations at the consensus level for half of the TF patients and this figure rises to 75% if we look at the emergence of RAP or RAS. A hypothesis could be that since the post-treatment samples were taken after 12 weeks of follow-up, most of the NS5B mutations are likely to have disappeared as they are supposed to decrease the fitness of the protein. It is also possible that resistance arose due to undescribed mutations or combinations associated with DAA resistance. Interestingly, the sample for which we best explained the treatment failure was our “positive” control: sample HCV04 from subtype 1a. In this sample, we found one RAS at consensus after treatment and two in the low-frequency variants after treatment. This confirms that resistance mutations have been well characterized for GT1, but that further efforts are needed for GT6.

We observed a large reduction in nucleotide diversity after treatment in TF patients which suggests good compliance. The observation of bottleneck in the nucleotide diversity was previously done in adapting populations of HIV (Pennings, Kryazhimskiy, and Wakeley 2014) or in HCV (R. A. Bull et al. 2011; Gutiérrez, Michalakis, and Blanc 2012; da Silva et al. 2017). This shows that despite only a few RASs observed post-treatment in the low-frequency variants, there is a clear effect of treatment on the viral populations. It is a further argument in favor of treatment resistance occurring primarily in response to treatment.

In this work, we have studied a limited number of genetically diverse sequences, since belonging to various subtypes of GT6. Even though positions associated with resistance were previously characterized, on GT6 this is still an ongoing work. Thus, the RAPs we have proposed here, either at baseline or newly acquired, have not been shown to confer resistance in GT6 and require

experimental or structural validation. We believe that studying larger datasets could provide more insights to identify RASs in GT6.

Overall, we found more signs of possible resistance on NS5A than on NS5B. Notably, it has been previously shown that NS5A resistance mutations could be observed several months after stopping treatment whereas in NS5B RASs tend to rapidly reverse towards the wild-type amino acid (Hedskog et al. 2015; D. Wyles, Dvory-Sobol, et al. 2017; D. Wyles, Mangia, et al. 2017). Follow-up with multiple time points throughout treatment and after cessation of treatment would have been ideal to track changes in the genetic composition of the viruses.

The results on low-frequency variants should be taken with caution as only one RT/NGS run was made per sample and some of the low-frequency variants could be sequencing errors. This is particularly true for the post-treatment sample HCV11, which exhibits a coverage ranging between 100 and 971 on the NS5A, NS5B region and for which we identified the post-treatment RAP T54A on NS5A. Moreover, short-read NGS makes it more difficult to detect association of mutations in the low-frequency variants.

Finally, this work provides interesting insights into the resistance of several HCV GT6 subtypes in terms of consensus sequences, minor variants, and viral diversity. We hope that with the sequencing of more samples of GT6 combined with directed mutagenesis experiments, these leads will be further developed.

Material and Methods

Patient cohort

Between September 2016 and June 2019, 9,158 HCV-infected patients initiated a DAA regimen containing Sofosbuvir (SOF) and Daclatasvir (DCV) in the framework of the pilot program set up by MSF. The treatment efficacy was high, with 97.2% of patients achieving SVR12. In contrast, 2.8% of patients experienced treatment failure (TF) (Zhang et al. 2020). In the present study, we selected a set of 21 patients who were infected with HCV GT6 and who had TF (Nouhin et al. 2019). Plasma samples were obtained at two time points: (i) before treatment (pre-treatment) and (ii) at 12 weeks after stopping treatment (post-treatment). Additionally, we included pre-treatment samples obtained from 30 patients successfully treated who achieved SVR12, as controls.

Resistance associated substitutions

We focused on non-structural proteins (NS) 5A and NS5B respectively targeted by DCV and SOF. RASs in response to DCV and SOF have been described on positions 24, 28, 30, 31, 32, 38, 54, 58, 62, 92 and 93 for NS5A and positions 96, 159, 282, 289, 316, 320 and 321 for NS5B (Bunchorntavakul, Chavalitdhamrong, and Tanwandee 2013; Z. Li et al. 2017). Our analysis focused on HCV GT6. A drug RAS in HCV GT6 was defined as mutations that have previously been shown to reduce the susceptibility of HCV GT6 for that specific drug. Considering the lack of data for genotype 6 and its genetic diversity, we chose to retain all of the previously unknown mutations observed in any of the resistance-associated positions and defined it as a resistance-associated polymorphism (RAP). Only mutations that have previously been shown to decrease response to the treatment in GT6, were considered as RASs.

Library preparation and metagenomic sequencing

RNA was extracted from 140ul of plasma using the QiaAmp Viral RNA kit (Qiagen), followed by TurboDNase treatment (Ambion) and purification using Agencourt RNAClean XP beads (Beckman Coulter Genomics). We next depleted host ribosomal RNA (rRNA) using custom DNA probes and RNAse H treatment as previously described (Matranga et al. 2014). rRNA-depleted samples were purified using Agencourt RNAClean XP beads and eluted in 10ul of nuclease-free water (Ambion). RNA was converted into double-stranded cDNA in two steps. First, RNA was reverse transcribed using random primers and SuperScript IV (Invitrogen). Second-strand cDNA was generated using E. coli DNA ligase, RNAse H, and DNA polymerase (New England Biolabs) and purified using Agencourt AMPure XP beads. Libraries were then prepared using the Nextera XT DNA Library Prep Kit (Illumina). The size and quality of the libraries were checked on a Bioanalyzer instrument (Agilent), then pooled together and sequenced on an Illumina NextSeq500 (2 × 75 cycles or 2 x 150 cycles).

Genome assembly

Whole-genome assembly was done with de novo assembly and mapping to a reference genome. We used Metaspades (Nurk et al. 2017) for the de novo assembly. Annotation of the reconstructed contigs was done with DIAMOND Blastx (Buchfink, Xie, and Huson 2015) on the NCBI non-redundant database. The reconstructed HCV contigs were then used as reference for the mapping. In the case where they did not cover the entire genome, we filled the uncovered areas with a reference sequence (from RefSeq NCBI database) of the subtype identified with DIAMOND Blastx.

Mapping of the reads on the reference sequence (either fully reconstructed de novo or a hybrid sequence between de novo contigs and reference sequence from RefSeq) was then done using the QIAGEN CLC Genomics Workbench (https://digitalinsights.qiagen.com). Paired-end fastq files were first trimmed with Trimmomatic (Bolger, Lohse, and Usadel 2014) to remove Illumina adapters and low-quality reads. Then we mapped the reads presenting 70% similarity and 70% coverage with the reference sequence and extract consensus as new reference. On this new reference, we mapped the reads with 90% similarity and 90% coverage and extract again the consensus which we consider as the consensus sequence of the reads in our sample. The mapped reads were finally extracted with SAMtools (Li et al. 2009) and sorted for low variants identification and analysis. We visually checked the coherence of the mapping using IGV (Robinson et al. 2011) and eventually adapt the stringency of the mapping (similarity, coverage) accordingly. We ran RDP4 (Martin et al. 2015) using genotype reference sequences from the Los Alamos database (https://hcv.lanl.gov/content/sequence/NEWALIGN/align.html) to seek recombination events in the newly reconstructed genomes and did not find any.

Phylogeny

The phylogeny was reconstructed using the 65 complete sequences newly assembled and adding one sequence from each known subtype and sequences from undefined subtypes. Nucleotidic sequences were retrieved from ICTV and ViPR databases. We aligned them using MAFFT and constructed the tree with iqtree, *options: -m MFP -wsr -bb 1000* (Nguyen et al. 2015). Pairwise distances were calculated with Goalign (Lemoine and Gascuel 2021). The visualization was

obtained with Itol (Letunic et Bork 2019). Alignment and reconstructed tree are provided in the supplementary files “align.fa” and “align_tree.nwk”.

Variants detection at the consensus level

Variants at the consensus level were identified with Geno2Pheno (Kalaghatgi et al. 2016). We analyzed with Geno2Pheno the consensus sequences of the pre- and post-treatment samples for patients with TF as well as sequences from successfully treated patients. We added in our analysis GT6 reference sequences from treatment naïve patients to increase the number of control samples (6p: n=2, 6r: n=5, 6q: n=7, 6e: n=10, 6xf: n=2). The whole results from Geno2Pheno are given in the supplementary file “geno2pheno_results.tsv”.

We retained as interesting all mutations on NS5A or NS5B observed in the samples from patients experiencing TF, which were not found in the sensitive sequences of the same subtype. Those mutations can be observed either before (baseline mutations) or after treatment (breakthrough mutations).

Low-frequency variants and diversity index

We used Ivar (Grubaugh et al. 2019) and Lofreq (Wilm et al. 2012) to call minor variants from the bam files. We focused on variants with a frequency higher than 1% and found on resistance-associated positions. Intra-host viral diversity was compared between samples using an unbiased metric (Zhao et Illingworth 2019). In our analysis, we focused on variants with frequencies above 1% both among pre- and post-treatment samples. We only examined here positions associated with resistance which are detailed in the “Resistance associated substitutions” section. Low-frequency variants are noted with the format *XposY*, X being the major variant found at consensus and Y the minor variant, for the given position *pos*.

Supplementary materials

Supplementary materials can be found in Annexes section 6.

4.3 CONCLUSIONS ET PERSPECTIVES

Dans ce chapitre, je me suis intéressée à la détection de DRMs avec une approche différente de celle de ConDor. J'ai cherché à caractériser de possibles DRMs dans un jeu de données constitué d'échantillons de VHC prélevés avant et après administration d'un traitement antiviral. Ce travail présente une grande part d'exploratoire étant donné que le génotype 6 du VHC est encore peu caractérisé, que ce soit en termes de diversité (certains de nos échantillons sont de sous-type inconnu) ou au niveau des DRMs. Les résultats que nous discutons dans l'article seront donc à compléter par des analyses ultérieures.

Outre la recherche de résistances, ce travail présente l'intérêt de venir enrichir le nombre de génomes complets disponibles pour le génotype 6 du VHC avec l'assemblage de 65 nouveaux génomes complets (45 en comptant un génome par paire). Ce travail explore également la question de la diversité virale intra-hôte. A ce propos, nous avons montré que si l'impact du traitement ne se voyait pas forcément au niveau des séquences consensus, il se traduisait par une réduction de la diversité génétique pour la quasi-totalité des échantillons.

Etant donné la petite taille de notre jeu de données, nous ne sommes pas dans un contexte de recherche de convergence. En effet, la plupart des mutations que nous avons identifiées comme possiblement associées à de la résistance dans nos échantillons sont uniques. Une perspective intéressante à ce travail serait ainsi la recherche de DRMs sur de plus grands jeux de données. Cela permettrait entre autres de pouvoir appliquer différentes méthodes de détection dont ConDor. Notons que pour le VHC, il risque d'être plus difficile de caractériser la résistance sur des données échantillonnées en conditions réelles plutôt qu'en conditions expérimentales. En effet, dans le cas du VHC, les échantillons post-traitement sont prélevés après arrêt total du traitement pendant plusieurs semaines. Dans le cas du VIH, le traitement n'est normalement jamais interrompu et les pressions de sélection conduisant à l'émergence des DRMs sont donc constantes. Dans ce sens, la caractérisation des DRMs pour le VHC en condition naturelle se révèle sans doute un exercice plus complexe.

5 CONCLUSION GÉNÉRALE

L'objectif principal de cette thèse était de détecter des mutations convergentes dans les génomes viraux. Je me suis ainsi intéressée aux mutations de résistance qui sont particulièrement bien caractérisées chez le VIH-1 et un exemple très clair de convergence. Je me suis également intéressée à la détection de mutations de résistance chez le VHC. A travers ces deux projets, il a été possible d'étudier l'effet d'un traitement antiviral au niveau des séquences consensus mais aussi au niveau de la diversité virale intra-hôte.

Les deux projets de cette thèse posaient des questions de recherche voisines mais la structure des données et les caractéristiques des deux virus étudiés étaient différentes. Ainsi, là où chez le VIH-1 les DRMs sont similaires entre les différents sous-types, elles peuvent être différentes chez le VHC. Cela peut s'expliquer par la plus grande diversité génétique du VHC qui circule dans les populations humaines depuis plusieurs centaines d'années. D'autre part, en cas d'échec thérapeutique pour le VIH-1, les patients sont toujours sous traitement conduisant à un maintien des DRMs dans la population virale. A contrario, après arrêt du traitement, les DRMs dans le VHC ne sont pas forcément conservées au niveau de la séquence consensus. L'étude de la résistance au niveau des variants minoritaires chez le VHC prend ainsi toute son importance.

La mise au point de la méthode de détection de convergence ConDor m'a amenée à explorer d'autres organismes que les virus en recherchant de la convergence chez des plantes et des poissons. Les génomes de ces organismes sont plus complexes que ceux des virus avec la présence d'introns ou de duplications de gènes. Pour autant, la caractérisation des mutations convergentes associées à un phénotype convergent s'est révélée similaire au VIH-1. Par cet aspect, la convergence moléculaire est répandue chez une très grande diversité d'organismes.

En développant ConDor, et en particulier sa composante émergence, il est apparu que les modèles d'évolution markoviens sous-estimaient la convergence dans les vraies données conduisant à de nombreux faux positifs. J'ai pu observer les mêmes résultats dans les données simulées si le modèle d'évolution n'est plus le vrai modèle. Ces observations confirment l'importance de la composante corrélation de ConDor qui permet de tester si l'apparition d'une mutation est dépendante d'un phénotype. Dans un certain nombre de nos analyses, la composante corrélation seule s'est révélée avoir de bonnes performances. Cela était d'autant plus vrai que le phénotype est un bon prédicteur des mutations de convergence. Dans ce cas particulier, il serait intéressant de voir si le nombre d'émergences observées d'une mutation, couplé à la composante corrélation, ne serait pas un indicateur de convergence évolutive suffisant.

L'importance de caractériser la convergence évolutive chez les virus et leur adaptation à de nouvelles contraintes semble évidente à l'heure de la pandémie causée par le SARS-CoV-2. Nos analyses préliminaires sur les séquences de SARS-CoV-2 ont toutefois révélé que du chemin restait à faire pour l'étude d'organismes dont les séquences sont peu divergentes. De la convergence moléculaire a été caractérisée dans les lignées 501Y du SARS-CoV-2 (Martin et al. 2021) en combinant l'augmentation de la fréquence de certaines mutations au cours du temps dans différentes lignées de 501Y avec des méthodes de sélection positive (MEME et FEL). Cependant, comme nous l'avons vu dans les analyses complémentaires du VIH-1, ces méthodes de sélection positive ne sont pas forcément les plus appropriées pour la détection de convergence.

Notamment, MEME est adapté à la détection de sélection épisodique de diversification (Murrell, Wertheim, et al. 2012) et FEL ne permet pas de détecter de la sélection épisodique (Pond et Frost 2005). Des modèles d'évolution de codons adaptés à la détection de convergence ou de sélection positive directionnelle pourraient ainsi être développés.

A travers le prisme de la convergence moléculaire, je me suis intéressée à la résistance chez les virus et notamment le VIH-1 et le VHC. Des questions similaires pourraient également être soulevées pour les bactéries dont le sujet de la résistance devient de plus en plus préoccupant. La résistance chez les bactéries s'opère par des mécanismes différents de ceux des virus avec notamment les transferts de plasmide. Une application directe des méthodes présentées ici apparait donc plus difficile.

Dans cette thèse je ne me suis pas intéressée aux indels ni à des réarrangements génomiques plus grands que des mutations ponctuelles. Pourtant chez de nombreux virus ces réarrangements peuvent conduire à de la résistance, des échappements immunitaires, etc. De même, l'émergence répétée de mutations en coévolution ou par épistasie pourrait être une piste pour la détection de mutations accessoires ou compensatrices. Ces mutations viennent en effet compenser une réduction de fitness de mutations présentant un avantage par ailleurs (typiquement les mutations de résistance). Les mutations compensatrices peuvent également émerger en l'absence d'une mutation à compenser et sont donc plus faiblement corrélées à un phénotype convergent. De la même manière, certaines résistances sont dues à la combinaison de plusieurs mutations. Une poursuite des travaux de cette thèse pourrait consister à coupler la recherche de convergence évolutive et la coévolution, deux questions biologiques qui semblent intimement liées.

6 ANNEXES

MATÉRIEL SUPPLÉMENTAIRE DE L'ARTICLE "GENOMIC VARIATIONS ASSOCIATED WITH DRUG RESISTANCE IN HCV GENOTYPE 6 INFECTED PATIENTS FAILING DAA-BASED THERAPY".

Table S2: *characteristics of viral genomes from patients with TF*
nt: nucleotide, ORF: open reading frame. In bold are sequences with depth sequencing bellow 10,000x.

Subtype	ID	Length (nt)	ORF length (nt)	Coverage sample A (x)	Coverage sample B (x)	VL (LogIU/mL) sample A	VL (LogIU/mL) sample B	Length NS5A (nt)	Length NS5B (nt)
6e	HCV01	9438	9069	60557.92	**3426.42**	4.3	4.5	1365	1773
	HCV06	9440	9069	148196.1	94787.14	5.8	6.4	1365	1773
	HCV10	9469	9069	61884.24	21668.88	7.1	6.7	1365	1773
	HCV30	9468	9069	28251.78	30206.28	6.4	5.6	1365	1773
6p	HCV22	9464	9060	57285.72	61665.06	5.8	6.2	1365	1773
	HCV33	9454	9051	56982.04	28696.88	5.0	6.3	1356	1773
6q	HCV23	9461	9051	59913.37	38167.82	5.6	5.8	1356	1773
6r	HCV11	9429	9060	14026.43	**351.21**	4.8	4.0	1365	1773
	HCV14	9446	9051	**9004.83**	33520.93	5.9	6.0	1356	1773
	HCV15	9453	9051	33797.11	30810.78	7.0	6.1	1356	1773
	HCV16	9453	9051	56792.07	73904.52	6.6	6.4	1356	1773
	HCV17	9452	9051	12760.05	**6436.02**	5.7	5.5	1356	1773
	HCV24	9453	9051	x	29366.78	6.3	5.9	1356	1773
	HCV29	9391	9051	24946.47	26027.09	5.8	4.4	1356	1773
	HCV32	9402	9057	14405.47	x	6.8	5.4	1365	1773
6s	HCV27	9391	9051	45867.51	x	6.7	7.1	1353	1773
6xc	HCV03	9444	9069	x	15111.31	5.8	5.8	1356	1773
6xf	HCV05	9447	9069	25353.24	**417.38**	4.2	6.2	1365	1773
6	HCV18	9474	9048	73113.37	72693.22	5.3	6.4	1356	1773
1a	HCV04	9469	9036	37913.78	26414.18	6.4	5.9	1344	1773
Average		9444.9	9056.7	45613.97	32981.77	5.86	5.83	1358.85	1773

Table S2: *characteristics of viral genomes from patients successfully treated*
nt: nucleotide, ORF: open reading frame

Subtype	ID	Length (nt)	ORF length (nt)	Coverage	VL (LogIU/mL)	Length NS5A (nt)	Length NS5B (nt)
6e	HCV108	9407	9066	67732.74	7.2	1365	1773
	HCV119	9464	9063	39360.47	7.28	1365	1773
	HCV121	9467	9066	63389.04	7.3	1365	1773
	HCV122	9467	9066	31348.28	7.18	1365	1773
	HCV124	9464	9063	19191.69	7.32	1365	1773
	HCV125	9470	9069	38697.91	7.2	1365	1773
6p	HCV101	9535	9048	27863	7.28	1356	1773
	HCV102	9535	9048	26295.3	7.17	1356	1773
	HCV103	9434	9048	17377.54	6.69	1356	1773
	HCV117	9534	9048	25396.78	6.74	1356	1773
	HCV120	9404	9048	19612.14	6.73	1356	1773
	HCV126	9422	9048	11285.75	6.4	1356	1773
6q	HCV107	9403	9048	177607.5	6.77	1356	1773
	HCV111	9400	9045	34374.86	6.78	1353	1773
	HCV116	9339	9048	35742.68	6.81	1356	1773
	HCV118	9454	9048	23591.51	6.91	1356	1773
	HCV127	9383	9048	14889.49	6.33	1356	1773
6r	HCV104	9449	9048	24281.44	6.68	1356	1773
	HCV109	9446	9048	25820.47	7	1356	1773
	HCV110	9449	9048	13838.90	6.89	1356	1773
	HCV112	9450	9048	40480.01	7.14	1356	1773
	HCV123	9448	9048	33276.62	6.67	1356	1773
6s	HCV113	9443	9048	11297.24	6.89	1353	1773
	HCV114	9421	9048	69531.17	6.96	1353	1773
6xc	HCV129	9500	9066	95701.93	6.94	1356	1773
	HCV130	9466	9063	106677.1	7.09	1356	1773
6xf	HCV128	9418	9054	24815.87	6.77	1356	1773
6	HCV106	9472	9051	**8474.55**	**4.93**	1359	1773
6	HCV115	9402	9048	**6469.91**	**5.75**	1356	1773
Average		9446.41	9052.96	39117.99	6.89	1357.65	1773

Table S3: *Mutations found by Geno2Pheno on samples HCV24(B) and HCV03(B). XposY[1]: mutation found on a resistance-associated position.*

Subtype	Sample ID	Mutations on NS5A	Mutations on NS5B
6r	HCV24(B)	R41Q, Q125H, A197T, L199V	A73V, R81K, T149V, K307N, A335S, L336V, P349H, S553A
6xc	HCV03(B)	Y54H[1], V75A, N124S	N117K, S255A, S487A, V564M

We were unable to reconstruct the pretreatment sequences for these 2 samples. We identified several mutations with Geno2Pheno (relative to the closest reference sequence to Geno2Pheno) that were not found in control samples of the same subtype. The Y54H mutation in NS5A is noted as a RAP but we cannot say whether it was present at baseline or a response to treatment.

Table S4: *Polymorphisms found at baseline on resistance-associated positions in NS5A for successfully treated patients or patients with TF.[1]: polymorphisms found only in patients experiencing TF.*

	24	**28**	**30**	**31**	**32**	**38**	**54**	**58**	**62**	**92**	**93**
6e	K/R	V/M	S	L	P	S	C/Y/H[1]	P	Q/R/S/T/L[1]	A	T
6p	K	V/M	S	L	P	S	Y/H	P	M/V/E[1]	A	T/S
6q	K	V/M	S	L	P	S	H	P	N/Q/H	A	T
6r	K	T/A/K[1]	A	L	P	S	T/S	P	V/A	A	T
6s	K	ML	S	L	P	S	H	P	E	A	T
6xc	K	V/M	A	L/M	P	S	Y/H[1]	P	Q	A	T
6xf	K	V	A	L	P	S	N	P	A	A	T
1a	K	M	R	L	P	S	H	H	E	A	Y

Of the eleven positions noted as associated with resistance, seven have at least one amino acid shared by all the studied subtypes of GT6 (positions 24,31,32,38,58,92,93). Positions 54 and 62 show a high degree of variability both within the same subtype and between samples of different subtypes. Polymorphisms marked ([1]) are found only in samples from patients that did not achieve SVR12. They have not been demonstrated to confer resistance to Daclatasvir in genotype 6 or other HCV genotypes. Only RAP Q62L was shown to reduce susceptibility to Daclatasvir in genotype 3 (Sorbo et al. 2018).

Table S5: *Polymorphisms found at baseline on resistance-associated positions in NS5B for successfully treated patients or patients with TF.*

	96	**159**	**282**	**289**	**316**	**320**	**321**
6e	S	L	S	L	C	L	V
6p	S	L	S	L	C	L	V
6q	S	L	S	L	C	L	V
6r	S	L	S	L	C	L	V
6s	S	L	S	L	C	L	V
6xc	S	L	S	L	C	L	V
6xf	S	L	S	L	C	L	V
1a	S	L	S	C	C	L	V

All sequences of genotype 6 (associated with TF or control) share the same amino acid for a given resistance-associated position at baseline. Resistance-associated positions in NS5B are highly conserved at baseline and show no polymorphism even between different subtypes.

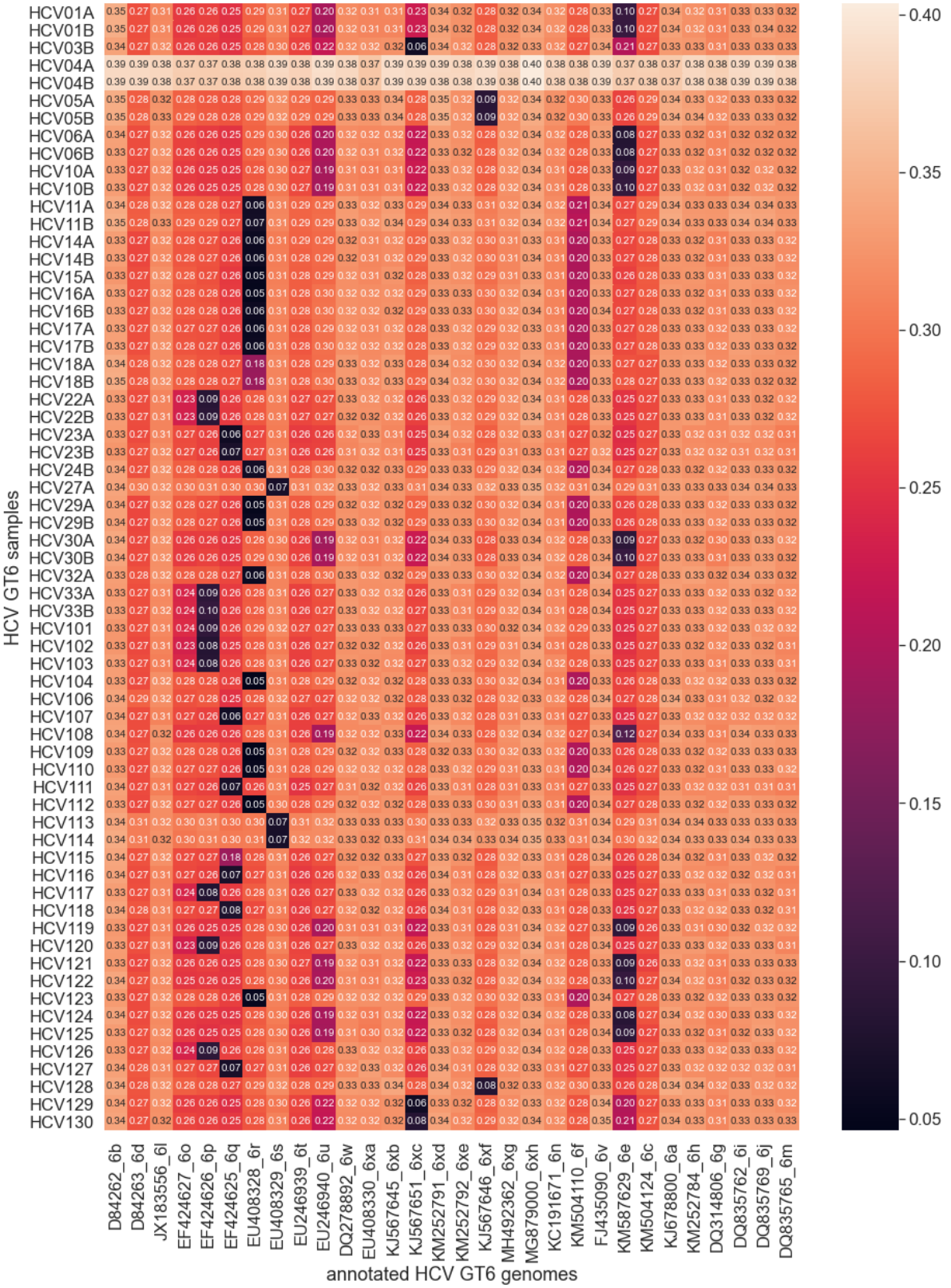


***Figure S1**: **Heatmap of the pairwise distance between newly sequenced genomes from patients infected with HCV GT6 and annotated subtypes from GT6**. The color bar represents the pairwise distance. A distance of 0.05 equals a similarity of 0.95.*

All samples present at least one pairwise distance below 0.15 with reference samples from GT6 but samples HCV04, HCV18, HCV106, and HCV115. HCV18, HCV106, and HCV115 are of undefined subtype of GT6 and HCV04 is from subtype 1a.

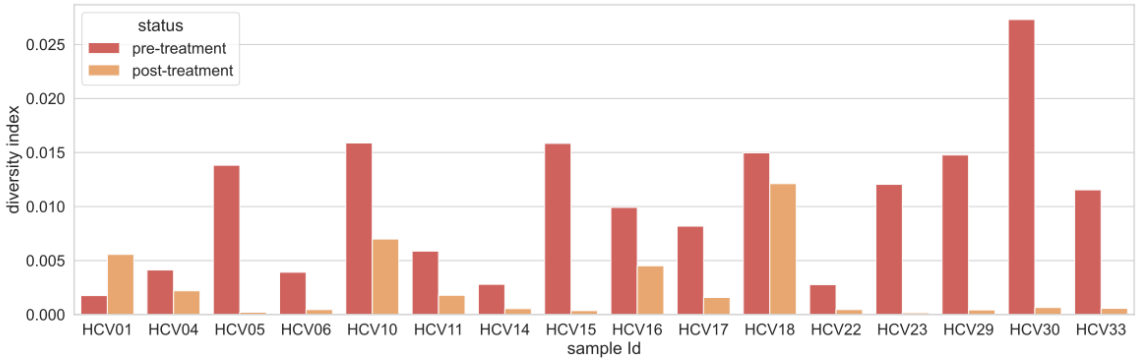


Figure S2: diversity index by samples at pre-treatment and post-treatment.

All samples but HCV01 exhibited a decrease in nucleotide diversity after treatment. The decrease is significant for all samples.

BIBLIOGRAPHIE

« A Simple Method for Estimating and Testing Minimum-Evolution Trees ». 1992. *Molecular Biology and Evolution*, septembre. https://doi.org/10.1093/oxfordjournals.molbev.a040771.

Abram, Michael E., Andrea L. Ferris, Wei Shao, W. Gregory Alvord, et Stephen H. Hughes. 2010. « Nature, Position, and Frequency of Mutations Made in a Single Cycle of HIV-1 Replication ». *Journal of Virology* 84 (19): 9864-78. https://doi.org/10.1128/JVI.00915-10.

Agrawal, Anurag A. 2017. « Toward a Predictive Framework for Convergent Evolution: Integrating Natural History, Genetic Mechanisms, and Consequences for the Diversity of Life ». *The American Naturalist* 190 (S1): S1-12. https://doi.org/10.1086/692111.

Alessandri-Gradt, Elodie, Fabienne De Oliveira, Marie Leoz, Véronique Lemee, David L. Robertson, Felix Feyertag, Paul-Alain Ngoupo, Philippe Mauclere, François Simon, et Jean-Christophe Plantier. 2018. « HIV-1 Group P Infection: Towards a Dead-End Infection? » *AIDS* 32 (10): 1317-22. https://doi.org/10.1097/QAD.0000000000001791.

Alizon, Samuel, et Christophe Fraser. 2013. « Within-host and between-host evolutionary rates across the HIV-1 genome ». *Retrovirology* 10 (1): 49. https://doi.org/10.1186/1742-4690-10-49.

Almeida, June D. 1963. « A Classification of Virus Particles Based on Morphology ». *Canadian Medical Association Journal* 89 (16): 787-98.

Almeida, June D., et A. P. Waterson. 1970. « Some implications of a morphologically oriented classification of viruses ». *Archiv Fur Die Gesamte Virusforschung* 32 (1): 66-72. https://doi.org/10.1007/BF01241521.

Anthony, Simon J., Jonathan H. Epstein, Kris A. Murray, Isamara Navarrete-Macias, Carlos M. Zambrana-Torrelio, Alexander Solovyov, Rafael Ojeda-Flores, et al. 2013. « A Strategy To Estimate Unknown Viral Diversity in Mammals ». *mBio* 4 (5): e00598-13. https://doi.org/10.1128/mBio.00598-13.

Arbuckle, Kevin, Cheryl M. Bennett, et Michael P. Speed. 2014. « A Simple Measure of the Strength of Convergent Evolution ». *Methods in Ecology and Evolution* 5 (7): 685-93. https://doi.org/10.1111/2041-210X.12195.

Arbuckle, Kevin, et Michael P. Speed. 2016. « Analysing Convergent Evolution: A Practical Guide to Methods ». In *Evolutionary Biology*, 23-36. Springer, Cham. https://doi.org/10.1007/978-3-319-41324-2_2.

Arcia, David, Liliana Acevedo-Sáenz, María Teresa Rugeles, et Paula A. Velilla. 2017. « Role of CD8+ T Cells in the Selection of HIV-1 Immune Escape Mutations ». *Viral Immunology* 30 (1): 3-12. https://doi.org/10.1089/vim.2016.0095.

Arendt, Jeff, et David Reznick. 2008. « Convergence and parallelism reconsidered: what have we learned about the genetics of adaptation? » *Trends in Ecology & Evolution* 23 (1): 26-32. https://doi.org/10.1016/j.tree.2007.09.011.

Arslan, Defne, Matthieu Legendre, Virginie Seltzer, Chantal Abergel, et Jean-Michel Claverie. 2011. « Distant Mimivirus relative with a larger genome highlights the fundamental features of Megaviridae ». *Proceedings of the National Academy of Sciences of the United States of America* 108 (42): 17486-91. https://doi.org/10.1073/pnas.1110889108.

Arul, Albert B., et Renã A. S. Robinson. 2019. « Sample Multiplexing Strategies in Quantitative Proteomics ». *Analytical chemistry* 91 (1): 178-89. https://doi.org/10.1021/acs.analchem.8b05626.

Ateto, Abdulrahman. 2014. *Bioinformatics for Beginners, Genes, Genome, Molecular Evolution, Databases and Analytical Tools.*

Baechlein, Christine, Nicole Fischer, Adam Grundhoff, Malik Alawi, Daniela Indenbirken, Alexander Postel, Anna Lena Baron, et al. 2015. « Identification of a Novel Hepacivirus in Domestic Cattle from Germany ». *Journal of Virology* 89 (14): 7007-15. https://doi.org/10.1128/JVI.00534-15.

Bailey, Susan F., François Blanquart, Thomas Bataillon, et Rees Kassen. 2017. « What Drives Parallel Evolution? » *BioEssays* 39 (1): 1-9. https://doi.org/10.1002/bies.201600176.

Bairoch, Amos, et Rolf Apweiler. 2000. « The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000 ». *Nucleic Acids Research* 28 (1): 45-48.

Barker, Daniel, et Mark Pagel. 2005. « Predicting Functional Gene Links from Phylogenetic-Statistical Analyses of Whole Genomes ». *PLOS Computational Biology* 1 (1): e3. https://doi.org/10.1371/journal.pcbi.0010003.

Barre-Sinoussi, F., J. C. Chermann, F. Rey, M. T. Nugeyre, S. Chamaret, J. Gruest, C. Dauguet, et al. 1983. « Isolation of a T-Lymphotropic Retrovirus from a Patient at Risk for Acquired Immune Deficiency Syndrome (AIDS) ». *Science* 220 (4599): 868-71. https://doi.org/10.1126/science.6189183.

Bawden, F. C., N. W. Pirie, J. D. Bernal, et I. Fankuchen. 1936. « Liquid Crystalline Substances from Virus-Infected Plants ». *Nature* 138 (3503): 1051-52. https://doi.org/10.1038/1381051a0.

Beachboard, Dia C, et Stacy M Horner. 2016. « Innate Immune Evasion Strategies of DNA and RNA Viruses ». *Current Opinion in Microbiology*, Host-microbe interactions: parasites/fungi/viruses, 32 (août): 113-19. https://doi.org/10.1016/j.mib.2016.05.015.

Bedhomme, Stéphanie, Guillaume Lafforgue, et Santiago F. Elena. 2012. « Multihost Experimental Evolution of a Plant RNA Virus Reveals Local Adaptation and Host-Specific Mutations ». *Molecular Biology and Evolution* 29 (5): 1481-92. https://doi.org/10.1093/molbev/msr314.

Beerenwinkel, Niko, Martin Däumer, Mark Oette, Klaus Korn, Daniel Hoffmann, Rolf Kaiser, Thomas Lengauer, Joachim Selbig, et Hauke Walter. 2003. « Geno2pheno: estimating phenotypic drug resistance from HIV-1 genotypes ». *Nucleic Acids Research* 31 (13): 3850-55.

Beijerinck, M. W. 1898. « Ueber ein Contagium vivum fluidum als Ursache der Fleckenkrankheit der Tabaksblätter ». J. Müller. https://scholar.google.com/scholar_lookup?title=Ueber+ein+Contagium+vivum+fluidum+als+Ursache+der+Fleckenkrankheit+der+Tabaksbla%CC%88tter&author=Beijerinck%2C+M.+W.+%28Martinus+Willem%29&publication_year=1898.

Benson, Dennis A., Mark Cavanaugh, Karen Clark, Ilene Karsch-Mizrachi, David J. Lipman, James Ostell, et Eric W. Sayers. 2013. « GenBank ». *Nucleic Acids Research* 41 (Database issue): D36-42. https://doi.org/10.1093/nar/gks1195.

Bertels, Frederic, Christine Leemann, Karin J. Metzner, et Roland Regoes. 2019. « Parallel Evolution of HIV-1 in a Long-Term Experiment ». *Molecular Biology and Evolution*, juin. https://doi.org/10.1093/molbev/msz155.

Bertels, Frederic, Karin J. Metzner, et Roland Regoes. 2021. « Convergent evolution as an indicator for selection during acute HIV-1 infection ». *Peer Community Journal* 1. https://doi.org/10.24072/pcjournal.6.

Besnard, Guillaume, A. Muthama Muasya, Flavien Russier, Eric H. Roalson, Nicolas Salamin, et Pascal-Antoine Christin. 2009. « Phylogenomics of C4 Photosynthesis in Sedges (Cyperaceae): Multiple Appearances and Genetic Convergence ». *Molecular Biology and Evolution* 26 (8): 1909-19. https://doi.org/10.1093/molbev/msp103.

Bhattacharya, Tanmoy, Marcus Daniels, David Heckerman, Brian Foley, Nicole Frahm, Carl Kadie, Jonathan Carlson, et al. 2007. « Founder Effects in the Assessment of HIV

Polymorphisms and HLA Allele Associations ». *Science* 315 (5818): 1583-86. https://doi.org/10.1126/science.1131528.

Blach, Sarah, Stefan Zeuzem, Michael Manns, Ibrahim Altraif, Ann-Sofi Duberg, David H Muljono, Imam Waked, et al. 2017. « Global Prevalence and Genotype Distribution of Hepatitis C Virus Infection in 2015: A Modelling Study ». *The Lancet Gastroenterology & Hepatology* 2 (3): 161-76. https://doi.org/10.1016/S2468-1253(16)30181-9.

Bläsing, Oliver E., Peter Westhoff, et Per Svensson. 2000. « Evolution of C4 Phosphoenolpyruvate Carboxylase InFlaveria, a Conserved Serine Residue in the Carboxyl-Terminal Part of the Enzyme Is a Major Determinant for C4-Specific Characteristics* ». *Journal of Biological Chemistry* 275 (36): 27917-23. https://doi.org/10.1074/jbc.M909832199.

Blassel, Luc, Anna Tostevin, Christian Julian Villabona-Arenas, Martine Peeters, Stéphane Hué, Olivier Gascuel, et On behalf of the UK HIV Drug Resistance Database. 2021. « Using Machine Learning and Big Data to Explore the Drug Resistance Landscape in HIV ». *PLOS Computational Biology* 17 (8): e1008873. https://doi.org/10.1371/journal.pcbi.1008873.

Blassel, Luc, Anna Zhukova, Christian J Villabona-Arenas, Katherine E Atkins, Stéphane Hué, et Olivier Gascuel. 2021. « Drug Resistance Mutations in HIV: New Bioinformatics Approaches and Challenges ». *Current Opinion in Virology* 51 (décembre): 56-64. https://doi.org/10.1016/j.coviro.2021.09.009.

Blumberg, Baruch S., et Harvey J. Alter. 1965. « A “New” Antigen in Leukemia Sera ». *JAMA* 191 (7): 541-46. https://doi.org/10.1001/jama.1965.03080070025007.

Bolger, Anthony M., Marc Lohse, et Bjoern Usadel. 2014. « Trimmomatic: a flexible trimmer for Illumina sequence data ». *Bioinformatics* 30 (15): 2114-20. https://doi.org/10.1093/bioinformatics/btu170.

Borman, Andrew M., Sylvie Paulous, et François Clavel. 1996. « Resistance of human immunodeficiency virus type 1 to protease inhibitors: selection of resistance mutations in the presence and absence of the drug ». *Journal of General Virology* 77 (3): 419-26. https://doi.org/10.1099/0022-1317-77-3-419.

Borrow, Persephone, Hanna Lewicki, Xiping Wei, Marc S. Horwitz, Nancy Peffer, Heather Meyers, Jay A. Nelson, et al. 1997. « Antiviral Pressure Exerted by HIV-l-Specific Cytotoxic T Lymphocytes (CTLs) during Primary Infection Demonstrated by Rapid Selection of CTL Escape Virus ». *Nature Medicine* 3 (2): 205-11. https://doi.org/10.1038/nm0297-205.

Boucher, Charles A. B., Eithne O’Sullivan, Jan W. Mulder, Chitra Ramautarsing, Paul Kellam, Graham Darby, Joep M. A. Lange, Jaap Goudsmit, et Brendan A. Larder. 1992. « Ordered Appearance of Zidovudine Resistance Mutations during Treatment of 18 Human Immunodeficiency Virus-Positive Subjects ». *The Journal of Infectious Diseases* 165 (1): 105-10. https://doi.org/10.1093/infdis/165.1.105.

Bowen, David G., et Christopher M. Walker. 2005. « Mutational escape from CD8+ T cell immunity : HCV evolution, from chimpanzees to man ». *Journal of Experimental Medicine* 201 (11): 1709-14. https://doi.org/10.1084/jem.20050808.

Boyer, Mickaël, Natalya Yutin, Isabelle Pagnier, Lina Barrassi, Ghislain Fournous, Leon Espinosa, Catherine Robert, et al. 2009. « Giant Marseillevirus Highlights the Role of Amoebae as a Melting Pot in Emergence of Chimeric Microorganisms ». *Proceedings of the National Academy of Sciences of the United States of America* 106 (51): 21848-53. https://doi.org/10.1073/pnas.0911354106.

Boyer, Paul L., Catherine A. Rehm, Michael C. Sneller, JoAnn Mican, Margaret R. Caplan, Robin Dewar, Andrea L. Ferris, et al. 2022. « A Combination of Amino Acid Mutations Leads to Resistance to Multiple Nucleoside Analogs in Reverse Transcriptases from HIV-1 Subtypes B and C ». *Antimicrobial Agents and Chemotherapy* 66 (1): e01500-21. https://doi.org/10.1128/AAC.01500-21.

Bridgham, Jamie T. 2016. « Predicting the Basis of Convergent Evolution ». *Science* 354 (6310): 289-289. https://doi.org/10.1126/science.aai7394.

Brik, Ashraf, et Chi-Huey Wong. 2003. « HIV-1 Protease: Mechanism and Drug Discovery ». *Organic & Biomolecular Chemistry* 1 (1): 5-14. https://doi.org/10.1039/B208248A.

Bruhl, Jeremy, et Karen Wilson. 2007. « Towards a Comprehensive Survey of C3 and C4 Photosynthetic Pathways in Cyperaceae ». *Aliso* 23 (1): 99-148. https://doi.org/10.5642/aliso.20072301.11.

Buchfink, Benjamin, Chao Xie, et Daniel H. Huson. 2015. « Fast and Sensitive Protein Alignment Using DIAMOND ». *Nature Methods* 12 (1): 59-60. https://doi.org/10.1038/nmeth.3176.

Bull, J. J., M. R. Badgett, H. A. Wichman, J. P. Huelsenbeck, D. M. Hillis, A. Gulati, C. Ho, et I. J. Molineux. 1997. « Exceptional Convergent Evolution in a Virus ». *Genetics* 147 (4): 1497-1507.

Bull, J. J., R. Sanjuán, et C. O. Wilke. 2007. « Theory of Lethal Mutagenesis for Viruses ». *Journal of Virology* 81 (6): 2930-39. https://doi.org/10.1128/JVI.01624-06.

Bull, Rowena A., Fabio Luciani, Kerensa McElroy, Silvana Gaudieri, Son T. Pham, Abha Chopra, Barbara Cameron, et al. 2011. « Sequential Bottlenecks Drive Viral Evolution in Early Acute Hepatitis C Virus Infection ». *PLOS Pathogens* 7 (9): e1002243. https://doi.org/10.1371/journal.ppat.1002243.

Bunchorntavakul, Chalermrat, Disaya Chavalitdhamrong, et Tawesak Tanwandee. 2013. « Hepatitis C genotype 6: A concise review and response-guided therapy proposal ». *World Journal of Hepatology* 5 (9): 496-504. https://doi.org/10.4254/wjh.v5.i9.496.

Callaway, Ewen. 2022. « Why Does the Omicron Sub-Variant Spread Faster than the Original? » *Nature* 602 (7898): 556-57. https://doi.org/10.1038/d41586-022-00471-2.

Cassan, Elodie, Anne-Muriel Arigon-Chifolleau, Jean-Michel Mesnard, Antoine Gross, et Olivier Gascuel. 2016. « Concomitant Emergence of the Antisense Protein Gene of HIV-1 and of the Pandemic ». *Proceedings of the National Academy of Sciences* 113 (41): 11537-42. https://doi.org/10.1073/pnas.1605739113.

Castoe, Todd A., A. P. Jason de Koning, Hyun-Min Kim, Wanjun Gu, Brice P. Noonan, Gavin Naylor, Zhi J. Jiang, Christopher L. Parkinson, et David D. Pollock. 2009. « Evidence for an Ancient Adaptive Episode of Convergent Molecular Evolution ». *Proceedings of the National Academy of Sciences* 106 (22): 8986-91. https://doi.org/10.1073/pnas.0900233106.

Centers for Disease Control (CDC). 1981. « Pneumocystis Pneumonia--Los Angeles ». *MMWR. Morbidity and Mortality Weekly Report* 30 (21): 250-52.

Chabrol, Olivier, Manuela Royer-Carenzi, Pierre Pontarotti, et Gilles Didier. 2018. « Detecting the Molecular Basis of Phenotypic Convergence ». *Methods in Ecology and Evolution* 9 (11): 2170-80. https://doi.org/10.1111/2041-210X.13071.

Chai, Simin, Ran Tian, Xinghua Rong, Guiting Li, Bingyao Chen, Wenhua Ren, Shixia Xu, et Guang Yang. 2020. « Evidence of Echolocation in the Common Shrew from Molecular Convergence with Other Echolocating Mammals ». *Zoological Studies* 59: e4. https://doi.org/10.6620/ZS.2020.59-4.

Chang, Jia-Ming, Paolo Di Tommaso, et Cedric Notredame. 2014. « TCS: A New Multiple Sequence Alignment Reliability Measure to Estimate Alignment Accuracy and Improve Phylogenetic Tree Reconstruction ». *Molecular Biology and Evolution* 31 (6): 1625-37. https://doi.org/10.1093/molbev/msu117.

Chang, Pengxiang, Joshua E. Sealy, Jean-Remy Sadeyen, Sushant Bhat, Deimante Lukosaityte, Yipeng Sun, et Munir Iqbal. 2020. « Immune Escape Adaptive Mutations in the H7N9 Avian Influenza Hemagglutinin Protein Increase Virus Replication Fitness and Decrease Pandemic Potential ». *Journal of Virology* 94 (19): e00216-20. https://doi.org/10.1128/JVI.00216-20.

Cheng, Guofeng, Yang Tian, Brian Doehle, Betty Peng, Amoreena Corsa, Yu-Jen Lee, Ruoyu Gong, et al. 2016. « In Vitro Antiviral Activity and Resistance Profile Characterization of the

Hepatitis C Virus NS5A Inhibitor Ledipasvir ». *Antimicrobial Agents and Chemotherapy* 60 (3): 1847-53. https://doi.org/10.1128/AAC.02524-15.

Chevaliez, Stéphane, et Jean-Michel Pawlotsky. 2006. « HCV Genome and Life Cycle ». In *Hepatitis C Viruses: Genomes and Molecular Biology*, édité par Seng-Lai Tan. Norfolk (UK): Horizon Bioscience. http://www.ncbi.nlm.nih.gov/books/NBK1630/.

Chevin, Luis-Miguel, Guillaume Martin, et Thomas Lenormand. 2010. « Fisher's Model and the Genomics of Adaptation: Restricted Pleiotropy, Heterogenous Mutation, and Parallel Evolution ». *Evolution* 64 (11): 3213-31. https://doi.org/10.1111/j.1558-5646.2010.01058.x.

Choi, Bina, Manish C. Choudhary, James Regan, Jeffrey A. Sparks, Robert F. Padera, Xueting Qiu, Isaac H. Solomon, et al. 2020. « Persistence and Evolution of SARS-CoV-2 in an Immunocompromised Host ». *New England Journal of Medicine* 383 (23): 2291-93. https://doi.org/10.1056/NEJMc2031364.

Choo, Qui-Lim, George Kuo, Amy J. Weiner, Lacy R. Overby, Daniel W. Bradley, et Michael Houghton. 1989. « Isolation of a cDNA cLone Derived from a Blood-Borne Non-A, Non-B Viral Hepatitis Genome ». *Science* 244 (4902): 359-62. https://doi.org/10.1126/science.2523562.

Christin, Pascal-Antoine, Nicolas Salamin, A. Muthama Muasya, Eric H. Roalson, Flavien Russier, et Guillaume Besnard. 2008. « Evolutionary Switch and Genetic Convergence on RbcL Following the Evolution of C4 Photosynthesis ». *Molecular Biology and Evolution* 25 (11): 2361-68. https://doi.org/10.1093/molbev/msn178.

Christin, Pascal-Antoine, Nicolas Salamin, Vincent Savolainen, Melvin R. Duvall, et Guillaume Besnard. 2007. « C4 Photosynthesis Evolved in Grasses via Parallel Adaptive Genetic Changes ». *Current Biology* 17 (14): 1241-47. https://doi.org/10.1016/j.cub.2007.06.036.

Christin, Pascal-Antoine, Daniel M. Weinreich, et Guillaume Besnard. 2010. « Causes and evolutionary significance of genetic convergence ». *Trends in Genetics* 26 (9): 400-405. https://doi.org/10.1016/j.tig.2010.06.005.

Clarke, James, Hai-Chen Wu, Lakmal Jayasinghe, Alpesh Patel, Stuart Reid, et Hagan Bayley. 2009. « Continuous Base Identification for Single-Molecule Nanopore DNA Sequencing ». *Nature Nanotechnology* 4 (4): 265-70. https://doi.org/10.1038/nnano.2009.12.

Clavel, F., D. Guétard, F. Brun-Vézinet, S. Chamaret, M. A. Rey, M. O. Santos-Ferreira, A. G. Laurent, C. Dauguet, C. Katlama, et C. Rouzioux. 1986. « Isolation of a New Human Retrovirus from West African Patients with AIDS ». *Science (New York, N.Y.)* 233 (4761): 343-46. https://doi.org/10.1126/science.2425430.

Clavel, François, et Allan J. Hance. 2004. « HIV Drug Resistance ». *New England Journal of Medicine* 350 (10): 1023-35. https://doi.org/10.1056/NEJMra025195.

Clutter, Dana S., Michael R. Jordan, Silvia Bertagnolio, et Robert W. Shafer. 2016. « HIV-1 Drug Resistance and Resistance Testing ». *Infection, Genetics and Evolution: Journal of Molecular Epidemiology and Evolutionary Genetics in Infectious Diseases* 46 (décembre): 292-307. https://doi.org/10.1016/j.meegid.2016.08.031.

Cock, Peter J. A., Tiago Antao, Jeffrey T. Chang, Brad A. Chapman, Cymon J. Cox, Andrew Dalke, Iddo Friedberg, et al. 2009. « Biopython: freely available Python tools for computational molecular biology and bioinformatics ». *Bioinformatics* 25 (11): 1422-23. https://doi.org/10.1093/bioinformatics/btp163.

Cohen, Myron S., Ying Q. Chen, Marybeth McCauley, Theresa Gamble, Mina C. Hosseinipour, Nagalingeswaran Kumarasamy, James G. Hakim, et al. 2011. « Prevention of HIV-1 Infection with Early Antiretroviral Therapy ». *New England Journal of Medicine* 365 (6): 493-505. https://doi.org/10.1056/NEJMoa1105243.

Collaboration, The HIV-CAUSAL. 2010. « The Effect of Combined Antiretroviral Therapy on the Overall Mortality of HIV-Infected Individuals ». *AIDS* 24 (1): 123-37. https://doi.org/10.1097/QAD.0b013e3283324283.

Corey, Lawrence, Chris Beyrer, Myron S. Cohen, Nelson L. Michael, Trevor Bedford, et Morgane Rolland. 2021. « SARS-CoV-2 Variants in Patients with Immunosuppression ». *New England Journal of Medicine* 385 (6): 562-66. https://doi.org/10.1056/NEJMsb2104756.

Crandall, K. A., C. R. Kelsey, H. Imamichi, H. C. Lane, et N. P. Salzman. 1999. « Parallel Evolution of Drug Resistance in HIV: Failure of Nonsynonymous/Synonymous Substitution Rate Ratio to Detect Selection ». *Molecular Biology and Evolution* 16 (3): 372-82. https://doi.org/10.1093/oxfordjournals.molbev.a026118.

Crill, W D, H A Wichman, et J J Bull. 2000. « Evolutionary Reversals During Viral Adaptation to Alternating Hosts ». *Genetics* 154 (1): 27-37. https://doi.org/10.1093/genetics/154.1.27.

Cuevas, José M., Santiago F. Elena, et Andrés Moya. 2002. « Molecular Basis of Adaptive Convergence in Experimental Populations of RNA Viruses ». *Genetics* 162 (2): 533-42.

Cuevas, José M., Ron Geller, Raquel Garijo, José López-Aldeguer, et Rafael Sanjuán. 2015. « Extremely High Mutation Rate of HIV-1 In Vivo ». *PLoS Biology* 13 (9): e1002251. https://doi.org/10.1371/journal.pbio.1002251.

Cummins, Carla, Alisha Ahamed, Raheela Aslam, Josephine Burgin, Rajkumar Devraj, Ossama Edbali, Dipayan Gupta, et al. 2022. « The European Nucleotide Archive in 2021 ». *Nucleic Acids Research* 50 (D1): D106-10. https://doi.org/10.1093/nar/gkab1051.

Davies, K. T. J., J. A. Cotton, J. D. Kirwan, E. C. Teeling, et S. J. Rossiter. 2012. « Parallel Signatures of Sequence Evolution among Hearing Genes in Echolocating Mammals: An Emerging Model of Genetic Convergence ». *Heredity* 108 (5): 480-89. https://doi.org/10.1038/hdy.2011.119.

Dayhoff, Margaret, Schwartz, et Orcutt. 1978. « A Model of Evolutionary Change in Proteins ». In *Atlas of Protein Sequence and Structure*, Natl. Biomed. Res. Found., Washington DC, 345-52.

Deaton, Aimée M., et Adrian Bird. 2011. « CpG islands and the regulation of transcription ». *Genes & Development* 25 (10): 1010-22. https://doi.org/10.1101/gad.2037511.

Delsuc, Frédéric, et Emmanuel Douzery. 2004a. « Les méthodes probabilistes en phylogénie moléculaire: (1) Les modèles d'évolution des séquences et le maximum de vraisemblance », 17.

Delsuc, Frédéric, et Emmanuel J P Douzery. 2004b. « Les méthodes probabilistes en phylogénie moléculaire: (2) L'approche bayésienne », 13.

Denison, Mark R, Rachel L Graham, Eric F Donaldson, Lance D Eckerle, et Ralph S Baric. 2011. « Coronaviruses ». *RNA Biology* 8 (2): 270-79. https://doi.org/10.4161/rna.8.2.15013.

Desai, Michael M., et Daniel S. Fisher. 2007. « Beneficial Mutation–Selection Balance and the Effect of Linkage on Positive Selection ». *Genetics* 176 (3): 1759-98. https://doi.org/10.1534/genetics.106.067678.

Desper, Richard, et Olivier Gascuel. 2002. « Fast and Accurate Phylogeny Reconstruction Algorithms Based on the Minimum-Evolution Principle ». *Journal of Computational Biology: A Journal of Computational Molecular Cell Biology* 9 (5): 687-705. https://doi.org/10.1089/106652702761034136.

———. 2004. « Theoretical Foundation of the Balanced Minimum Evolution Method of Phylogenetic Inference and Its Relationship to Weighted Least-Squares Tree Fitting ». *Molecular Biology and Evolution* 21 (3): 587-98. https://doi.org/10.1093/molbev/msh049.

Dietz, Julia, Simone Susser, Johannes Vermehren, Kai-Henrik Peiffer, Georgios Grammatikos, Annemarie Berger, Peter Ferenci, et al. 2018. « Patterns of Resistance-Associated Substitutions in Patients With Chronic HCV Infection Following Treatment With Direct-Acting Antivirals ». *Gastroenterology* 154 (4): 976-988.e4. https://doi.org/10.1053/j.gastro.2017.11.007.

Ditmarsch, Dave van, Kerry E. Boyle, Hassan Sakhtah, Jennifer E. Oyler, Carey D. Nadell, Éric Déziel, Lars E. P. Dietrich, et Joao B. Xavier. 2013. « Convergent evolution of

hyperswarming leads to impaired biofilm formation in pathogenic bacteria ». *Cell reports* 4 (4): 697-708. https://doi.org/10.1016/j.celrep.2013.07.026.

Dobler, Susanne, Safaa Dalla, Vera Wagschal, et Anurag A. Agrawal. 2012. « Community-Wide Convergent Evolution in Insect Adaptation to Toxic Cardenolides by Substitutions in the Na,K-ATPase ». *Proceedings of the National Academy of Sciences of the United States of America* 109 (32): 13040-45. https://doi.org/10.1073/pnas.1202111109.

Domingo, E., et J. J. Holland. 1997. « RNA VIRUS MUTATIONS AND FITNESS FOR SURVIVAL ». *Annual Review of Microbiology* 51 (1): 151-78. https://doi.org/10.1146/annurev.micro.51.1.151.

Donnell, Deborah, Jared M. Baeten, James Kiarie, Katherine K. Thomas, Wendy Stevens, Craig R. Cohen, James McIntyre, Jairam R. Lingappa, et Connie Celum. 2010. « Heterosexual HIV-1 Transmission after Initiation of Antiretroviral Therapy: A Prospective Cohort Analysis ». *The Lancet* 375 (9731): 2092-98. https://doi.org/10.1016/S0140-6736(10)60705-2.

Doolittle Russell F. 1994. « Convergent evolution: the need to be explicit », janvier.

Dorp, Lucy van, Mislav Acman, Damien Richard, Liam P. Shaw, Charlotte E. Ford, Louise Ormond, Christopher J. Owen, et al. 2020. « Emergence of Genomic Diversity and Recurrent Mutations in SARS-CoV-2 ». *Infection, Genetics and Evolution*, mai, 104351. https://doi.org/10.1016/j.meegid.2020.104351.

Drake, J. W. 1993. « Rates of Spontaneous Mutation among RNA Viruses ». *Proceedings of the National Academy of Sciences of the United States of America* 90 (9): 4171-75. https://doi.org/10.1073/pnas.90.9.4171.

Drake, J W, B Charlesworth, D Charlesworth, et J F Crow. 1998. « Rates of spontaneous mutation. » *Genetics* 148 (4): 1667-86.

Drexler, Jan Felix, Victor Max Corman, Marcel Alexander Müller, Alexander N. Lukashev, Anatoly Gmyl, Bruno Coutard, Alexander Adam, et al. 2013. « Evidence for Novel Hepaciviruses in Rodents ». Édité par David Wang. *PLoS Pathogens* 9 (6): e1003438. https://doi.org/10.1371/journal.ppat.1003438.

Drexler, Jan Felix, Victor Max Corman, Marcel Alexander Müller, Gael Darren Maganga, Peter Vallo, Tabea Binger, Florian Gloza-Rausch, et al. 2012. « Bats Host Major Mammalian Paramyxoviruses ». *Nature Communications* 3 (1): 796. https://doi.org/10.1038/ncomms1796.

Dudas, Gytis, Luiz Max Carvalho, Trevor Bedford, Andrew J. Tatem, Guy Baele, Nuno R. Faria, Daniel J. Park, et al. 2017. « Virus Genomes Reveal Factors That Spread and Sustained the Ebola Epidemic ». *Nature* 544 (7650): 309-15. https://doi.org/10.1038/nature22040.

Due, Ong The, Ammarin Thakkinstian, Montarat Thavorncharoensap, Abhasnee Sobhonslidsuk, Olivia Wu, Nguyen Khanh Phuong, et Usa Chaikledkaew. 2020. « Cost-Utility Analysis of Direct-Acting Antivirals for Treatment of Chronic Hepatitis C Genotype 1 and 6 in Vietnam ». *Value in Health* 23 (9): 1180-90. https://doi.org/10.1016/j.jval.2020.03.018.

Duffy, Siobain. 2018. « Why are RNA virus mutation rates so damn high? » *PLoS Biology* 16 (8): e3000003. https://doi.org/10.1371/journal.pbio.3000003.

Duffy, Siobain, Laura A. Shackelton, et Edward C. Holmes. 2008. « Rates of Evolutionary Change in Viruses: Patterns and Determinants ». *Nature Reviews Genetics* 9 (4): 267-76. https://doi.org/10.1038/nrg2323.

Dunn, Glynis, Dimitra Klapsa, Thomas Wilton, Lindsay Stone, Philip D. Minor, et Javier Martin. 2015. « Twenty-Eight Years of Poliovirus Replication in an Immunodeficient Individual: Impact on the Global Polio Eradication Initiative ». *PLoS Pathogens* 11 (8): e1005114. https://doi.org/10.1371/journal.ppat.1005114.

Edgar, Robert C. 2004. « MUSCLE: multiple sequence alignment with high accuracy and high throughput ». *Nucleic Acids Research* 32 (5): 1792-97. https://doi.org/10.1093/nar/gkh340.

Edgar, Robert C., Jeff Taylor, Victor Lin, Tomer Altman, Pierre Barbera, Dmitry Meleshko, Dan Lohr, et al. 2022. « Petabase-Scale Sequence Alignment Catalyses Viral Discovery ». *Nature* 602 (7895): 142-47. https://doi.org/10.1038/s41586-021-04332-2.

Efron, B. 1979. « Bootstrap Methods: Another Look at the Jackknife ». *The Annals of Statistics* 7 (1): 1-26. https://doi.org/10.1214/aos/1176344552.

Efron, Bradley, Elizabeth Halloran, et Susan Holmes. 1996. « Bootstrap Confidence Levels for Phylogenetic Trees ». *Proceedings of the National Academy of Sciences* 93 (23): 13429-13429.

Ehleringer, James R., Thure E. Cerling, et Brent R. Helliker. 1997. « C4 Photosynthesis, Atmospheric CO2, and Climate ». *Oecologia* 112 (3): 285-99. https://doi.org/10.1007/s004420050311.

Eid, John, Adrian Fehr, Jeremy Gray, Khai Luong, John Lyle, Geoff Otto, Paul Peluso, et al. 2009. « Real-Time DNA Sequencing from Single Polymerase Molecules ». *Science* 323 (5910): 133-38. https://doi.org/10.1126/science.1162986.

Eigen, Manfred. 2002. « Error catastrophe and antiviral strategy ». *Proceedings of the National Academy of Sciences of the United States of America* 99 (21): 13374-76. https://doi.org/10.1073/pnas.212514799.

El-Attar, L. M. R., J. A. Mitchell, H. Brooks Brownlie, S. L. Priestnall, et J. Brownlie. 2015. « Detection of Non-Primate Hepaciviruses in UK Dogs ». *Virology* 484 (octobre): 93-102. https://doi.org/10.1016/j.virol.2015.05.005.

Elena, Santiago F., et Rafael Sanjuán. 2005. « Adaptive Value of High Mutation Rates of RNA Viruses: Separating Causes from Consequences ». *Journal of Virology* 79 (18): 11555-58. https://doi.org/10.1128/JVI.79.18.11555-11558.2005.

Elliott, Paul, David Haw, Haowei Wang, Oliver Eales, Caroline E. Walters, Kylie E. C. Ainslie, Christina Atchison, et al. 2021. « Exponential Growth, High Prevalence of SARS-CoV-2, and Vaccine Effectiveness Associated with the Delta Variant ». *Science*, novembre. https://doi.org/10.1126/science.abl9551.

Escalera-Zamudio, Marina, Michael Golden, Bernardo Gutiérrez, Julien Thézé, Jeremy Russell Keown, Loic Carrique, Thomas A. Bowden, et Oliver G. Pybus. 2020. « Parallel Evolution in the Emergence of Highly Pathogenic Avian Influenza A Viruses ». *Nature Communications* 11 (1): 5511. https://doi.org/10.1038/s41467-020-19364-x.

Evans, Jason, Luke Sheneman, et James Foster. 2006. « Relaxed Neighbor Joining: A Fast Distance-Based Phylogenetic Tree Construction Method ». *Journal of Molecular Evolution* 62 (6): 785-92. https://doi.org/10.1007/s00239-005-0176-2.

Falade-Nwulia, Oluwaseun, Catalina Suarez-Cuervo, David R. Nelson, Michael W. Fried, Jodi B. Segal, et Mark S. Sulkowski. 2017. « Oral Direct-Acting Agent Therapy for Hepatitis C Virus Infection ». *Annals of internal medicine* 166 (9): 637-48. https://doi.org/10.7326/M16-2575.

Faria, Nuno R., Andrew Rambaut, Marc A. Suchard, Guy Baele, Trevor Bedford, Melissa J. Ward, Andrew J. Tatem, et al. 2014. « The early spread and epidemic ignition of HIV-1 in human populations ». *Science* 346 (6205): 56-61. https://doi.org/10.1126/science.1256739.

Feinstone, Stephen M., Albert Z. Kapikian, et Robert H. Purcell. 1973. « Hepatitis A: Detection by Immune Electron Microscopy of a Viruslike Antigen Associated with Acute Illness ». *Science* 182 (4116): 1026-28. https://doi.org/10.1126/science.182.4116.1026.

Feinstone, Stephen M., Albert Z. Kapikian, Robert H. Purcell, Harvey J. Alter, et Paul V. Holland. 1975. « Transfusion-Associated Hepatitis Not Due to Viral Hepatitis Type A or B ». *New England Journal of Medicine* 292 (15): 767-70. https://doi.org/10.1056/NEJM197504102921502.

Felsenstein, Joseph. 1981. « Evolutionary Trees from DNA Sequences: A Maximum Likelihood Approach ». *Journal of Molecular Evolution* 17 (6): 368-76. https://doi.org/10.1007/BF01734359.

———. 1985. « Confidence Limits on Phylogenies: An Approach Using the Bootstrap ». *Evolution* 39 (4): 783-91. https://doi.org/10.2307/2408678.

———. 2003. *Inferring Phylogenies*. Oxford, New York: Oxford University Press.

Feng, Da-Fei, et Russell F. Doolittle. 1987. « Progressive Sequence Alignment as a Prerequisitetto Correct Phylogenetic Trees ». *Journal of Molecular Evolution* 25 (4): 351-60. https://doi.org/10.1007/BF02603120.

Filée, Jonathan, Noëlle Pouget, et Mick Chandler. 2008. « Phylogenetic evidence for extensive lateral acquisition of cellular genes by Nucleocytoplasmic large DNA viruses ». *BMC Evolutionary Biology* 8 (novembre): 320. https://doi.org/10.1186/1471-2148-8-320.

Fischl, M. A., D. D. Richman, M. H. Grieco, M. S. Gottlieb, P. A. Volberding, O. L. Laskin, J. M. Leedom, J. E. Groopman, D. Mildvan, et R. T. Schooley. 1987. « The Efficacy of Azidothymidine (AZT) in the Treatment of Patients with AIDS and AIDS-Related Complex. A Double-Blind, Placebo-Controlled Trial ». *The New England Journal of Medicine* 317 (4): 185-91. https://doi.org/10.1056/NEJM198707233170401.

Fitch, W M, J M Leiter, X Q Li, et P Palese. 1991. « Positive Darwinian evolution in human influenza A viruses. » *Proceedings of the National Academy of Sciences of the United States of America* 88 (10): 4270-74.

Fitch, Walter M. 1971. « Toward Defining the Course of Evolution: Minimum Change for a Specific Tree Topology ». *Systematic Biology* 20 (4): 406-16. https://doi.org/10.1093/sysbio/20.4.406.

Fleischmann, W. Robert. 1996. « Viral Genetics ». In *Medical Microbiology*, édité par Samuel Baron, 4th éd. Galveston (TX): University of Texas Medical Branch at Galveston. http://www.ncbi.nlm.nih.gov/books/NBK8439/.

Foll, Matthieu, Oscar E. Gaggiotti, Josephine T. Daub, Alexandra Vatsiou, et Laurent Excoffier. 2014. « Widespread Signals of Convergent Adaptation to High Altitude in Asia and America ». *The American Journal of Human Genetics* 95 (4): 394-407. https://doi.org/10.1016/j.ajhg.2014.09.002.

Foll, Matthieu, Yu-Ping Poh, Nicholas Renzette, Anna Ferrer-Admetlla, Claudia Bank, Hyunjin Shim, Anna-Sapfo Malaspinas, et al. 2014. « Influenza Virus Drug Resistance: A Time-Sampled Population Genetics Perspective ». *PLoS Genetics* 10 (2): e1004185. https://doi.org/10.1371/journal.pgen.1004185.

Foote, Andrew D., Yue Liu, Gregg W. C. Thomas, Tomáš Vinař, Jessica Alföldi, Jixin Deng, Shannon Dugan, et al. 2015. « Convergent Evolution of the Genomes of Marine Mammals ». *Nature Genetics* 47 (3): 272-75. https://doi.org/10.1038/ng.3198.

Forni, Diego, Rachele Cagliani, Chiara Pontremoli, Uberto Pozzoli, Jacopo Vertemara, Luca De Gioia, Mario Clerici, et Manuela Sironi. 2018. « Evolutionary Analysis Provides Insight Into the Origin and Adaptation of HCV ». *Frontiers in Microbiology* 0. https://doi.org/10.3389/fmicb.2018.00854.

Frankel, Alan D., et John A. T. Young. 1998. « HIV-1: Fifteen Proteins and an RNA ». *Annual Review of Biochemistry* 67 (1): 1-25. https://doi.org/10.1146/annurev.biochem.67.1.1.

Fridell, Robert A., Dike Qiu, Lourdes Valera, Chunfu Wang, Ronald E. Rose, et Min Gao. 2011. « Distinct Functions of NS5A in Hepatitis C Virus RNA Replication Uncovered by Studies with the NS5A Inhibitor BMS-790052 ». *Journal of Virology* 85 (14): 7312-20. https://doi.org/10.1128/JVI.00253-11.

Fried, Michael W., Mitchell L. Shiffman, K. Rajender Reddy, Coleman Smith, George Marinos, Fernando L. Gonçales, Dieter Häussinger, et al. 2002. « Peginterferon Alfa-2a plus Ribavirin for Chronic Hepatitis C Virus Infection ». *The New England Journal of Medicine* 347 (13): 975-82. https://doi.org/10.1056/NEJMoa020047.

Friedrich, Thomas C., Christopher A. Frye, Levi J. Yant, David H. O'Connor, Nancy A. Kriewaldt, Meghan Benson, Lara Vojnov, et al. 2004. « Extraepitopic Compensatory Substitutions Partially Restore Fitness to Simian Immunodeficiency Virus Variants That Escape from an Immunodominant Cytotoxic-T-Lymphocyte Response ». *Journal of Virology* 78 (5): 2581-85. https://doi.org/10.1128/JVI.78.5.2581-2585.2004.

Frost, Simon D. W., Monique Nijhuis, Rob Schuurman, Charles A. B. Boucher, et Andrew J. Leigh Brown. 2000. « Evolution of Lamivudine Resistance in Human Immunodeficiency Virus Type 1-Infected Individuals: the Relative Roles of Drift and Selection ». *Journal of Virology* 74 (14): 6262-68. https://doi.org/10.1128/JVI.74.14.6262-6268.2000.

Fry, Bryan G., Kim Roelants, Donald E. Champagne, Holger Scheib, Joel D. A. Tyndall, Glenn F. King, Timo J. Nevalainen, et al. 2009. « The Toxicogenomic Multiverse: Convergent Recruitment of Proteins into Animal Venoms ». *Annual Review of Genomics and Human Genetics* 10: 483-511. https://doi.org/10.1146/annurev.genom.9.081307.164356.

Gago, Selma, Santiago F. Elena, Ricardo Flores, et Rafael Sanjuán. 2009. « Extremely High Mutation Rate of a Hammerhead Viroid ». *Science* 323 (5919): 1308-1308. https://doi.org/10.1126/science.1169202.

Garcia-Diaz, Miguel, et Katarzyna Bebenek. 2007. « Multiple functions of DNA polymerases ». *Critical reviews in plant sciences* 26 (2): 105-22. https://doi.org/10.1080/07352680701252817.

Gardy, Jennifer, Nicholas J. Loman, et Andrew Rambaut. 2015. « Real-time digital pathogen surveillance — the time is now ». *Genome Biology* 16 (1): 155. https://doi.org/10.1186/s13059-015-0726-x.

Gascuel, Olivier. 2005. *Mathematics of Evolution and Phylogeny*. OUP Oxford.

Gascuel, Olivier, et Mike Steel. 2006. « Neighbor-Joining Revealed ». *Molecular Biology and Evolution* 23 (11): 1997-2000. https://doi.org/10.1093/molbev/msl072.

———. 2007. *Reconstructing Evolution: New Mathematical and Computational Advances*. OUP Oxford.

Gaudieri, S., A. Rauch, L.P. Park, E. Freitas, S. Herrmann, G. Jeffrey, W. Cheng, et al. 2006. « Evidence of Viral Adaptation to HLA Class I-Restricted Immune Pressure in Chronic Hepatitis C Virus Infection ». *Journal of Virology* 80 (22): 11094-104. https://doi.org/10.1128/JVI.00912-06.

Gemaque, Bernard Salame, Alex Junior Souza de Souza, Manoel do Carmo Pereira Soares, Andreza Pinheiro Malheiros, Andrea Lima Silva, Max Moreira Alves, Michele Soares Gomes-Gouvêa, et al. 2014. « Hepacivirus Infection in Domestic Horses, Brazil, 2011–2013 ». *Emerging Infectious Diseases* 20 (12): 2180-82. https://doi.org/10.3201/eid2012.140603.

Goldman, N., et Z. Yang. 1994. « A Codon-Based Model of Nucleotide Substitution for Protein-Coding DNA Sequences ». *Molecular Biology and Evolution* 11 (5): 725-36. https://doi.org/10.1093/oxfordjournals.molbev.a040153.

Goldstein, Richard A., Stephen T. Pollard, Seena D. Shah, et David D. Pollock. 2015. « Nonadaptive Amino Acid Convergence Rates Decrease over Time ». *Molecular Biology and Evolution* 32 (6): 1373-81. https://doi.org/10.1093/molbev/msv041.

Gompel, Nicolas, et Benjamin Prud'homme. 2009. « The Causes of Repeated Genetic Evolution ». *Developmental Biology*, Special Section: Evolution of Developmental Regulatory Systems, 332 (1): 36-47. https://doi.org/10.1016/j.ydbio.2009.04.040.

Goodwin, Sara, John D. McPherson, et W. Richard McCombie. 2016. « Coming of Age: Ten Years of next-Generation Sequencing Technologies ». *Nature Reviews Genetics* 17 (6): 333-51. https://doi.org/10.1038/nrg.2016.49.

Gorbalenya, Alexander E., Luis Enjuanes, John Ziebuhr, et Eric J. Snijder. 2006. « Nidovirales: Evolving the largest RNA virus genome ». *Virus Research* 117 (1): 17-37. https://doi.org/10.1016/j.virusres.2006.01.017.

Gottwein, Judith M., Long V. Pham, Lotte S. Mikkelsen, Lubna Ghanem, Santseharay Ramirez, Troels K. H. Scheel, Thomas H. R. Carlsen, et Jens Bukh. 2018. « Efficacy of NS5A Inhibitors Against Hepatitis C Virus Genotypes 1–7 and Escape Variants ». *Gastroenterology* 154 (5): 1435-48. https://doi.org/10.1053/j.gastro.2017.12.015.

Gould, S.J. 1989. « Wonderful Life. The Burgess Shale and the Nature of History. » *New York, NY: W. W. Norton & Co.*, 347. https://doi.org/10.3366/anh.1991.18.1.138.

Gower, Erin, Chris Estes, Sarah Blach, Kathryn Razavi-Shearer, et Homie Razavi. 2014. « Global Epidemiology and Genotype Distribution of the Hepatitis C Virus Infection ». *Journal of Hepatology* 61 (1 Suppl): S45-57. https://doi.org/10.1016/j.jhep.2014.07.027.

Grubaugh, Nathan D., Karthik Gangavarapu, Joshua Quick, Nathaniel L. Matteson, Jaqueline Goes De Jesus, Bradley J. Main, Amanda L. Tan, et al. 2019. « An amplicon-based sequencing framework for accurately measuring intrahost virus diversity using PrimalSeq and iVar ». *Genome Biology* 20 (1): 8. https://doi.org/10.1186/s13059-018-1618-7.

Guglielmini, Julien, Anthony C. Woo, Mart Krupovic, Patrick Forterre, et Morgan Gaia. 2019. « Diversification of Giant and Large Eukaryotic DsDNA Viruses Predated the Origin of Modern Eukaryotes ». *Proceedings of the National Academy of Sciences* 116 (39): 19585-92. https://doi.org/10.1073/pnas.1912006116.

Gutierrez, Bernardo, Marina Escalera-Zamudio, et Oliver G Pybus. 2019. « Parallel molecular evolution and adaptation in viruses ». *Current Opinion in Virology* 34 (février): 90-96. https://doi.org/10.1016/j.coviro.2018.12.006.

Gutiérrez, Serafín, Yannis Michalakis, et Stéphane Blanc. 2012. « Virus Population Bottlenecks during Within-Host Progression and Host-to-Host Transmission ». *Current Opinion in Virology*, Virus evolution / Antivirals and resistance, 2 (5): 546-55. https://doi.org/10.1016/j.coviro.2012.08.001.

Hadfield, James, Colin Megill, Sidney M Bell, John Huddleston, Barney Potter, Charlton Callender, Pavel Sagulenko, Trevor Bedford, et Richard A Neher. 2018. « Nextstrain: real-time tracking of pathogen evolution ». *Bioinformatics* 34 (23): 4121-23. https://doi.org/10.1093/bioinformatics/bty407.

Hammond, Jennifer, Brendan A Larder, Raymond F Schinazi, et John W Mellors. 1998. « Mutations in Retroviral Genes Associated with Drug Resistance », n° 20: 43.

Han, Bin, Ross Martin, Simin Xu, Aiyappa Parvangada, Evguenia S. Svarovskaia, Hongmei Mo, et Hadas Dvory-Sobol. 2019. « Sofosbuvir Susceptibility of Genotype 1 to 6 HCV from DAA-Naïve Subjects ». *Antiviral Research* 170 (octobre): 104574. https://doi.org/10.1016/j.antiviral.2019.104574.

Harris, Charles R., K. Jarrod Millman, Stéfan J. van der Walt, Ralf Gommers, Pauli Virtanen, David Cournapeau, Eric Wieser, et al. 2020. « Array Programming with NumPy ». *Nature* 585 (7825): 357-62. https://doi.org/10.1038/s41586-020-2649-2.

Harvey, William T., Alessandro M. Carabelli, Ben Jackson, Ravindra K. Gupta, Emma C. Thomson, Ewan M. Harrison, Catherine Ludden, et al. 2021. « SARS-CoV-2 Variants, Spike Mutations and Immune Escape ». *Nature Reviews Microbiology* 19 (7): 409-24. https://doi.org/10.1038/s41579-021-00573-0.

Hasegawa, Masami, Hirohisa Kishino, et Taka-aki Yano. 1985. « Dating of the Human-Ape Splitting by a Molecular Clock of Mitochondrial DNA ». *Journal of Molecular Evolution* 22 (2): 160-74. https://doi.org/10.1007/BF02101694.

Hashimoto, Masao, Mitsutaka Kitano, Kazutaka Honda, Hirokazu Koizumi, Sachi Dohki, Shinichi Oka, et Masafumi Takiguchi. 2010. « Selection of Escape Mutation by Pol154-162-Specific Cytotoxic T Cells among Chronically HIV-1-Infected HLA-B*5401-Positive Individuals ». *Human Immunology* 71 (2): 123-27. https://doi.org/10.1016/j.humimm.2009.10.015.

Hedskog, C., H. Dvory-Sobol, V. Gontcharova, R. Martin, W. Ouyang, B. Han, E. J. Gane, et al. 2015. « Evolution of the HCV Viral Population from a Patient with S282T Detected at

Relapse after Sofosbuvir Monotherapy ». *Journal of Viral Hepatitis* 22 (11): 871-81. https://doi.org/10.1111/jvh.12405.

Hedskog, Charlotte, Bandita Parhy, Silvia Chang, Stefan Zeuzem, Christophe Moreno, Stephen D. Shafran, Sergio M. Borgia, et al. 2019. « Identification of 19 Novel Hepatitis C Virus Subtypes-Further Expanding HCV Classification ». *Open Forum Infectious Diseases* 6 (3): ofz076. https://doi.org/10.1093/ofid/ofz076.

Hemelaar, Joris. 2012. « The Origin and Diversity of the HIV-1 Pandemic ». *Trends in Molecular Medicine* 18 (3): 182-92. https://doi.org/10.1016/j.molmed.2011.12.001.

Hemelaar, Joris, Ramyiadarsini Elangovan, Jason Yun, Leslie Dickson-Tetteh, Shona Kirtley, Eleanor Gouws-Williams, Peter D. Ghys, et al. 2020. « Global and Regional Epidemiology of HIV-1 Recombinants in 1990–2015: A Systematic Review and Global Survey ». *The Lancet HIV* 7 (11): e772-81. https://doi.org/10.1016/S2352-3018(20)30252-6.

Hendrix, Roger W., Margaret C. M. Smith, R. Neil Burns, Michael E. Ford, et Graham F. Hatfull. 1999. « Evolutionary relationships among diverse bacteriophages and prophages: All the world's a phage ». *Proceedings of the National Academy of Sciences* 96 (5): 2192-97. https://doi.org/10.1073/pnas.96.5.2192.

Henikoff, S, et J G Henikoff. 1992. « Amino acid substitution matrices from protein blocks. » *Proceedings of the National Academy of Sciences of the United States of America* 89 (22): 10915-19.

Hill, Jason, Erik D. Enbody, Mats E. Pettersson, C. Grace Sprehn, Dorte Bekkevold, Arild Folkvord, Linda Laikre, Gunnar Kleinau, Patrick Scheerer, et Leif Andersson. 2019. « Recurrent Convergent Evolution at Amino Acid Residue 261 in Fish Rhodopsin ». *Proceedings of the National Academy of Sciences* 116 (37): 18473-78. https://doi.org/10.1073/pnas.1908332116.

Hillis, David M., et James J. Bull. 1993. « An Empirical Test of Bootstrapping as a Method for Assessing Confidence in Phylogenetic Analysis ». *Systematic Biology* 42 (2): 182-92. https://doi.org/10.1093/sysbio/42.2.182.

Hlaing, Naomi Khaing Than, Gayatri Nangia, Kyaw Thet Tun, Sithu Lin, Moe Zaw Maung, Khin Thuzar Myint, A. Mi Mi Kyaw, et al. 2019. « High Sustained Virologic Response in Genotypes 3 and 6 with Generic NS5A Inhibitor and Sofosbuvir Regimens in Chronic HCV in Myanmar ». *Journal of Viral Hepatitis* 26 (10): 1186-99. https://doi.org/10.1111/jvh.13133.

Hodcroft, Emma B., Moira Zuber, Sarah Nadeau, Timothy G. Vaughan, Katharine H. D. Crawford, Christian L. Althaus, Martina L. Reichmuth, et al. 2021. « Spread of a SARS-CoV-2 Variant through Europe in the Summer of 2020 ». *Nature* 595 (7869): 707-12. https://doi.org/10.1038/s41586-021-03677-y.

Hoenen, Thomas, Allison Groseth, Kyle Rosenke, Robert J. Fischer, Andreas Hoenen, Seth D. Judson, Cynthia Martellaro, et al. 2016. « Nanopore Sequencing as a Rapidly Deployable Ebola Outbreak Tool ». *Emerging Infectious Diseases* 22 (2): 331-34. https://doi.org/10.3201/eid2202.151796.

Holm, Sture. 1979. « A Simple Sequentially Rejective Multiple Test Procedure ». *Scandinavian Journal of Statistics* 6 (2): 65-70.

Holmes, E C, L Q Zhang, P Simmonds, C A Ludlam, et A J Brown. 1992. « Convergent and divergent sequence evolution in the surface envelope glycoprotein of human immunodeficiency virus type 1 within a single infected patient. » *Proceedings of the National Academy of Sciences of the United States of America* 89 (11): 4835-39.

Holmes, Edward C., et Sebastián Duchêne. 2019. « Can Sequence Phylogenies Safely Infer the Origin of the Global Virome? » *mBio* 10 (2): e00289-19. https://doi.org/10.1128/mBio.00289-19.

Holmes, Edward C., Stephen A. Goldstein, Angela L. Rasmussen, David L. Robertson, Alexander Crits-Christoph, Joel O. Wertheim, Simon J. Anthony, et al. 2021. « The Origins of SARS-

CoV-2: A Critical Review ». *Cell* 184 (19): 4848-56. https://doi.org/10.1016/j.cell.2021.08.017.

Houldcroft, Charlotte J., Mathew A. Beale, et Judith Breuer. 2017. « Clinical and Biological Insights from Viral Genome Sequencing ». *Nature Reviews Microbiology* 15 (3): 183-92. https://doi.org/10.1038/nrmicro.2016.182.

Hu, Yibo, Qi Wu, Shuai Ma, Tianxiao Ma, Lei Shan, Xiao Wang, Yonggang Nie, et al. 2017. « Comparative Genomics Reveals Convergent Evolution between the Bamboo-Eating Giant and Red Pandas ». *Proceedings of the National Academy of Sciences* 114 (5): 1081-86. https://doi.org/10.1073/pnas.1613870114.

Huerta-Cepas, Jaime, François Serra, et Peer Bork. 2016. « ETE 3: Reconstruction, Analysis, and Visualization of Phylogenomic Data ». *Molecular Biology and Evolution* 33 (6): 1635-38. https://doi.org/10.1093/molbev/msw046.

« Human Immunodeficiency Virus (HIV) ». 2016. *Transfusion Medicine and Hemotherapy* 43 (3): 203-22. https://doi.org/10.1159/000445852.

Hunter, John D. 2007. « Matplotlib: A 2D Graphics Environment ». *Computing in Science Engineering* 9 (3): 90-95. https://doi.org/10.1109/MCSE.2007.55.

Ingram, Travis, et D.Luke Mahler. 2013. « SURFACE: Detecting Convergent Evolution from Comparative Data by Fitting Ornstein-Uhlenbeck Models with Stepwise Akaike Information Criterion ». *Methods in Ecology and Evolution* 4 (5): 416-25. https://doi.org/10.1111/2041-210X.12034.

Ishikawa, Sohta A., Anna Zhukova, Wataru Iwasaki, et Olivier Gascuel. 2019. « A Fast Likelihood Method to Reconstruct and Visualize Ancestral Scenarios ». *Molecular Biology and Evolution*. https://doi.org/10.1093/molbev/msz131.

Ivanowsky, D. 1892. « Ueber die Mosaikkrankheit der Tabakspflanze ». *St Petersb Acad Imp Sci Bul* 35: 67-70.

Jacobson, Ira M. 2016. « The HCV Treatment Revolution Continues: Resistance Considerations, Pangenotypic Efficacy, and Advances in Challenging Populations ». *Gastroenterology & Hepatology* 12 (10 Suppl 4): 1-11.

Jain, Miten, Sergey Koren, Karen H Miga, Josh Quick, Arthur C Rand, Thomas A Sasani, John R Tyson, et al. 2018. « Nanopore sequencing and assembly of a human genome with ultra-long reads ». *Nature Biotechnology* 36 (4): 338-45. https://doi.org/10.1038/nbt.4060.

Jefferson, Tom, Mark Jones, Peter Doshi, Elizabeth A. Spencer, Igho Onakpoya, et Carl J. Heneghan. 2014. « Oseltamivir for Influenza in Adults and Children: Systematic Review of Clinical Study Reports and Summary of Regulatory Comments ». *BMJ* 348 (avril): g2545. https://doi.org/10.1136/bmj.g2545.

Jones, David T., William R. Taylor, et Janet M. Thornton. 1992. « The Rapid Generation of Mutation Data Matrices from Protein Sequences ». *Bioinformatics* 8 (3): 275-82. https://doi.org/10.1093/bioinformatics/8.3.275.

Joy, Jeffrey B., Richard H. Liang, Rosemary M. McCloskey, T. Nguyen, et Art F. Y. Poon. 2016. « Ancestral Reconstruction ». *PLOS Computational Biology* 12 (7): e1004763. https://doi.org/10.1371/journal.pcbi.1004763.

Jubin, R. 2001. « Hepatitis C IRES: Translating Translation into a Therapeutic Target ». *Current Opinion in Molecular Therapeutics* 3 (3): 278-87.

Jukes, Thomas H., et Charles R. Cantor. 1969. « Evolution of Protein Molecules ». In *Mammalian Protein Metabolism*, 21-132. Elsevier. https://doi.org/10.1016/B978-1-4832-3211-9.50009-7.

Kai, Yugo, Hayato Hikita, Naoki Morishita, Kazuhiro Murai, Tasuku Nakabori, Sadaharu Iio, Hideki Hagiwara, et al. 2017. « Baseline quasispecies selection and novel mutations contribute to emerging resistance-associated substitutions in hepatitis C virus after direct-acting antiviral treatment ». *Scientific Reports* 7 (janvier). https://doi.org/10.1038/srep41660.

Kaiser, Laurent, John-David Aubert, Jean-Claude Pache, Christelle Deffernez, Thierry Rochat, Jorge Garbino, Werner Wunderli, et al. 2006. « Chronic Rhinoviral Infection in Lung Transplant Recipients ». *American Journal of Respiratory and Critical Care Medicine* 174 (12): 1392-99. https://doi.org/10.1164/rccm.200604-489OC.

Kalaghatgi, Prabhav, Anna Maria Sikorski, Elena Knops, Daniel Rupp, Saleta Sierra, Eva Heger, Maria Neumann-Fraune, et al. 2016. « Geno2pheno[HCV] - A Web-Based Interpretation System to Support Hepatitis C Treatment Decisions in the Era of Direct-Acting Antiviral Agents ». *PloS One* 11 (5): e0155869. https://doi.org/10.1371/journal.pone.0155869.

Kalyaanamoorthy, Subha, Bui Quang Minh, Thomas K. F. Wong, Arndt von Haeseler, et Lars S. Jermiin. 2017. « ModelFinder: Fast Model Selection for Accurate Phylogenetic Estimates ». *Nature Methods* 14 (6): 587-89. https://doi.org/10.1038/nmeth.4285.

Kanhayuwa, Lakkhana, Ioly Kotta-Loizou, Selin Özkan, A. Patrick Gunning, et Robert H. A. Coutts. 2015. « A novel mycovirus from Aspergillus fumigatus contains four unique dsRNAs as its genome and is infectious as dsRNA ». *Proceedings of the National Academy of Sciences* 112 (29): 9100-9105. https://doi.org/10.1073/pnas.1419225112.

Karim, Salim S. Abdool, et Quarraisha Abdool Karim. 2021. « Omicron SARS-CoV-2 Variant: A New Chapter in the COVID-19 Pandemic ». *The Lancet* 398 (10317): 2126-28. https://doi.org/10.1016/S0140-6736(21)02758-6.

Kass, Robert E., et Adrian E. Raftery. 1995. « Bayes Factors ». *Journal of the American Statistical Association* 90 (430): 773-95. https://doi.org/10.2307/2291091.

Kati, Warren, Gennadiy Koev, Michelle Irvin, Jill Beyer, Yaya Liu, Preethi Krishnan, Thomas Reisch, et al. 2015. « In Vitro Activity and Resistance Profile of Dasabuvir, a Nonnucleoside Hepatitis C Virus Polymerase Inhibitor ». *Antimicrobial Agents and Chemotherapy* 59 (3): 1505-11. https://doi.org/10.1128/AAC.04619-14.

Katoh, Kazutaka, Kazuharu Misawa, Kei-ichi Kuma, et Takashi Miyata. 2002. « MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform ». *Nucleic Acids Research* 30 (14): 3059-66. https://doi.org/10.1093/nar/gkf436.

Katoh, Kazutaka, et Daron M. Standley. 2013. « MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability ». *Molecular Biology and Evolution* 30 (4): 772-80. https://doi.org/10.1093/molbev/mst010.

Kausche, Gustav A., Edgar Pfankuch, et Helmut Ruska. 1939. « Die Sichtbarmachung von pflanzlichem virus im Übermikroskop ». *Naturwissenschaften* 27 (18): 292-99.

Kaverin, Nikolai V., Irina A. Rudneva, Natalia A. Ilyushina, Natalia L. Varich, Aleksandr S. Lipatov, Yuri A. Smirnov, Elena A. Govorkova, Asya K. Gitelman, Dmitri K. Lvov, et Robert G. Webster. 2002. « Structure of Antigenic Sites on the Haemagglutinin Molecule of H5 Avian Influenza Virus and Phenotypic Variation of Escape Mutants ». *The Journal of General Virology* 83 (Pt 10): 2497-2505. https://doi.org/10.1099/0022-1317-83-10-2497.

Keele, Brandon F., Fran Van Heuverswyn, Yingying Li, Elizabeth Bailes, Jun Takehisa, Mario L. Santiago, Frederic Bibollet-Ruche, et al. 2006. « Chimpanzee Reservoirs of Pandemic and Nonpandemic HIV-1 ». *Science* 313 (5786): 523-26. https://doi.org/10.1126/science.1126531.

Kellam, Paul, Charles A. B. Boucher, Jolanda M. G. H. Tijnagel, et Brendan A.YR Larder. 1994. « Zidovudine treatment results in the selection of human immunodeficiency virus type 1 variants whose genotypes confer increasing levels of drug resistance ». *Journal of General Virology* 75 (2): 341-51. https://doi.org/10.1099/0022-1317-75-2-341.

Kilmarx, Peter H. 2009. « Global Epidemiology of HIV ». *Current Opinion in HIV and AIDS* 4 (4): 240-46. https://doi.org/10.1097/COH.0b013e32832c06db.

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): 111-20.

Kimura, Motoo. 1977. « Preponderance of Synonymous Changes as Evidence for the Neutral Theory of Molecular Evolution ». *Nature* 267 (5608): 275-76. https://doi.org/10.1038/267275a0.

King, Jacqueline, Timm Harder, Martin Beer, et Anne Pohlmann. 2020. « Rapid multiplex MinION nanopore sequencing workflow for Influenza A viruses ». *BMC Infectious Diseases* 20 (1): 648. https://doi.org/10.1186/s12879-020-05367-y.

Kluge, Arnold G., et James S. Farris. 1969. « Quantitative Phyletics and the Evolution of Anurans ». *Systematic Biology* 18 (1): 1-32. https://doi.org/10.1093/sysbio/18.1.1.

Kluge, Silvia F., Katharina Mack, Shilpa S. Iyer, François M. Pujol, Anke Heigele, Gerald H. Learn, Shariq M. Usmani, et al. 2014. « Nef Proteins of Epidemic HIV-1 Group O Strains Antagonize Human Tetherin ». *Cell Host & Microbe* 16 (5): 639-50. https://doi.org/10.1016/j.chom.2014.10.002.

Koonin, Eugene V., Valerian V. Dolja, Mart Krupovic, Arvind Varsani, Yuri I. Wolf, Natalya Yutin, F. Murilo Zerbini, et Jens H. Kuhn. 2020. « Global Organization and Proposed Megataxonomy of the Virus World ». *Microbiology and Molecular Biology Reviews: MMBR* 84 (2): e00061-19. https://doi.org/10.1128/MMBR.00061-19.

Korber, Bette, Will M. Fischer, Sandrasegaram Gnanakaran, Hyejin Yoon, James Theiler, Werner Abfalterer, Nick Hengartner, et al. 2020. « Tracking Changes in SARS-CoV-2 Spike: Evidence That D614G Increases Infectivity of the COVID-19 Virus ». *Cell* 182 (4): 812-827.e19. https://doi.org/10.1016/j.cell.2020.06.043.

Kuhn, Jens H., Yiming Bao, Sina Bavari, Stephan Becker, Steven Bradfute, J. Rodney Brister, Alexander A. Bukreyev, et al. 2013. « Virus nomenclature below the species level: a standardized nomenclature for natural variants of viruses assigned to the family Filoviridae ». *Archives of Virology* 158 (1): 301-11. https://doi.org/10.1007/s00705-012-1454-0.

Kuhn, Jens H., Yuri I. Wolf, Mart Krupovic, Yong-Zhen Zhang, Piet Maes, Valerian V. Dolja, et Eugene V. Koonin. 2019. « Classify Viruses — the Gain Is Worth the Pain ». *Nature* 566 (7744): 318-20. https://doi.org/10.1038/d41586-019-00599-8.

Kuiken, Carla, Bette Korber, et Robert W. Shafer. 2003. « HIV Sequence Databases ». *AIDS reviews* 5 (1): 52-61.

Kuiken, Thijs, Edward C. Holmes, John McCauley, Guus F. Rimmelzwaan, Catherine S. Williams, et Bryan T. Grenfell. 2006. « Host Species Barriers to Influenza Virus Infections ». *Science* 312 (5772): 394-97. https://doi.org/10.1126/science.1122818.

Kuo, G., Q.-L. Choo, H.J. Alter, G.L. Gitnick, A.G. Redeker, R.H. Purcell, T. Miyamura, et al. 1989. « An Assay for Circulating Antibodies to a Major Etiologic Virus of Human Non-A, Non-B Hepatitis ». *Science* 244 (4902): 362-64. https://doi.org/10.1126/science.2496467.

La Scola, Bernard, Stéphane Audic, Catherine Robert, Liang Jungang, Xavier de Lamballerie, Michel Drancourt, Richard Birtles, Jean-Michel Claverie, et Didier Raoult. 2003. « A Giant Virus in Amoebae ». *Science (New York, N.Y.)* 299 (5615): 2033. https://doi.org/10.1126/science.1081867.

La Scola, Bernard, Christelle Desnues, Isabelle Pagnier, Catherine Robert, Lina Barrassi, Ghislain Fournous, Michèle Merchat, et al. 2008. « The Virophage as a Unique Parasite of the Giant Mimivirus ». *Nature* 455 (7209): 100-104. https://doi.org/10.1038/nature07218.

Lam, Angela M., Christine Espiritu, Shalini Bansal, Holly M. Micolochick Steuer, Congrong Niu, Veronique Zennou, Meg Keilman, et al. 2012. « Genotype and Subtype Profiling of PSI-7977 as a Nucleotide Inhibitor of Hepatitis C Virus ». *Antimicrobial Agents and Chemotherapy* 56 (6): 3359-68. https://doi.org/10.1128/AAC.00054-12.

Lambisia, Arnold W., Khadija S. Mohammed, Timothy O. Makori, Leonard Ndwiga, Maureen W. Mburu, John M. Morobe, Edidah O. Moraa, et al. 2022. « Optimization of the SARS-CoV-2 ARTIC Network V4 Primers and Whole Genome Sequencing Protocol ». *Frontiers in Medicine* 9. https://doi.org/10.3389/fmed.2022.836728.

Land, S. S. Fr ø, P. Jenum, C. F. Lindboe, K. W. Wefring, P. J. Linnestad, et T. Böhmer. 1988. « HIV-1 INFECTION IN NORWEGIAN FAMILY BEFORE 1970 ». *The Lancet* 331 (8598): 1344-45. https://doi.org/10.1016/S0140-6736(88)92164-2.

Larder, B. A., G. Darby, et D. D. Richman. 1989. « HIV with Reduced Sensitivity to Zidovudine (AZT) Isolated during Prolonged Therapy ». *Science (New York, N.Y.)* 243 (4899): 1731-34. https://doi.org/10.1126/science.2467383.

Larter, Maximilian, Amy Dunbar-Wallis, Andrea E Berardi, et Stacey D Smith. 2018. « Convergent Evolution at the Pathway Level: Predictable Regulatory Changes during Flower Color Transitions ». *Molecular Biology and Evolution* 35 (9): 2159-69. https://doi.org/10.1093/molbev/msy117.

Lawitz, Eric, Alessandra Mangia, David Wyles, Maribel Rodriguez-Torres, Tarek Hassanein, Stuart C. Gordon, Michael Schultz, et al. 2013. « Sofosbuvir for Previously Untreated Chronic Hepatitis C Infection ». *New England Journal of Medicine* 368 (20): 1878-87. https://doi.org/10.1056/NEJMoa1214853.

Le, Si Quang, Cuong Cao Dang, et Olivier Gascuel. 2012. « Modeling Protein Evolution with Several Amino Acid Replacement Matrices Depending on Site Rates ». *Molecular Biology and Evolution* 29 (10): 2921-36. https://doi.org/10.1093/molbev/mss112.

Le, Si Quang, et Olivier Gascuel. 2008. « An Improved General Amino Acid Replacement Matrix ». *Molecular Biology and Evolution* 25 (7): 1307-20. https://doi.org/10.1093/molbev/msn067.

Le, Si Quang, Olivier Gascuel, et Nicolas Lartillot. 2008. « Empirical profile mixture models for phylogenetic reconstruction ». *Bioinformatics* 24 (20): 2317-23.

Le, Si Quang, Nicolas Lartillot, et Olivier Gascuel. 2008. « Phylogenetic mixture models for proteins ». *Philosophical Transactions of the Royal Society B: Biological Sciences* 363 (1512): 3965-76. https://doi.org/10.1098/rstb.2008.0180.

Le, Vinh Sy, Cuong Cao Dang, et Quang Si Le. 2017. « Improved Mitochondrial Amino Acid Substitution Models for Metazoan Evolutionary Studies ». *BMC Evolutionary Biology* 17 (1): 136-136. https://doi.org/10.1186/s12862-017-0987-y.

Lee, Jun-Hoe, Kevin M. Lewis, Timothy W. Moural, Bogdan Kirilenko, Barbara Borgonovo, Gisa Prange, Manfred Koessl, Stefan Huggenberger, ChulHee Kang, et Michael Hiller. 2018. « Building Superfast Muscles: Insights from Molecular Parallelism in Fast-Twitch Muscle Proteins in Echolocating Mammals ». *BioRxiv*, janvier, 244566. https://doi.org/10.1101/244566.

Lefkowitz, Elliot J, Donald M Dempsey, Robert Curtis Hendrickson, Richard J Orton, Stuart G Siddell, et Donald B Smith. 2018. « Virus taxonomy: the database of the International Committee on Taxonomy of Viruses (ICTV) ». *Nucleic Acids Research* 46 (Database issue): D708-17. https://doi.org/10.1093/nar/gkx932.

Lefort, Vincent, Richard Desper, et Olivier Gascuel. 2015. « FastME 2.0: A Comprehensive, Accurate, and Fast Distance-Based Phylogeny Inference Program ». *Molecular Biology and Evolution* 32 (10): 2798-2800. https://doi.org/10.1093/molbev/msv150.

Lemey, Philippe, Andrew Rambaut, Trevor Bedford, Nuno Faria, Filip Bielejec, Guy Baele, Colin A. Russell, et al. 2014. « Unifying Viral Genetics and Human Transportation Data to Predict the Global Transmission Dynamics of Human Influenza H3N2 ». *PLoS Pathogens* 10 (2): e1003932. https://doi.org/10.1371/journal.ppat.1003932.

Lemey, Philippe, Marco Salemi, et Anne-Mieke Vandamme, éd. 2009. *The Phylogenetic Handbook: A Practical Approach to Phylogenetic Analysis and Hypothesis Testing*. 2e éd. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511819049.

Lemoine, F., J.-B. Domelevo Entfellner, E. Wilkinson, D. Correia, M. Dávila Felipe, T. De Oliveira, et O. Gascuel. 2018. « Renewing Felsenstein's Phylogenetic Bootstrap in the Era of Big Data ». *Nature* 556 (7702): 452-56. https://doi.org/10.1038/s41586-018-0043-0.

Lemoine, Frédéric, et Olivier Gascuel. 2021a. « Gotree/Goalign : Toolkit and Go API to Facilitate the Development of Phylogenetic Workflows ». *BioRxiv*, juin, 2021.06.09.447704. https://doi.org/10.1101/2021.06.09.447704.

———. 2021b. « Gotree/Goalign: toolkit and Go API to facilitate the development of phylogenetic workflows ». *NAR Genomics and Bioinformatics* 3 (3). https://doi.org/10.1093/nargab/lqab075.

Lenormand, Thomas, Luis-Miguel Chevin, et Thomas Bataillon. 2016. « Parallel evolution: what does it (not) tell us and why is it (still) interesting? » In .

Leslie, A. J., K. J. Pfafferott, P. Chetty, R. Draenert, M. M. Addo, M. Feeney, Y. Tang, et al. 2004. « HIV Evolution: CTL Escape Mutation and Reversion after Transmission ». *Nature Medicine* 10 (3): 282-89. https://doi.org/10.1038/nm992.

Letunic, Ivica, et Peer Bork. 2019. « Interactive Tree Of Life (iTOL) v4: recent updates and new developments ». *Nucleic Acids Research* 47 (W1): W256-59. https://doi.org/10.1093/nar/gkz239.

Li, Chunhua, Eleanor Barnes, Paul N. Newton, Yongshui Fu, Manivanh Vongsouvath, Paul Klenerman, Hiroaki Okamoto, Kenji Abe, Oliver G. Pybus, et Ling Lu. 2015. « An Expanded Taxonomy of Hepatitis C Virus Genotype 6: Characterization of 22 New Full-Length Viral Genomes ». *Virology* 476 (février): 355-63. https://doi.org/10.1016/j.virol.2014.12.025.

Li, Gang, Jinhong Wang, Stephen J. Rossiter, Gareth Jones, James A. Cotton, et Shuyi Zhang. 2008. « The Hearing Gene Prestin Reunites Echolocating Bats ». *Proceedings of the National Academy of Sciences of the United States of America* 105 (37): 13959-64. https://doi.org/10.1073/pnas.0802097105.

Li, Heng, Bob Handsaker, Alec Wysoker, Tim Fennell, Jue Ruan, Nils Homer, Gabor Marth, Goncalo Abecasis, Richard Durbin, et 1000 Genome Project Data Processing Subgroup. 2009. « The Sequence Alignment/Map Format and SAMtools ». *Bioinformatics (Oxford, England)* 25 (16): 2078-79. https://doi.org/10.1093/bioinformatics/btp352.

Li, Hongye, et Marilyn J. Roossinck. 2004. « Genetic Bottlenecks Reduce Population Variation in an Experimental RNA Virus Population ». *Journal of Virology* 78 (19): 10582-87. https://doi.org/10.1128/JVI.78.19.10582-10587.2004.

Li, Ying, Zhen Liu, Peng Shi, et Jianzhi Zhang. 2010. « The Hearing Gene Prestin Unites Echolocating Bats and Whales ». *Current Biology: CB* 20 (2): R55-56. https://doi.org/10.1016/j.cub.2009.11.042.

Li, Zhanyi, Ying Liu, Ying Zhang, Xiaoqiong Shao, Qiumin Luo, Xiaoyan Guo, Guoli Lin, Qingxian Cai, Zhixin Zhao, et Yutian Chong. 2017. « Naturally Occurring Resistance-Associated Variants to Hepatitis C Virus Direct-Acting Antiviral Agents in Treatment-Naive HCV Genotype 6a-Infected Patients ». *BioMed Research International* 2017. https://doi.org/10.1155/2017/9849823.

Liang, Xiao-Zhen, Bernett T. K. Lee, et Sek-Man Wong. 2002. « Covariation in the Capsid Protein of Hibiscus Chlorotic Ringspot Virus Induced by Serial Passaging in a Host That Restricts Movement Leads to Avirulence in Its Systemic Host ». *Journal of Virology* 76 (23): 12320-24. https://doi.org/10.1128/JVI.76.23.12320-12324.2002.

Lindboe, C. F., S. S. Frøland, K. W. Wefring, P. J. Linnestad, T. Bøhmer, A. Foerster, et A. C. Løken. 1986. « Autopsy Findings in Three Family Members with a Presumably Acquired Immunodeficiency Syndrome of Unknown Etiology ». *Acta Pathologica, Microbiologica, Et Immunologica Scandinavica. Section A, Pathology* 94 (2): 117-23. https://doi.org/10.1111/j.1699-0463.1986.tb02973.x.

Liu, Yang, James A. Cotton, Bin Shen, Xiuqun Han, Stephen J. Rossiter, et Shuyi Zhang. 2010. « Convergent Sequence Evolution between Echolocating Bats and Dolphins ». *Current Biology* 20 (2): R53-54. https://doi.org/10.1016/j.cub.2009.11.058.

Liu, Zhen, Fei-Yan Qi, Dong-Ming Xu, Xin Zhou, et Peng Shi. 2018. « Genomic and Functional Evidence Reveals Molecular Insights into the Origin of Echolocation in Whales ». *Science Advances* 4 (10): eaat8821. https://doi.org/10.1126/sciadv.aat8821.

Liu, Zhen, Fei-Yan Qi, Xin Zhou, Hai-Qing Ren, et Peng Shi. 2014. « Parallel Sites Implicate Functional Convergence of the Hearing Gene Prestin among Echolocating Mammals ». *Molecular Biology and Evolution* 31 (9): 2415-24. https://doi.org/10.1093/molbev/msu194.

Loeffler, Frosch, et P. Frosch. 1898. « Report of the commission for research on foot-and-mouth disease ». *Zent. Bakt. Parasitkde. Abt. I* 23: 371-91.

Loewe, Laurence, et William G. Hill. 2010. « The population genetics of mutations: good, bad and indifferent ». *Philosophical Transactions of the Royal Society B: Biological Sciences* 365 (1544): 1153-67. https://doi.org/10.1098/rstb.2009.0317.

Longdon, Ben, Michael A. Brockhurst, Colin A. Russell, John J. Welch, et Francis M. Jiggins. 2014. « The Evolution and Genetics of Virus Host Shifts ». *PLoS Pathogens* 10 (11): e1004395. https://doi.org/10.1371/journal.ppat.1004395.

Longdon, Ben, Jonathan P. Day, Joel M. Alves, Sophia C. L. Smith, Thomas M. Houslay, John E. McGonigle, Lucia Tagliaferri, et Francis M. Jiggins. 2018. « Host Shifts Result in Parallel Genetic Changes When Viruses Evolve in Closely Related Species ». *PLOS Pathogens* 14 (4): e1006951. https://doi.org/10.1371/journal.ppat.1006951.

Lontok, Erik, Patrick Harrington, Anita Howe, Tara Kieffer, Johan Lennerstrand, Oliver Lenz, Fiona McPhee, et al. 2015. « Hepatitis C Virus Drug Resistance–Associated Substitutions: State of the Art Summary ». *Hepatology* 62 (5): 1623-32. https://doi.org/10.1002/hep.27934.

Losos, Jonathan B. 2011. « Convergence, Adaptation, and Constraint ». *Evolution* 65 (7): 1827-40. https://doi.org/10.1111/j.1558-5646.2011.01289.x.

Louis, John M., G. Marius Clore, et Angela M. Gronenborn. 1999. « Autoprocessing of HIV-1 Protease Is Tightly Coupled to Protein Folding ». *Nature Structural Biology* 6 (9): 868-75. https://doi.org/10.1038/12327.

Lu, Bin, Hong Jin, et Jinzhong Fu. 2020. « Molecular Convergent and Parallel Evolution among Four High-Elevation Anuran Species from the Tibetan Region ». *BMC Genomics* 21 (1): 839. https://doi.org/10.1186/s12864-020-07269-4.

Lucas, Michaela, URS Karrer, ANDREW Lucas, et Paul Klenerman. 2001. « Viral escape mechanisms – escapology taught by viruses ». *International Journal of Experimental Pathology* 82 (5): 269-86. https://doi.org/10.1046/j.1365-2613.2001.00204.x.

Lwoff, A.YR 1957. 1957. « The Concept of Virus ». *Microbiology* 17 (2): 239-53. https://doi.org/10.1099/00221287-17-2-239.

Macdonald, Andrew, et Mark Harris. 2004. « Hepatitis C Virus NS5A: Tales of a Promiscuous Protein ». *The Journal of General Virology* 85 (Pt 9): 2485-2502. https://doi.org/10.1099/vir.0.80204-0.

Maddison, David R., et Wayne P. Maddison. 2000. *Macclade 4: Analysis of Phylogeny and Character Evolution*. Sunderland, MA: Sinauer Associates Inc.

Malinsky, Milan, Richard J. Challis, Alexandra M. Tyers, Stephan Schiffels, Yohey Terai, Benjamin P. Ngatunga, Eric A. Miska, Richard Durbin, Martin J. Genner, et George F. Turner. 2015. « Genomic Islands of Speciation Separate Cichlid Ecomorphs in an East African Crater Lake ». *Science (New York, N.Y.)* 350 (6267): 1493-98. https://doi.org/10.1126/science.aac9927.

Manns, M. P., J. G. McHutchison, S. C. Gordon, V. K. Rustgi, M. Shiffman, R. Reindollar, Z. D. Goodman, K. Koury, M. Ling, et J. K. Albrecht. 2001. « Peginterferon Alfa-2b plus Ribavirin Compared with Interferon Alfa-2b plus Ribavirin for Initial Treatment of Chronic Hepatitis C: A Randomised Trial ». *Lancet (London, England)* 358 (9286): 958-65. https://doi.org/10.1016/s0140-6736(01)06102-5.

Marcelin, Anne-Geneviève, Philippe Flandre, Juliette Pavie, Nathalie Schmidely, Marc Wirden, Olivier Lada, Dan Chiche, Jean-Michel Molina, et Vincent Calvez. 2005. « Clinically Relevant Genotype Interpretation of Resistance to Didanosine ». *Antimicrobial Agents and Chemotherapy* 49 (5): 1739-44. https://doi.org/10.1128/AAC.49.5.1739-1744.2005.

Marcellin, P. 1999. « Hepatitis C: The Clinical Spectrum of the Disease ». *Journal of Hepatology* 31 Suppl 1: 9-16. https://doi.org/10.1016/s0168-8278(99)80368-7.

Marcovitz, Amir, Yatish Turakhia, Heidi I. Chen, Michael Gloudemans, Benjamin A. Braun, Haoqing Wang, et Gill Bejerano. 2019. « A Functional Enrichment Test for Molecular Convergent Evolution Finds a Clear Protein-Coding Signal in Echolocating Bats and Whales ». *Proceedings of the National Academy of Sciences of the United States of America* 116 (42): 21094-103. https://doi.org/10.1073/pnas.1818532116.

Markov, Peter V., Jacques Pepin, Eric Frost, Sylvie Deslandes, Annie-Claude Labbé, et Oliver G. Pybus. 2009. « Phylogeography and Molecular Epidemiology of Hepatitis C Virus Genotype 2 in Africa ». *The Journal of General Virology* 90 (Pt 9): 2086-96. https://doi.org/10.1099/vir.0.011569-0.

Martin, Darren P., Ben Murrell, Michael Golden, Arjun Khoosal, et Brejnev Muhire. 2015. « RDP4: Detection and analysis of recombination patterns in virus genomes ». *Virus Evolution* 1 (1). https://doi.org/10.1093/ve/vev003.

Martin, Darren P, Steven Weaver, Houriiyah Tegally, James Emmanuel San, Stephen D. Shank, Eduan Wilkinson, Alexander G. Lucaci, et al. 2021. « The emergence and ongoing convergent evolution of the SARS-CoV-2 N501Y lineages ». *Cell* 184 (20): 5189-5200.e7. https://doi.org/10.1016/j.cell.2021.09.003.

Martin, Darren P, Steven Weaver, Houryiah Tegally, Emmanuel James San, Stephen D Shank, Eduan Wilkinson, Jennifer Giandhari, et al. 2021. « The Emergence and Ongoing Convergent Evolution of the N501Y Lineages Coincides with a Major Global Shift in the SARS-CoV-2 Selective Landscape ». Preprint. Infectious Diseases (except HIV/AIDS). https://doi.org/10.1101/2021.02.23.21252268.

Marx, P A, P G Alcabes, et E Drucker. 2001. « Serial human passage of simian immunodeficiency virus by unsterile injections and the emergence of epidemic human immunodeficiency virus in Africa. » *Philosophical Transactions of the Royal Society of London. Series B* 356 (1410): 911-20. https://doi.org/10.1098/rstb.2001.0867.

Matranga, Christian B., Kristian G. Andersen, Sarah Winnicki, Michele Busby, Adrianne D. Gladden, Ryan Tewhey, Matthew Stremlau, et al. 2014. « Enhanced Methods for Unbiased Deep Sequencing of Lassa and Ebola RNA Viruses from Clinical and Biological Samples ». *Genome Biology* 15 (11): 519. https://doi.org/10.1186/PREACCEPT-1698056557139770.

McCutchan, Francine E. 2006. « Global Epidemiology of HIV ». *Journal of Medical Virology* 78 (S1): S7-12. https://doi.org/10.1002/jmv.20599.

McKinney, Wes. 2010. « Data Structures for Statistical Computing in Python ». In , 56-61. Austin, Texas. https://doi.org/10.25080/Majora-92bf1922-00a.

McPhee, Fiona, Joseph Ueland, Vincent Vellucci, Scott Bowden, William Sievert, et Nannan Zhou. 2019. « Impact of Preexisting Hepatitis C Virus Genotype 6 NS3, NS5A, and NS5B Polymorphisms on the In Vitro Potency of Direct-Acting Antiviral Agents ». *Antimicrobial Agents and Chemotherapy* 63 (4). https://doi.org/10.1128/AAC.02205-18.

Meyer, Justin R., Devin T. Dobias, Joshua S. Weitz, Jeffrey E. Barrick, Ryan T. Quick, et Richard E. Lenski. 2012. « Repeatability and Contingency in the Evolution of a Key Innovation in Phage Lambda ». *Science (New York, N.y.)* 335 (6067): 428-32. https://doi.org/10.1126/science.1214449.

Meyer, Peter R, Suzanne E Matsuura, A. Mohsin Mian, Antero G So, et Walter A Scott. 1999. « A Mechanism of AZT Resistance: An Increase in Nucleotide-Dependent Primer Unblocking

by Mutant HIV-1 Reverse Transcriptase ». *Molecular Cell* 4 (1): 35-43. https://doi.org/10.1016/S1097-2765(00)80185-9.

Micallef, J. M., J. M. Kaldor, et G. J. Dore. 2006. « Spontaneous Viral Clearance Following Acute Hepatitis C Infection: A Systematic Review of Longitudinal Studies ». *Journal of Viral Hepatitis* 13 (1): 34-41. https://doi.org/10.1111/j.1365-2893.2005.00651.x.

Miller, Craig R., Anna C. Nagel, LuAnn Scott, Matt Settles, Paul Joyce, et Holly A. Wichman. 2016. « Love the One You're with: Replicate Viral Adaptations Converge on the Same Phenotypic Change ». *PeerJ* 4: e2227. https://doi.org/10.7717/peerj.2227.

Minskaia, Ekaterina, Tobias Hertzig, Alexander E. Gorbalenya, Valérie Campanacci, Christian Cambillau, Bruno Canard, et John Ziebuhr. 2006. « Discovery of an RNA Virus 3'->5' Exoribonuclease That Is Critically Involved in Coronavirus RNA Synthesis ». *Proceedings of the National Academy of Sciences of the United States of America* 103 (13): 5108-13. https://doi.org/10.1073/pnas.0508200103.

Montagutelli, X, Frédéric Lemoine, F Donati, F Touret, Bourret J, Prot M, Munier S, et al. 2022. « Rapid Characterization of a Delta-Omicron SARS-CoV-2 Recombinant Detected in Europe », avril. https://doi.org/10.21203/rs.3.rs-1502293/v1.

Moore, Corey B., Mina John, Ian R. James, Frank T. Christiansen, Campbell S. Witt, et Simon A. Mallal. 2002. « Evidence of HIV-1 Adaptation to HLA-Restricted Immune Responses at a Population Level ». *Science (New York, N.Y.)* 296 (5572): 1439-43. https://doi.org/10.1126/science.1069660.

Moradpour, Darius, et François Penin. 2013. « Hepatitis C Virus Proteins: From Structure to Function ». In *Hepatitis C Virus: From Molecular Virology to Antiviral Therapy*, édité par Ralf Bartenschlager, 113-42. Current Topics in Microbiology and Immunology. Berlin, Heidelberg: Springer. https://doi.org/10.1007/978-3-642-27340-7_5.

Morel, Benoit, Pierre Barbera, Lucas Czech, Ben Bettisworth, Lukas Hübner, Sarah Lutteropp, Dora Serdari, et al. 2021. « Phylogenetic Analysis of SARS-CoV-2 Data Is Difficult ». *Molecular Biology and Evolution* 38 (5): 1777-91. https://doi.org/10.1093/molbev/msaa314.

Morel, Marie, Frédéric Lemoine, et Olivier Gascuel. 2021. « Sensitive Detection of Site-Wise Convergent Evolution in Large Protein Alignments with ConDor ». bioRxiv. https://doi.org/10.1101/2021.06.30.450558.

Murrell, Ben, Sasha Moola, Amandla Mabona, Thomas Weighill, Daniel Sheward, Kosakovsky Pond, Sergei L, et Konrad Scheffler. 2013. « FUBAR: A Fast, Unconstrained Bayesian AppRoximation for Inferring Selection ». *Molecular Biology and Evolution* 30 (5): 1196-1205. https://doi.org/10.1093/molbev/mst030.

Murrell, Ben, Tulio de Oliveira, Chris Seebregts, Sergei L. Kosakovsky Pond, Konrad Scheffler, et on behalf of the Southern African Treatment and Resistance Network (SATuRN) Consortium. 2012. « Modeling HIV-1 Drug Resistance as Episodic Directional Selection ». *PLOS Computational Biology* 8 (5): e1002507. https://doi.org/10.1371/journal.pcbi.1002507.

Murrell, Ben, Joel O. Wertheim, Sasha Moola, Thomas Weighill, Konrad Scheffler, et Sergei L. Kosakovsky Pond. 2012. « Detecting Individual Sites Subject to Episodic Diversifying Selection ». *PLOS Genetics* 8 (7): e1002764. https://doi.org/10.1371/journal.pgen.1002764.

Muschick, Moritz, Adrian Indermaur, et Walter Salzburger. 2012. « Convergent Evolution within an Adaptive Radiation of Cichlid Fishes ». *Current Biology* 22 (24): 2362-68. https://doi.org/10.1016/j.cub.2012.10.048.

Muse, S. V., et B. S. Gaut. 1994. « A Likelihood Approach for Comparing Synonymous and Nonsynonymous Nucleotide Substitution Rates, with Application to the Chloroplast Genome ». *Molecular Biology and Evolution* 11 (5): 715-24. https://doi.org/10.1093/oxfordjournals.molbev.a040152.

Mushegian, A. R. 2020. « Are There 1031 Virus Particles on Earth, or More, or Fewer? » *Journal of Bacteriology* 202 (9): e00052-20. https://doi.org/10.1128/JB.00052-20.

Nagler, F. P. O., et Geoffrey Rake. 1948. « The Use of the Electron Microscope in Diagnosis of Variola, Vaccinia, and Varicella ». *Journal of Bacteriology* 55 (1): 45-51.

Narahari, Shobha, Abida Juwle, Subhankar Basak, et Dhananjaya Saranath. 2009. « Prevalence and Geographic Distribution of Hepatitis C Virus Genotypes in Indian Patient Cohort ». *Infection, Genetics and Evolution: Journal of Molecular Epidemiology and Evolutionary Genetics in Infectious Diseases* 9 (4): 643-45. https://doi.org/10.1016/j.meegid.2009.04.001.

Needleman, S. B., et C. D. Wunsch. 1970. « A General Method Applicable to the Search for Similarities in the Amino Acid Sequence of Two Proteins ». *Journal of Molecular Biology* 48 (3): 443-53. https://doi.org/10.1016/0022-2836(70)90057-4.

Nei, Masatoshi. 2005. « Selectionism and Neutralism in Molecular Evolution ». *Molecular Biology and Evolution* 22 (12): 2318-42. https://doi.org/10.1093/molbev/msi242.

Nelemans, Tessa, et Marjolein Kikkert. 2019. « Viral Innate Immune Evasion and the Pathogenesis of Emerging RNA Virus Infections ». *Viruses* 11 (10): 961. https://doi.org/10.3390/v11100961.

Nguyen, Lam-Tung, Heiko A. Schmidt, Arndt von Haeseler, et Bui Quang Minh. 2015. « IQ-TREE: A Fast and Effective Stochastic Algorithm for Estimating Maximum-Likelihood Phylogenies ». *Molecular Biology and Evolution* 32 (1): 268-74. https://doi.org/10.1093/molbev/msu300.

Nguyen, Thuy, Sepideh Akhavan, Fabienne Caby, Luminita Bonyhay, Lucile Larrouy, Anne Gervais, Pascal Lebray, et al. 2019. « Net Emergence of Substitutions at Position 28 in NS5A of Hepatitis C Virus Genotype 4 in Patients Failing Direct-Acting Antivirals Detected by next-Generation Sequencing ». *International Journal of Antimicrobial Agents* 53 (1): 80-83. https://doi.org/10.1016/j.ijantimicag.2018.09.010.

Nickle, David C., Laura Heath, Mark A. Jensen, Peter B. Gilbert, James I. Mullins, et Sergei L. Kosakovsky Pond. 2007. « HIV-Specific Probabilistic Models of Protein Evolution ». *PLOS ONE* 2 (6): e503. https://doi.org/10.1371/journal.pone.0000503.

Nielsen, Rasmus, et Ziheng Yang. 1998. « Likelihood Models for Detecting Positively Selected Amino Acid Sites and Applications to the HIV-1 Envelope Gene ». *Genetics* 148 (3): 929-36. https://doi.org/10.1093/genetics/148.3.929.

Nijhuis, Monique, Steven Deeks, et Charles Boucher. 2001. « Implications of Antiretroviral Resistance on Viral Fitness ». *Current Opinion in Infectious Diseases* 14 (1): 23-28.

Nouhin, Janin, Momoko Iwamoto, Sophearot Prak, Jean-Philippe Dousset, Kerya Phon, Seiha Heng, Alexandra Kerleguer, et al. 2019. « Molecular Epidemiology of Hepatitis C Virus in Cambodia during 2016–2017 ». *Scientific Reports* 9 (1): 7314. https://doi.org/10.1038/s41598-019-43785-4.

Novella, Isabel S, Esteban Domingo, et John J Holland. 1995. « Quasispecies Evolution'. Implications for Vaccine and Drug Strategies », 7.

Nozawa, Masafumi, Yoshiyuki Suzuki, et Masatoshi Nei. 2009. « Reliabilities of identifying positive selection by the branch-site and the site-prediction methods ». *Proceedings of the National Academy of Sciences* 106 (16): 6700-6705. https://doi.org/10.1073/pnas.0901855106.

Nurk, Sergey, Dmitry Meleshko, Anton Korobeynikov, et Pavel A. Pevzner. 2017. « metaSPAdes: a new versatile metagenomic assembler ». *Genome Research* 27 (5): 824-34. https://doi.org/10.1101/gr.213959.116.

O'Reilly, Joseph E., Mark N. Puttick, Luke Parry, Alastair R. Tanner, James E. Tarver, James Fleming, Davide Pisani, et Philip C. J. Donoghue. 2016. « Bayesian Methods Outperform Parsimony but at the Expense of Precision in the Estimation of Phylogeny from Discrete

Morphological Data ». *Biology Letters* 12 (4): 20160081. https://doi.org/10.1098/rsbl.2016.0081.

Organisation mondiale de la Santé. 2016. « Stratégie mondiale du secteur de la santé contre le VIH 2016-2021: vers l'élimination du SIDA ». WHO/HIV/2016.05. Organisation mondiale de la Santé. https://apps.who.int/iris/handle/10665/250576.

Orland, Jennie R., Teresa L. Wright, et Stewart Cooper. 2001. « Acute Hepatitis C ». *Hepatology* 33 (2): 321-27. https://doi.org/10.1053/jhep.2001.22112.

Pagel, Mark. 1994. « Detecting correlated evolution on phylogenies: a general method for the comparative analysis of discrete characters ». *Proceedings of the Royal Society of London. Series B: Biological Sciences* 255 (1342): 37-45. https://doi.org/10.1098/rspb.1994.0006.

Pagel, Mark, et Andrew Meade. 2006. « Bayesian Analysis of Correlated Evolution of Discrete Characters by Reversible-Jump Markov Chain Monte Carlo. » *The American Naturalist* 167 (6): 808-25. https://doi.org/10.1086/503444.

Palella, Frank J., Kathleen M. Delaney, Anne C. Moorman, Mark O. Loveless, Jack Fuhrer, Glen A. Satten, Diane J. Aschman, et Scott D. Holmberg. 1998. « Declining Morbidity and Mortality among Patients with Advanced Human Immunodeficiency Virus Infection ». *New England Journal of Medicine* 338 (13): 853-60. https://doi.org/10.1056/NEJM199803263381301.

Paolucci, Stefania, Marta Premoli, Stefano Novati, Roberto Gulminetti, Renato Maserati, Giorgio Barbarini, Paolo Sacchi, et al. 2017. « Baseline and Breakthrough Resistance Mutations in HCV Patients Failing DAAs ». *Scientific Reports* 7 (1): 16017. https://doi.org/10.1038/s41598-017-15987-1.

Paradis, Emmanuel, Julien Claude, et Korbinian Strimmer. 2004. « APE: Analyses of Phylogenetics and Evolution in R language ». *Bioinformatics* 20 (2): 289-90. https://doi.org/10.1093/bioinformatics/btg412.

Parker, Joe, Georgia Tsagkogeorga, James A. Cotton, Yuan Liu, Paolo Provero, Elia Stupka, et Stephen J. Rossiter. 2013. « Genome-Wide Signatures of Convergent Evolution in Echolocating Mammals ». *Nature* 502 (7470): 228-31. https://doi.org/10.1038/nature12511.

Parto, Sahar, et Nicolas Lartillot. 2017. « Detecting Consistent Patterns of Directional Adaptation Using Differential Selection Codon Models ». *BMC Evolutionary Biology* 17 (1). https://doi.org/10.1186/s12862-017-0979-y.

———. 2018. « Molecular Adaptation in Rubisco: Discriminating between Convergent Evolution and Positive Selection Using Mechanistic and Classical Codon Models ». *PLOS ONE* 13 (2): e0192697. https://doi.org/10.1371/journal.pone.0192697.

Parvez, Mohammad K., et Shama Parveen. 2017. « Evolution and Emergence of Pathogenic Viruses: Past, Present, and Future ». *Intervirology* 60 (1-2): 1-7. https://doi.org/10.1159/000478729.

Pawlotsky, Jean-Michel. 2016. « Hepatitis C Virus Resistance to Direct-Acting Antiviral Drugs in Interferon-Free Regimens ». *Gastroenterology* 151 (1): 70-86. https://doi.org/10.1053/j.gastro.2016.04.003.

Pawlotsky, Jean-Michel, Francesco Negro, Alessio Aghemo, Marina Berenguer, Olav Dalgard, Geoffrey Dusheiko, Fiona Marra, Massimo Puoti, et Heiner Wedemeyer. 2018. « EASL Recommendations on Treatment of Hepatitis C 2018 ». *Journal of Hepatology* 69 (2): 461-511. https://doi.org/10.1016/j.jhep.2018.03.026.

Payne, Alexander, Nadine Holmes, Vardhman Rakyan, et Matthew Loose. 2019. « BulkVis: a graphical viewer for Oxford nanopore bulk FAST5 files ». *Bioinformatics* 35 (13): 2193-98. https://doi.org/10.1093/bioinformatics/bty841.

Peacock, Thomas P., Rebekah Penrice-Randal, Julian A. Hiscox, et Wendy S. Barclay. 2021. « SARS-CoV-2 one year on: evidence for ongoing viral adaptation ». *The Journal of General Virology* 102 (4): 001584. https://doi.org/10.1099/jgv.0.001584.

Pennings, Pleuni S. 2013. « HIV Drug Resistance: Problems and Perspectives ». *Infectious Disease Reports* 5 (Suppl 1): e5. https://doi.org/10.4081/idr.2013.s1.e5.

Pennings, Pleuni S., Sergey Kryazhimskiy, et John Wakeley. 2014. « Loss and Recovery of Genetic Diversity in Adapting Populations of HIV ». *PLOS Genetics* 10 (1): e1004000. https://doi.org/10.1371/journal.pgen.1004000.

Pépin, Jacques. 2021. *The Origins of AIDS*. Cambridge University Press.

Perales, Celia, Qian Chen, Maria Eugenia Soria, Josep Gregori, Damir Garcia-Cehic, Leonardo Nieto-Aponte, Lluis Castells, et al. 2018. « Baseline hepatitis C virus resistance-associated substitutions present at frequencies lower than 15% may be clinically significant ». *Infection and Drug Resistance* 11 (novembre): 2207-10. https://doi.org/10.2147/IDR.S172226.

Perelson, A. S., A. U. Neumann, M. Markowitz, J. M. Leonard, et D. D. Ho. 1996. « HIV-1 Dynamics in Vivo: Virion Clearance Rate, Infected Cell Life-Span, and Viral Generation Time ». *Science (New York, N.Y.)* 271 (5255): 1582-86. https://doi.org/10.1126/science.271.5255.1582.

Pérez-Losada, Marcos, Miguel Arenas, Juan Carlos Galán, Ferran Palero, et Fernando González-Candelas. 2015. « Recombination in viruses: Mechanisms, methods of study, and evolutionary consequences ». *Infection, Genetics and Evolution* 30 (mars): 296-307. https://doi.org/10.1016/j.meegid.2014.12.022.

Perrière, G., et Céline Brochier-Armanet. 2010. *Concepts et méthodes en phylogénie moléculaire*. Édité par Springer. https://hal.archives-ouvertes.fr/hal-02091278.

Petruzziello, Arnolfo, Samantha Marigliano, Giovanna Loquercio, Anna Cozzolino, et Carmela Cacciapuoti. 2016. « Global epidemiology of hepatitis C virus infection: An up-date of the distribution and circulation of hepatitis C virus genotypes ». *World Journal of Gastroenterology* 22 (34): 7824-40. https://doi.org/10.3748/wjg.v22.i34.7824.

Peyerl, Fred W., Heidi S. Bazick, Michael H. Newberg, Dan H. Barouch, Joseph Sodroski, et Norman L. Letvin. 2004. « Fitness Costs Limit Viral Escape from Cytotoxic T Lymphocytes at a Structurally Constrained Epitope ». *Journal of Virology* 78 (24): 13901-10. https://doi.org/10.1128/JVI.78.24.13901-13910.2004.

Pham, Long V., Santseharay Ramirez, Judith M. Gottwein, Ulrik Fahnøe, Yi-Ping Li, Jannie Pedersen, et Jens Bukh. 2018. « HCV Genotype 6a Escape From and Resistance to Velpatasvir, Pibrentasvir, and Sofosbuvir in Robust Infectious Cell Culture Models ». *Gastroenterology* 154 (8): 2194-2208.e12. https://doi.org/10.1053/j.gastro.2018.02.017.

Philippe, Nadège, Matthieu Legendre, Gabriel Doutre, Yohann Couté, Olivier Poirot, Magali Lescot, Defne Arslan, et al. 2013. « Pandoraviruses: Amoeba Viruses with Genomes up to 2.5 Mb Reaching That of Parasitic Eukaryotes ». *Science (New York, N.Y.)* 341 (6143): 281-86. https://doi.org/10.1126/science.1239181.

Pickett, Brett E., Eva L. Sadat, Yun Zhang, Jyothi M. Noronha, R. Burke Squires, Victoria Hunt, Mengya Liu, et al. 2012. « ViPR: an open bioinformatics database and analysis resource for virology research ». *Nucleic Acids Research* 40 (Database issue): D593-98. https://doi.org/10.1093/nar/gkr859.

Pinsky, Benjamin A., Samantha Mix, Judy Rowe, Sheryl Ikemoto, et Ellen J. Baron. 2010. « Long-Term Shedding of Influenza A Virus in Stool of Immunocompromised Child ». *Emerging Infectious Diseases* 16 (7): 1165-67. https://doi.org/10.3201/eid1607.091248.

Pizzorno, Andrés, Yacine Abed, et Guy Boivin. 2011. « Influenza Drug Resistance ». *Seminars in Respiratory and Critical Care Medicine* 32 (4): 409-22. https://doi.org/10.1055/s-0031-1283281.

Ploegh, Hidde L. 1998. « Viral Strategies of Immune Evasion ». *Science* 280 (5361): 248-53. https://doi.org/10.1126/science.280.5361.248.

Pond, Kosakovsky, Sergei L, et Simon D. W. Frost. 2005. « Not So Different After All: A Comparison of Methods for Detecting Amino Acid Sites Under Selection ». *Molecular Biology and Evolution* 22 (5): 1208-22. https://doi.org/10.1093/molbev/msi105.

Pond, Sergei Kosakovsky, et Spencer V. Muse. 2005. « Site-to-Site Variation of Synonymous Substitution Rates ». *Molecular Biology and Evolution* 22 (12): 2375-85. https://doi.org/10.1093/molbev/msi232.

Pond, Sergei L. Kosakovsky, et Simon D. W. Frost. 2005. « Datamonkey: Rapid Detection of Selective Pressure on Individual Sites of Codon Alignments ». *Bioinformatics (Oxford, England)* 21 (10): 2531-33. https://doi.org/10.1093/bioinformatics/bti320.

Pond, Sergei L. Kosakovsky, Simon D. W. Frost, Zehava Grossman, Michael B. Gravenor, Douglas D. Richman, et Andrew J. Leigh Brown. 2006. « Adaptation to Different Human Populations by HIV-1 Revealed by Codon-Based Analyses ». *PLOS Computational Biology* 2 (6): e62. https://doi.org/10.1371/journal.pcbi.0020062.

Pond, Sergei L. Kosakovsky, Simon D. W. Frost, et Spencer V. Muse. 2005. « HyPhy: hypothesis testing using phylogenies ». *Bioinformatics* 21 (5): 676-79. https://doi.org/10.1093/bioinformatics/bti079.

Pond, Sergei L. Kosakovsky, Ben Murrell, et Art F. Y. Poon. 2012. « Evolution of Viral Genomes: Interplay Between Selection, Recombination, and Other Forces ». *Evolutionary Genomics*, 239-72. https://doi.org/10.1007/978-1-61779-585-5_10.

Popovic, Mikulas, M. G. Sarngadharan, Elizabeth Read, et Robert C. Gallo. 1984. « Detection, Isolation, and Continuous Production of Cytopathic Retroviruses (HTLV-III) from Patients with AIDS and Pre-AIDS ». *Science*, mai. https://www.science.org/doi/abs/10.1126/science.6200935.

Projecto-Garcia, Joana, Chandrasekhar Natarajan, Hideaki Moriyama, Roy E. Weber, Angela Fago, Zachary A. Cheviron, Robert Dudley, Jimmy A. McGuire, Christopher C. Witt, et Jay F. Storz. 2013. « Repeated elevational transitions in hemoglobin function during the evolution of Andean hummingbirds ». *Proceedings of the National Academy of Sciences* 110 (51): 20669-74. https://doi.org/10.1073/pnas.1315456110.

Pupko, Tal, Itsik Pe, Ron Shamir, et Dan Graur. 2000. « A Fast Algorithm for Joint Reconstruction of Ancestral Amino Acid Sequences ». *Molecular Biology and Evolution* 17 (6): 890-96. https://doi.org/10.1093/oxfordjournals.molbev.a026369.

Pybus, Oliver G., Eleanor Barnes, Rachel Taggart, Philippe Lemey, Peter V. Markov, Bouachan Rasachak, Bounkong Syhavong, et al. 2009. « Genetic History of Hepatitis C Virus in East Asia ». *Journal of Virology* 83 (2): 1071-82. https://doi.org/10.1128/JVI.01501-08.

Pybus, Oliver G., et Julien Thézé. 2016. « Hepacivirus Cross-Species Transmission and the Origins of the Hepatitis C Virus ». *Current Opinion in Virology* 16 (février): 1-7. https://doi.org/10.1016/j.coviro.2015.10.002.

Quan, Phenix-Lan, Cadhla Firth, Juliette M. Conte, Simon H. Williams, Carlos M. Zambrana-Torrelio, Simon J. Anthony, James A. Ellison, et al. 2013. « Bats are a major natural reservoir for hepaciviruses and pegiviruses ». *Proceedings of the National Academy of Sciences* 110 (20): 8194-99. https://doi.org/10.1073/pnas.1303037110.

Quick, Joshua, Nicholas J. Loman, Sophie Duraffour, Jared T. Simpson, Ettore Severi, Lauren Cowley, Joseph Akoi Bore, et al. 2016. « Real-Time, Portable Genome Sequencing for Ebola Surveillance ». *Nature* 530 (7589): 228-32. https://doi.org/10.1038/nature16996.

Quiñones-Mateu, Miguel E, et Eric J Arts. 2002. « Fitness of Drug Resistant HIV-1: Methodology and Clinical Implications ». *Drug Resistance Updates* 5 (6): 224-33. https://doi.org/10.1016/S1368-7646(02)00123-1.

Quniñones-Mateu, Miguel E., Santiago Avila, Gustavo REYES-TERAN, et Miguel A. MARTINEZ. 2014. « Deep Sequencing: Becoming a Critical Tool in Clinical Virology ». *Journal of*

clinical virology : the official publication of the Pan American Society for Clinical Virology 61 (1): 9-19. https://doi.org/10.1016/j.jcv.2014.06.013.

Rambaut, Andrew, Edward C. Holmes, Áine O'Toole, Verity Hill, John T. McCrone, Christopher Ruis, Louis du Plessis, et Oliver G. Pybus. 2020. « A Dynamic Nomenclature Proposal for SARS-CoV-2 Lineages to Assist Genomic Epidemiology ». *Nature Microbiology* 5 (11): 1403-7. https://doi.org/10.1038/s41564-020-0770-5.

Raoult, Didier, Stéphane Audic, Catherine Robert, Chantal Abergel, Patricia Renesto, Hiroyuki Ogata, Bernard La Scola, Marie Suzan, et Jean-Michel Claverie. 2004. « The 1.2-Megabase Genome Sequence of Mimivirus ». *Science* 306 (5700): 1344-50. https://doi.org/10.1126/science.1101485.

Raoult, Didier, et Patrick Forterre. 2008. « Redefining Viruses: Lessons from Mimivirus ». *Nature Reviews. Microbiology* 6 (4): 315-19. https://doi.org/10.1038/nrmicro1858.

Ratner, Lee, William Haseltine, Roberto Patarca, Kenneth J. Livak, Bruno Starcich, Steven F. Josephs, Ellen R. Doran, et al. 1985. « Complete Nucleotide Sequence of the AIDS Virus, HTLV-III ». *Nature* 313 (6000): 277-84. https://doi.org/10.1038/313277a0.

Ray, Stuart C., Liam Fanning, Xiao-Hong Wang, Dale M. Netski, Elizabeth Kenny-Walsh, et David L. Thomas. 2005. « Divergent and Convergent Evolution after a Common-Source Outbreak of Hepatitis C Virus ». *The Journal of Experimental Medicine* 201 (11): 1753-59. https://doi.org/10.1084/jem.20050122.

Reagan, R. L., et A. L. Brueckner. 1952. « Morphological Observations by Electron Microscopy of the Lansing Strain of Poliomyelitis Virus after Propagation in the Swiss Albino Mouse ». *Texas Reports on Biology and Medicine* 10 (2): 425-28.

Remold, Susanna K., Andrew Rambaut, et Paul E. Turner. 2008. « Evolutionary Genomics of Host Adaptation in Vesicular Stomatitis Virus ». *Molecular Biology and Evolution* 25 (6): 1138-47. https://doi.org/10.1093/molbev/msn059.

Rey, Carine, Laurent Guéguen, Marie Sémon, et Bastien Boussau. 2018. « Accurate Detection of Convergent Amino-Acid Evolution with PCOC ». *Molecular Biology and Evolution* 35 (9): 2296-2306. https://doi.org/10.1093/molbev/msy114.

Rey, Carine, Philippe Veber, Guéguen Laurent, Lartillot, Nicolas, Sémon, Marie, et Boussau, Bastien. 2019. « Detecting adaptive convergent amino acid evolution ». *Philosophical Transactions of the Royal Society B: Biological Sciences* 374 (1777): 20180234. https://doi.org/10.1098/rstb.2018.0234.

Rhee, S.-Y. 2003. « Human Immunodeficiency Virus Reverse Transcriptase and Protease Sequence Database ». *Nucleic Acids Research* 31 (1): 298-303. https://doi.org/10.1093/nar/gkg100.

Rice, Alan M, Atahualpa Castillo Morales, Alexander T Ho, Christine Mordstein, Stefanie Mühlhausen, Samir Watson, Laura Cano, Bethan Young, Grzegorz Kudla, et Laurence D Hurst. 2021. « Evidence for Strong Mutation Bias toward, and Selection against, U Content in SARS-CoV-2: Implications for Vaccine Design ». *Molecular Biology and Evolution* 38 (1): 67-83. https://doi.org/10.1093/molbev/msaa188.

Robertson, D. L., J. P. Anderson, J. A. Bradac, J. K. Carr, B. Foley, R. K. Funkhouser, F. Gao, et al. 2000. « HIV-1 Nomenclature Proposal ». *Science* 288 (5463): 55-55. https://doi.org/10.1126/science.288.5463.55d.

Robinson, James T., Helga Thorvaldsdóttir, Wendy Winckler, Mitchell Guttman, Eric S. Lander, Gad Getz, et Jill P. Mesirov. 2011. « Integrative Genomics Viewer ». *Nature Biotechnology* 29 (1): 24-26. https://doi.org/10.1038/nbt.1754.

Robson, Fran, Khadija Shahed Khan, Thi Khanh Le, Clément Paris, Sinem Demirbag, Peter Barfuss, Palma Rocchi, et Wai-Lung Ng. 2020. « Coronavirus RNA Proofreading: Molecular Basis and Therapeutic Targeting ». *Molecular Cell* 79 (5): 710-27. https://doi.org/10.1016/j.molcel.2020.07.027.

Rodger, Alison J., Valentina Cambiano, Tina Bruun, Pietro Vernazza, Simon Collins, Olaf Degen, Giulio Maria Corbelli, et al. 2019. « Risk of HIV Transmission through Condomless Sex in Serodifferent Gay Couples with the HIV-Positive Partner Taking Suppressive Antiretroviral Therapy (PARTNER): Final Results of a Multicentre, Prospective, Observational Study ». *The Lancet* 393 (10189): 2428-38. https://doi.org/10.1016/S0140-6736(19)30418-0.

Rokas, Antonis, et Sean B. Carroll. 2008. « Frequent and Widespread Parallel Evolution of Protein Sequences ». *Molecular Biology and Evolution* 25 (9): 1943-53. https://doi.org/10.1093/molbev/msn143.

Ronquist, Fredrik. 2004. « Bayesian Inference of Character Evolution ». *Trends in Ecology & Evolution* 19 (9): 475-81. https://doi.org/10.1016/j.tree.2004.07.002.

Rosenblum, Erica Bree, Christine E. Parent, et Erin E. Brandt. 2014. « The Molecular Basis of Phenotypic Convergence ». *Annual Review of Ecology, Evolution, and Systematics* 45 (1): 203-26. https://doi.org/10.1146/annurev-ecolsys-120213-091851.

Sackman, Andrew M., Lindsey W. McGee, Anneliese J. Morrison, Jessica Pierce, Jeremy Anisman, Hunter Hamilton, Stephanie Sanderbeck, Cayla Newman, et Darin R. Rokyta. 2017. « Mutation-Driven Parallel Evolution during Viral Adaptation ». *Molecular Biology and Evolution* 34 (12): 3243-53. https://doi.org/10.1093/molbev/msx257.

Sahlin, Kristoffer, et Paul Medvedev. 2021. « Error Correction Enables Use of Oxford Nanopore Technology for Reference-Free Transcriptome Analysis ». *Nature Communications* 12 (1): 2. https://doi.org/10.1038/s41467-020-20340-8.

Saitou, N., et M. Nei. 1987. « The Neighbor-Joining Method: A New Method for Reconstructing Phylogenetic Trees. » *Molecular Biology and Evolution* 4 (4): 406-25. https://doi.org/10.1093/oxfordjournals.molbev.a040454.

Salimi Alizei, Elahe, Maike Hofmann, Robert Thimme, et Christoph Neumann-Haefelin. 2021. « Mutational Escape from Cellular Immunity in Viral Hepatitis: Variations on a Theme ». *Current Opinion in Virology* 50 (octobre): 110-18. https://doi.org/10.1016/j.coviro.2021.08.002.

Sandaa, Ruth-Anne, Mikal Heldal, Tonje Castberg, Runar Thyrhaug, et Gunnar Bratbak. 2001. « Isolation and Characterization of Two Viruses with Large Genome Size Infecting Chrysochromulina Ericina (Prymnesiophyceae) and Pyramimonas Orientalis (Prasinophyceae) ». *Virology* 290 (2): 272-80. https://doi.org/10.1006/viro.2001.1161.

Sanger, F., S. Nicklen, et A. R. Coulson. 1977. « DNA sequencing with chain-terminating inhibitors ». *Proceedings of the National Academy of Sciences of the United States of America* 74 (12): 5463-67.

Sanjuán, Rafael, Miguel R. Nebot, Nicola Chirico, Louis M. Mansky, et Robert Belshaw. 2010. « Viral Mutation Rates ». *Journal of Virology* 84 (19): 9733-48. https://doi.org/10.1128/JVI.00694-10.

Sarrazin, Christoph. 2021. « Treatment Failure with DAA Therapy: Importance of Resistance ». *Journal of Hepatology* 74 (6): 1472-82. https://doi.org/10.1016/j.jhep.2021.03.004.

Sarrazin, Christoph, Hadas Dvory-Sobol, Evguenia S. Svarovskaia, Brian P. Doehle, Phillip S. Pang, Shu-Min Chuang, Julie Ma, et al. 2016. « Prevalence of Resistance-Associated Substitutions in HCV NS5A, NS5B, or NS3 and Outcomes of Treatment With Ledipasvir and Sofosbuvir ». *Gastroenterology* 151 (3): 501-512.e1. https://doi.org/10.1053/j.gastro.2016.06.002.

Sauter, Daniel, Michael Schindler, Anke Specht, Wilmina N. Landford, Jan Münch, Kyeong-Ae Kim, Jörg Votteler, et al. 2009. « Tetherin-Driven Adaptation of Vpu and Nef Function and the Evolution of Pandemic and Nonpandemic HIV-1 Strains ». *Cell Host & Microbe* 6 (5): 409-21. https://doi.org/10.1016/j.chom.2009.10.004.

Schultz, Anne-Kathrin, Ingo Bulla, Mariama Abdou-Chekaraou, Emmanuel Gordien, Burkhard Morgenstern, Fabien Zoulim, Paul Dény, et Mario Stanke. 2012. « jpHMM:

recombination analysis in viruses with circular genomes such as the hepatitis B virus ». *Nucleic Acids Research* 40 (W1): W193-98. https://doi.org/10.1093/nar/gks414.

Shenge, Juliet A., Georgina N. Odaibo, et David O. Olaleye. 2019. « Phylogenetic analysis of hepatitis C virus among HIV/ HCV co-infected patients in Nigeria ». *PLoS ONE* 14 (2): e0210724. https://doi.org/10.1371/journal.pone.0210724.

Shinde, Vivek, Carolyn B. Bridges, Timothy M. Uyeki, Bo Shu, Amanda Balish, Xiyan Xu, Stephen Lindstrom, et al. 2009. « Triple-Reassortant Swine Influenza A (H1) in Humans in the United States, 2005–2009 ». *New England Journal of Medicine* 360 (25): 2616-25. https://doi.org/10.1056/NEJMoa0903812.

Sievers, Fabian, et Desmond G. Higgins. 2021. « The Clustal Omega Multiple Alignment Package ». *Methods in Molecular Biology (Clifton, N.J.)* 2231: 3-16. https://doi.org/10.1007/978-1-0716-1036-7_1.

Silva, Rafael Alves da, Isabel Maria Vicente Guedes de Carvalho, Renata Prandini Adum de Matos, Lilian Hiromi Tomonari Yamasaki, Cíntia Bittar, Paula Rahal, et Ana Carolina Gomes Jardim. 2017. « Evidence of Bottleneck Effect on Hepatitis C Virus Transmission between a Couple under Interferon Based Therapy ». *Infection, Genetics and Evolution* 47 (janvier): 87-93. https://doi.org/10.1016/j.meegid.2016.11.012.

Simmonds, P, L Q Zhang, F McOmish, P Balfe, C A Ludlam, et A J Brown. 1991. « Discontinuous sequence change of human immunodeficiency virus (HIV) type 1 env sequences in plasma viral and lymphocyte-associated proviral populations in vivo: implications for models of HIV pathogenesis. » *Journal of Virology* 65 (11): 6266-76.

Simmonds, Peter. 2013. « The Origin of Hepatitis C Virus ». In *Hepatitis C Virus: From Molecular Virology to Antiviral Therapy*, édité par Ralf Bartenschlager, 1-15. Current Topics in Microbiology and Immunology. Berlin, Heidelberg: Springer. https://doi.org/10.1007/978-3-642-27340-7_1.

Simmons, Ruth A, Christian B Willberg, et Klenerman Paul. 2013. « Immune Evasion by Viruses ». In *ELS*. John Wiley & Sons, Ltd. https://doi.org/10.1002/9780470015902.a0024790.

Simon, François, Philippe Mauclère, Pierre Roques, Ibtissam Loussert-Ajaka, Michaela C. Müller-Trutwin, Sentob Saragosti, Marie Claude Georges-Courbot, Françoise Barré-Sinoussi, et Françoise Brun-Vézinet. 1998. « Identification of a New Human Immunodeficiency Virus Type 1 Distinct from Group M and Group O ». *Nature Medicine* 4 (9): 1032-37. https://doi.org/10.1038/2017.

Simon-Loriere, Etienne, et Edward C. Holmes. 2011. « Why Do RNA Viruses Recombine? » *Nature Reviews Microbiology* 9 (8): 617-26. https://doi.org/10.1038/nrmicro2614.

Sims, David, Ian Sudbery, Nicholas E. Ilott, Andreas Heger, et Chris P. Ponting. 2014. « Sequencing Depth and Coverage: Key Considerations in Genomic Analyses ». *Nature Reviews Genetics* 15 (2): 121-32. https://doi.org/10.1038/nrg3642.

Sivakumaran, Shanya, Felix Agakov, Evropi Theodoratou, James G. Prendergast, Lina Zgaga, Teri Manolio, Igor Rudan, Paul McKeigue, James F. Wilson, et Harry Campbell. 2011. « Abundant Pleiotropy in Human Complex Diseases and Traits ». *The American Journal of Human Genetics* 89 (5): 607-18. https://doi.org/10.1016/j.ajhg.2011.10.004.

Sluis-Cremer, Nicolas, Michael R. Jordan, Kelly Huber, Carole L. Wallis, Silvia Bertagnolio, John W. Mellors, Neil T. Parkin, et P. Richard Harrigan. 2014. « E138A in HIV-1 Reverse Transcriptase Is More Common in Subtype C than B: Implications for Rilpivirine Use in Resource-Limited Settings ». *Antiviral Research* 107 (juillet): 31-34. https://doi.org/10.1016/j.antiviral.2014.04.001.

Smith, Derek J., Alan S. Lapedes, Jan C. de Jong, Theo M. Bestebroer, Guus F. Rimmelzwaan, Albert D. M. E. Osterhaus, et Ron A. M. Fouchier. 2004. « Mapping the Antigenic and Genetic Evolution of Influenza Virus ». *Science (New York, N.Y.)* 305 (5682): 371-76. https://doi.org/10.1126/science.1097211.

Smith, Donald B, Jens Bukh, Carla Kuiken, A Scott Muerhoff, Charles M Rice, Jack T Stapleton, et Peter Simmonds. 2014. « Expanded Classification of Hepatitis C Virus Into 7 Genotypes and 67 Subtypes: Updated Criteria and Genotype Assignment Web Resource ». *Hepatology (Baltimore, Md.)* 59 (1): 318-27. https://doi.org/10.1002/hep.26744.

Smith, T. F., et M. S. Waterman. 1981. « Identification of Common Molecular Subsequences ». *Journal of Molecular Biology* 147 (1): 195-97. https://doi.org/10.1016/0022-2836(81)90087-5.

Snijder, Eric J., Peter J. Bredenbeek, Jessika C. Dobbe, Volker Thiel, John Ziebuhr, Leo L. M. Poon, Yi Guan, Mikhail Rozanov, Willy J. M. Spaan, et Alexander E. Gorbalenya. 2003. « Unique and Conserved Features of Genome and Proteome of SARS-Coronavirus, an Early Split-off From the Coronavirus Group 2 Lineage ». *Journal of Molecular Biology* 331 (5): 991-1004. https://doi.org/10.1016/S0022-2836(03)00865-9.

Sorbo, Maria C., Valeria Cento, Velia C. Di Maio, Anita Y. M. Howe, Federico Garcia, Carlo F. Perno, et Francesca Ceccherini-Silberstein. 2018. « Hepatitis C virus drug resistance associated substitutions and their clinical relevance: Update 2018 ». *Drug Resistance Updates* 37 (mars): 17-39. https://doi.org/10.1016/j.drup.2018.01.004.

Soria, María Eugenia, Josep Gregori, Qian Chen, Damir García-Cehic, Meritxell Llorens, Ana I. de Ávila, Nathan M. Beach, et al. 2018. « Pipeline for specific subtype amplification and drug resistance detection in hepatitis C virus ». *BMC Infectious Diseases* 18 (1): 446. https://doi.org/10.1186/s12879-018-3356-6.

Soubrier, Julien, Mike Steel, Michael S.Y. Lee, Clio Der Sarkissian, Stéphane Guindon, Simon Y.W. Ho, et Alan Cooper. 2012. « The Influence of Rate Heterogeneity among Sites on the Time Dependence of Molecular Rates ». *Molecular Biology and Evolution* 29 (11): 3345-58. https://doi.org/10.1093/molbev/mss140.

Spady, Tyrone C., Ole Seehausen, Ellis R. Loew, Rebecca C. Jordan, Thomas D. Kocher, et Karen L. Carleton. 2005. « Adaptive Molecular Evolution in the Opsin Genes of Rapidly Speciating Cichlid Species ». *Molecular Biology and Evolution* 22 (6): 1412-22. https://doi.org/10.1093/molbev/msi137.

Ssemwanga, Deogratius, Raphael W. Lihana, Chinenye Ugoji, Alash'le Abimiku, John Nkengasong, Patrick Dakum, et Nicaise Ndembi. 2015. « Update on HIV-1 Acquired and Transmitted Drug Resistance in Africa ». *AIDS Reviews* 17 (1): 3-20.

St. Clair, M. H., J. L. Martin, G. Tudor-Williams, M. C. Bach, C. L. Vavro, D. M. King, P. Kellam, S. D. Kemp, et B. A. Larder. 1991. « Resistance to ddI and Sensitivity to AZT Induced by a Mutation in HIV-1 Reverse Transcriptase ». *Science* 253 (5027): 1557-59. https://doi.org/10.1126/science.1716788.

Stanley, W. M. 1935. « Isolation of a Crystalline Protein Possessing the Properties of Tobacco-Mosaic Virus ». *Science*, juin. https://doi.org/10.1126/science.81.2113.644.

Stayton, C. Tristan. 2015. « The Definition, Recognition, and Interpretation of Convergent Evolution, and Two New Measures for Quantifying and Assessing the Significance of Convergence ». *Evolution* 69 (8): 2140-53. https://doi.org/10.1111/evo.12729.

Steel, John, Anice C. Lowen, Samira Mubareka, et Peter Palese. 2009. « Transmission of Influenza Virus in a Mammalian Host Is Increased by PB2 Amino Acids 627K or 627E/701N ». *PLOS Pathogens* 5 (1): e1000252. https://doi.org/10.1371/journal.ppat.1000252.

Steiner, Cynthia C., Holger Römpler, Linda M. Boettger, Torsten Schöneberg, et Hopi E. Hoekstra. 2009. « The Genetic Basis of Phenotypic Convergence in Beach Mice: Similar Pigment Patterns but Different Genes ». *Molecular Biology and Evolution* 26 (1): 35-45. https://doi.org/10.1093/molbev/msn218.

Stern, David L. 2013. « The Genetic Causes of Convergent Evolution ». *Nature Reviews. Genetics* 14 (11): 751-64. https://doi.org/10.1038/nrg3483.

Stern, David L., et Virginie Orgogozo. 2009. « Is Genetic Evolution Predictable? » *Science* 323 (5915): 746-51. https://doi.org/10.1126/science.1158997.

Stern, David L., Virginie Orgogozo, et M. Rausher. 2008. « The Loci of Evolution: How Predictable is Genetic Evolution? » *Evolution* 62 (9): 2155-77. https://doi.org/10.1111/j.1558-5646.2008.00450.x.

Stewart, Caro-Beth, James W. Schilling, et Allan C. Wilson. 1987. « Adaptive Evolution in the Stomach Lysozymes of Foregut Fermenters ». *Nature* 330 (6146): 401-4. https://doi.org/10.1038/330401a0.

Stoltzfus, Arlin. 2006. « Mutationism and the Dual Causation of Evolutionary Change ». *Evolution & Development* 8 (3): 304-17. https://doi.org/10.1111/j.1525-142X.2006.00101.x.

Stoltzfus, Arlin, et David M. McCandlish. 2017. « Mutational Biases Influence Parallel Adaptation ». *Molecular Biology and Evolution* 34 (9): 2163-72. https://doi.org/10.1093/molbev/msx180.

Storz, Jay F. 2016. « Causes of molecular convergence and parallelism in protein evolution ». *Nature reviews. Genetics* 17 (4): 239-50. https://doi.org/10.1038/nrg.2016.11.

Storz, Jay F., Chandrasekhar Natarajan, Anthony V. Signore, Christopher C. Witt, David M. McCandlish, et Arlin Stoltzfus. 2019. « The role of mutation bias in adaptive molecular evolution: insights from convergent changes in protein function ». *Philosophical Transactions of the Royal Society B: Biological Sciences* 374 (1777): 20180238. https://doi.org/10.1098/rstb.2018.0238.

Studier, J A, et K J Keppler. 1988. « A note on the neighbor-joining algorithm of Saitou and Nei. » *Molecular Biology and Evolution* 5 (6): 729-31. https://doi.org/10.1093/oxfordjournals.molbev.a040527.

Sugawara, Tohru, Yohey Terai, Hiroo Imai, George F. Turner, Stephan Koblmüller, Christian Sturmbauer, Yoshinori Shichida, et Norihiro Okada. 2005. « Parallelism of Amino Acid Changes at the RH1 Affecting Spectral Sensitivity among Deep-Water Cichlids from Lakes Tanganyika and Malawi ». *Proceedings of the National Academy of Sciences* 102 (15): 5448-53. https://doi.org/10.1073/pnas.0405302102.

Sun, Cong, Yin-Feng Kang, Yuan-Tao Liu, Xiang-Wei Kong, Hui-Qin Xu, Dan Xiong, Chu Xie, et al. 2022. « Parallel Profiling of Antigenicity Alteration and Immune Escape of SARS-CoV-2 Omicron and Other Variants ». *Signal Transduction and Targeted Therapy* 7 (1): 1-10. https://doi.org/10.1038/s41392-022-00910-6.

Susko, Edward, Chris Field, Christian Blouin, et Andrew J. Roger. 2003. « Estimation of Rates-across-Sites Distributions in Phylogenetic Substitution Models ». *Systematic Biology* 52 (5): 594-603. https://doi.org/10.1080/10635150390235395.

Suzuki, Y., et T. Gojobori. 1999. « A Method for Detecting Positive Selection at Single Amino Acid Sites ». *Molecular Biology and Evolution* 16 (10): 1315-28. https://doi.org/10.1093/oxfordjournals.molbev.a026042.

Suzuki, Yoshiyuki. 2004. « New Methods for Detecting Positive Selection at Single Amino Acid Sites ». *Journal of Molecular Evolution* 59 (1): 11-19. https://doi.org/10.1007/s00239-004-2599-6.

Svensson, Per, Oliver E Bläsing, et Peter Westhoff. 2003. « Evolution of C4 Phosphoenolpyruvate Carboxylase ». *Archives of Biochemistry and Biophysics* 414 (2): 180-88. https://doi.org/10.1016/S0003-9861(03)00165-6.

Swanson, Kara W., David M. Irwin, et Allan C. Wilson. 1991. « Stomach Lysozyme Gene of the Langur Monkey: Tests for Convergence and Positive Selection ». *Journal of Molecular Evolution* 33 (5): 418-25. https://doi.org/10.1007/BF02103133.

Tabor, Edward, JacquesA. Drucker, JayH. Hoofnagle, Milton April, RobertJ. Gerety, LeonardB. Seeff, DanielR. Jackson, LewellysF. Barker, et Geronima Pineda-Tamondong. 1978. « TRANSMISSION OF NON-A, NON-B HEPATITIS FROM MAN TO CHIMPANZEE ». *The Lancet*, Originally published as Volume 1, Issue 8062, 311 (8062): 463-66. https://doi.org/10.1016/S0140-6736(78)90132-0.

Takehisa, Jun, Matthias H. Kraus, Ahidjo Ayouba, Elizabeth Bailes, Fran Van Heuverswyn, Julie M. Decker, Yingying Li, et al. 2009. « Origin and Biology of Simian Immunodeficiency Virus in Wild-Living Western Gorillas ». *Journal of Virology* 83 (4): 1635-48. https://doi.org/10.1128/JVI.02311-08.

Tamuri, Asif U, Nick Goldman, et Mario dos Reis. 2014. « A Penalized-Likelihood Method to Estimate the Distribution of Selection Coefficients from Phylogenetic Data ». *Genetics* 197 (1): 257-71. https://doi.org/10.1534/genetics.114.162263.

Tamuri, Asif U, Mario dos Reis, et Richard A Goldstein. 2012. « Estimating the Distribution of Selection Coefficients from Phylogenetic Data Using Sitewise Mutation-Selection Models ». *Genetics* 190 (3): 1101-15. https://doi.org/10.1534/genetics.111.136432.

Tamuri, Asif U., Mario dos Reis, Alan J. Hay, et Richard A. Goldstein. 2009. « Identifying Changes in Selective Constraints: Host Shifts in Influenza ». *PLOS Computational Biology* 5 (11): e1000564. https://doi.org/10.1371/journal.pcbi.1000564.

Tateno, Y., T. Imanishi, S. Miyazaki, K. Fukami-Kobayashi, N. Saitou, H. Sugawara, et T. Gojobori. 2002. « DNA Data Bank of Japan (DDBJ) for genome scale research in life science ». *Nucleic Acids Research* 30 (1): 27-30.

Tavaré, Simon et Miura. 1986. « Some Probabilistic and Statistical Problems in the Analysis of DNA Sequences ». In *Lectures on mathematics in the life sciences*, Providence, 17:57-86.

Tegally, Houriiyah, Eduan Wilkinson, Christian L. Althaus, Marta Giovanetti, James Emmanuel San, Jennifer Giandhari, Sureshnee Pillay, et al. 2021. « Rapid Replacement of the Beta Variant by the Delta Variant in South Africa ». https://doi.org/10.1101/2021.09.23.21264018.

The UniProt Consortium. 2021. « UniProt: the universal protein knowledgebase in 2021 ». *Nucleic Acids Research* 49 (D1): D480-89. https://doi.org/10.1093/nar/gkaa1100.

Thimme, Robert, Volker Lohmann, et Friedemann Weber. 2006. « A Target on the Move: Innate and Adaptive Immune Escape Strategies of Hepatitis C Virus ». *Antiviral Research* 69 (3): 129-41. https://doi.org/10.1016/j.antiviral.2005.12.001.

Thomas, Gregg W. C., et Matthew W. Hahn. 2015. « Determining the Null Model for Detecting Adaptive Convergence from Genomic Data: A Case Study Using Echolocating Mammals ». *Molecular Biology and Evolution* 32 (5): 1232-36. https://doi.org/10.1093/molbev/msv013.

Thomas, Gregg W. C., Matthew W. Hahn, et Yoonsoo Hahn. 2017. « The Effects of Increasing the Number of Taxa on Inferences of Molecular Convergence ». *Genome Biology and Evolution* 9 (1): 213-21. https://doi.org/10.1093/gbe/evw306.

Thompson, J D, D G Higgins, et T J Gibson. 1994. « CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. » *Nucleic Acids Research* 22 (22): 4673-80.

Timm, Joerg, Georg M. Lauer, Daniel G. Kavanagh, Isabelle Sheridan, Arthur Y. Kim, Michaela Lucas, Thillagavathie Pillay, et al. 2004. « CD8 Epitope Escape and Reversion in Acute HCV Infection ». *The Journal of Experimental Medicine* 200 (12): 1593-1604. https://doi.org/10.1084/jem.20041006.

Tommaso, Paolo Di, Maria Chatzou, Evan W. Floden, Pablo Prieto Barja, Emilio Palumbo, et Cedric Notredame. 2017. « Nextflow Enables Reproducible Computational Workflows ». *Nature Biotechnology*, avril. https://doi.org/10.1038/nbt.3820.

Tosi, Michael F. 2005. « Innate Immune Responses to Infection ». *Journal of Allergy and Clinical Immunology* 116 (2): 241-49. https://doi.org/10.1016/j.jaci.2005.05.036.

Trifonov, Vladimir, Hossein Khiabanian, et Raul Rabadan. 2009. « Geographic Dependence, Surveillance, and Origins of the 2009 Influenza A (H1N1) Virus ». *New England Journal of Medicine* 361 (2): 115-19. https://doi.org/10.1056/NEJMp0904572.

Troupin, Cécile, Laurent Dacheux, Marion Tanguy, Claude Sabeta, Hervé Blanc, Christiane Bouchier, Marco Vignuzzi, Sebastián Duchene, Edward C. Holmes, et Hervé Bourhy.

2016. « Large-Scale Phylogenomic Analysis Reveals the Complex Evolutionary History of Rabies Virus in Multiple Carnivore Hosts ». *PLoS Pathogens* 12 (12): e1006041. https://doi.org/10.1371/journal.ppat.1006041.

Tsetsarkin, Konstantin A., Dana L. Vanlandingham, Charles E. McGee, et Stephen Higgs. 2007. « A Single Mutation in Chikungunya Virus Affects Vector Specificity and Epidemic Potential ». *PLOS Pathogens* 3 (12): e201. https://doi.org/10.1371/journal.ppat.0030201.

Turner, Dan, Bluma Brenner, et Mark A. Wainberg. 2003. « Multiple Effects of the M184V Resistance Mutation in the Reverse Transcriptase of Human Immunodeficiency Virus Type 1 ». *Clinical and Diagnostic Laboratory Immunology* 10 (6): 979-81. https://doi.org/10.1128/CDLI.10.6.979-981.2003.

Ujvari, Beata, Nicholas R. Casewell, Kartik Sunagar, Kevin Arbuckle, Wolfgang Wüster, Nathan Lo, Denis O'Meally, et al. 2015. « Widespread Convergence in Toxin Resistance by Predictable Molecular Evolution ». *Proceedings of the National Academy of Sciences* 112 (38): 11911-16. https://doi.org/10.1073/pnas.1511706112.

UNAIDS. 2016. « Global AIDS update 2016 ». *Geneva: UNAIDS*.

Vallari, Ana, Vera Holzmayer, Barbara Harris, Julie Yamaguchi, Charlotte Ngansop, Florence Makamche, Dora Mbanya, et al. 2011. « Confirmation of Putative HIV-1 Group P in Cameroon ». *Journal of Virology* 85 (3): 1403-7. https://doi.org/10.1128/JVI.02005-10.

Van Rooyen, C. E., et G. D. Scott. 1948. « Smallpox Diagnosis with Special Reference to Electron Microscopy ». *Canadian Journal of Public Health = Revue Canadienne De Sante Publique* 39 (12): 467-77.

Vergidis, Paschalis I., Matthew E. Falagas, et Davidson H. Hamer. 2009. « Meta-Analytical Studies on the Epidemiology, Prevention, and Treatment of Human Immunodeficiency Virus Infection ». *Infectious Disease Clinics of North America*, Meta-analysis in Infectious Diseases, 23 (2): 295-308. https://doi.org/10.1016/j.idc.2009.01.013.

Vidal, Nicole, Martine Peeters, Claire Mulanga-Kabeya, Nzila Nzilambi, David Robertson, Wantabala Ilunga, Hurogo Sema, Kazadi Tshimanga, Beni Bongo, et Eric Delaporte. 2000. « Unprecedented Degree of Human Immunodeficiency Virus Type 1 (HIV-1) Group M Genetic Diversity in the Democratic Republic of Congo Suggests that the HIV-1 Pandemic Originated in Central Africa ». *Journal of Virology* 74 (22): 10498-507. https://doi.org/10.1128/JVI.74.22.10498-10507.2000.

Vignuzzi, Marco, et Stephen Higgs. 2017. « The Bridges and Blockades to Evolutionary Convergence on the Road to Predicting Chikungunya Virus Evolution ». *Annual Review of Virology* 4 (1): 181-200. https://doi.org/10.1146/annurev-virology-101416-041757.

Villabona-Arenas, Christian Julian, Nicole Vidal, Emilande Guichet, Laetitia Serrano, Eric Delaporte, Olivier Gascuel, et Martine Peeters. 2016. « In-Depth Analysis of HIV-1 Drug Resistance Mutations in HIV-Infected Individuals Failing First-Line Regimens in West and Central Africa ». *AIDS* 30 (17): 2577-89. https://doi.org/10.1097/QAD.0000000000001233.

Virtanen, Pauli, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, et al. 2020. « SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python ». *Nature Methods* 17: 261-72. https://doi.org/10.1038/s41592-019-0686-2.

Volkow, Patricia, Antonio Marin Lopez, et Indiana Torres. 1997. « Plasma Trade and the HIV Epidemic ». *The Lancet* 349 (9048): 327-28. https://doi.org/10.1016/S0140-6736(05)62826-7.

Wain, Louise V., Elizabeth Bailes, Frederic Bibollet-Ruche, Julie M. Decker, Brandon F. Keele, Fran Van Heuverswyn, Yingying Li, et al. 2007. « Adaptation of HIV-1 to Its Human Host ». *Molecular Biology and Evolution* 24 (8): 1853-60. https://doi.org/10.1093/molbev/msm110.

Wain-Hobson, Simon, Pierre Sonigo, Olivier Danos, Stewart Cole, et Marc Alizon. 1985. « Nucleotide Sequence of the AIDS Virus, LAV ». *Cell* 40 (1): 9-17. https://doi.org/10.1016/0092-8674(85)90303-4.

Walker, A., K. Skibbe, E. Steinmann, S. Pfaender, T. Kuntzen, D.A. Megger, S. Groten, et al. 2016. « Distinct Escape Pathway by Hepatitis C Virus Genotype 1a from a Dominant CD8+ T Cell Response by Selection of Altered Epitope Processing ». *Journal of Virology* 90 (1): 33-42. https://doi.org/10.1128/JVI.01993-15.

Walker, Andreas, Sandra Filke, Nadine Lübke, Martin Obermeier, Rolf Kaiser, Dieter Häussinger, Jörg Timm, et Hans H. Bock. 2017. « Detection of a genetic footprint of the sofosbuvir resistance-associated substitution S282T after HCV treatment failure ». *Virology Journal* 14 (juin). https://doi.org/10.1186/s12985-017-0779-4.

Walker, Peter J., Stuart G. Siddell, Elliot J. Lefkowitz, Arcady R. Mushegian, Donald M. Dempsey, Bas E. Dutilh, Balázs Harrach, et al. 2019. « Changes to Virus Taxonomy and the International Code of Virus Classification and Nomenclature Ratified by the International Committee on Taxonomy of Viruses (2019) ». *Archives of Virology* 164 (9): 2417-29. https://doi.org/10.1007/s00705-019-04306-w.

Wallace, Iain M., Orla O'Sullivan, Desmond G. Higgins, et Cedric Notredame. 2006. « M-Coffee: combining multiple sequence alignment methods with T-Coffee ». *Nucleic Acids Research* 34 (6): 1692-99. https://doi.org/10.1093/nar/gkl091.

Wallace, Robert G., Hoangminh Hodac, Richard H. Lathrop, et Walter M. Fitch. 2007. « A Statistical Phylogeography of Influenza A H5N1 ». *Proceedings of the National Academy of Sciences of the United States of America* 104 (11): 4473-78. https://doi.org/10.1073/pnas.0700435104.

Wargo, Andrew R, et Gael Kurath. 2012. « Viral fitness: definitions, measurement, and current insights ». *Current Opinion in Virology* 2 (5): 538-45. https://doi.org/10.1016/j.coviro.2012.07.007.

Weaver, Scott C., Naomi L. Forrester, Jianying Liu, et Nikos Vasilakis. 2021. « Population bottlenecks and founder effects: implications for mosquito-borne arboviral emergence ». *Nature Reviews. Microbiology* 19 (3): 184-95. https://doi.org/10.1038/s41579-020-00482-8.

Webby, Rj, et Robert G. Webster. 2001. « Emergence of Influenza A Viruses ». *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences* 356 (1416). https://doi.org/10.1098/rstb.2001.0997.

Weigang, Sebastian, Jonas Fuchs, Gert Zimmer, Daniel Schnepf, Lisa Kern, Julius Beer, Hendrik Luxenburger, et al. 2021. « Within-Host Evolution of SARS-CoV-2 in an Immunosuppressed COVID-19 Patient as a Source of Immune Escape Variants ». *Nature Communications* 12 (1): 6405. https://doi.org/10.1038/s41467-021-26602-3.

Wenger, Aaron M., Paul Peluso, William J. Rowell, Pi-Chuan Chang, Richard J. Hall, Gregory T. Concepcion, Jana Ebler, et al. 2019. « Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome ». *Nature biotechnology* 37 (10): 1155-62. https://doi.org/10.1038/s41587-019-0217-9.

Wensing, Annemarie M., Vincent Calvez, Francesca Ceccherini-Silberstein, Charlotte Charpentier, Huldrych F. Günthard, Roger Paredes, Robert W. Shafer, et Douglas D. Richman. 2019. « 2019 Update of the Drug Resistance Mutations in HIV-1 ». *Topics in Antiviral Medicine* 27 (3): 111-21.

Wichman, H. A., M. R. Badgett, L. A. Scott, C. M. Boulianne, et J. J. Bull. 1999. « Different Trajectories of Parallel Evolution During Viral Adaptation ». *Science* 285 (5426): 422-24. https://doi.org/10.1126/science.285.5426.422.

Wilkinson, Mark D., Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, et al. 2016. « The FAIR Guiding Principles for scientific

data management and stewardship ». *Scientific Data* 3 (mars): 160018. https://doi.org/10.1038/sdata.2016.18.

Wilm, Andreas, Pauline Poh Kim Aw, Denis Bertrand, Grace Hui Ting Yeo, Swee Hoe Ong, Chang Hua Wong, Chiea Chuen Khor, Rosemary Petric, Martin Lloyd Hibberd, et Niranjan Nagarajan. 2012. « LoFreq: a sequence-quality aware, ultra-sensitive variant caller for uncovering cell-population heterogeneity from high-throughput sequencing datasets ». *Nucleic Acids Research* 40 (22): 11189-201. https://doi.org/10.1093/nar/gks918.

Wittkopp, Patricia J., Barry L. Williams, Jayne E. Selegue, et Sean B. Carroll. 2003. « Drosophila Pigmentation Evolution: Divergent Genotypes Underlying Convergent Phenotypes ». *Proceedings of the National Academy of Sciences of the United States of America* 100 (4): 1808-13. https://doi.org/10.1073/pnas.0336368100.

Wolf, Yuri I., Darius Kazlauskas, Jaime Iranzo, Adriana Lucía-Sanz, Jens H. Kuhn, Mart Krupovic, Valerian V. Dolja, et Eugene V. Koonin. 2018. « Origins and Evolution of the Global RNA Virome ». *MBio* 9 (6): e02329-18. https://doi.org/10.1128/mBio.02329-18.

Wolkowicz, Roland, et Moselio Schaechter. 2008. « What Makes a Virus a Virus? » *Nature Reviews Microbiology* 6 (8): 643-643. https://doi.org/10.1038/nrmicro1858-c1.

World Health Organization. 2016. « Global Health Sector Strategy on Viral Hepatitis 2016-2021. Towards Ending Viral Hepatitis ». *Global Health Sector Strategy on Viral Hepatitis 2016-2021. Towards Ending Viral Hepatitis*. https://apps.who.int/iris/handle/10665/246177.

World Health Organization, Global Fund, et US Centers for Disease Control and Prevention. 2017. *HIV Drug Resistance Report 2017*. Geneva: World Health Organization. https://apps.who.int/iris/handle/10665/255896.

Wu, Dong-Bo, Wei Jiang, Yong-Hong Wang, Bin Chen, Meng-Lan Wang, Ya-Chao Tao, En-Qiang Chen, et Hong Tang. 2019. « Safety and Efficacy of Sofosbuvir-Based Direct-Acting Antiviral Regimens for Hepatitis C Virus Genotype 6 in Southwest China: Real-World Experience of a Retrospective Study ». *Journal of Viral Hepatitis* 26 (3): 316-22. https://doi.org/10.1111/jvh.13033.

Wyles, David, Hadas Dvory-Sobol, Evguenia S. Svarovskaia, Brian P. Doehle, Ross Martin, Nezam H. Afdhal, Kris V. Kowdley, et al. 2017. « Post-Treatment Resistance Analysis of Hepatitis C Virus from Phase II and III Clinical Trials of Ledipasvir/Sofosbuvir ». *Journal of Hepatology* 66 (4): 703-10. https://doi.org/10.1016/j.jhep.2016.11.022.

Wyles, David L., et Anne F. Luetkemeyer. 2017. « Understanding Hepatitis C Virus Drug Resistance: Clinical Implications for Current and Future Regimens ». *Topics in Antiviral Medicine* 25 (3): 103-9.

Wyles, David, Alessandra Mangia, Wendy Cheng, Stephen Shafran, Christian Schwabe, Wen Ouyang, Charlotte Hedskog, et al. 2017. « Long-term persistence of HCV NS5A resistance-associated substitutions after treatment with the HCV NS5A inhibitor, ledipasvir, without sofosbuvir ». *Antiviral therapy* 23 (3): 229-38. https://doi.org/10.3851/IMP3181.

Xiang, Dan, Xuejuan Shen, Zhiqing Pu, David M Irwin, Ming Liao, et Yongyi Shen. 2018. « Convergent Evolution of Human-Isolated H7N9 Avian Influenza A Viruses ». *The Journal of Infectious Diseases* 217 (11): 1699-1707. https://doi.org/10.1093/infdis/jiy082.

Xu, Shaohua, Jiayan Wang, Zixiao Guo, Ziwen He, et Suhua Shi. 2020. « Genomic Convergence in the Adaptation to Extreme Environments ». *Plant Communications* 1 (6): 100117. https://doi.org/10.1016/j.xplc.2020.100117.

Xu, Simin, Brian Doehle, Sonal Rajyaguru, Bin Han, Ona Barauskas, Joy Feng, Jason Perry, et al. 2017. « In vitro selection of resistance to sofosbuvir in HCV replicons of genotype 1 to 6 ». *Antiviral Therapy* 22 (7): 587-97. https://doi.org/10.3851/IMP3149.

Yang, Z. 1997. « PAML: A Program Package for Phylogenetic Analysis by Maximum Likelihood ». *Computer Applications in the Biosciences: CABIOS* 13 (5): 555-56.

———. 1998. « Likelihood Ratio Tests for Detecting Positive Selection and Application to Primate Lysozyme Evolution ». *Molecular Biology and Evolution* 15 (5): 568-73. https://doi.org/10.1093/oxfordjournals.molbev.a025957.

———. 2000. « Maximum Likelihood Estimation on Large Phylogenies and Analysis of Adaptive Evolution in Human Influenza Virus A ». *Journal of Molecular Evolution* 51 (5): 423-32. https://doi.org/10.1007/s002390010105.

Yang, Ziheng. 2007. « PAML 4: Phylogenetic Analysis by Maximum Likelihood ». *Molecular Biology and Evolution* 24 (8): 1586-91. https://doi.org/10.1093/molbev/msm088.

Yang, Ziheng, et Rasmus Nielsen. 2002. « Codon-Substitution Models for Detecting Molecular Adaptation at Individual Sites along Specific Lineages ». *Molecular Biology and Evolution* 19 (6): 908-17. https://doi.org/10.1093/oxfordjournals.molbev.a004148.

Ye, Jing, Bibo Zhu, Zhen F. Fu, Huanchun Chen, et Shengbo Cao. 2013. « Immune Evasion Strategies of Flaviviruses ». *Vaccine* 31 (3): 461-71. https://doi.org/10.1016/j.vaccine.2012.11.015.

Yebra, Gonzalo, Dan Frampton, Tiziano Gallo Cassarino, Jade Raffle, Jonathan Hubb, R. Bridget Ferns, Laura Waters, et al. 2018. « A High HIV-1 Strain Variability in London, UK, Revealed by Full-Genome Analysis: Results from the ICONIC Project ». *PLOS ONE* 13 (2): e0192081. https://doi.org/10.1371/journal.pone.0192081.

Yi, MinKyung, et Stanley M. Lemon. 2003. « 3' Nontranslated RNA Signals Required for Replication of Hepatitis C Virus RNA ». *Journal of Virology* 77 (6): 3557-68. https://doi.org/10.1128/jvi.77.6.3557-3568.2003.

Yokoyama, Shozo. 2008. « Evolution of Dim-Light and Color Vision Pigments ». *Annual Review of Genomics and Human Genetics* 9 (1): 259-82. https://doi.org/10.1146/annurev.genom.9.081307.164228.

Yokoyama, Shozo, Takashi Tada, Huan Zhang, et Lyle Britt. 2008. « Elucidation of Phenotypic Adaptations: Molecular Analyses of Dim-Light Vision Proteins in Vertebrates ». *Proceedings of the National Academy of Sciences of the United States of America* 105 (36): 13480-85. https://doi.org/10.1073/pnas.0802426105.

Yokoyama, Shozo, Hui Yang, et William T. Starmer. 2008. « Molecular Basis of Spectral Tuning in the Red- and Green-Sensitive (M/LWS) Pigments in Vertebrates ». *Genetics* 179 (4): 2037-43. https://doi.org/10.1534/genetics.108.090449.

Yooseph, Shibu, Granger Sutton, Douglas B Rusch, Aaron L Halpern, Shannon J Williamson, Karin Remington, Jonathan A Eisen, et al. 2007. « The Sorcerer II Global Ocean Sampling Expedition: Expanding the Universe of Protein Families ». *PLoS Biology* 5 (3): e16. https://doi.org/10.1371/journal.pbio.0050016.

Zhang, Jianzhi. 2003. « Parallel Functional Changes in the Digestive RNases of Ruminants and Colobines by Divergent Amino Acid Substitutions ». *Molecular Biology and Evolution* 20 (8): 1310-17. https://doi.org/10.1093/molbev/msg143.

———. 2006. « Parallel Adaptive Origins of Digestive RNases in Asian and African Leaf Monkeys ». *Nature Genetics* 38 (7): 819-23. https://doi.org/10.1038/ng1812.

Zhang, Jianzhi, et Sudhir Kumar. 1997. « Detection of convergent and parallel evolution at the amino acid sequence level. » *Molecular biology and evolution* 14 (5): 527-36.

Zhang, Jianzhi, Rasmus Nielsen, et Ziheng Yang. 2005. « Evaluation of an Improved Branch-Site Likelihood Method for Detecting Positive Selection at the Molecular Level ». *Molecular Biology and Evolution* 22 (12): 2472-79. https://doi.org/10.1093/molbev/msi237.

Zhang, Meiwen, Daniel O'Keefe, Momoko Iwamoto, Kimchamroeun Sann, Antharo Kien, Vithurneat Hang, Cecile Brucker, et al. 2020. « High Sustained Viral Response Rate in Patients with Hepatitis C Using Generic Sofosbuvir and Daclatasvir in Phnom Penh, Cambodia ». *Journal of Viral Hepatitis* 27 (9): 886-95. https://doi.org/10.1111/jvh.13311.

Zhao, Lei, et Christopher J R Illingworth. 2019. « Measurements of Intrahost Viral Diversity Require an Unbiased Diversity Metric ». *Virus Evolution* 5 (1). https://doi.org/10.1093/ve/vey041.

Zhen, Ying, Matthew L. Aardema, Edgar M. Medina, Molly Schumer, et Peter Andolfatto. 2012. « Parallel Molecular Evolution in an Herbivore Community ». *Science* 337 (6102): 1634-37. https://doi.org/10.1126/science.1226630.

Zhukova, Anna, Luc Blassel, Frédéric Lemoine, Marie Morel, Jakub Voznica, et Olivier Gascuel. 2021. « Origin, evolution and global spread of SARS-CoV-2 ». *Comptes Rendus. Biologies* 344 (1): 57-75. https://doi.org/10.5802/crbiol.29.

Zou, Z., et J. Zhang. 2015a. « No Genome-Wide Protein Sequence Convergence for Echolocation ». *Molecular Biology and Evolution* 32 (5): 1237-41. https://doi.org/10.1093/molbev/msv014.

———. 2015b. « Are Convergent and Parallel Amino Acid Substitutions in Protein Evolution More Prevalent Than Neutral Expectations? » *Molecular Biology and Evolution* 32 (8): 2085-96. https://doi.org/10.1093/molbev/msv091.

Zwart, Mark P., et Santiago F. Elena. 2015. « Matters of Size: Genetic Bottlenecks in Virus Infection and Their Potential Impact on Evolution ». *Annual Review of Virology* 2 (1): 161-79. https://doi.org/10.1146/annurev-virology-100114-055135.

TABLE DES MATIÈRES

Résumé i
Abstract ii
Remerciements iii
Liste des figures vii
Liste des tableaux ix
Liste des annexes ix
1 Introduction Générale 1
1.1 Généralités sur les virus 1
1.1.1 Définition et historique 1
1.1.1.1 Découverte des virus 1
1.1.1.2 Définition des virus 2
1.1.1.3 Classification des virus 3
1.1.2 Evolution et propagation des virus 5
1.1.2.1 Les mutations comme moteurs de l'évolution virale. 6
1.1.2.2 L'importance de la recombinaison dans l'émergence de nouvelles souches 8
1.1.2.3 Le réassortiment chez les virus segmentés. 8
1.1.2.4 La population virale 9
1.1.2.5 Les facteurs qui influencent l'évolution virale 10
1.2 Outils et méthodes bioinformatiques pour l'étude de l'évolution moléculaire 12
1.2.1 Production et analyse des séquences virales 12
1.2.1.1 Séquençage et production des données. 12
1.2.1.2 Obtention d'une séquence consensus 14
1.2.1.3 Les bases de données génomiques 15
1.2.2 Analyse comparative des génomes viraux 16
1.2.2.1 Comparaison de séquences et homologie 16
1.2.2.2 Alignement 16
1.2.2.3 Algorithmes d'alignement 17
1.2.3 Modèles d'évolution 17
1.2.3.1 Distance observée ou p-distance 17
1.2.3.2 Modèles de Markov en temps continu. 18
1.2.3.2.1 Définition et caractéristiques 18
1.2.3.2.2 Modèles markoviens pour l'évolution de séquences ADN 19
1.2.3.2.3 Modèles markoviens pour l'évolution de séquences protéiques 20

1.2.3.3 Modèles de mélange pour l'évolution de séquences protéiques 20
1.2.3.4 Vitesses d'évolution 21
1.2.4 Inférence phylogénétique 21
1.2.4.1 Arbre phylogénétique 21
1.2.4.2 Méthodes de distance 23
1.2.4.3 Méthodes basées sur les caractères 23
1.2.4.4 Robustesse des phylogénies 25
1.2.5 Reconstruction de caractères ancestraux 26
1.2.6 Sélection positive ou négative 27
1.3 VIH, VHC et mutations de résistance 30
1.3.1 Le VIH 30
1.3.1.1 Découverte du VIH et manifestations cliniques 30
1.3.1.2 Diversité génétique du VIH 31
1.3.1.3 Structure et génome 33
1.3.1.4 Le gène pol 34
1.3.1.5 Mécanismes d'action des antiviraux 35
1.3.1.6 Mutations de résistance aux traitements 36
1.3.1.7 Implications des mutations de résistance 37
1.3.2 Le VHC 38
1.3.2.1 Découverte du VHC et manifestations cliniques 38
1.3.2.2 Diversité génétique du VHC 38
1.3.2.3 Structure et Génome 39
1.3.2.4 Mécanismes d'action des antiviraux et résistance 40
1.3.2.4.1 Inhibiteurs de NS5A et résistance 40
1.3.2.4.2 Inhibiteurs de la polymérase et résistance 41
2 Etude de la convergence évolutive chez les virus 42
2.1 La convergence évolutive : un signe d'adaptation ? 42
2.1.1 Quand l'évolution se répète 42
2.1.2 Les bases génétiques des convergences phénotypiques 43
2.1.3 Comment expliquer que l'évolution se répète ? 45
2.1.3.1 Pressions de sélection et sélection naturelle 46
2.1.3.2 Biais de mutation 46
2.1.3.3 Biais de fixation, pléiotropie et épistasie 47
2.1.3.4 Taille de population et taux d'évolution 48
2.1.4 Convergences moléculaires de premier plan et d'arrière-plan 49
2.2 La convergence moléculaire est un phénomène répandu chez les virus 50

2.2.1 La nature des virus facilite l'émergence de convergences moléculaires 50
2.2.2 Convergences observées en conditions expérimentales 50
2.2.3 Convergences observées en milieu naturel 51
2.2.3.1 Adaptation à de nouvelles espèces hôtes 52
2.2.3.2 Echappement au système immunitaire de l'hôte 52
2.2.3.3 Résistance 53
2.3 Quelles méthodes permettent d'étudier la convergence au niveau moléculaire 55
2.3.1 Recherche manuelle des mutations 55
2.3.2 Association phénotype-génotype 56
2.3.3 Sélection positive directionnelle 56
2.3.4 Topologies et autres signaux phylogénétiques 57
2.3.5 Détection de mutations spécifiques aux espèces cibles 59
2.3.5.1 Définition stricte 59
2.3.5.2 Changements de profils 60
2.3.5.3 Index de convergence 61
2.3.6 Nécessité d'une nouvelle méthode 61
3 Développement d'une méthode pour détecter la convergence moléculaire 63
3.1 Une approche par simulation et corrélation 63
3.1.1 Des simulations pour estimer le nombre d'émergence d'une mutation. 64
3.1.2 Une mesure de corrélation pour estimer si une mutation émerge indépendamment du phénotype convergent ou non. 66
3.1.3 Implémentation dans le logiciel ConDor 66
3.2 Article : Accurate detection of Convergent Mutations in Protein Alignments with ConDor 66
3.3 Analyses complémentaires 92
3.3.1 Performances sur simulations 92
3.3.1.1 Génération des données simulées 92
3.3.1.2 Résultats 93
3.3.1.3 Conclusions 95
3.3.2 Sensibilité de la composante émergence 96
3.3.3 Sélection positive 100
3.4 Limites : étude du SARS-CoV-2 102
3.5 Conclusions 104
3.6 Perspectives 105
4 Etude de la résistance aux traitements chez le VHC 106
4.1 Présentation des données et du problème 106

4.2 Article : Genomic variations associated with drug resistance in HCV genotype 6 infected patients failing DAA-based therapy. 107
4.3 Conclusions et Perspectives 120
5 Conclusion Générale 121
6 Annexes 123
Matériel supplémentaire de l'article "Genomic variations associated with drug resistance in HCV genotype 6 infected patients failing DAA-based therapy". 123
Bibliographie 130
Table des Matières 165

Résumé en français

Mots clés : Convergence évolutive, Evolution moléculaire, Phylogénétique, Sélection, Adaptation, VIH, VHC, Mutations de résistance, Variants mineurs.

Les génomes des virus à ARN présentent un taux d'évolution parmi les plus élevés, ce qui leur permet de s'adapter rapidement en cas de changement d'environnement. Au sein de leurs hôtes, ces virus existent généralement sous la forme d'une population de mutants dont les génomes sont distincts, bien qu'apparentés. Cette caractéristique présente un avantage évolutif important car parmi la multitude de variants, il y a de plus grande chance que certains soient adaptés à un nouvel environnement. Par exemple, les virus à ARN sont connus pour être capables de changer d'hôte, d'échapper à la reconnaissance du système immunitaire ou de résister aux traitements antiviraux. Ainsi, les virus à ARN sont surreprésentés parmi les agents pathogènes émergents et ré-émergents (anciennement connus mais dont l'incidence augmente de manière inattendue), et leur évolution rapide contribue probablement à leur risque d'émergence élevé, ce qui constitue un défi majeur pour leur contrôle.
La résistance aux traitements antiviraux peut s'avérer un véritable problème dans le contrôle de certains virus pour lesquels il n'existe pas de vaccin. D'une part, certains traitements peuvent devenir inefficaces, ce qui se caractérise par la réplication du virus et pouvant entrainer le décès de l'hôte. D'autre part, des patients traités peuvent avoir une charge virale suffisante pour transmettre des virus déjà résistants, réduisant ainsi l'éventail des thérapies envisageables en première intention.
Durant cette thèse j'ai exploré sous deux angles différents la question de l'adaptation des virus aux traitements antiviraux : 1) l'identification de convergence moléculaire (émergence répétée et indépendante d'une même mutation) dans de grands jeux de données de séquences (plusieurs milliers de séquences), et 2) l'étude des variations génomiques au sein de populations virales soumises à un traitement antiviral. Pour cela, j'ai développé une méthode de détection de convergence évolutive dans les séquences protéiques en suivant l'hypothèse que la présence de mutations de convergence était un indicateur des pressions de sélection sur les séquences virales. J'ai également étudié l'évolution de la diversité virale intra-hôte lorsque les virus sont soumis à un traitement antiviral. Les organismes étudiés pour ce projet sont le virus de l'immunodéficience humaine (VIH) et le virus de l'hépatite C (VHC), tous deux responsables d'affections longue durée, traitées par traitements antiviraux et responsables de millions d'infections à travers le monde.

Résumé en anglais / Abstract

Key words: Evolutionary convergence, Molecular evolution, Phylogenetics, Selection, Adaptation, HIV, HCV, Resistance mutations, Minor variants.

The genomes of RNA viruses have some of the highest rates of evolution, allowing them to adapt rapidly to changing environments. Within their hosts, these viruses generally exist as a population of mutants with distinct, but related genomes. This is an important evolutionary advantage because, among the multitude of variants, there is a higher chance that some will be adapted to a new environment. For example, RNA viruses are known to be able to switch hosts, evade recognition by the immune system or resist antiviral treatments. Thus, RNA viruses are over-represented among emerging and re-emerging pathogens (previously known but unexpectedly increasing in incidence), and their rapid evolution probably contributes to their high risk of emergence, representing a major challenge for their control.
Resistance to antiviral treatments can be a real problem in the control of some viruses for which there is no vaccine. On the one hand, some treatments can become ineffective, which is characterised by the replication of the virus and can lead to the death of the host. On the other hand, treated patients may have a sufficient viral load to transmit already resistant viruses, thus reducing the range of possible first-line therapies.
During this thesis, I explored the question of viral adaptation to antiviral treatments from two different angles: 1) the identification of molecular convergence (repeated and independent emergence of the same mutation) in large sequence datasets (several thousand sequences), and 2) the study of genomic variations within viral populations subjected to antiviral treatment. For this purpose, I developed a method for detecting evolutionary convergence in protein sequences under the assumption that the presence of convergent mutations was an indicator of selection pressures on viral sequences. I also studied the evolution of intra-host viral diversity when viruses are subjected to antiviral treatment. The organisms studied for this project are the human immunodeficiency virus (HIV) and the hepatitis C virus (HCV), both of which are responsible for long-term illnesses, treated with antiviral therapy, and responsible for millions of infections worldwide.