Évolution du VIH : méthodes, modèles et algorithmes

Matthieu Jung

UNIVERSITE MONTPELLIER II

SCIENCES ET TECHNIQUES DU LANGUEDOC

THÈSE

Présentée au Laboratoire d'Informatique, de Robotique et de Microélectronique de Montpellier

pour obtenir le grade de

DOCTEUR DE L'UNIVERSITE MONTPELLIER II

Discipline Bioinformatique
Formation Doctorale Informatique
École Doctorale Information Structure Système (I2S)

Évolution du VIH : méthodes, modèles et algorithmes

par

Matthieu JUNG

Soutenue le 21 mai 2012 devant le jury composé de :

M. Olivier GASCUEL	Directeur de recherche, CNRS/LIRMM, Montpellier	Directeur de thèse
M^{me} Martine PEETERS	Directeur de recherche, IRD, Montpellier	Co-directeur de thèse
M. Alain GUÉNOCHE	Directeur de recherche, IML, Marseille	Rapporteur
M^{me} Marie-Laure CHAIX	Maître de conférences, HDR, Université Paris Descartes	Rapporteur
M. Denis FARGETTE	Directeur de recherche, IRD, Montpellier	Invité

RESUME

La donnée de séquences nucléotidiques permet d'inférer des arbres phylogénétiques, ou phylogénies, qui décrivent leurs liens de parenté au cours de l'évolution. Associer à ces séquences leur date de prélèvement ou leur pays de collecte, permet d'inférer la localisation temporelle ou spatiale de leurs ancêtres communs. Ces données et procédures sont très utilisées pour les séquences de virus et, notamment, celles du virus de l'immunodéficience humaine (VIH), afin d'en retracer l'histoire épidémique à la surface du globe et au cours du temps. L'utilisation de séquences échantillonnées à des moments différents (ou hétérochrones) sert aussi à estimer leur taux de substitution, qui caractérise la vitesse à laquelle elles évoluent.

Les méthodes les plus couramment utilisées pour ces différentes tâches sont précises, mais lourdes en temps de calcul car basées sur des modèles complexes, et ne peuvent traiter que quelques centaines de séquences. Devant le nombre croissant de séquences disponibles dans les bases de données, souvent plusieurs milliers pour une étude donnée, le développement de méthodes rapides et efficaces devient indispensable. Nous présentons une méthode de distances, *Ultrametric Least Squares*, basée sur le principe des moindres carrés, souvent utilisé en phylogénie, qui permet d'estimer le taux de substitution d'un ensemble de séquences hétérochrones, dont on déduit ensuite facilement les dates des spéciations ancestrales. Nous montrons que le critère à optimiser est parabolique par morceaux et proposons un algorithme efficace pour trouver l'optimum global.

L'utilisation de séquences échantillonnées en des lieux différents permet aussi de retracer les chaînes de transmission d'une épidémie. Dans ce cadre, nous utilisons la totalité des séquences disponibles (~3 500) du sous-type C du VIH-1, responsable de près de 50% des infections mondiales au VIH-1, pour estimer ses principaux flux migratoires à l'échelle mondiale, ainsi que son origine géographique. Des outils novateurs, basés sur le principe de parcimonie combiné avec différents critères statistiques, sont utilisés afin de synthétiser et interpréter l'information contenue dans une grande phylogénie représentant l'ensemble des séquences étudiées. Enfin, l'origine géographique et temporelle de ce variant (VIH-1 C) au Sénégal est précisément explorée lors d'une seconde étude, portant notamment sur les hommes ayant des rapports sexuels avec des hommes.

MOTS-CLEFS : Moindres carrés, optimisation, estimation statistique, horloge moléculaire, taux de substitution, épidémiologie moléculaire, origine du VIH-1 sous-type C.

TITLE: Evolution of HIV: methods, models and algorithms

ABSTRACT

Nucleotide sequences data enable the inference of phylogenetic trees, or phylogenies, describing their evolutionary relationships during evolution. Combining these sequences with their sampling date or country of origin, allows inferring the temporal or spatial localization of their common ancestors. These data and methods are widely used with viral sequences, and particularly with human immunodeficiency virus (HIV), to trace the viral epidemic history over time and throughout the globe. Using sequences sampled at different points in time (or heterochronous) is also a mean to estimate their substitution rate, which characterizes the speed of evolution.

The most commonly used methods to achieve these tasks are accurate, but are computationally heavy since they are based on complex models, and can only handle few hundreds of sequences. With an increasing number of sequences available in the databases, often several thousand for a given study, the development of fast and accurate methods becomes essential. Here, we present a new distance-based method, named Ultrametric Least Squares, which is based on the principle of least squares (very popular in phylogenetics) to estimate the substitution rate of a set of heterochronous sequences and the dates of their most recent common ancestors. We demonstrate that the criterion to be optimized is piecewise parabolic, and provide an efficient algorithm to find the global minimum.

Using sequences sampled at different locations also helps to trace transmission chains of an epidemic. In this respect, we used all available sequences (~3,500) of HIV-1 subtype C, responsible for nearly 50% of global HIV-1 infections, to estimate its major migratory flows on a worldwide scale and its geographic origin. Innovative tools, based on the principle of parsimony, combined with several statistical criteria were used to synthesize and interpret information in a large phylogeny representing all the studied sequences. Finally, the temporal and geographical origins of the HIV-1 subtype C in Senegal were further explored and more specifically for men who have sex with men.

KEY WORDS: Least squares, optimization, statistical estimation, molecular clock, substitution rate, molecular epidemiology, origin of HIV-1 subtype C.

DISCIPLINE

Bioinformatique

LABORATOIRES

Laboratoire d'Informatique de Robotique et de Microélectronique de Montpellier (LIRMM)
161 rue ADA, 34392 Montpellier cedex 5, France
Institut de Recherche pour le Développement (IRD)
911 avenue Agropolis, BP 64501, 34394 Montpellier cedex 5, France

Remerciements

Je tiens d'abord à remercier Monsieur Alain GUÉNOCHE, directeur de recherche au CNRS, Madame Marie-Laure CHAIX, maître de conférences à l'Université Paris Descartes, et Monsieur Denis FARGETTE, directeur de recherche à l'IRD, pour l'honneur qu'ils me font d'avoir accepté de juger mes travaux de thèse.

Je remercie chaleureusement mes directeurs de thèse Monsieur Olivier GASCUEL et Madame Martine PEETERS pour m'avoir encadré et soutenu dans cette thèse. Je ne saurais trop leur dire à quel point je suis reconnaissant envers eux pour leur gentillesse, leur patience, leur bienveillance, leurs conseils et leurs nombreux encouragements.

Je remercie tout autant François CHEVENET, Denis FARGETTE, Thu Hien TO et Nicole VIDAL qui ont travaillé avec moi sur certains projets, su répondre à mes questions, et qui m'ont aidé, chacun à leur façon, à améliorer la qualité de mon travail par des regards critiques et constructifs.

J'adresse une pensé particulière pour les membres et les anciens membres des équipes « Méthodes et Algorithmes pour la Bioinformatique » et « Diversité génétique du VIH ; émergence des rétrovirus et autres pathogènes » que j'ai côtoyé quotidiennement et avec lesquels j'ai passé de bons moments. Je n'oublie pas les membres de l'équipe « Recherche Opérationnelle » qui m'ont soutenu au moment de la rédaction de ce tapuscrit.

Enfin, je tiens également à remercier les personnes qui ont contribué à l'élaboration de ce mémoire, dont, une nouvelle fois, mes directeurs de thèse Olivier GASCUEL et Martine PEETERS, mais aussi Thu Hien TO, François CHEVENET, Aurélie SCHAETZEL et pour avoir comblé mes lacunes en anglais Lucie ÉTIENNE et Fabio PARDI. Enfin, un grand merci à Nicole VIDAL pour avoir contribué, plus que sa part, à l'écriture de certaines parties et à la relecture intensive de ce mémoire.

Table des matières

Avant-propos

Cette thèse pluridisciplinaire a été co-financée par l'Université Montpellier 2 et la Région Languedoc-Roussillon, puis sur fonds propres par les équipes « Méthodes et Algorithmes pour la Bioinformatique » (MAB) et « Diversité génétique du VIH ; émergence des rétrovirus et autres pathogènes » dont j'ai fait partie.

Durant la première année, ma thèse s'est déroulée au Laboratoire d'Informatique, de Robotique et de Microélectronique de Montpellier (LIRMM) sous la direction d'Olivier GASCUEL, responsable de l'équipe MAB, où j'ai développé mes connaissances théoriques en phylogénie moléculaire parallèlement au développement de la méthode de distances *Ultrametic Least Squares* (ULS) et au développement d'autres méthodes de distances d'estimation de taux de substitution (sUPGMA, TREBLE, *Pairwise-Distance*, etc.).

Au début de la seconde année, et sous la co-direction de Martine PEETERS, responsable de l'équipe « Diversité génétique du VIH », je me suis initié à la problématique du VIH à l'Institut de Recherche pour le Développement (IRD), où Nicole VIDAL m'a appris à manier les outils bioinformatiques régulièrement utilisés par les membres de l'équipe (Clustal W, NJPlot, PhyML, etc.). Après cela, ma thèse s'est principalement déroulée à mi-temps entre le LIRMM et l'IRD où nous avons mis en place un projet commun visant à retracer l'histoire épidémiologique (géographique et temporelle) du sous-type C du VIH-1 en Afrique, et simultanément celle des hommes ayant des rapports sexuels avec des hommes au Sénégal. L'utilisation d'un outil, PhyloType, développé par François CHEVENET, et pour lequel j'ai apporté ma contribution par des regards critiques et par le développement d'un logiciel permettant d'enraciner les phylogénies de différentes manières, a permis de retracer les principaux flux de cette épidémie en Afrique. L'origine temporelle et les flux migratoires de cette épidémie ont aussi été recherchés (sans résultat probant) avec le logiciel bayésien BEAST, sur lequel Denis FARGETTE a répondu à mes nombreuses questions et interrogations.

Au début de ma troisième année, nous avons donc décidé de scinder cette étude en nous focalisant, d'abord sur l'origine géographique et temporelle du sous-type C du VIH-1 au Sénégal à l'aide d'outils déjà publiés, mais en limitant le nombre de séquences étudiées à cause d'un temps de calcul prohibitif dû à certains logiciels utilisés, comme BEAST par exemple ; ces travaux ont fait l'objet d'une publication dans le journal PLoS One (Jung *et al*, 2012). Dans un second temps, nous nous sommes intéressés à retracer l'histoire épidémiologique de ce variant au niveau mondial, mais en utilisant des outils informatiques demandant peu de temps de calcul, comme ceux basés sur le principe de parcimonie. Lors de cette dernière étude nous avons développé et adapté divers indices permettant de synthétiser l'information contenue dans de grandes phylogénies.

Enfin, à la fin de ma troisième année, avec l'arrivée au LIRMM d'une post-doctorante, Thu Hien TO, travaillant aussi sur les méthodes de distances d'estimation de taux de substitution et de datation (en supposant un taux par branche), nous avons finalisé le jeu de données simulées à partir duquel sont obtenus les résultats de la méthode ULS présentés dans cette thèse.

Introduction

Depuis la découverte du virus de l'immunodéficience humaine (VIH) en 1983, il y a près de 30 ans, la recherche fondamentale et la recherche clinique ont contribué à la compréhension de la biologie du virus, des interactions virus-hôte, de la diversité génétique du virus ainsi qu'à l'élaboration de stratégies thérapeutiques efficaces, malgré l'absence regrettée d'un vaccin préventif ou d'un traitement éradiquant totalement le virus chez une personne infectée (Barré-Sinoussi, 2010; Wainberg & Jeang, 2008; Weiss, 2008; Gallo, 2006). Une des difficultés majeures à l'élaboration d'une médication universelle est la capacité du virus à échapper très rapidement aux pressions immunitaires ou pharmacologiques par la production d'une multitude de variants génétiquement distincts. De ce fait, le VIH présente non seulement une grande diversité génétique inter-hôte, mais aussi intra-hôte (Taylor *et al*, 2008). Certains projets scientifiques s'attachent d'ailleurs à surveiller l'apparition de nouveaux variants génétiques, ou de nouvelles mutations de résistance aux thérapies antirétrovirales, au sein d'une population ou dans une région géographique donnée, afin d'adapter au mieux les traitements et les tests de dépistage (Hemelaar *et al*, 2011). Pour ce faire, ces études doivent systématiquement séquencer le génome du virus, ou une portion de celui-ci, et analyser les séquences obtenues à l'aide d'outils informatiques. À la fin de ces projets, les séquences sont généralement mises à disposition du public et peuvent donc être récupérées dans des bases de données biologiques, notamment dans celle maintenue par le laboratoire national de Los Alamos, spécifique au VIH. Ces séquences génétiques, éparpillées dans le temps et l'espace au moment de leur prélèvement, sont une véritable source d'information pour les études d'épidémiologie moléculaire.

Les séquences échantillonnées à des moments différents peuvent servir à estimer la vitesse évolutive du VIH. De nombreuses applications biologiques en découlent comme, par exemple, la reconnaissance de gènes devant être ciblés par les traitements antirétroviraux. En effet, les gènes conservés sont essentiels au cycle réplicatif viral et ont donc une vitesse évolutive plus faible par rapport à celle des autres gènes. L'estimation de cette vitesse évolutive est aussi nécessaire pour dater l'origine d'une épidémie locale ou mondiale. C'est dans ce contexte que Korber *et al*. (2000), et

d'autres, ont estimé la date de l'ancêtre commun aux souches du VIH responsables de la pandémie actuelle au début du xx^{e} siècle. Pour que cela soit possible, il faut toutefois que les séquences génétiques entre deux temps de collecte présentent une accumulation significative de mutations (Drummond *et al*, 2003b) et, dans ce cas, ce procédé ne peut être appliqué aux organismes évolués, comme l'homme (excepté avec de l'ADN ancien). Les méthodes généralement utilisées pour ce genre d'étude sont précises, mais lourdes en temps de calcul ; elles ne permettent donc pas de traiter simultanément un grand nombre de séquences nucléotidiques (au plus quelques centaines) et ne considèrent donc seulement qu'une partie de l'information disponible. Mais devant le nombre croissant de telles séquences nucléotidiques, le besoin d'une méthode rapide et efficace, pouvant traiter un grand nombre de séquences, se fait sentir. Dans cette thèse, nous proposons une méthode de distances (approche alternative aux méthodes probabilistes couramment utilisées), *Ultrametric Least Squares*, qui permet d'estimer la vitesse évolutive à partir d'un ensemble important (plusieurs milliers voire dizaine de milliers) de séquences échantillonnées dans le temps.

Les séquences échantillonnées en des lieux différents peuvent fournir d'autres informations, comme, par exemple, la région géographique à l'origine de la diffusion d'une épidémie au niveau local ou mondial (Holmes, 2008, 2004). Cela ne présente pas qu'un intérêt documentaire, puisque cette donnée peut aider à comprendre la manière dont une nouvelle épidémie a émergé au sein de populations données, en observant leurs coutumes et leur environnement, et d'y apporter des solutions afin d'éviter d'autres émergences. À l'aide de ces séquences, il est aussi possible d'identifier les flux migratoires du virus ainsi que les chaînes de transmission majeures ou mineures (par exemple, de quel pays vers quel pays, de quelle population vers quelle population ou de quel individu vers quel individu). Les études moléculaires de ce genre sont très nombreuses et existent pour la plupart des variants génétiques principaux ou secondaires du VIH (Chen *et al*, 2011; Faria *et al*, 2011; Shen *et al*, 2011; Véras *et al*, 2011a). Nous y apportons une contribution supplémentaire dans cette thèse, avec une étude ayant pour objectif de déterminer l'origine géographique et temporelle de l'épidémie du sous-type C (variant génétique responsable de près de 50% des infections mondiales au VIH de type 1) au Sénégal, à l'aide d'outils phylogénétiques classiquement utilisés. Une seconde contribution, plus novatrice, est faite lors d'une étude qui vise à déterminer les principaux flux migratoires de ce variant à l'échelle mondiale, ainsi que son origine géographique, à l'aide de nouveaux outils informatiques développés pour l'occasion.

Plan de la thèse

Cette thèse est composée de six chapitres et une annexe. Les trois premiers chapitres présentent les connaissances nécessaires à la compréhension des trois derniers chapitres qui eux décrivent les travaux effectués au cours de la thèse.

Le premier chapitre est une introduction à la phylogénie moléculaire. Les concepts de base, ainsi que les méthodes et algorithmes classiquement utilisés afin d'inférer une phylogénie à partir de données moléculaires, y sont décrits. Nous y présentons les bases de données biologiques, essentielles aux études moléculaires, et particulièrement la base de données du laboratoire national de Los Alamos, spécifique aux VIH et SIV, à partir desquelles on peut récupérer des séquences nucléotidiques. Puis nous discutons de l'étape d'alignement qui consiste à positionner chaque nucléotide d'un ensemble de séquences homologues, dérivant d'un même nucléotide ancestral, en regard les uns des autres. Cette étape est le fondement de toutes analyses de phylogénie moléculaire, et de sa justesse dépendent fortement les méthodes d'inférence phylogénétique. De là, nous distinguons deux catégories de méthodes d'inférence phylogénique, comme les méthodes de caractères, basées sur l'alignement, où l'on retrouve les principes de parcimonie et probabiliste (vraisemblance et bayésien, par exemple PhyML et MrBayes) et les méthodes de distances (UPGMA, NJ, FastME, etc.) qui elles se basent sur une matrice de distances, contenant les distances évolutives entre paires de séquences. Ces méthodes, brièvement exposées, utilisent des modèles d'évolution (comme GTR, HKY ou F84), afin d'estimer, au mieux, la distance évolutive qui sépare les séquences depuis leur divergence de leur ancêtre commun, ainsi que des méthodes statistiques afin d'évaluer la fiabilité des reconstructions phylogénétiques (*bootstrap*, aLRT, etc.). Enfin, nous présentons des algorithmes de parcimonie (ACCTRAN, DELTRAN et DOWNPASS) qui permettent, à partir d'une phylogénie enracinée et d'annotations associées aux feuilles, de reconstruire les annotations ancestrales, ainsi que la méthode du *shuffling* qui permet de tester la significativité statistique des résultats.

Le deuxième chapitre présente différentes méthodes de distances qui permettent d'estimer la vitesse évolutive, ou taux de substitution, à partir de données moléculaires échantillonnées dans le temps (hétérochrones). Ces méthodes font les hypothèses du modèle *Single Rate Dated Tips* (SRDT), à taux de substitution constant (à travers le temps) et uniforme (à travers les lignées), mais d'autres modèles d'horloges moléculaires sont aussi exposés, comme les modèles *Multiple Rates Dated Tips* (MRDT), où les taux de substitution varient entre intervalles de temps, ou *Different Rate* (DR) qui suppose un taux de substitution différent par branche. Les méthodes sUPGMA, une approche par moindres carrés, et TREBLE, une approche par triplets, y sont décrites, ainsi que les régressions linaires *Pairwise-Distance* et *Root-to-tip* ; cette dernière est souvent utilisée pour sa simplicité, sa rapidité et sa capacité à estimer en même temps la date de l'ancêtre commun. Nous présentons aussi

deux autres méthodes de distances qui font des hypothèses supplémentaires à celles du modèle SRDT, comme la méthode Langley-Fitch, qui considère un arbre enraciné et la méthode *TreeRate*, qui nécessite l'intervention de l'utilisateur afin de définir deux groupes de séquences à partir desquels le taux est estimé. Enfin, nous présentons rapidement deux méthodes probabilistes, estimant toujours le taux de substitution sous le modèle SRDT, *TipDate*, une méthode de maximum de vraisemblance, et BEAST, une méthode bayésienne qui est actuellement la référence dans le domaine, principalement à cause des multiples possibilités qu'elle offre.

Le troisième chapitre fait un point sur l'épidémiologie moléculaire et les origines zoonotiques du VIH. La nomenclature associée aux différents variants génétiques du VIH (groupe, sous-type, sous-sous-type, forme recombinante circulante [CRF] ou unique [URF]) est exposée, ainsi que sa diversité génétique. En effet, il existe deux types de VIH (VIH-1 et VIH-2), quatre groupes pour le VIH-1 (M, N, O et P) et huit groupes pour le VIH-2 (A à H). Ces groupes sont à chaque fois le résultat d'une transmission inter-espèce d'un virus infectant les singes d'Afrique à l'homme (anthropozoonose), mais seul le groupe M du VIH-1 est responsable de la pandémie actuelle. La distribution géographique dans le monde entier des souches du groupe M, présentant une grande diversité génétique (9 sous-types, A à D, F à H, J et K, et 51 CRF), ainsi que celle des autres variants génétiques sont discutées. Les origines géographiques et temporelles des différents groupes du VIH sont exposées (par exemple, l'épidémie du groupe M est datée au début du XX^{e} siècle et son réservoir se situe au sud-est du Cameroun, bien que son épicentre soit en République Démocratique du Congo), ainsi que, brièvement, les facteurs probables à l'origine de ces transmissions inter-espèce (consommation de viande de brousse, domestication des singes, etc.). Enfin, nous présentons les causes biologiques (sélection naturelle, multiplication rapide, etc.) et les conséquences de cette diversité génétique (tests de diagnostic, médications, vaccin, etc.), ainsi que les facteurs sociologiques liés à l'expansion de l'épidémie (guerre, mondialisation, tourisme, groupes à risque, etc.).

Le quatrième chapitre expose la méthode de distances *Ultrametric Least Squares* qui permet d'estimer la vitesse évolutive d'un gène ou d'un organisme, à partir d'un ensemble de séquences échantillonnées dans le temps et sous l'hypothèse d'une horloge moléculaire stricte. Pour cela, cette méthode corrige les distances génétiques par l'ajout d'un facteur correctif, proportionnel au taux de substitution à estimer, aux souches anciennes afin de les voir comme contemporaines. Puis elle minimise un critère basé sur le principe des moindres carrés (souvent utilisé en phylogénie avec les méthodes de distances) qui mesure l'ultramétricité d'une distance. Nous verrons que ce critère a le comportement d'une fonction parabolique par morceaux et proposons un algorithme efficace en $O(n^3 \log n)$, où n est le nombre de séquences, pour en trouver le minimum. Une méthode d'échantillonnage permet de borner cette complexité, et cela sans perte de précision. Cette méthode

est ensuite adaptée à l'estimation de plusieurs taux de substitution : un pour chaque intervalle de temps obtenu entre deux dates d'échantillonnage consécutives ou un par lignage (horloges moléculaires locales). La précision d'estimation de cette approche est ensuite comparée à celle des autres méthodes de distances (sUPGMA, TREBLE, *Root-to-tip* et *Pairwise-Distance*) et à celle de la méthode probabiliste de référence BEAST. Les résultats montrent, qu'en moyenne, ULS est la méthode la plus performante avec des matrices de distances ou des arbres FastME, mais est équivalente à la régression linéaire *Root-to-tip* sur des arbres PhyML. Enfin, la confrontation avec BEAST montre que la méthode ULS est meilleure ou équivalente à BEAST.

Le cinquième chapitre présente le résumé détaillé d'une étude, sur l'origine géographique et temporelle du sous-type C du VIH-1 au Sénégal, particulièrement chez les hommes ayant des rapports sexuels avec des hommes (MSM), à partir d'outils phylogénétiques déjà publiés ; cette étude a fait l'objet d'une publication dans le journal PLoS One (l'article en anglais est joint à la suite du chapitre). Une grande phylogénie (3 081 séquences) construite avec PhyML, nous a permis d'identifier les séquences épidémiologiquement proches de celles du Sénégal. Puis, un second arbre de maximum de vraisemblance (PhyML) et un arbre bayésien (MrBayes) sont calculés mais uniquement avec les séquences du Sénégal et celles identifiées comme proches. Ces derniers arbres montrent de multiples introductions de ce variant, provenant de l'Afrique australe et de l'Afrique de l'est, au sein de la population générale sénégalaise. Les souches isolées chez les MSM forment un cluster, suggérant une introduction unique de ce variant, suivie d'une diffusion efficace (évènement fondateur), provenant de l'Afrique australe (probablement de Zambie). Les analyses temporelles sont faites avec le logiciel BEAST, sous différents modèles d'horloges moléculaires, et datent l'ancêtre commun aux souches de la population générale au début des années soixante-dix, et celui des MSM environ dix ans après, au début des années quatre-vingt.

Le sixième chapitre explique les méthodes que nous avons développé et auxquelles nous avons contribué, afin de décrire les flux migratoires mondiaux et l'origine géographique de l'épidémie du sous-type C du VIH-1, en se basant sur une phylogénie (PhyML) comprenant plus de 3 600 souches, et sur la donnée des pays de collecte associés à chaque séquence. Trois indices, basés sur les transitions entre pays, obtenues par parcimonie, et associés à des sorties graphiques appropriées, permettent de synthétiser l'information contenue dans la phylogénie. Un de ces indices mesure la dispersion des feuilles associées au même pays dans la phylogénie (forment-elles un clade ? sont-elles regroupées ? sont-elles éparpillées ?). Cette mesure peut être décomposée en une normalisation du nombre de transitions entrantes et sortantes. Les deux autres indices mesurent les flux migratoires (de quel pays ? vers quel pays ?) et la symétrie des échanges (y a-t-il autant de transitions d'un pays

donné vers un autre que l'inverse ? le flux est-il unidirectionnel ?). Ensuite nous avons utilisé le logiciel PhyloType, auquel j'ai contribué pour une part, afin de retracer les chaînes de transmission majeures du sous-type C du VIH-1, reflets d'évènements fondateurs probables. De cette étude, nous avons pu identifier la Zambie comme étant l'épicentre de l'épidémie du sous-type C du VIH-1. Nous avons aussi identifié les principaux flux migratoires déjà connus (comme le lien épidémiologique entre le Brésil et le Burundi, celui entre l'Éthiopie et l'Israël, etc.), d'autres nouveaux (comme deux introductions d'origine géographique différente en Éthiopie, probablement le résultat de l'observation du sous-cluster C'), ainsi que certaines contradictions avec des résultats déjà publiés (comme le lien épidémiologique entre l'Inde et l'Afrique du Sud que nous observons plutôt avec la Zambie). L'Annexe A contient des résultats supplémentaires concernant cette étude.

Enfin, une conclusion générale rappelle les principaux apports scientifiques de ces travaux de thèse, ainsi que les perspectives ouvertes.

Chapitre 1

Bagage de phylogénie moléculaire

Nous discutons brièvement des concepts de base de la phylogénie moléculaire. Les bases de données biologiques qui mettent à disposition les séquences nucléotidiques, et notamment la base de données sur le VIH du laboratoire national de Los Alamos, sont présentées. Puis nous discutons de l'étape d'alignement, essentielle à toutes analyses de phylogénie moléculaire, qui consiste à mettre en regard les nucléotides de chaque séquence homologue, dérivant d'un même nucléotide ancestral. Nous présentons ensuite, les principaux modèles d'évolution nucléotidique (GTR, HKY, F84, etc.) permettant d'estimer la distance évolutive qui sépare les séquences depuis leur divergence de leur ancêtre commun. Les méthodes de distances d'inférence phylogénétique, comme UPGMA, NJ ou FastME sont rapidement exposées, tout comme les méthodes de parcimonie et les méthodes probabilistes (PhyML, MrBayes). Enfin, nous présentons les algorithmes de parcimonie ACCTRAN, DELTRAN et DOWNPASS qui permettent d'inférer les annotations ancestrales d'une phylogénie enracinée à partir d'annotations aux feuilles. La méthode du shuffling *qui permet d'en dégager la significativité statistique est aussi présentée.*

Sommaire

1.1 Introduction

La phylogénie est une discipline scientifique qui étudie les « parentés entre différents êtres vivants en vue de comprendre l'évolution des organismes vivants »[1]. Les premières phylogénies (Charles DARWIN, 1809-1882 ; Ernest HAECKEL, 1834-1919) se basaient sur des caractères morphologiques, anatomiques et/ou physiologiques afin de comparer les organismes vivants et d'étudier leur parenté. Mais lorsqu'il s'agit de comparer des organismes bactériens ou viraux ces critères de comparaison atteignent leur limite.

Depuis le développement de la biologie moléculaire et la découverte de l'ADN (acide désoxyribonucléique) comme support de l'hérédité dans les années cinquante, de nouveaux caractères sont utilisés comme source d'information pour l'inférence de phylogénies : les séquences de macromolécules (ADN, ARN et protéines). Les premières études phylogénétiques essentiellement basées sur des séquences protéiques remontent au début des années soixante et donnent ainsi naissance à une nouvelle branche de la phylogénie : la phylogénie moléculaire. Mais ce n'est que vers la fin des années soixante-dix, avec le développement de techniques spécifiques permettant de séquencer des fragments d'ADN à grande échelle et à faible coût que la phylogénie moléculaire connait un essor grandissant. En particulier parce que cette discipline est très utilisée en génomique fonctionnelle, science qui étudie le rôle des gènes.

La phylogénie moléculaire est aussi très utilisée par les épidémiologistes car elle permet de mettre en évidence des liens entre différentes souches virales, liens qui reflètent des chaînes de transmission. Un exemple souvent cité car c'est le premier qui utilise des outils de phylogénie moléculaire dans un cadre médico-légal, est celui d'un dentiste de Floride, séropositif, qui est suspecté être la source de contamination de quelques uns de ses patients (Ou *et al*, 1992). Les indices ayant menés à cette hypothèse proviennent d'une patiente atteinte du syndrome de l'immunodéficience acquise (SIDA) mais pour laquelle aucune situation de contamination n'a pu clairement être identifiée, hormis deux interventions chirurgicales venant de son dentiste. Pour confirmer un éventuel lien épidémiologique, des souches virales ont été prélevées chez le dentiste, chez la patiente, ainsi que chez six autres patients qui ont séroconverti pendant l'enquête ; par ailleurs, trente-cinq souches virales provenant d'individus locaux ont été rajoutées comme souches témoins. L'analyse phylogénétique de toutes ces souches virales a révélé que la souche collectée chez le dentiste est phylogénétiquement très proche de celles collectées chez ses patients, confirmant ainsi la source de contamination. Mais le mode de contamination reste indéterminé. De nombreux autres exemples comme celui-là sont disponibles dans la littérature, Leitner et Fitch (1999) en commentent d'autres.

[1] Source Wikipédia.

Dans ce chapitre, nous présentons brièvement les différentes méthodes d'inférence phylogénétique. Mais avant cela, nous présentons les bases de données biologiques, véritables sources d'information pour les études moléculaires, puis l'étape d'alignement, fondamentale à toute analyse phylogénétique. Enfin, nous terminerons ce chapitre par l'exposé de quelques méthodes de parcimonie permettant de reconstruire les annotations ancestrales (par exemple des régions géographiques) à partir d'une phylogénie et des annotations associées aux feuilles de cette phylogénie qui représentent les souches virales de l'alignement. Des compléments d'information peuvent être trouvés dans les ouvrages de Lemey *et al*. (2009b) ou celui de Felsenstein (2003).

1.2 Bases de données biologiques

Les études de phylogénie moléculaire sont souvent basées sur des séquences nucléotidiques. Pour être facilement accessibles, et pour faciliter le traitement de l'information, les séquences nucléotidiques obtenues par les biologistes sont stockées dans des bases de données. Ces bases de données fournissent aussi une pléthore d'outils pour manipuler ou analyser les séquences, mais aussi des informations supplémentaires sur chacune d'elles. Ces informations, ou annotations, sont très utiles car elles renseignent sur l'organisme de collecte, les propriétés de la séquence, les auteurs, etc., permettant ainsi de cibler les recherches dans ces bases.

Il existe de nombreuses bases de données biologiques mais la plupart sont spécifiques à un organisme, une fonction, etc. Toutefois, il existe trois bases de données principales :

- EMBL-Bank (*European Molecular Biology Laboratory*), maintenue par EMBL-EBI (*European Bioinformatics Institute*) à Hinxton au Royaume-Uni ;
- GenBank, maintenue par NCBI (*National Center for Biotechnology Information*) à Bethesda aux États-Unis ;
- DDBJ (*DNA Data Bank of Japan*), maintenue par NIG/CIB (*National Institute of Genetics*, *Center for Information Biology*) à Mishima au Japon.

Ces trois bases de données collaborent ensemble afin de partager les nouvelles soumissions ou les éventuelles mises à jour. L'ensemble des séquences nucléotidiques publiées y est donc accessible. Chaque séquence soumise se voit attribuer un numéro d'accession unique (qui reste le même quelle que soit la base de données) et qui permet de désigner, sans ambiguïté, les séquences dans la littérature. Par convention, les séquences nucléotidiques sont stockées sous le format de l'ADN, mais les bases de données contiennent aussi des séquences d'ARN (acide ribonucléique). Dans ce cas, ces dernières sont codées avec un « T », qui signifie la thymine, à la place d'un « U », pour designer l'uracile.

Dans nos études, nous utilisons la base de données spécifique au VIH maintenue par le laboratoire national de Los Alamos : *HIV Databases* (www.hiv.lanl.gov). Elle met à disposition un grand nombre de séquences nucléotidiques du VIH de type 1 (VIH-1), du VIH de type 2 (VIH-2) et même du SIV (*simian immunodeficency virus*), virus analogue au VIH mais infectant naturellement les singes d'Afrique. Mise à jour périodiquement, elle contient toutes les séquences soumises dans GenBank, avec un décalage de quelques mois sur les dernières entrées de GenBank. En revanche, les séquences sont annotées avec plus d'informations que celles disponibles via GenBank, comme l'origine géographique de collecte, l'année d'isolation, le sous-type d'appartenance, le groupe à risque de l'individu chez lequel elle est prélevée, etc. Ces informations sont récupérées dans les publications correspondantes aux séquences par les gestionnaires de la base de données. De plus, le site internet propose une interface de recherche conviviale, ergonomique et adaptée aux particularités du VIH et du SIV. Il est ainsi très facile d'obtenir des séquences sur une région précise du génome, provenant d'un même pays ou d'un même continent, isolées chez un patient avec un facteur à risque particulier, etc. Des outils sont aussi mis à disposition et permettent le traitement spécifique de séquences du VIH/SIV, comme, par exemple, *Sequence Locator* qui permet de retrouver les coordonnées de début et de fin d'une séquence sur le génome de référence (HXB2 pour le VIH et SIVmm239 pour le SIV).

Malgré le soin apporté au classement et au référencement des séquences, ces bases de données peuvent contenir des informations erronées. Il revient à l'utilisateur de vérifier la justesse des informations.

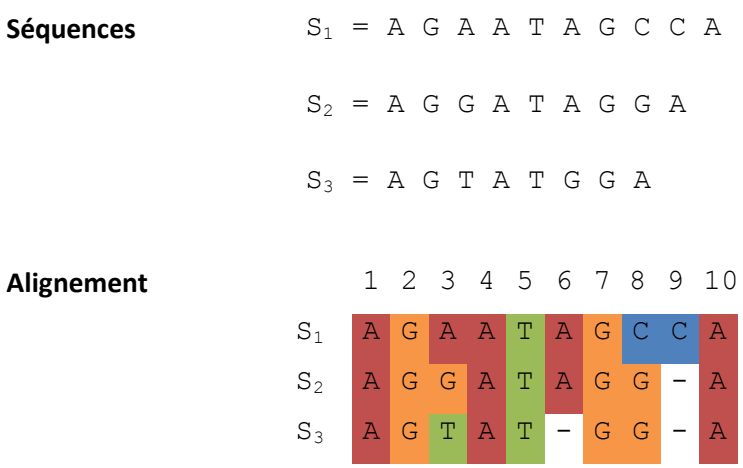
1.3 L'alignement, une étape indispensable

L'alignement de séquences nucléotidiques est une étape clef des études de phylogénie moléculaire. Cette étape ne peut se faire qu'avec des séquences homologues, c'est-à-dire des séquences nucléotidiques partageant un même ancêtre commun, puisqu'elle consiste à identifier, pour chaque séquence, les nucléotides dérivant du même nucléotide ancestral et à les positionner en regard. Le résultat de cette étape est l'obtention d'une matrice, appelée alignement, où chaque ligne correspond à une séquence et où chaque colonne, appelée site, contient les nucléotides dérivés d'un même nucléotide ancestral (Figure 1).

Dans certaines séquences de l'alignement des gaps (ou *indels*) ont pu être introduits. Ils correspondent aux phénomènes biologiques d'insertions (ajout d'un ou de plusieurs nucléotides) ou de délétions (perte d'un ou de plusieurs nucléotides) qui se sont produits au cours de l'évolution. Toutefois, l'utilisation de gaps dans un alignement doit être faite avec parcimonie. Ainsi, un bon alignement est défini comme un alignement qui contient le moins d'évènements de mutation possibles,

avec des pondérations différentes pour les différents évènements mutationnels (substitution, insertion, délétion, ouverture de gap, prolongation de gap, etc.).

Figure 1. Exemple d'alignement de séquences.
L'alignement du bas est un alignement possible résultant des trois séquences du haut. Les positions 1, 2, 4, 5, 7 et 10 ne présentent aucune modification. La position 3 présente deux substituions et la position 8 une substitution pour la séquence S_1. La position 6 présente une délétion pour la séquence S_3 et la position 9 une insertion pour la séquences S_1. D'autres interprétations de l'alignement sont possibles mais elles impliquent davantage d'évènements de mutation.
L'exemple est extrait de Caraux et al. (1995).



Comme l'alignement est la base de toutes méthodes de phylogénie moléculaire, il est indispensable d'avoir un alignement d'une qualité optimale afin d'inférer des phylogénies fiables. Dans le cas contraire, elles peuvent contenir des erreurs ou être aberrantes. C'est pour cela que les biologistes ôtent de l'alignement les sites les plus incertains, comme ceux contenant des gaps ou les parties trop divergentes (souvent en début ou en fin de l'alignement).

Des méthodes automatisées existent pour résoudre des alignements. La plus simple concerne l'alignement entre deux séquences en se basant sur la distance d'édition (ou distance de Levenshtein). Cette distance mesure la similarité entre deux mots. Pour cela, elle calcule le nombre minimum de remplacements (ou substitutions), de délétions ou d'insertions nécessaires pour transformer un mot en l'autre. Rappelons que les séquences nucléotidiques peuvent être vues comme des mots sur l'alphabet génétique $\mathcal{A} = \{A, C, G, T\}$. Un algorithme simple de programmation quadratique permet de calculer la distance d'édition en $O(n \times m)$, où n et m sont les longueurs respectives des deux séquences. Néanmoins cet algorithme calcule uniquement la distance (ou le score) de l'alignement optimal. Un algorithme supplémentaire est nécessaire afin d'en déduire l'alignement, il se fait en $O(n + m)$ en réutilisant le tableau construit lors du calcul de la distance d'édition. Lorsque l'on souhaite aligner plus de deux séquences simultanément, le problème devient très vite complexe. Il est bien sûr possible d'adapter l'algorithme précédent dans le cas de plusieurs séquences, mais la complexité devient alors exponentielle sur le nombre de séquences, et l'application sur plus de

quatre ou cinq séquences est inenvisageable. Pour contrer ce problème, des heuristiques sont proposées mais elles ne permettent pas de résoudre avec exactitude le problème de l'alignement. Les biologistes utilisent donc ces heuristiques afin d'obtenir une base convenable de l'alignement, puis le modifient manuellement avec des logiciels d'éditions.

De nombreux programmes sont disponibles pour résoudre le problème d'alignement multiple de séquences. Une liste exhaustive est trouvée dans Lemey *et al*. (2009b). Dans nos études, seul le logiciel MAFFT (Katoh *et al*, 2005) est utilisé car il a été démontré qu'il est l'un des plus performants (Thompson *et al*, 2011).

1.4 Modèles d'évolution moléculaire

La distance évolutive entre deux séquences nucléotidiques est définie comme « le nombre moyen de substitutions par site s'étant produites depuis que ces séquences ont divergé de leur ancêtre commun » (Perrière & Brochier-Armanet, 2010). Pour calculer la distance évolutive qui sépare deux séquences dans l'alignement, une approche simpliste consisterait à compter le nombre de dissemblances (c'est-à-dire le nombre de sites différents) et de le diviser par la longueur de l'alignement. Cette distance évolutive est appelée p-distance (exprimée en substitutions par site) et correspond à la distance observée entre les deux séquences, et non à la distance évolutive réelle. En effet, imaginons qu'entre deux séquences données, et sur un site donné, les nucléotides A et G sont observés. La p-distance comptabilise une substitution, car c'est ce qui est observé. Mais si la base A est remplacée par la base T, puis par la base G, il y a eu deux évènements de substitutions réelles, mais toujours une substitution observée. Donc la p-distance sous-estime la distance évolutive réelle, puisque de nombreuses substitutions cachées ont pu se produire.

Des modèles d'évolution sont donc proposés pour estimer au mieux la distance évolutive réelle à partir de la distance évolutive observée. Ces modèles font les hypothèses simplificatrices suivantes :

- les séquences évoluent uniquement avec un processus de substitution nucléotidique, c'est-à-dire que les évènements d'insertion et de délétion ne sont pas pris en compte ;
- les sites de l'alignement sont indépendants les uns des autres, c'est-à-dire que les évènements évolutifs d'un site n'ont aucune influence et ne sont pas influencés par les évènements évolutifs des autres sites de l'alignement ;
- le processus d'évolution est markovien d'ordre 1, c'est-à-dire que l'état futur d'un site ne dépend que de son état actuel et non des états passés précédents ;
- le processus d'évolution est identiquement distribué, c'est-à-dire qu'il est le même quel que soit le site de l'alignement ;

- le processus d'évolution est homogène, c'est-à-dire qu'il ne varie pas au cours du temps, il est donc applicable à toutes les branches de la phylogénie ;
- le processus d'évolution est stationnaire, c'est-à-dire que les probabilités d'observer une base particulière sont celles attendues à l'état d'équilibre (atteint lorsque les séquences ont évolué après un temps infini). Ces probabilités sont donc les mêmes pour toutes les séquences de l'alignement, puisqu'après un temps infini les séquences sont supposées avoir une composition en base identique ;
- au plus une mutation peut se produire dans un temps infinitésimal, c'est-à-dire qu'il ne peut y avoir plus d'une substitution simultanément.

Même si ces hypothèses simplifient fortement les modèles d'évolution, les résultats obtenus sont jugés largement acceptables par les biologistes.

Plusieurs modèles sont proposés afin de simuler le processus d'évolution. Chacun fait des hypothèses différentes en ce qui concerne les fréquences d'apparition des nucléotides et les taux de substitution (probabilité de passer d'un nucléotide i vers un nucléotide j). Le modèle le plus général est le modèle *general time reversible* (GTR) (Lanave *et al*, 1984) qui suppose une fréquence d'apparition différente pour chacun des quatre nucléotides et un taux de substitution relatif différent pour chacune des deux transitions (A ↔ T et C ↔ G) et des quatre transversions possibles (A ↔ C, A ↔ G, T ↔ C et T ↔ G) (dans les modèles réversibles comme GTR, le taux de substitution relatif [ou échangeabilité] de $i \rightarrow j$ est supposé le même que celui de $j \rightarrow i$). Comme il existe deux relations linéaires, une entre les fréquences d'apparition (somme à 1) des nucléotides et l'autre entre les taux de substitution (facteur de normalisation), le modèle GTR a huit paramètres libres (4 fréquences + 6 taux symétriques – 2 relations linéaires). Le modèle de Jukes et Cantor (1969), abrégé JC69, quant à lui, est le moins général. Il suppose que les fréquences d'apparition sont toutes égales et que les taux de substitution sont tous identiques. Il n'a donc aucun paramètre libre. La Figure 2 liste les modèles d'évolution les plus utilisés et les classe suivant leurs paramètres libres.

Les modèles d'évolution décrits ci-dessus supposent que tous les sites évoluent de façon identique, c'est-à-dire que les taux de substitution sont identiques quels que soit les sites de l'alignement. C'est une hypothèse fausse qui peut amener à des estimations biaisées si elle est fortement contredite par les données étudiées. Une alternative est d'ajouter un paramètre qui permet de faire varier les taux de substitution en fonction des sites suivant une loi gamma (Yang, 1994, 1993; Jin & Nei, 1990). Dans la littérature, l'ajout d'une loi gamma au modèle d'évolution considéré est noté +G ou +Γ. Généralement on n'est pas capable, pour des raisons mathématiques, d'utiliser la loi gamma continue standard (l'exception est le calcul des distances évolutives entre paires de sé-

quences, où cette loi est utilisable directement) ; on utilise alors une discrétisation de la loi gamma, souvent à 4, 6 ou 8 catégories, notée +Γ4, +Γ6 ou +Γ8.

Figure 2. Liste des modèles d'évolution.
Tableau récapitulant les différents modèles d'évolution généralement employés, organisés en fonction de leur supposition sur les différents paramètres. Les chiffres entre parenthèse indiquent le nombre de paramètres libres du modèle. Le modèle le plus général est le modèle GTR (Lanave *et al*, 1984) et le plus restreint JC69 (Jukes & Cantor, 1969). Les autres modèles (TN93 (Tamura & Nei, 1993), K2P (Kimura, 1980), HKY85 (Hasegawa *et al*, 1985) et F81 (Felsenstein, 1981)) sont des modèles intermédiaires. Le modèle K2P est parfois appelé K80. Le modèle HKY85 est aussi décrit par Felsenstein (1993) mais avec une formulation différente (d'où l'emploi des deux noms).

		Fréquence des nucléotides	
		Identique	**Différente**
Taux relatif (symétrique) de substitution	6 taux relatifs différents	pas utilisé	GTR (8)
	3 taux relatifs différents (transversions et les deux transitions)	pas utilisé	TN93 (5)
	2 taux différents (transitions contre transversions)	K2P (1)	HKY85 et F84 (4)
	1 taux (tous les taux sont égaux)	JC69 (0)	F81 (3)

Ces modèles d'évolutions supposent aussi que tous les sites de l'alignement sont variables, même lorsque le même nucléotide est présent sur chacune des séquences. Ces sites sont peut-être très conservés et évoluent donc à une vitesse nettement différente que celle des sites voisins. Pour supposer qu'une fraction de sites peut être invariable ou varier avec un taux de substitution très nettement inférieur à ceux des autres sites, un paramètre supplémentaire, qui mesure la proportion de sites invariants, est ajouté aux modèles d'évolution (Waddell & Penny, 1996; Gu *et al*, 1995). Dans les articles, les modèles rajoutant cette catégorie (potentielle) de sites invariants sont notés +I.

Pour les séquences codantes, il est parfois préférable d'utiliser des modèles d'évolution protéique (donc, à partir de séquences protéiques) plutôt que des modèles d'évolution nucléotidique. Les séquences protéiques sont plus conservées que les séquences nucléotidiques puisqu'elles ne prennent en compte que les substitutions non synonymes, et de ce fait, elles sont plus adaptées à la comparaison de séquences très divergentes. Les modèles JTT (Jones *et al*, 1992) et WAG (Whelan & Goldman, 2001) sont parmi les modèles d'évolution protéique les plus connus et les plus utilisés, mais il existe une variété de modèle d'évolution protéique, en raison des différentes pressions de sélection subies sur telle ou telle protéine. À cet effet, Dimmic *et al*. (2002) proposent le modèle rtREV qui est adapté à l'évolution et aux pressions de sélection de la transcriptase inverse des rétrovirus. Les modèles d'évolution protéique ne sont pas utilisés dans nos travaux puisque la région génomique utilisée (*pol*) reste très conservée à l'intérieur d'un même sous-type.

1.5 Méthodes d'inférence phylogénétique

Les méthodes d'inférence phylogénétique peuvent être divisées en deux catégories : les méthodes de distances et les méthodes de caractères qui comprennent les méthodes basées sur des modèles probabilistes d'évolution, cités ci-dessus. Les méthodes de distances ont pour base de calcul, non pas un alignement, mais une matrice contenant les distances évolutives entre paires de séquences, tandis que les méthodes de caractères se réfèrent constamment à l'alignement. Remarquons qu'en phylogénie moléculaire, les distances entre paires de séquences sont calculées à partir de l'alignement et en utilisant un modèle d'évolution. Mais ces distances peuvent très bien provenir d'une autre source d'information, comme, par exemple, des données morphologiques. Les méthodes de distances sont bien connues pour être rapides en temps de calcul, tandis que les méthodes probabilistes, plus lentes, sont généralement bien plus précises.

Avant d'aborder ces différentes méthodes, nous exposons la représentation d'une phylogénie.

1.5.1 Arbre phylogénétique

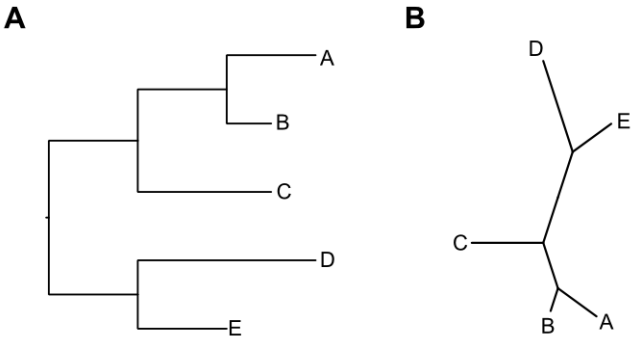
Une phylogénie est représentée graphiquement par un arbre, c'est-à-dire qu'elle est composée de nœuds internes, de nœuds externes (ou feuilles) et de branches. Les nœuds externes représentent les séquences étudiées, parfois appelées OTU (*operational taxonomic unit*), et les nœuds internes représentent des ancêtres communs hypothétiques, parfois appelées HTU (*hypothetical taxonomic unit*). Lorsque chaque nœud interne est adjacent à exactement trois branches, la phylogénie est dite binaire et les nœuds internes des bifurcations. La plupart des phylogénies sont binaires, mais cela n'est pas une généralité. Lorsque la phylogénie n'est pas binaire, les nœuds internes sont appelés des multifurcations. Par la suite, une phylogénie sera toujours considérée binaire.

Un groupe d'OTU est dit monophylétique si aucun autre OTU ne partage leur ancêtre commun, nous disons qu'ils forment un clade dans la phylogénie. Si des OTU additionnels sont inclus dans ce clade, ils sont paraphylétiques. En principe, une phylogénie est valuée, c'est-à-dire que chaque branche a une valeur qui représente le nombre de substitutions par site. Ainsi, en lisant une phylogénie il est possible de connaître la distance évolutive qui sépare deux OTU, elle correspond à la longueur du plus court chemin qui les sépare.

Une phylogénie peut être enracinée ou non (Figure 3). Contrairement à une phylogénie non enracinée, une phylogénie enracinée indique le sens du processus d'évolution, c'est-à-dire le sens de l'écoulement du temps. Plusieurs méthodes existent afin d'enraciner une phylogénie. La plus utilisée est sans doute l'ajout d'un ou de plusieurs OTU, appelés *outgroup*, qui sont connus pour être les OTU

distants et monophylétiques par rapport au group d'intérêt ou *ingroup*. Le nœud racine est alors placé sur la branche qui sépare l'*outgroup* de l'*ingroup*.

Figure 3. Différence entre phylogénie enracinée et non enracinée.
La figure A représente une phylogénie enracinée, tandis que la figure B une phylogénie non enracinée. Les deux phylogénies ont la même topologie, mais la phylogénie de la figure A est obtenue en ayant supposée que les OTU D et E réfèrent à un *outgroup*, le nœud racine a donc pu y être placé.



Dans le cas d'une phylogénie non enracinée, il y a exactement $2n - 3$ branches internes, où n est le nombre d'OTU, et $n - 2$ nœuds internes. Si elle est enracinée, il faut considérer une branche supplémentaire et un nœud supplémentaire.

Le nombre de topologies possibles croît exponentiellement en fonction du nombre d'OTU. Pour n OTU, il existe

$$B(n) = \frac{(2n-3)!}{2^{n-2}(n-2)!}$$

phylogénies enracinées. Le nombre de topologies non enracinées pour n OTU est égal au nombre de topologies enracinées pour $n - 1$ OTU. Avec 6 OTU nous avons donc 954 topologies possibles de phylogénies enracinées et avec 9 OTU les deux millions sont dépassés. Il devient donc vite impossible de parcourir l'ensemble des topologies au fur et à mesure que le nombre d'OTU considéré grandit. Des heuristiques sont proposées comme alternative à ce problème. Citons les algorithmes *nearest-neighbor interchange* (NNI), *subtree pruning and regrafting* (SPR) et *tree bisection and reconnection* (TBR) qui sont largement utilisés pour explorer un sous-espace de l'ensemble des arbres possibles.

1.5.2 Méthodes de distances

Les méthodes de distances essayent de faire correspondre un arbre (une phylogénie) à une matrice de distances. Les distances de la matrice sont généralement obtenues à partir d'un alignement et sont symétriques, c'est-à-dire que pour tout x et y, $d(x, y) = d(y, x)$. Par la suite nous noterons d_{xy} au lieu de $d(x, y)$. Les méthodes de distances peuvent être scindées en deux groupes : les méthodes agglomératives et les méthodes optimisant un critère.

1.5.2.1 Les méthodes agglomératives

Les méthodes agglomératives utilisent un algorithme pour construire pas à pas une phylogénie à partir d'une matrice de distances. Généralement, elles agglomèrent deux OTU x et y, répondant à un critère agglomératif précis, en un nouvel OTU u, puis calculent une nouvelle matrice de distances où les OTU x et y sont remplacés par l'OTU u. Le même procédé est de nouveau répété sur la dernière matrice de distances et ceci jusqu'à l'agglomération de tous les OTU. La phylogénie est ainsi calculée.

La méthode la plus connue et la plus ancienne est UPGMA (*unweighted-pair group method with arithmetic means*) (Sokal & Michener, 1958). Cette méthode est rarement utilisée car elle fait l'hypothèse d'une horloge moléculaire stricte (Zuckerkandl & Pauling, 1965, 1962). L'hypothèse de l'horloge moléculaire stricte stipule que l'évolution est un processus constant et uniforme. Faire cette hypothèse sur une phylogénie, c'est admettre que chaque feuille se situe à égale distance de la racine. C'est une hypothèse très forte qui nécessite souvent une justification lors de son utilisation. En revanche, si la matrice de distances satisfait la condition d'ultramétricité, alors l'algorithme UPGMA construit la phylogénie optimale et, de plus, elle est enracinée. Dans le cas contraire, cet algorithme n'est plus utilisé. Une matrice de distances est ultramétrique si pour tout triplet x, y et z, la condition d'ultramétricité (ou condition des trois points)

$$d_{xy} \leq \max\{d_{xz}, d_{zy}\}$$

est vérifiée. Cela signifie que deux des trois distances d_{xy}, d_{xz} et d_{zy} sont égales et maximales.

Pour surpasser l'hypothèse de l'horloge moléculaire stricte faite par l'algorithme UPGMA, une autre méthode agglomérative est suggérée, il s'agit de *neighbor-joining* (NJ) (Studier & Keppler, 1988; Saitou & Nei, 1987). NJ est l'une des méthodes de distances les plus utilisées et l'est encore aujourd'hui (Ye *et al*, 2011), même si de nombreuses variantes visant à améliorer cet algorithme sont proposées. Citons entre autre BIONJ (Gascuel, 1997), *generalized neighbor-joining* (Pearson *et al*, 1999), *weighted neighbor-joining* (Bruno *et al*, 2000), etc. Cette méthode construit une phylogénie non enracinée et lorsque la matrice de distances est additive, la phylogénie est optimale ou exacte. Une matrice de distances est additive si la condition des quatre points (Buneman, 1971) est satisfaite, c'est-à-dire que pour tout x, y, z et t, nous avons

$$d_{xy} + d_{zt} \leq \max\{d_{xz} + d_{yt}, d_{xt} + d_{yz}\}.$$

Cette condition implique que les deux plus grandes sommes sont égales. Seules les matrices additives peuvent aboutir à une phylogénie non enracinée, telle que la distance fournie en entrée entre deux OTU x et y est strictement égale à la somme des longueurs de branches sur le chemin reliant x et y

dans la phylogénie. La condition d'ultramétricité implique la condition des quatre points, donc si la matrice de distances est ultramétrique (elle est donc aussi additive) alors l'algorithme NJ construit la phylogénie optimale et la racine se trouve sur le point équidistant à chaque feuille.

1.5.2.2 Les méthodes optimisant un critère

Les méthodes qui optimisent un critère d'optimalité explorent l'espace des arbres à l'aide d'heuristiques, puis choisissent la meilleure phylogénie suivant le critère d'optimalité.

Pour les méthodes de distances, deux genres de critères d'optimalité sont utilisés. Le premier utilise l'approche standard des moindres carrés (Fitch & Margoliash, 1967). Il choisit la phylogénie qui minimise la somme des différences au carré entre la distance mesurée (celle de la matrice de distances) sur une paire d'OTU et la distance qui sépare ces deux OTU dans la phylogénie. La phylogénie résultante est celle qui contient les distances de chemin entre chaque paire d'OTU les plus proches possibles de celles contenues dans la matrice de distances initiale. Le programme FITCH (Felsenstein, 1989) utilise ce critère pour inférer une phylogénie. Le deuxième critère, bien différent du précédent, consiste à trouver l'arbre d'évolution minimum (*minimum evolution*, ME) (Kidd & Sgaramella-Zonta, 1971), c'est-à-dire celui qui minimise la somme des longueurs de branches, celles-ci étant estimées par moindres carrés à partir de la matrice de distances. Le logiciel FastME (Desper & Gascuel, 2002) utilise ce critère dans sa version « balancée » afin de proposer la meilleure phylogénie possible. Par la suite, il a été montré que l'algorithme NJ minimise également ce même critère d'évolution minimum balancé (Gascuel & Steel, 2006).

Diverses études ont montré que cette approche est remarquablement précise et rapide, avec des algorithmes en $O(n^3)$ ou moins pour construire un arbre initial et le modifier itérativement par mouvements NNI et SPR (Vinh & von Haeseler, 2005; Desper & Gascuel, 2004, 2002).

1.5.3 Méthodes de caractères

Les méthodes de caractères regroupent toutes celles qui se basent sur un alignement pour inférer une phylogénie. Comme les méthodes de distances, elles peuvent être scindées en deux catégories. La première regroupe les méthodes qui ne sont pas basées sur un modèle d'évolution explicite, comme la parcimonie. La seconde regroupe les méthodes basées sur un modèle d'évolution explicite. Ces dernières sont actuellement les plus employées par les biologistes, bien que lourdes en temps de calcul, en raison de leur fiabilité.

Les méthodes de parcimonie parcourent l'espace des phylogénies possibles et choisissent celles qui minimisent la quantité de changement évolutionnaire, c'est-à-dire celles qui expliquent l'alignement avec le moins de substitutions possibles (Fitch, 1971; Farris, 1970; Kluge & Farris, 1969).

En ce sens, elles rappellent le critère d'évolution minimum. Le point de vu philosophique derrière ce critère est que les hypothèses les plus simples sont souvent préférables aux plus compliquées. De ce fait, ces méthodes n'utilisent pas de modèle d'évolution à proprement parler. Aujourd'hui les méthodes de parcimonie sont de moins en moins utilisées pour inférer des phylogénies, mais elles restent utilisées pour inférer des annotations ancestrales à partir d'annotations contemporaines et d'une phylogénie précédemment calculée (cf. section 1.6). Dans ce cadre, où il n'existe souvent pas de modèles bien étudiés et incontestables, elles offrent une grande simplicité associée à des algorithmes rapides.

Les méthodes du maximum de vraisemblance (*maximum likelihood*) sont des méthodes probabilistes qui calculent une probabilité conditionnelle (la vraisemblance) exprimant le fait d'observer l'alignement suivant un modèle d'évolution particulier et une phylogénie particulière. Depuis l'introduction de ce principe en phylogénie en 1981 par Joseph FELSENSTEIN (Felsenstein, 1981), son utilisation est devenue de plus en plus populaire, en particulier grâce aux avancés algorithmiques et technologiques qui réduisent leur temps de calculs. Leur but est de choisir la phylogénie (un parcours, généralement heuristique, de l'espace des topologies est nécessaire) et les paramètres du modèle d'évolution (qui peuvent être estimés en même temps que la recherche de la topologie) qui maximisent la vraisemblance, contrairement aux méthodes de distances où c'est le plus souvent l'utilisateur qui choisit les valeurs des paramètres du modèle d'évolution. Outre le fait de se baser sur un modèle d'évolution, un avantage des méthodes de vraisemblance et de distances par rapport aux méthodes de parcimonie est de converger vers la vraie phylogénie au fur et à mesure que la quantité d'information en entrée augmente, et sous la condition que le modèle évolutif choisi soit le vrai modèle auquel obéissent les données (Felsenstein, 1978). Le logiciel PhyML (Guindon *et al.*, 2010; Guindon & Gascuel, 2003) utilise ce principe pour inférer des phylogénies. C'est l'un des logiciels les plus utilisés dans ce domaine et il le sera aussi dans nos études. D'autres logiciels sont aussi disponibles comme RAxML (Stamatakis, 2006) ou DNAML (Felsenstein, 1989).

Une dernière classe de méthodes de caractères existe. Il s'agit des méthodes bayésiennes, très proches des méthodes de vraisemblance. Ces méthodes sont fondées sur le théorème de BAYES (1702-1761) qui fut publié à titre posthume en 1763. Ce théorème combine la probabilité *a priori* d'un arbre $P(T)$ avec la vraisemblance $P(D|T)$ d'observer les données D (qui incluent l'alignement, les paramètres du modèle d'évolution et les longueurs de branches) sachant la topologie T pour en déduire la probabilité *a posteriori* de T, $P(T|D)$, par

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

La probabilité $P(D)$ est une constante de normalisation définie comme la somme des numérateurs pour toutes les topologies possibles : $P(D) = \sum_T P(D|T)P(T)$. L'objectif est alors de maximiser la probabilité *a posteriori* $P(T|D)$. En pratique, le calcul de la probabilité $P(D)$ est trop complexe et on utilise des algorithmes de type MCMC (*Markov Chain Monte Carlo*), et la variante Metropolis-Hastings (Metropolis *et al*, 1953; Hastings, 1970), qui permettent de s'affranchir du calcul de $P(D)$. Ces algorithmes effectuent un parcours au hasard de l'espace des arbres, mais à l'aide d'une marche guidée, jusqu'à un état d'équilibre. Puis un consensus des arbres obtenus est calculé après la suppression des premiers arbres calculés avant l'état d'équilibre (*burn-in*). Plus le nombre d'arbres parcourus est grand, meilleure est l'approximation. Généralement, cet algorithme nécessite au minimum un million de générations (nombre d'arbres parcourus) et reste réservé pour des phylogénies ne dépassant pas quelques centaines d'OTU. La différence fondamentale entre les méthodes bayésiennes par rapport aux méthodes de vraisemblance est que les paramètres suivent une distribution donnée *a priori* par l'utilisateur. Ceci en fait une méthode assez controversée puisque les résultats sont modifiables suivant les choix effectués, sans que ces choix puissent être guidés par des principes rigoureux. Le logiciel MrBayes (Ronquist & Huelsenbeck, 2003), utilisé dans nos études, lorsque cela est possible, permet d'inférer des arbres phylogénétiques sous ce principe. C'est aussi un des programmes de phylogénie les plus utilisés à l'heure actuelle.

1.5.4 Fiabilité des phylogénies

Différentes méthodes statistiques permettent de tester la fiabilité des arbres phylogénétiques, en particulier celle des branches internes. Ces techniques sont systématiquement utilisées lors de l'inférence d'arbres phylogénétiques. La plus répandue est celle du *bootstrap* (Felsenstein, 1985). Cette technique utilise le ré-échantillonnage aléatoire pour générer un grand nombre (souvent 1 000) d'alignements bruités. Ces derniers sont construits sur la base de l'alignement de départ. Un alignement bruité est constitué d'une succession de sites (autant que pour l'alignement de base) choisis aléatoirement (avec remise) parmi ceux de l'alignement de départ. Ainsi, certains sites de l'alignement initial peuvent apparaître plusieurs fois, tandis que d'autres jamais. Des phylogénies sont ensuite calculées, sur la base de ces alignements bruités, avec la même méthode et les mêmes paramètres que ceux utilisés pour calculer la phylogénie initiale. Le support statistique attribué à chaque clade de la phylogénie de départ correspond au nombre de fois où ce clade est trouvé dans les réplicas bruités. Plus le signal phylogénétique est fort, plus le support *bootstrap* est élevé. Cette méthode est l'une des plus employée et il est communément admis qu'un support de *bootstrap* supérieur à 80% est statistiquement fiable. Cependant, elle a un inconvénient majeur. Si le temps de calcul nécessaire pour inférer la phylogénie initiale est de x unités de temps, alors $1\,000 \times x$ unités de temps sont nécessaires pour estimer les supports de branches. Par exemple, si l'inférence d'une

phylogénie dure 20 minutes, il faut $20 \times 1\,000$ minutes (soit presque 14 jours) pour calculer les supports de branches associés. En pratique, les supports sont souvent calculés avec 100 réplicas (ce qui revient à un peu plus de un jour de calcul avec l'exemple précédent).

Pour augmenter la vitesse de calcul des méthodes probabilistes, d'autres tests statistiques sont proposés. Les méthodes de vraisemblance utilisent le test *approximate likelihood-ratio test* (aLRT) (Anisimova & Gascuel, 2006) qui estime les supports de chaque branche à l'aide de la seconde meilleure phylogénie parmi les deux différentes phylogénies obtenues par permutation des quatre branches adjacentes à la branche d'intérêt (deux mouvements NNI). En général, les supports sont jugés significatifs à partir de 80-90%. Quant aux méthodes bayésiennes, elles utilisent les arbres générés par la méthode MCMC pour en déduire les supports de branches ou probabilités postérieures. Toutefois, cette dernière méthode a tendance à surestimer les vrais supports de branches (Douady *et al*, 2003), et c'est pour cela que seules des probabilités postérieures supérieures à 95% ou proches de 100% sont jugées statistiquement fiables.

1.6 Reconstruire l'évolution de caractères

Outre le fait d'informer sur les relations de parenté, les phylogénies trouvent de nombreuses autres applications. Certaines sont exposées dans l'introduction de ce chapitre, d'autres dans les chapitres suivants de ce mémoire. Dans cette section, nous introduisons quelques concepts pour reconstruire l'évolution de caractères à partir d'une phylogénie.

Ici, un caractère est un ensemble d'annotations (ou états) qui sont capables d'évoluer de l'une vers l'autre (Maddison & Maddison, 2003). Par exemple, l'annotation « yeux bleus » est cohérente avec les annotations « yeux verts » et « yeux marrons » et il est supposé que l'on peut passer de l'une vers l'autre et vice-versa. Ce caractère, « couleur des yeux », est un caractère discret car la transition d'une annotation à une autre s'opère en une fois et (généralement) non graduellement. Toutefois, les caractères évoluant continuellement (comme la taille ou le poids) peuvent aussi être vus comme des caractères discrets en considérant un intervalle de valeur comme une annotation (Maddison & Maddison, 2003). Les zones géographiques et les pays sont des caractères discrets couramment utilisés, qui feront l'objet d'études dans cette thèse, pour déterminer les flux épidémiques du sous-type C du VIH-1 à la surface du globe.

La reconstruction d'annotations ancestrales cherche à déterminer quelles sont les annotations des HTU à partir : (1) des annotations associées aux OTU et (2) d'une phylogénie enracinée. La phylogénie doit nécessairement être enracinée afin d'orienter le cours de l'évolution. Dans le cas discret, les annotations ancestrales ne peuvent prendre que des valeurs parmi les annotations assignées aux

OTU. Il est donc nécessaire de considérer des OTU pertinents. Ainsi, nous pouvons déjà énoncer trois hypothèses fondamentales sur lesquelles se basent toutes les méthodes de reconstruction de caractères (Omland, 1999) :

- la phylogénie utilisée est la « vraie » phylogénie ;
- tous les OTU pertinents sont dans la phylogénie ;
- les états sont correctement assignés aux OTU, et cela sans erreur possible.

Même si ces trois hypothèses sont parfaitement respectées, les méthodes de reconstruction de caractères ancestraux ne garantissent pas la fiabilité des résultats. Elles suivent des principes divers que nous décrivons maintenant.

Dans le cas de caractères discrets, les méthodes de reconstruction de caractères ancestraux généralement utilisées sont basées sur le principe de parcimonie qui choisit la reconstruction impliquant le moins de changements de caractère le long de l'arbre phylogénétique. Des méthodes plus élaborées existent, certaines sont basées sur le principe du maximum de vraisemblance (Schluter *et al*, 1997) et d'autres sur des approches bayésiennes (Schultz & Churchill, 1999), mais elles nécessitent un modèle probabiliste de transition entre annotations afin d'inférer les annotations ancestrales. L'élaboration d'un tel modèle n'est pas chose aisée et les erreurs de jugement peuvent produire des résultats biaisés. Certaines méthodes récemment développées peuvent auto-estimer les paramètres du modèle probabiliste mais le temps de calcul en est considérablement rallongé (Lemey *et al*, 2009a). Même si les méthodes de parcimonie n'utilisent pas de modèle probabiliste, il est toutefois possible d'attribuer des contraintes ou des poids sur les transitions afin d'en privilégier certaines par rapport à d'autres. On distingue ainsi plusieurs principes de parcimonie (Maddison & Maddison, 2003). Dans le cas de caractères continus, les méthodes de parcimonie sont rarement utilisées car elles n'emploient pas l'information contenue dans les longueurs de branches, information généralement nécessaire pour reconstruire correctement les caractères ancestraux continus. Par la suite, nous nous focaliserons sur les méthodes de parcimonie et leur utilisation sur des caractères discrets, avec en ligne de mire les annotations géographiques qui seront étudiées au Chapitre 6 sur l'épidémie du VIH-1 sous-type C à l'échelle mondiale.

Les méthodes de parcimonie font trois hypothèses supplémentaires à celles émises précédemment (Omland, 1999) :

- les transformations sont identiquement probables quelle que soit la branche, c'est-à-dire que les longueurs de branches ne sont pas prises en compte ;

- les taux d'évolution d'un état vers un autre doivent être relativement lents, et on suppose qu'au plus une transition a lieu sur chaque branche de l'arbre ;
- les coûts de transformations sont symétriques, c'est-à-dire que la probabilité de gagner ou de perdre un état est identique.

Il existe plusieurs méthodes de parcimonie permettant de reconstruire les annotations ancestrales. La méthode de parcimonie la plus utilisée est la parcimonie non ordonnée (Fitch, 1971; Hartigan, 1973), c'est-à-dire que la transition d'un état vers n'importe quel autre état a le même coût. On parle de parcimonie de Fitch. Pour la calculer et déterminer les états ancestraux on utilise deux étapes successives. Une première étape, appelée UPPASS, parcourt l'arbre des feuilles jusqu'à la racine en assignant à chaque nœud ancestral l'information relative aux seuls nœuds fils. La Figure 4 décrit les étapes de l'algorithme UPPASS. À la fin de cette étape, les états assignés aux nœuds ancestraux ne sont pas forcément les plus parcimonieux, puisqu'ils sont calculés uniquement avec l'information des nœuds inclus dans le clade sous-jacent et ne considère donc pas l'information de toute la phylogénie. Ils sont simplement les plus parcimonieux par rapport au clade sous-jacent. Ainsi, le seul nœud qui y fait exception est le nœud racine, puisque l'information de toute la phylogénie sert à le calculer.

Figure 4. Algorithme UPPASS.
L'algorithme UPPASS est un algorithme récursif de type « *postorder* », dans lequel le calcul à proprement parler se fait après les appels récursifs. Il utilise un paramètre qui est un nœud (initialisé avec la racine de la phylogénie) et calcule les états associées aux nœuds ancestraux. N est le nœud courant, G et D ses fils gauche et droit et P le nœud père. $S(X)$ est l'ensemble des états associés au nœud X.

Entrée : N un nœud
1. **si** N est une feuille **alors**
2. $S(N) =$ état associé à N
3. **sinon**
4. UPPASS(G)
5. UPPASS(D)
6. **si** $S(G) \cap S(D) = \emptyset$ **alors**
7. $S(N) = S(G) \cup S(D)$
9. **sinon**
10. $S(N) = S(G) \cap S(D)$
11. **fin si**
12. **fin si**

Une deuxième étape, appelée DOWNPASS (Maddison & Maddison, 2003), parcourt l'arbre de la racine aux feuilles en considérant pour chaque nœud l'information des nœuds adjacents. Comme la racine contient déjà la valeur la plus parcimonieuse, cette deuxième phase commence avec les nœuds fils du nœud racine. La Figure 5 décrit les étapes de l'algorithme DOWNPASS. Après

l'application de cet algorithme, chaque nœud interne contient les valeurs finales issues de l'information de tous les nœuds et feuilles de la phylogénie. Cette information nous dit essentiellement quels sont les états du nœud qui appartiennent à au moins un scénario optimal de parcimonie. Il faut noter que toutes les combinaisons d'annotations ancestrales ainsi obtenues ne correspondent pas à un tel scénario optimal, loin de là. Nous y reviendrons plus loin.

Figure 5. Algorithme DOWNPASS.
L'algorithme DOWNPASS vient en complément de l'algorithme UPPASS pour intégrer l'information de toute la phylogénie sur chacun des nœuds internes. N est le nœud courant, G et D ses fils gauche et droit et P le nœud père. $S(X)$ est l'ensemble des états associés au nœud X. C'est un algorithme récursif de type « *preorder* », dans lequel le calcul à proprement parler se fait avant les appels récursifs.

Entrée : N un nœud
1. **si** N n'est pas une feuille **alors**
2. **si** N n'est pas la racine **alors**
3. $V = S(P) \cap S(G) \bigcap \mathrm{S(D)}$
4. **si** $V = \emptyset$ **alors**
5. $V = (S(P) \cap S(G)) \cup (S(P) \cap S(D)) \cup (S(G) \cap S(D))$
6. **fin si**
7. **si** $V = \emptyset$ **alors**
8. $V = S(P) \cup S(G) \cup S(D)$
9. **fin si**
10. $S(N) = V$
11. **fin si**
12. DOWNPASS(G)
13. DOWNPASS(D)
14. **fin si**

Après l'utilisation de l'algorithme DOWNPASS, des ambiguïtés peuvent se produire au niveau des nœuds internes, c'est-à-dire qu'un nœud interne peut être associé à plus d'une annotation, signifiant que la parcimonie hésite entre plusieurs solutions. Ces ambiguïtés peuvent être gênantes lors de l'estimation du nombre de transitions (nombre de branches ayant des annotations différentes à ses extrémités), pratique courante dans ce genre d'étude. Il existe plusieurs possibilités afin de diminuer ces ambiguïtés. Les deux plus connues sont les algorithmes ACCTRAN (*accelerated transformation*) (Farris, 1970) et DELTRAN (*delayed transformation*) (Swofford & Maddison, 1987). Ces deux algorithmes font des hypothèses différentes en ce qui concerne le choix final des états de chaque nœud interne. La méthode ACCTRAN force les changements d'états à se produire le plus près possible de la racine et donc à favoriser les transformations reverses, tandis que la méthode DELTRAN force les changements d'états à se produire le plus près possible des feuilles et donc à favoriser les transformations parallèles. La Figure 6 décrit l'algorithme ACCTRAN, proposé conjointement par Farris et Fitch, et la Figure 7 l'algorithme DELTRAN. Il faut bien noter que l'algorithme ACCTRAN remplace

l'algorithme DOWNPASS, alors que l'algorithme DELTRAN vient en complément de l'algorithme DOWNPASS, après son exécution.

Figure 6. Algorithme ACCTRAN.
L'algorithme ACCTRAN remplace l'algorithme DOWNPASS afin de diminuer les ambiguïtés de valeurs au niveau des nœuds internes de la phylogénie. L'annotation d'un nœud privilégie les informations venant des fils plutôt que du père, et va « pousser » les changements vers la racine. N est le nœud courant, G et D ses fils gauche et droit. $S(X)$ est l'ensemble des états associés au nœud X.

Entrée : N un nœud
1. **si** N n'est pas une feuille **alors**
2. **si** $S(N) \cap S(G) \neq \emptyset$ **alors**
3. $S(G) = S(N) \cap S(G)$
4. **sinon**
5. $S(G)$ est inchangé et contient l'information issue de ses fils après UPPASS
6. **fin si**
7. **si** $S(N) \cap S(D) \neq \emptyset$ **alors**
8. $S(D) = S(N) \cap S(D)$
9. **sinon**
10. S(D) est inchangé et contient l'information issue de ses fils après UPPASS
11. **fin si**
12. ACCTRAN(G)
13. ACCTRAN(D)
14. **fin si**

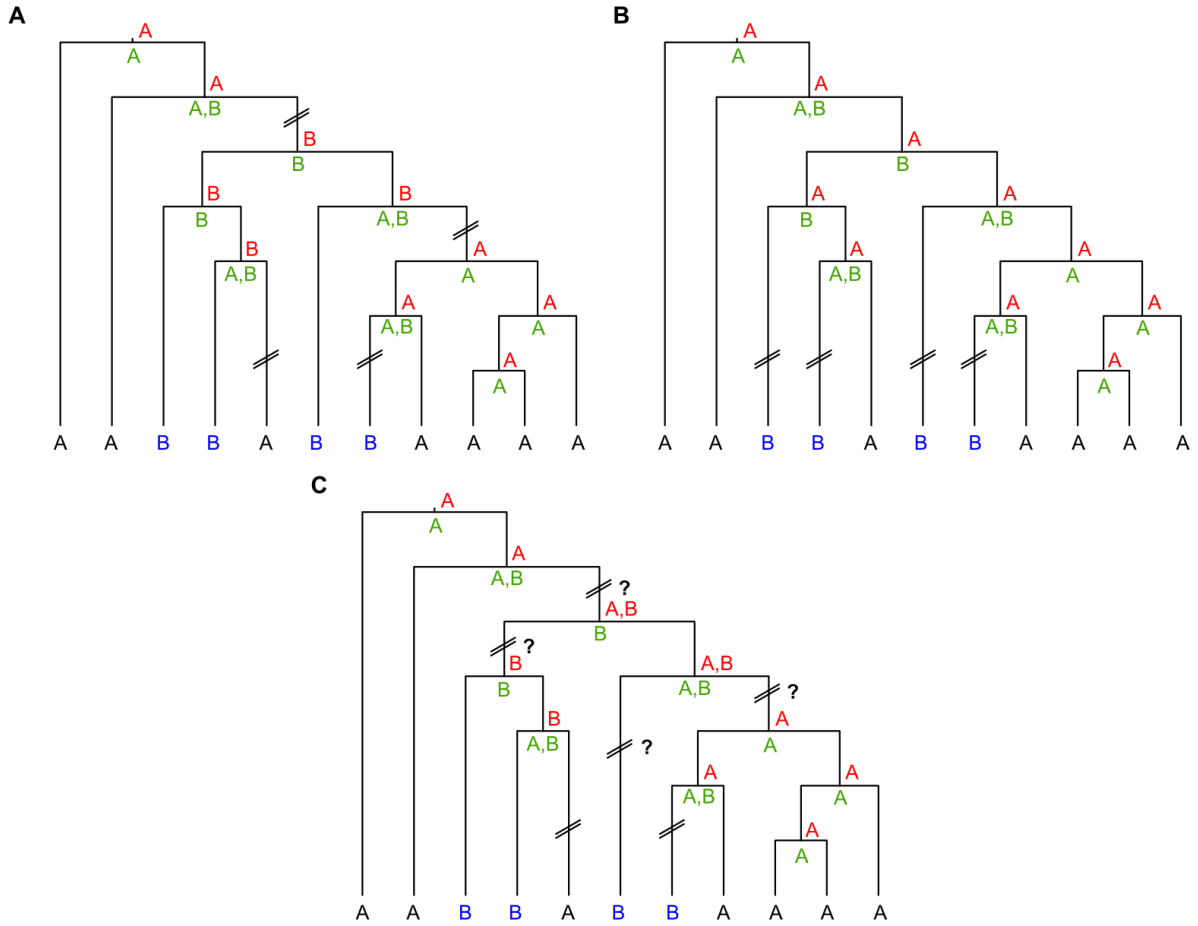
Figure 7. Algorithme DELTRAN.
L'algorithme DELTRAN vient en supplément de l'algorithme DOWNPASS après son exécution afin de diminuer les ambiguïtés de valeurs au niveau des nœuds internes de la phylogénie. L'annotation d'un nœud privilégie ainsi les informations venant du père, et « pousse » les changements vers les feuilles. N est le nœud courant, G et D ses fils gauche et droit et P le nœud père. $S(X)$ est l'ensemble des états associés au nœud X. Comme on se place après DOWNPASS, $S(N)$ contient déjà tous les états les plus parcimonieux, contrairement à ACCTRAN où $S(G)$ et $S(D)$ ne contiennent que les informations issues des clades de racine G et D.

Entrée : N un nœud
1. **si** N n'est pas une feuille **et** N n'est pas la racine **alors**
2. **si** $S(P) \cap S(N) \neq \emptyset$ **alors**
3. $S(N) = S(P) \cap S(N)$
4. **fin si**
5. DELTRAN(G)
6. DELTRAN(D)
7. **fin si**

Les algorithmes ACCTRAN et DELTRAN ne résolvent pas forcément toutes les ambiguïtés des nœuds internes, c'est-à-dire qu'après l'application de l'un ou l'autre, des nœuds internes peuvent encore être ambigus, en particulier lorsque le nœud racine l'est. Dans ce cas, lors de l'estimation du nombre de transitions, certains auteurs ne considèrent pas ces nœuds (Nakano *et al*, 2004) ou alors ils calculent le nombre moyen de transitions parmi toutes les reconstructions les plus parcimonieuses

possibles (Salemi *et al*, 2008). La Figure 8 montre l'application des algorithmes ACCTRAN, DELTRAN et DOWNPASS sur une phylogénie exemple.

Figure 8. Exemple d'application des algorithmes ACCTRAN, DELTRAN et DOWNPASS.
La figure A montre les résultats de l'application d'ACCTRAN, la figure B ceux de DELTRAN et la figure C ceux de DOWNPASS sur une même phylogénie. Les résultats de l'algorithme UPPASS sont indiqués en vert, ceux d'ACCTRAN, DELTRAN ou DOWNPASS en rouge. Les barres obliques indiquent les branches où des transformations ont lieux. Les reconstructions coûtent chacune quatre transitions, mais la reconstruction avec DOWNPASS hésite entre deux scénarios possibles.



Le plus utilisé de ces deux algorithmes semble être ACCTRAN ou, si DELTRAN est choisi, les résultats avec ACCTRAN sont souvent présentés en complément (Agnarsson & Miller, 2008), l'idée étant que ces deux algorithmes constituent deux extrêmes et que la « vérité se situe entre les deux » (cf. ci-après). Cela est du à un commentaire de De Pinna (1991) qui argumente sur le fait que les transformations reverses sont préférables aux transformations parallèles. Mais aucune preuve formelle ne démontre qu'ACCTRAN serait mieux que DELTRAN ou vice-versa. L'utilisation de l'un ou l'autre dépend en réalité largement du caractère étudié. En effet, lorsque l'on considère des caractères morphologiques, on favorise les séquences acquisitions – perte (par exemple d'ailes fonctionnelles) plutôt que l'invention multiple de caractères. En ce sens, la parcimonie ACCTRAN est préférable puisqu'elle force les changements à se produire le plus près possible de la racine et donc défavorise les mutations parallèles par rapport aux évènements reverses. Mais lorsque des caractères épidémiolo-

giques sont considérés, comme par exemple des lieux géographiques, il est plus facile de penser qu'une épidémie s'intensifie dans un lieu donné avant de se diffuser à partir de celui-ci, avec des transmissions multiples. Dans ce cas, l'algorithme DELTRAN est le plus approprié puisqu'il force les changements (de lieux) à se produire le plus près possible des feuilles.

En considérant l'espace de toutes les reconstructions les plus parcimonieuses possibles (*most parsimonious reconstruction*, MPR), c'est-à-dire celles qui minimisent le nombre de changement d'états, ainsi qu'une relation d'ordre sur cet espace, Minaka (1993) a montré que les algorithmes ACCTRAN et DELTRAN sont les deux bornes de cet espace. Ainsi, si les résultats d'ACCTRAN et de DELTRAN sont identiques, il en est de même pour toutes les autres MPR.

Plusieurs méthodes statistiques existent afin de tester la fiabilité des reconstructions de caractères ancestraux par parcimonie. Elles sont toutes basées sur des méthodes de Monte Carlo et comparent la quantité de transitions observées à celle de l'hypothèse nulle ou panmixie, dans laquelle il n'y aurait aucune corrélation entre la phylogénie et les annotations étudiées. Nous présentons ici la méthode de ré-échantillonnage aléatoire, ou *shuffling*, qui mélange les annotations des OTU et estime de nouveau, sous les mêmes conditions, la quantité de transitions. Typiquement ce procédé est répété un grand nombre de fois (1 000 ou 10 000). La quantité de transitions observées est alors comparée à la distribution des quantités obtenues aléatoirement afin de définir sa significativité statistique. Slatkin et Maddison (1989) semblent être les premiers à avoir utilisé cette procédure qui est maintenant un standard dans le domaine. Cette méthode de ré-échantillonnage aléatoire sera utilisée dans notre étude sur l'épidémie mondiale du VIH-1 sous-type C (Chapitre 6) pour établir les significativités statistiques de différents critères. Elle donne une vision plus complète et interprétable que l'approche consistant à soustraire au nombre de transitions observées le nombre de transitions attendues par hasard dans le modèle nul de panmixie (Nakano *et al*, 2004).

Chapitre 2

Méthodes de distances pour estimer le taux de substitution à partir d'un ensemble de séquences hétérochrones, application au virus de l'immunodéficience humaine (VIH)

La vitesse d'évolution (mesurée par le taux de substitution) des séquences est différente d'une espèce à l'autre. Ce taux peut être estimé à l'aide de séquences échantillonnées dans le temps, ou séquences hétérochrones, lorsque le nombre de substitutions accumulées entre ces séquences est significatif. Les virus sont des candidats idéals, car ils accumulent un nombre de substitutions important en seulement quelques années. L'utilisation de ce taux trouve de nombreuses applications biologiques, comme par exemple dater l'origine d'une épidémie ou d'une infection. Les méthodes présentées dans ce chapitre sont des méthodes de distances, rapides en temps de calcul, qui estiment le taux de substitution à l'aide de séquences hétérochrones uniquement, et en faisant l'hypothèse d'une horloge moléculaire stricte, comme, par exemple, TREBLE, sUPGMA ou encore les régressions linéaires Pairwise-Distance *et* Root-to-Tip*. Enfin, deux méthodes probabilistes, basées sur des principes différents, sont présentées succinctement.*

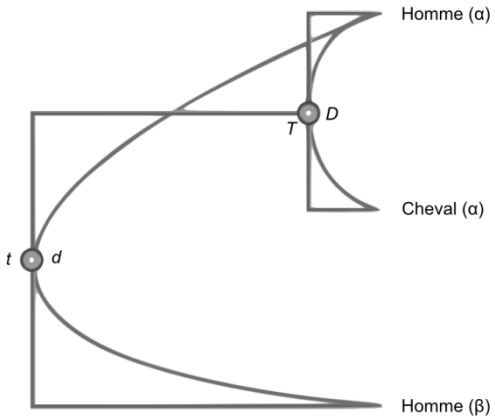
Sommaire

2.1 Introduction

Les organismes évolués, comme les mammifères, ont un processus de réplication de leur matériel génétique très sophistiqué, mais des erreurs de réplication surviennent souvent pendant ce processus (Reha-Krantz, 2010). Elles peuvent être dangereuses pour l'organisme car elles peuvent se produire sur un gène et le rendre inactif ou modifier sa fonction. Il existe certains mécanismes qui permettent de corriger ces erreurs, mais tous les organismes ne les possèdent pas (Roberts *et al*, 1988). Par exemple, ces mécanismes sont absents chez le virus de l'immunodéficience humaine (VIH) et donc, à l'intérieur d'un hôte, la population virale est constituée d'une multitude de variants génétiques changeant continuellement, quelque fois appelés des quasi-espèces (Domingo, 1998; Nowak, 1992). En partie pour cette raison, la vitesse d'évolution des organismes, identifiée par le taux de substitution, varie d'une espèce à l'autre. Cette vitesse est exprimée en nombre de substitutions par site et par unité de temps (généralement en années, jours ou générations).

En 1962 et 1965, Zuckerkandl et Pauling ont publié deux chapitres de livre fondamentaux sur la vitesse évolutive des protéines (Zuckerkandl & Pauling, 1965, 1962). Leur objectif était d'estimer la date de divergence de différentes globines. Pour cela, ils ont fait l'hypothèse d'une horloge moléculaire stricte, c'est-à-dire qu'ils ont supposé que la vitesse évolutive est constante au cours du temps et uniforme chez toutes les espèces étudiées. Cette hypothèse, ou une alternative, est essentielle pour estimer la vitesse évolutive (Kumar, 2005; Bromham & Penny, 2003). Par exemple, dans leur publication de 1962, Zuckerkandl et Pauling (1962) estimaient la date de divergence entre les hémoglobines α et β de l'homme. Ils disposaient des p-distances d et D définissant respectivement le nombre de différences observées entre les protéines de l'hémoglobine α et β de l'homme et le nombre de différences observées entre les protéines des hémoglobines α du cheval et de l'homme (Figure 9). Comme la vitesse évolutive est supposée constante et connaissant la date de divergence homme/cheval T sur la base d'estimations fossiles, ils ont estimé t à l'aide de la relation $d/t = D/T$. Ainsi, la date de divergence t entre les hémoglobines α et β de l'homme est estimée à $T(d/D)$. La vitesse évolutive dans cet exemple est donc égale à $D/2T$; le chiffre 2 venant du fait que la distance D correspond à la somme de la quantité évolutive séparant les deux espèces de leur ancêtre commun.

Figure 9. Illustration de la première utilisation d'une horloge moléculaire.
Dans leur papier de 1962, Zuckerkandl et Pauling (1962) estimaient la date de divergence t entre les hémoglobines α et β de l'homme. Pour cela, ils ont eu recourt à l'hypothèse de l'horloge moléculaire stricte qui stipule que la vitesse d'évolution est constante et uniforme. Comme la date de divergence T entre l'homme et le cheval était connue (d'après des estimations fossiles), ils ont pu estimer la date de divergence t sachant le nombre de substitutions D entre les séquences de l'hémoglobine α de l'homme et du cheval à l'aide de la relation $d/t = D/T$. Ainsi, la date t peut être estimée par $d(T/D)$.
Adaptation de Kumar (2005).



Dans cet exemple, les estimations du taux de substitution et de la date de divergence entre les hémoglobines α et β de l'homme n'étaient pas possibles sans l'information de la date de divergence entre les lignées de l'homme et du cheval. Autrement dit, les estimations nécessitent un point de calibration, limitant ainsi le nombre d'études similaires puisque les points de calibration sont généralement difficiles à obtenir et entachés d'erreurs. Toutefois, une autre source d'information temporelle peut servir à l'estimation du taux de substitution (et donc aux dates de divergence) : les dates d'échantillonnage des séquences. Mais pour qu'il soit possible d'estimer la vitesse évolutive à partir de séquences hétérochrones (séquences échantillonnées dans le temps ; à mettre en opposition avec les séquences isochrones, échantillonnées à la même date), il faut que l'accumulation de substitutions entre deux échantillons collectés à des moments différents soit significative. Les populations pour lesquelles des séquences hétérochrones peuvent être utilisées pour estimer le taux de substitution sont appelées des MEP (*measurably evolving populations*) (Drummond *et al*, 2003b). Ce terme désigne essentiellement des virus, organismes pour lesquels la vitesse évolutive est très importante et peut être mesurée à l'aide d'échantillons espacés dans le temps par seulement quelques années, comme le VIH ou le virus de la Dengue (Chen *et al*, 2011; Dunham & Holmes, 2007), ou, plus rare, des organismes dont on possède de l'ADN ancien (Lambert *et al*, 2002).

Les bases de données biologiques, et notamment celle du laboratoire national de Los Alamos sur le VIH, abondent en séquences hétérochrones. En effet, dans le cadre du VIH, le séquençage est une pratique routinière (Taylor *et al*, 2008) et, donc, des dates de prélèvement différentes sont associées aux séquences. Ces études renseignent généralement sur l'apparition de nouveaux sous-types ou formes recombinantes (Ng *et al*, 2011; Ibe *et al*, 2010), l'apparition de résistances aux traitements médicamenteux (Hanna & D'Aquila, 2001; Hirsch *et al*, 2000), l'apparition de nouvelles zoonoses

(Plantier *et al*, 2009; Damond *et al*, 2004) ou encore les stratégies de prévention comme, par exemple, la conception d'un vaccin (Gaschen *et al*, 2002). Les séquences hétérochrones, dont la quantité est en perpétuelle augmentation, sont donc des supports idéaux pour estimer le taux de substitution de virus et notamment celui du VIH.

La mesure du taux de substitution trouve de nombreuses applications biologiques. Par exemple, l'estimation de plusieurs taux de substitution différents au sein d'une même population est un indicateur dans la recherche de traitements efficaces contre les virus. Prenons le cas du VIH et supposons qu'un patient soit infecté par celui-ci. Des souches du VIH lui sont prélevées, et leur matériel génétique est séquencé en trois temps distincts t_0, t_1 et t_2 où t_1 représente la date à laquelle le patient a commencé un traitement contre le VIH, t_2 celle où le patient a été infecté et t_0 la date la plus récente, à laquelle le patient suit toujours son traitement (depuis t_1 donc). Pour pouvoir en déduire les taux de substitution, les intervalles de temps entre les dates d'échantillonnage doivent être suffisamment grands pour permettre une accumulation significative de substitutions. La comparaison entre les taux de substitution ω_2, correspondant à l'intervalle de temps où le patient n'a pas subi de traitement ($t_1 - t_2$), et ω_1, correspondant à l'intervalle de temps où le patient prend son traitement ($t_0 - t_1$), permet d'en déduire l'influence du traitement sur le virus. En effet, si la vitesse d'évolution du virus a subit une accélération ($\omega_1 > \omega_2$), alors le traitement est efficace contre la souche dominante du virus, car cette souche a tendance à disparaître pour en laisser apparaître de nouvelles, ayant une meilleure résistance au traitement, d'où une accélération de la vitesse évolutive. Dans le cas contraire ($\omega_1 \leq \omega_2$), le traitement n'a pas d'influence sur la souche dominante du virus. D'autres applications sont possibles, par exemple pour comparer les vitesses d'évolution des gènes les uns par rapport aux autres. Dans le cas de notre patient atteint par le VIH, cette pratique permettrait de savoir quels gènes le traitement doit cibler pour être efficace, c'est-à-dire ceux conservés car essentiels au virus (Hué *et al*, 2004). Le taux de substitution permet aussi de dater l'origine d'une épidémie ou d'une infection (Wertheim & Worobey, 2009; Korber *et al*, 2000), comme montré dans l'exemple du début. Les applications biologiques rendues possibles par la connaissance du taux de substitution sont donc nombreuses, et, grâce à l'accroissement considérable du nombre de séquences dans les bases de données biologiques, nous pouvons imaginer que certaines d'entre elles vont devenir routinières. Le besoin d'une méthode d'estimation précise et rapide se fait sentir, et pour ce faire les méthodes de distances ont de solides atouts, en raison de leur vitesse et de leur propriété de convergence asymptotique.

Nous présentons dans ce chapitre des méthodes de distances qui permettent d'estimer le taux de substitution à partir d'un ensemble de séquences hétérochrones, sans la connaissance de points de calibration, et sous l'hypothèse d'une horloge moléculaire stricte. Lorsqu'un (plusieurs) point(s) de

calibration existe(nt), l'information des temps de collecte n'est plus indispensable et l'estimation du taux de substitution ou des dates des ancêtres communs peut être faite pour n'importe quelle espèce, y compris à évolution lente (Xia & Yang, 2011; Sanderson, 1997). Nous présentons également deux méthodes probabilistes, chacune partant d'un principe différent (maximum de vraisemblance et bayésien), lourdes en temps de calcul, mais largement utilisées par la communauté scientifique. Mais avant cela, nous discutons de la différence entre taux de substitution synonyme et non synonyme, ainsi que des différents modèles d'horloge moléculaire, incluant notamment les horloges relâchées.

2.2 Taux de substitution synonyme et non synonyme

Dans la littérature, deux sortes de mutations sont distinguées : les mutations synonymes et les mutations non synonymes. Les mutations synonymes (ou silencieuses) sont des mutations qui n'induisent pas de changement d'acide aminé, tandis que les mutations non synonymes (non silencieuses) induisent un changement d'acide aminé. Cela est possible à cause de la redondance du code génétique. Par exemple, si la transversion C → A se produit en première position du codon GCC, codant une Alanine, alors ce codon sera traduit par une Thréonine, tandis que si elle se produit à la troisième position du codon, l'acide aminé traduit restera l'Alanine. De cette observation, découle deux taux de substitution différents : les taux de substitution synonyme et non synonyme et ils ne peuvent être estimés que sur les régions codantes du génome. Le taux de substitution synonyme (resp. non synonyme) est calculé à partir des seules mutations silencieuses (resp. non silencieuses). Généralement les mutations silencieuses se produisent sur le troisième nucléotide du codon et sont plus fréquentes que les mutations non silencieuses qui elles se produisent généralement sur les deux premiers nucléotides du codon (Gojobori *et al*, 1994, 1990). Lorsqu'aucun des deux termes (synonyme et non synonyme) n'est employé, le taux de substitution est calculé en comptant toutes les sortes de mutations (silencieuses ou non). Dans ce cas, il peut aussi être estimé sur les régions non codantes du génome.

2.3 Modèles d'horloge moléculaire

Il est communément admis que la vitesse d'évolution des séquences moléculaires n'est pas strictement uniforme et constante, mais qu'elle peut varier en fonction du temps (par exemple, lorsqu'une pression de sélection supplémentaire s'exerce sur un virus au moment du début d'un traitement) et/ou des lignées (Li & Tanimura, 1987). Ces variations ne sont pas considérées par le modèle d'horloge moléculaire stricte, mais s'en soustraire complètement est impossible. En effet, l'évolution est un processus complexe et la cause de plusieurs facteurs géographiques, géologiques, biologiques, sociologiques, etc. Imaginer une relation universelle entre la distance évolutive et le temps n'est

donc pas faisable (Bromham & Penny, 2003). Dans ce but, plusieurs modèles d'horloge moléculaire ont été proposés. Ils peuvent être regroupés en quatre catégories suivant une terminologie introduite par Rambaut (2000).

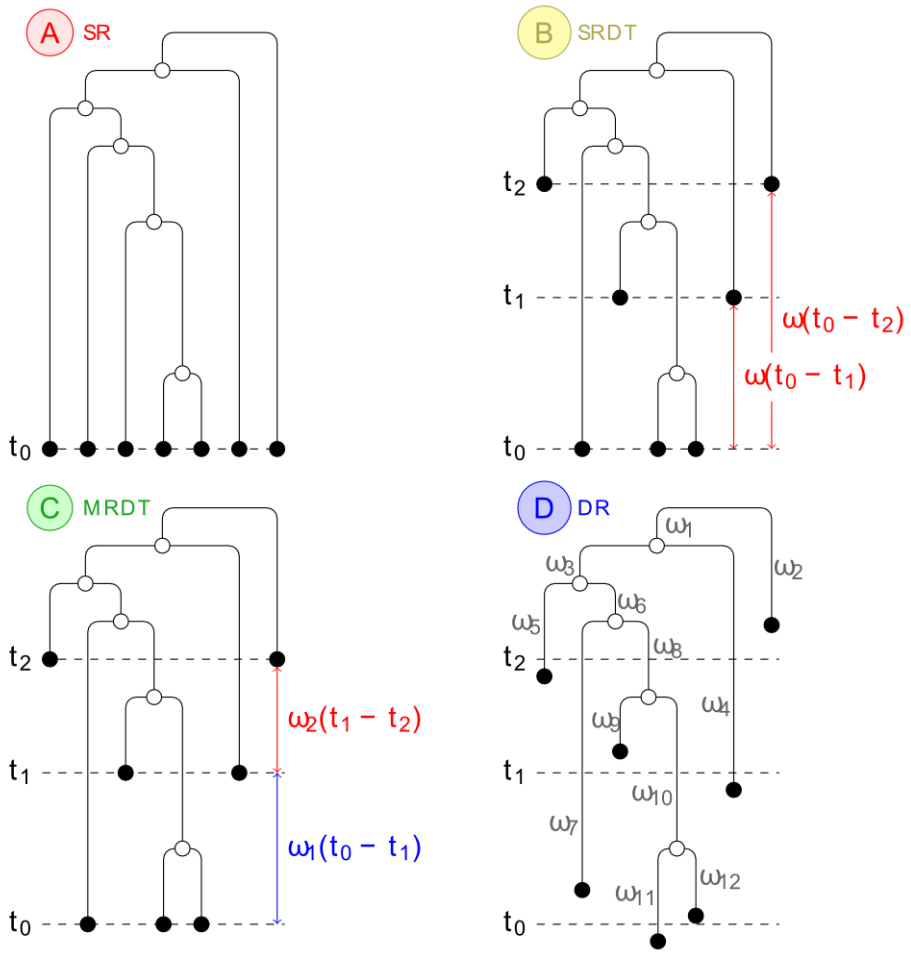
Le modèle *Single Rate* (SR) est le modèle standard (Figure 10A). Il fait l'hypothèse d'une horloge moléculaire stricte mais les séquences sont supposées être échantillonnées au même temps (séquences isochrones). Sinon les intervalles de temps qui séparent les dates de collecte doivent être négligeables par rapport à l'échelle de temps de l'arbre tout entier. Dans ce modèle, le taux de substitution peut uniquement être estimé à l'aide d'un (ou de plusieurs) point(s) de calibration (Xia & Yang, 2011).

Le modèle *Single Rate Dated Tips* (SRDT) fait toujours l'hypothèse d'une horloge moléculaire stricte, mais les séquences sont maintenant prélevées en des temps distincts (séquences hétérochrones) ; il est alors possible d'estimer le taux de substitution avec la connaissance des dates de collecte (Figure 10B) (Rambaut, 2000). Ce modèle est le plus couramment utilisé pour estimer le taux de substitution par des méthodes de distances.

Le modèle *Multiple Rates Dated Tips* (MRDT) suppose une horloge moléculaire relâchée par l'existence de plusieurs taux de substitution, un pour chaque intervalle de temps défini entre deux dates de prélèvement successives (Figure 10C) (Drummond *et al*, 2001). Ce modèle admet une approche alternative que nous distinguerons par la notation MRDT *alternative* (MRDTa). Ce dernier permet à l'utilisateur de choisir ses propres intervalles de temps. Notons toutefois qu'il est impossible d'estimer le taux de substitution lorsque le nombre d'intervalles de temps choisi par l'utilisateur est supérieur au nombre d'intervalles de temps obtenus avec les dates de collecte. De plus, comme les estimations des taux de substitution se font par rapport aux feuilles, il est nécessaire que chaque intervalle de temps contienne au moins une feuille. Donc le nombre maximum d'intervalle de temps est donnée par le nombre de dates de collecte moins un (un temps de collecte doit être utilisé comme référence). Typiquement, ce dernier modèle peut être utilisé pour connaître l'efficacité d'un traitement viral, en comparant sa vitesse évolutive avant le début du traitement et pendant celui-ci (cf. section 2.1).

Enfin, le modèle *Different Rate* (DR) suppose que chaque branche de l'arbre a un taux de substitution propre, ces taux pouvant être corrélés entre eux ou non (Figure 10D) (Rambaut, 2000; Felsenstein, 1981). Ce dernier modèle est le plus réaliste de tous, mais il est excessivement paramétré et insoluble en l'absence de corrélation ou contraintes fortes liant les taux. Les horloges moléculaires locales, c'est-à-dire des horloges moléculaires strictes spécifiques à certaines lignées, associées à une horloge moléculaire stricte globale, sont une variante à ce modèle (Yoder & Yang, 2000).

Figure 10. Illustrations des différents modèles d'horloge moléculaire.
La figure A montre le cas d'une phylogénie sous les contraintes du modèle SR (horloge moléculaire stricte et séquences isochrones). Cette phylogénie est ultramétrique, c'est-à-dire que toutes les séquences sont à égale distance de la racine. La figure B montre une phylogénie sous le modèle SRDT (horloge moléculaire stricte et séquences hétérochrones). La figure C une phylogénie sous le modèle MRDT (un taux de substitution par intervalle de temps entre dates de collecte successives et séquences hétérochrones) et la figure D une phylogénie sous le modèle DR (séquences hétérochrones avec un taux de substitution par branche ; dans cette figure l'écart à l'horloge reste faible).



2.4 Méthodes de distances estimant le taux de substitution sous le modèle SRDT

2.4.1 Premières méthodes

Les premières méthodes de distances permettant d'estimer la vitesse d'évolution sont relativement simples et s'appliquent généralement sur un groupe de deux à trois séquences au plus. À notre connaissance, Hahn *et al*. (1986) sont les premiers à avoir estimé le taux de substitution du VIH-1. Cette estimation est seulement faite à partir de deux séquences provenant d'un même patient, un enfant haïtien vivant en Floride et ayant eu une infection prénatale. Le taux de substitution $\hat{\omega}$ est estimé par la relation

$$\widehat{\omega} = \frac{\hat{d}}{2T}$$

où $\hat{d}$ est la distance évolutive estimée qui sépare les deux séquences, alors calculée sous le modèle JC69 (Jukes & Cantor, 1969), et T le temps écoulé depuis la divergence de leur ancêtre commun. Cette méthode a été préalablement décrite par Gojobori et Yokoyama (1985) mais appliquée à *Moloney murine sarcoma virus*, virus oncogène (pour les souris) de la même famille que le VIH-1. Bien que l'estimation du taux de substitution soit du même ordre de grandeur que celle admise aujourd'hui, plusieurs limites sont à relever. Premièrement, cette méthode suppose que le taux d'évolution est constant, c'est-à-dire que l'estimation du taux de substitution est faite sous l'hypothèse d'une horloge moléculaire stricte (Zuckerkandl & Pauling, 1962), hypothèse admise par de nombreuses autres méthodes, notamment par les méthodes de distances. Deuxièmement, la valeur du paramètre T ne peut être connue avec certitude, elle doit donc être estimée. Pour leurs séquences, Hahn *et al*. (1986) l'avaient estimée variant de une à cinq années. Ils proposaient alors un taux de substitution oscillant entre $1{,}58 \times 10^{-2}$ et $3{,}17 \times 10^{-3}$ substitutions par site et par année sur le gène *env* et entre $1{,}85 \times 10^{-3}$ et $3{,}70 \times 10^{-4}$ substitutions par site et par année sur le gène *gag*. Ces estimations sont donc très imprécises, car elles varient dans une fourchette de 1 à 5.

Pour contrer le problème dû à l'estimation de l'intervalle de temps entre le moment de divergence des séquences et le moment de collecte de celles-ci, nous devons utiliser des données temporelles connues. Li *et al*. (1988) proposent d'utiliser les dates de prélèvement des échantillons qui, elles, sont connues avec certitude. Pour les employer, nous devons toutefois utiliser une troisième séquence, servant d'*outgroup*, afin de mesurer la distance évolutive passée entre deux dates de prélèvement. En effet, le taux de substitution n'est pas égal à la distance évolutive entre deux échantillons divisée par l'intervalle de temps qui sépare leur date de prélèvement (Figure 11) (Drummond *et al.*, 2003). Cela produit une surestimation du taux de substitution, puisque la distance évolutive mesure le nombre de substitutions par site depuis leur divergence de leur ancêtre commun et qui a probablement existé bien avant leur date d'échantillonnage (Figure 11B). Notons que dans le cas où l'une des deux séquences est un ancêtre direct de l'autre, cette formule est exacte (Figure 11A), mais les cas sont rares.

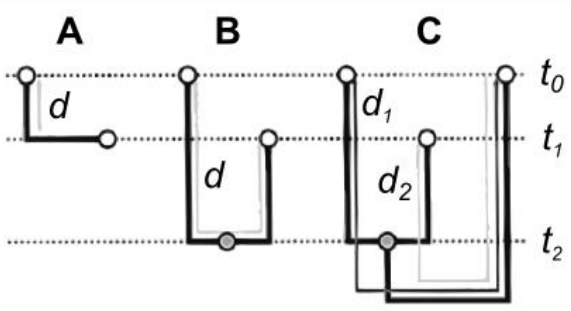
L'utilisation d'un *outgroup* permet donc d'obtenir la distance évolutive entre les deux dates de collecte (Figure 11C). Le choix de l'*outgroup* ne doit pas être fait au hasard, il doit être le plus proche possible des séquences d'intérêt afin d'obtenir une variance d'estimation faible. Soient trois séquences A, B et O où O réfère à l'*outgroup*. Les séquences A et B sont respectivement échantillonnées aux temps t_A et t_B, où t_A est plus récent que t_B, noté $t_B < t_A$, et $\hat{d}_{AO}$ et $\hat{d}_{BO}$ sont les distances

évolutives estimées (obtenues sous n'importe quel modèle) entre les séquences A et O, et, B et O respectivement. Alors le taux de substitution $\hat{\omega}$ vaut

$$\hat{\omega} = \frac{\hat{d}_{AO} - \hat{d}_{BO}}{t_A - t_B}.$$

En utilisant plusieurs séquences différentes comme *outgroup* et comme *ingroup*, dont notamment celles du jeune haïtien, Li *et al*. (1988) estiment un taux de substitution moyen à $5{,}9 \times 10^{-3}$ substitutions par site et par année sur le gène *env*. Avec cette méthode, Gojobori *et al*. (1994) estiment les taux de substitution synonyme et non synonyme du VIH-1 sur les gènes *env* et *gag*. Plusieurs souches y sont comparées et plusieurs estimations du taux de substitution synonyme et non synonyme sont présentées. En conclusion, ils retiennent que les taux de substitution synonyme et non synonyme sont respectivement de $26{,}0 \times 10^{-3}$ et $1{,}0 \times 10^{-3}$ substitutions par site et par année sur *gag* et respectivement de $35{,}5 \times 10^{-3}$ et $3{,}9 \times 10^{-3}$ substitutions par site et par année sur *env*. La différence entre les taux de substitution synonyme et non synonyme s'explique par le fait que les contraintes fonctionnelles appliquées sur le premier sont plus faibles que celles appliquées sur le second.

Figure 11. Relation entre distance évolutive et temps d'échantillonnage.
Schéma montrant la relation entre la distance évolutive et l'intervalle de temps qui sépare deux dates d'échantillonnage. Lorsqu'une souche est l'ancêtre commun d'une autre (figure A), la distance évolutive est proportionnelle au temps écoulé entre les deux dates de prélèvement et une estimation du taux de substitution est donnée en divisant la distance d par l'intervalle de temps $t_0 - t_1$, où t_0 est le temps le plus récent. Malheureusement, cela n'est pas le cas lorsqu'aucune des deux séquences n'est un ancêtre de l'autre (figure B). Dans ce cas, il est nécessaire d'utiliser un *outgroup* afin d'obtenir la distance évolutive $d = d_1 - d_2$ entre les deux temps de collecte t_0 et t_1 (figure C). Ainsi, le taux de substitution peut être estimé sur l'intervalle de temps entre t_0 et t_1 par $(d_1 - d_2)/(t_0 - t_1)$
Adaptation de Drummond *et al.* (2003a).



Bien que ces deux approches offrent des estimations cohérentes avec celles admises aujourd'hui (même ordre de grandeur), elles s'orientent vers une grande erreur type et ne peuvent être appliquées qu'à de petits jeux de données (Suzuki *et al*, 2000).

2.4.2 Les régressions linéaires simples

Le modèle de régression linéaire simple cherche à établir une relation linéaire entre une variable explicative $X = \{x_1, \cdots, x_n\}$ et une variable expliquée $Y = \{y_1, \cdots, y_n\}$, c'est-à-dire

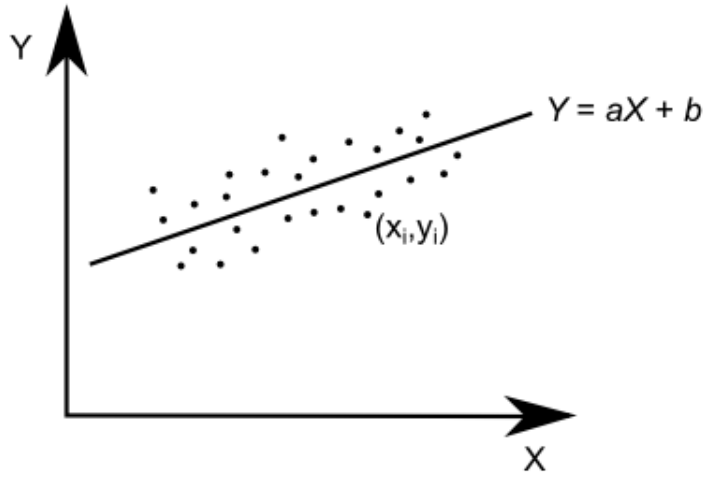
$$Y = aX + b + \varepsilon,$$

où les coefficients a et b sont les paramètres inconnus du modèle à estimer à l'aide des observations sur (X, Y). Le vecteur $\varepsilon = \{\varepsilon_1, \cdots, \varepsilon_n\}$ est le bruit associé au modèle (de moyenne nulle, c'est-à-dire $E[\varepsilon] = 0$), qui prend en compte le fait que la relation entre les variables X et Y n'est pratiquement jamais complètement expliquée par une droite. Afin de ne pas considérer cette erreur dans les notations, on note parfois

$$E[Y|X = x_i] = ax_i + b,$$

où $E[Y|X = x_i]$ représente la valeur moyenne de Y sachant la valeur x_i de X. Une régression linéaire peut être représentée par un graphique à deux dimensions sur lequel un nuage de points, de coordonnées (x_i, y_i), est approximé par une droite qui passe au plus près de ces points. Les coefficients de cette droite sont les paramètres a et b correspondant au modèle de régression linéaire.

Figure 12. Schéma représentant une régression linéaire.
Représentation graphique d'une régression linéaire. Chaque point (x_i,y_i) est représenté sur un graphique à deux dimensions et la droite qui passe au plus près de ces points est la régression linaire dont les coefficients (a et b) sont les paramètres du modèle.



L'estimation du taux de substitution à l'aide d'une régression linéaire ne peut être faite que sous le modèle SRDT, c'est-à-dire avec une horloge moléculaire stricte. Sous ce modèle, la variable Y est associée à la distance évolutive, la variable X au temps et le taux de substitution correspond donc au paramètre a. Sachant l'ensemble des points observés (temps, distance) le modèle cherche à établir une relation linéaire d'où découlera l'estimation du taux de substitution.

Une des faiblesses des modèles de régression linéaire est qu'ils supposent l'indépendance des observations (x_i, y_i) et donc, dans notre cas, des distances évolutives. Ce qui est faux puisque les séquences partagent une partie de leur histoire évolutive (Drummond *et al*, 2003a). Ce problème d'indépendance des données survient aussi dans plusieurs autres problèmes d'évolution, comme par

exemple dans les modèles d'évolution moléculaire qui supposent que les sites d'un alignement évoluent de manière indépendante (cf. Chapitre 1) (Morton & Clegg, 1995; Gutell *et al*, 1994). Les estimations résultant de ces méthodes doivent donc être interprétées avec précaution puisque l'utilisation de méthodes qui incorporent la notion d'indépendance peuvent induire des biais non prédictibles (Drummond *et al*, 2003a).

2.4.2.1 *Pairwise-Distance*

La régression linéaire *Pairwise-Distance* est introduite par Leitner et Albert (1999) dans le but de tester l'existence d'une horloge moléculaire stricte sur les gènes *env* et *gag* du VIH-1. Cette méthode se fonde sur un résultat de la génétique des populations qui dit qu'une population haploïde (resp. diploïde) de taille constante N_e partage un ancêtre commun à N_e générations dans le passé. Donc, deux séquences accumulent en moyenne $\Theta = 2N_e\omega_g$ (resp. $\Theta = 4N_e\omega_g$) mutations par site, où ω_g est le taux de substitution par site et par génération (Felsenstein, 2007; Rodrigo *et al*, 2007). Adapter ce résultat dans le cas où deux séquences i et j sont échantillonnées à des temps différents $t_i < t_j$, c'est-à-dire que t_j est plus récent que t_i, donne la relation linéaire

$$E[\hat{d}_{ij}] = \hat{\omega}(t_j - t_i) + \hat{\Theta},$$

où $\hat{\omega}$ est l'estimation du taux de substitution, $\hat{\Theta}$ une estimation de la diversité génétique des souches échantillonnées au temps t_i et $\hat{d}_{ij}$ la distance évolutive estimée entre les séquences i et j (Figure 13). Ainsi, la régression linéaire des variables $\hat{d}$ et des intervalles de temps d'échantillonnage fournit une estimation du taux de substitution ω et du paramètre Θ. La faiblesse de cette méthode est qu'elle suppose constante la distance génétique entre chaque paire de séquence prise au même temps, alors que celle-ci peut largement varier. Même si la méthode devient correcte lorsque le nombre de séquences est très important, elle est très largement sous-optimale dans la mesure où elle ignore totalement la phylogénie des séquences étudiées. Avec cette méthode, Leitner et Albert (1999) estiment le taux de substitution sur les gènes *gag* et *env* à $2{,}7 \pm 0{,}5 \times 10^{-3}$ substitutions par site et par année et à $6{,}7 \pm 2{,}1 \times 10^{-3}$ substitutions par site et par année respectivement.

2.4.2.2 *Root-to-tip*

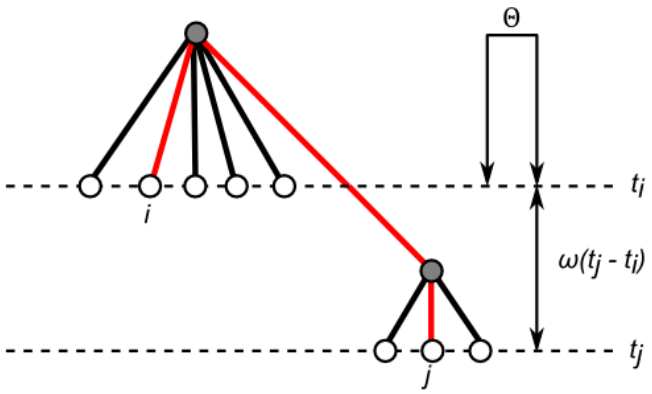
Cette méthode de régression linéaire est l'une des plus utilisées parce qu'elle permet d'estimer simultanément le taux de substitution ω et la date de l'ancêtre commun aux séquences t_{racine} (Drummond *et al*, 2003a). De ce fait, et contrairement à la régression *Pairwise-Distance*, cette méthode utilise une phylogénie enracinée des séquences étudiées, puis fait une régression linéaire entre les dates d'échantillonnage t_i de chaque séquence i avec la distance estimée $\hat{d}_{i,\text{racine}}$ qui sé-

pare la feuille représentant i de la racine (obtenue en additionnant les longueurs des branches de l'arbre situées sur le chemin de la feuille i jusqu'à la racine). Ainsi, le modèle linéaire (Figure 14) est

$$E\left[\hat{d}_{i,\text{racine}}\right] = \hat{\omega}(t_i - \hat{t}_{\text{racine}}),$$

où $\hat{\omega}$ et $\hat{t}_{\text{racine}}$ sont des estimations du taux de substitution et de la date de l'ancêtre commun aux séquences. L'intersection avec l'axe des abscisses donne l'estimation de t_{racine}, car, dans ce cas, on a $\hat{\omega}(t_i - \hat{t}_{\text{racine}}) = 0$, donc $t_i = \hat{t}_{\text{racine}}$ lorsque $\hat{\omega} \neq 0$. Avec cette méthode Korber *et al.* (2000) ont estimé, sur le gène *env*, la date de l'ancêtre commun aux souches appartenant au groupe du VIH-1 responsable de la pandémie actuelle (groupe M) à 1931 [1915-1941]. Leur estimation du taux de substitution est de $2{,}4 \times 10^{-3}$ $[1{,}8 \times 10^{-3}; 2{,}8 \times 10^{-3}]$ substitutions par site et par année. Sur le gène *gag*, ils estiment un taux de substitution à $1{,}9 \times 10^{-3}$ $[0{,}9 \times 10^{-3}$; $2{,}7 \times 10^{-3}]$ substitutions par site et par année et une date de l'ancêtre commun au VIH actuel à 1934 [1869 ; 1950].

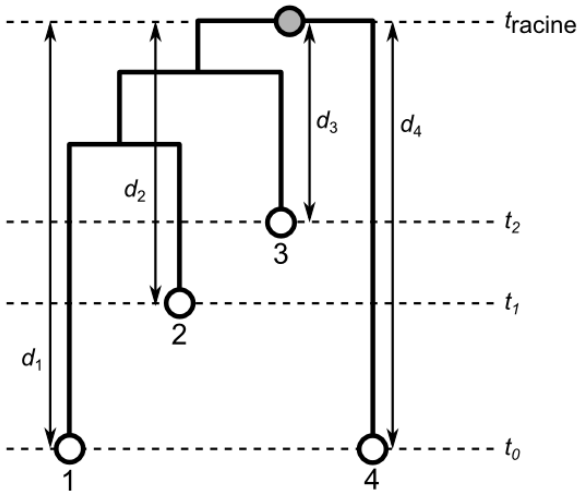
Figure 13. Modèle *Pairwise-Distance*.
Le modèle *Pairwise-Distance* suppose que la distance évolutive d_{ij}, séparant les souches i et j (en rouge), respectivement échantillonnées aux temps t_i et t_j (t_j est plus récent que t_i), est égale à la diversité génétique moyenne Θ entre chaque paire de séquences échantillonnées à t_i, plus la distance évolutive entre t_i et t_j (proportionnelle au taux de substitution ω à estimer). À savoir $d_{ij} = \omega(t_j - t_i) + \Theta$.



La connaissance de l'emplacement de la racine dans la phylogénie est donc primordiale pour utiliser cette méthode. Mais il est tout de même possible d'utiliser une phylogénie non enracinée. Dans ce cas, il est nécessaire de parcourir toutes les branches de la phylogénie afin de trouver l'emplacement optimal pour la racine. Par exemple, l'emplacement qui maximise le coefficient de corrélation de Pearson entre les dates de prélèvement t_i et les distances évolutives $\hat{d}_{i,\text{racine}}$, qui mesure la « qualité » de la régression linéaire. L'emplacement sur la phylogénie qui maximise ce coefficient est alors choisi comme racine et les paramètres sont estimés en fonction de cette racine. Cette méthode est mise en œuvre dans les versions antérieures à la version 1.3 du logiciel Path-O-Gen[2]. Depuis la version 1.3, Path-O-Gen localise l'emplacement optimal de la racine en minimisant la somme des résidus, c'est-à-dire l'écart des estimations à la droite de régression.

[2] http://tree.bio.ed.ac.uk/software/pathogen/

Figure 14. Modèle *Root-to-tip*.
Quatre souches (cercle blanc) sont échantillonnées à trois temps différents t_0, t_1 et t_2. Les valeurs d représentent la distance évolutive qui sépare chaque séquence i de la racine (cercle gris). Soit ω le taux de substitution, alors pour chaque séquence i, on a $d_i = \omega(t_i - t_{\text{racine}})$.



2.4.3 sUPGMA

Serial-Sample UPGMA (sUPGMA) est une méthode de distances d'inférence phylogénétique sous les hypothèses du modèle SRDT (Rodrigo *et al*, 2007; Drummond & Rodrigo, 2000). Elle est le prolongement de la méthode UPGMA (*unweighted pair grouping method with arithmetic means*, cf. Chapitre 1), qui est adaptée au modèle SR (Sokal & Michener, 1958). En effet, l'algorithme UPGMA construit une représentation où chaque feuille de l'arbre est à égale distance de la racine (cohérent si les séquences sont isochrones et si on suppose une horloge moléculaire stricte), c'est-à-dire une phylogénie ultramétrique ou un dendrogramme (Barthélemy & Guénoche, 1988). Or, le modèle SRDT implique que les feuilles sont échantillonnées à des dates différentes et doivent donc être à des distances différentes de la racine, en fonction de la date de collecte de ces dernières. Mais deux feuilles échantillonnées au même moment doivent se situer à la même distance de la racine. sUPGMA, mise en œuvre dans PEBBLE (Goode & Rodrigo, 2004), prend donc en considération les temps de collecte des feuilles dans le calcul de la phylogénie. Pour faire cela, quatre étapes principales sont nécessaires, sachant que la première étape est une méthode d'estimation du taux de substitution sous le modèle SRDT. Les autres étapes servent uniquement à calculer la phylogénie.

Estimation du taux de substitution

La première étape de la méthode sUPGMA consiste à estimer le taux de substitution relatif à l'ensemble des séquences. Soient p temps de prélèvement tels que le temps i est obtenu plus récemment que le temps $i+1$ ($i \in 1, \cdots p$). Soit $\hat{d}(m_i, n_j)$ la distance évolutive estimée entre la $i^{\text{ème}}$ séquence collectée à la date m et la $j^{\text{ème}}$ séquence collectée à la date n, avec $m \geq n$. Alors

$$\hat{d}(m_i, n_j) = \widehat{\Theta}_m + \widehat{\omega}(t_m - t_n) + \varepsilon_{m_i,n_j},$$

où t_k est la date du temps d'échantillonnage k, $\widehat{\omega}$ le taux de substitution à estimer et $\widehat{\Theta}_m$ la diversité génétique des séquences échantillonnées au temps m aussi à estimer. Les termes ε_{m_i,n_j} représentent les erreurs dues à l'estimation des distances évolutives. Il est possible d'exprimer ces équations à l'aide d'une notation matricielle. Soient D le vecteur contenant les estimations des distances évolutives, $\beta = \{\widehat{\Theta}_1, \cdots, \widehat{\Theta}_\mathrm{p}, \widehat{\omega}\}$ le vecteur des paramètres à estimer, et E le vecteur des erreurs, alors

$$D = M\beta + E$$

avec M la matrice telle que pour chaque ligne i et chaque colonne $j \leq p$

$$\left(M_{i,j}\right)_{m,n} = \begin{cases} 1 \text{ si } j = m \\ 0 \text{ sinon} \end{cases}$$

et $\left(M_{i,p+1}\right)_{m,n} = t_m - t_n$. Le vecteur des paramètres estimés $\beta = \{\widehat{\Theta}_1, \cdots, \widehat{\Theta}_\mathrm{p}, \widehat{\omega}\}$, qui minimise la somme des erreurs au carré $E^T E$, est alors donné par la méthode des moindres carrés :

$$\beta = (M^T M)^{-1} M^T D.$$

Cette méthode peut facilement être étendue au modèle MRDT (Drummond *et al*, 2001). Dans ce cas, il suffit de décomposer l'intervalle de temps $(t_m - t_n)$ en $(t_m - t_{m-1}) + \cdots + (t_{n+1} - t_n)$ et d'affecter à chaque intervalle de temps le taux de substitution correspondant. À l'inverse, une hypothèse simplificatrice est de supposer une diversité génétique constante quel que soit le temps d'échantillonnage. Cela revient à estimer qu'un seul paramètre Θ, au lieu d'un pour chaque temps de collecte. Dans ce cas, le modèle devient

$$\hat{d}_{ij} = \widehat{\Theta} + \widehat{\omega}(t_j - t_i)$$

et il est alors équivalent à la régression linéaire *Pairwise-Distance*.

Avec cette méthode les auteurs ont estimé le taux de substitution du VIH-1 sur des souches isolées chez un même patient, sur cinq temps d'échantillonnage couvrant 1 005 jours (Rodrigo *et al*, 1999). Leur estimation du taux de substitution sur le gène *env*, en considérant un paramètre Θ et un taux de substitution unique, est de $7{,}8 \times 10^{-6}$ $[-3{,}47 \times 10^{-6}; 3{,}87 \times 10^{-5}]$ substitutions par site et par jour. Ramenée à l'échelle des années, l'estimation est approximativement de 3×10^{-3} substitutions par site et par année.

Correction de la matrice de distances

Une fois le taux de substitution estimé, il est alors possible de corriger la matrice de distances $\hat{d}$ en ajoutant, à chaque distance estimée, la mesure manquante afin de voir i et j comme contemporains, c'est-à-dire que

$$\hat{d}'_{ij} = \hat{d}_{ij} + \hat{\omega}(t_0 - t_i) + \hat{\omega}(t_0 - t_j),$$

où t_i et t_j réfèrent au temps de collecte des souches i et j, et où le temps d'échantillonnage le plus récent est noté t_0. La mesure d' voit alors les séquences i et j comme contemporaines (c'est-à-dire échantillonnées au temps t_0).

Calcul de l'arbre à l'aide de UPGMA

Un arbre UPGMA ou WPGMA est calculé à partir de la mesure corrigée $\hat{d}'$ qui voit toutes les souches comme contemporaines, c'est-à-dire que toutes les souches doivent se situer à égale distance de la racine.

Modification de l'arbre UPGMA

L'arbre UPGMA ou WPGMA obtenu est ultramétrique, c'est-à-dire que toutes les feuilles sont à égale distance de la racine. Afin d'obtenir un arbre où chaque feuille collectée à un temps d'échantillonnage différent est à une distance différente de la racine, mais où toutes les feuilles d'un même temps d'échantillonnage sont à une même distance de la racine, il suffit de soustraire la mesure $\hat{\omega}(t_0 - t_i)$ à la longueur de la branche associée à la séquence i. De cette façon, la topologie obtenue respecte celle du modèle SRDT.

2.4.4 TREBLE

Tree and rate estimation by local evaluation (TREBLE) est une méthode estimant le taux de substitution à partir d'un ensemble de séquences hétérochrones et en faisant l'hypothèse d'une horloge moléculaire stricte, donc sous les hypothèses du modèle SRDT (Yang *et al*, 2007). Cette méthode utilise des triplets de séquences, c'est-à-dire que pour chaque triplet de séquences possible, vérifiant une certaine condition, un taux de substitution et sa variance sont estimés, puis elle calcule la moyenne des taux de substitution estimés sur chaque triplet, pondérée par l'inverse de la variance correspondante, afin d'obtenir un taux de substitution global, solution du problème.

Cette méthode part de l'observation que pour deux séquences données i et j, échantillonnées respectivement aux temps t_i et t_j, il existe une relation entre leur distance génétique estimée $\hat{d}_{ij}$, leur taux de substitution ω_{ij} et leur temps de collecte, telle que (Figure 15A)

$$\hat{d}_{ij} = \omega_{ij}(t_i + t_j - 2t_{ij}) + \varepsilon_{ij},$$

où t_{ij} réfère à la date de l'ancêtre commun aux souches i et j et ε_{ij} aux erreurs associées à l'estimation des distances évolutives $\hat{d}_{ij}$, négligées par la suite. Comme les paramètres ω_{ij} et t_{ij} sont inconnus, l'équation n'a pas de solution unique. Cet handicap peut être résolu en considérant une séquence supplémentaire k, échantillonnée au temps t_k, mais ayant une configuration topologique particulière avec les deux autres séquences i et j (Figure 15B). Considérant en plus cette troisième séquence et leur configuration géométrique, il est maintenant possible d'estimer le taux de substitution et les dates de leurs ancêtres communs par les équations

$$\hat{t}_{ik} = \hat{t}_{jk} = \frac{1}{2}\left[\frac{\hat{d}_{ik}t_j - \hat{d}_{jk}t_i}{\hat{d}_{ik} - \hat{d}_{jk}} + t_k\right],$$

$$\hat{t}_{ij} = \frac{1}{2}\left[(t_i + t_j) - \frac{\hat{d}_{ij}(t_i - t_j)}{(\hat{d}_{ik} - \hat{d}_{jk})}\right]$$

et

$$\widehat{\omega}_{ij|k} = \frac{\hat{d}_{ik} - \hat{d}_{jk}}{t_i - t_j}.$$

Le taux de substitution estimé (noté par $\widehat{\omega}$) est relatif aux séquences i et j sachant la séquence supplémentaire k (d'où la notation $\widehat{\omega}_{ij|k}$). Ce dernier dépend seulement des temps d'échantillonnage correspondants aux séquences i et j appelées « paire informative ». La séquence restante k est appelée *outgroup*. Cette formule est identique à celle proposée par Li *et al*. (1988), présentée à la section 2.4.1, c'est-à-dire qu'elle estime le taux de substitution à partir de la distance évolutive entre deux temps d'échantillonnage et, pour cette raison, elle nécessite l'utilisation d'un *outgroup*. Les dates des ancêtres communs étant aussi estimées, elles sont aussi notées $\hat{t}$. La topologie nécessaire pour de telles estimations n'est *a priori* pas connue et l'utilisation de triplets quelconques peut conduire à des estimations non valides. Ainisi, les estimations valides sont celles qui vérifient les conditions suivantes : $\widehat{\omega}_{ij|k} > 0$, $\hat{t}_{ij} \leq t_i$, $\hat{t}_{ij} \leq t_j$, $\hat{t}_{ik} \leq t_k$, $\hat{t}_{ik} \leq \hat{t}_{ij}$, en accord avec la configuration géométrique que doit présenter le triplet (Figure 15B).

Plusieurs sources d'erreur peuvent causer des biais dans l'estimation des taux de substitution $\widehat{\omega}_{ij|k}$. Par exemple, des erreurs inhérentes à l'estimation des distances évolutives dues aux substitutions cachées (cf. Chapitre 1). Afin d'augmenter la précision de l'estimation du taux de substitution global, Yang *et al*. (2007) proposent d'associer à chaque $\widehat{\omega}_{ij|k}$ une variance représentant la confiance associée à l'estimation. Suivant Rzhetsky et Nei (1995), la covariance entre les distances évolutives d_{ab} et d_{cd}, des séquences a, b, c et d, est égale à la variance de la longueur des branches partagées par les chemins reliant les séquences a à b et les séquences c à d, notée $d_{ab,cd}$. Soit

$$\mathrm{cov}(d_{ab}, d_{cd}) = \mathrm{var}(d_{ab,cd}).$$

Pour le triplet de la Figure 15B, il vient

$$\mathrm{cov}(d_{ik}, d_{jk}) = \mathrm{var}(d_{ik,jk}) = \mathrm{var}(d_{lk})$$

où l représente l'ancêtre commun des séquences i et j, c'est-à-dire celui au temps t_{ij}. Avec cette observation, la variance associée au taux de substitution estimé est

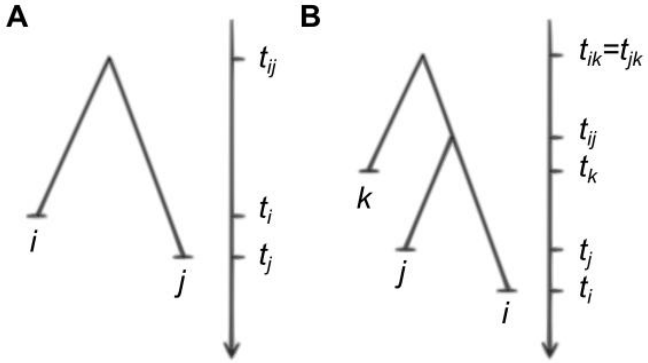
$$\begin{aligned} \mathrm{var}(\hat{\omega}_{ij|k}) &= \mathrm{var}\left(\frac{\hat{d}_{ik} - \hat{d}_{jk}}{t_i - t_j}\right) \\ &\approx \frac{\mathrm{var}(\hat{d}_{ik}) + \mathrm{var}(\hat{d}_{jk}) - 2\mathrm{cov}(\hat{d}_{ik}, \hat{d}_{jk})}{(t_i - t_j)^2} \\ &\approx \frac{\mathrm{var}(\hat{d}_{ij})}{(t_i - t_j)^2}. \end{aligned}$$

Remarquons que la variance est indépendante de l'*outgroup*. Elle est donc identique pour chaque paire informative quel que soit l'*outgroup* considéré. La plupart des modèles d'évolution moléculaire propose une formule analytique pour calculer la variance de la distance évolutive (Rzhetsky & Nei, 1995).Cependant, elle peut aussi être approximée par

$$\mathrm{var}(\hat{d}_{ij}) \approx \frac{\hat{d}_{ij}}{N},$$

où N est la longueur des séquences dans l'alignement (Gascuel, 2000; Bulmer, 1991).

Figure 15. Illustration et comportement d'une paire de séquence et d'un triplet de séquence.
La figure A montre deux séquences i et j respectivement échantillonnées aux temps t_i et t_j et divergent de leur ancêtre commun au temps t_{ij}. La figure B montre trois séquences i, j et k respectivement échantillonnées aux temps t_i, t_j et t_k. Les souches i et j divergent de leur ancêtre commun au temps t_{ij} et les souches k et i, ainsi que les souches k et j, divergent de leur ancêtre commun au temps $t_{ik} = t_{jk}$.
Adaptation de Yang *et al.* (2007).



Une fois la connaissance de toutes les paires informatives et des *outgroups* valides, TREBLE calcule pour chaque paire informative i et j une moyenne $\widehat{\omega}_{ij}$ des taux de substitution estimés avec chaque *outgroup*

$$\widehat{\omega}_{ij} = \frac{1}{|O_{ij}|} \sum_{k \in O_{ij}} \widehat{\omega}_{ij|k},$$

où O_{ij} est l'ensemble des *outgroups* retenus pour la paire informative i et j. Le taux de substitution global $\widehat{\omega}$ est alors donné par la moyenne pondérée de chaque $\widehat{\omega}_{ij}$

$$\widehat{\omega} = \frac{1}{W} \sum_{i,j} w_{ij} \widehat{\omega}_{ij},$$

avec $w_{ij} = 1/\mathrm{var}(\widehat{\omega}_{ij})$ et W= $\sum_{i,j} w_{ij}$.

Après avoir estimé le taux de substitution global, TREBLE propose de vérifier à nouveau la validité des *outgroups* associés à chaque paire informative mais en considérant cette fois-ci l'estimation globale du taux de substitution. Pour cela, il impose une contrainte sur les distances $\hat{d}_{jk}$, telle que

$$\widehat{\omega} t_i + \hat{d}_{jk} - \widehat{\omega} t_k - \hat{d}_{ij} > \varepsilon_{jk} - \varepsilon_{ij},$$

où les ε sont les erreurs provenant de l'estimation des distances évolutives. Vérifier cette contrainte, c'est vérifier que la moitié de la distance évolutive entre t_{ij} et t_{jk} est strictement positive, puisqu'on a l'égalité $\widehat{\omega} t_i + \hat{d}_{jk} - \widehat{\omega} t_k - \hat{d}_{ij} + 2\widehat{\omega}(t_{jk} - t_{ij}) = \varepsilon_{jk} - \varepsilon_{ij}$. Les erreurs ε sont des variables aléatoires inconnues distribuées suivant une loi normale d'espérance nulle (donc leur différence aussi). Ainsi, $(\varepsilon_{jk} - \varepsilon_{ij})/\sqrt{\mathrm{var}(\varepsilon_{jk} - \varepsilon_{ij})}$ suit une loi normale centrée réduite, avec $\mathrm{var}(\varepsilon_{jk} - \varepsilon_{ij}) \approx \mathrm{var}(\hat{d}_{ij})$ d'après les formules de variance ci-dessus. Soit Z_α la valeur correspondant du quantile α obtenue dans la table de la loi normale centrée réduite, avec α choisi par l'utilisateur. Alors la probabilité pour que $(\varepsilon_{jk} - \varepsilon_{ij}) < Z_\alpha \sqrt{\mathrm{var}(\hat{d}_{ij})}$ est $1 - \alpha$. Si $\widehat{\omega} t_i + \hat{d}_{jk} - \widehat{\omega} t_k - \hat{d}_{ij} > Z_\alpha \sqrt{\mathrm{var}(\hat{d}_{ij})}$, alors la probabilité pour que $\widehat{\omega} t_i + \hat{d}_{jk} - \widehat{\omega} t_k - \hat{d}_{ij} > (\varepsilon_{jk} - \varepsilon_{ij})$ est au moins $1 - \alpha$. Ainsi, les *outgroups* ne satisfaisant pas cette contrainte sont supprimés et la procédure complète est recommencée sans ces *outgroups*. Un nouveau taux de substitution global est alors estimé et ce dernier test répété. Et ceci jusqu'à stabilisation des *outgroups*.

Nous pouvons donc voir cette méthode comme une généralisation de la méthode proposée par Li *et al*. (1988), mais où les critères statistiques de sélection conservent uniquement les *outgroups* qui forment une configuration bien précise avec les paires de séquences informatives, donc ceux qui

permettent une bonne estimation du taux de substitution. Ces critères sont aussi une faiblesse de cette méthode parce qu'ils rejettent, en pratique, beaucoup de triplets et donc de l'information. Malgré les promesses de vitesse et de performance encourageantes, aucune application biologique concrète[3] n'a été faite avec cette méthode, outre celles des auteurs.

Une adaptation de cette méthode dans le cas où l'on considère trois clades différents dans la phylogénie, par exemple représentant chacun un sous-type différent au sein d'un même virus, est proposée par O'Brien *et al*. (2008). Cette dernière méthode estime la date de divergence entre deux clades (sachant le troisième) et le taux de substitution relatif aux deux clades considérés.

2.4.5 *TreeRate*

TreeRate est une méthode qui se base sur une phylogénie racinée pour estimer la distance génétique séparant deux groupes de séquences choisis par l'utilisateur (Maljkovic Berry *et al*, 2009, 2007). Elle permet aussi d'estimer le taux de substitution sous les hypothèses du modèle SRDT. Pour faire cela, l'utilisateur choisit préalablement deux collections de feuilles assignées respectivement au groupe T1 et T2 (Figure 16). Certaines feuilles peuvent être écartées de l'analyse. Elles sont alors considérées comme *outgroup*. Une moyenne $\bar{X}_1$ (respectivement $\bar{X}_2$) des distances de chaque feuille du groupe T1 (resp. T2) à la racine est calculée. Ainsi, la distance génétique qui sépare les deux groupes de feuilles est calculée comme la différence entre ces deux moyennes, soit $\Delta\hat{d} = \bar{X}_2 - \bar{X}_1$. Dès lors, il est possible d'estimer le taux de substitution ω entre ces deux groupes en s'aidant de Δt un intervalle de temps calculé à partir de $\bar{T}_1$ (resp. de $\bar{T}_2$) qui représente la moyenne des dates de collecte des feuilles de T1 (resp. T2), ainsi

$$\widehat{\omega} = \frac{\Delta\hat{d}}{\Delta t} = \frac{\bar{X}_2 - \bar{X}_1}{\bar{T}_2 - \bar{T}_1}.$$

Dans le cas d'une phylogénie non enracinée, *TreeRate* estime au préalable la position optimale de la racine suivant un test statistique. Plusieurs tests sont proposés par les auteurs, mais ils suggèrent que celui consistant à minimiser la somme des variances est le plus performant, c'est-à-dire à minimiser le terme

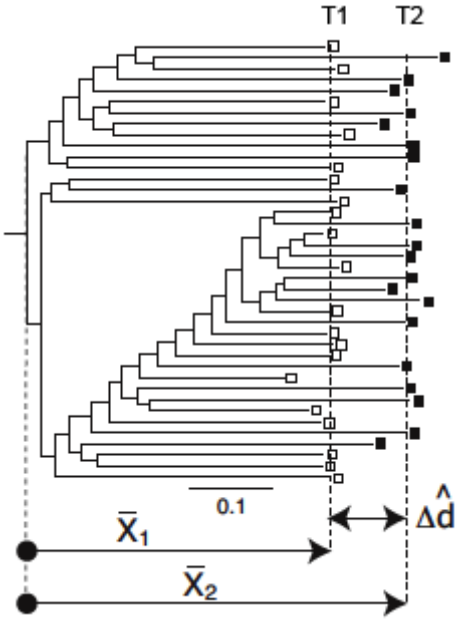
$$\sigma^2 = \left[\frac{1}{N_1}\sum_{i=1}^{N_1}\left(\hat{X}_i - \bar{X}_1\right)^2\right] + \left[\frac{1}{N_2}\sum_{j=1}^{N_2}\left(\hat{X}_j - \bar{X}_2\right)^2\right],$$

où N_i est le nombre de feuilles contenues dans le groupe Ti et X_i la distance de la feuille i à la racine. Cette méthode très simple n'est donc pas vraiment originale, puisque très proche de méthodes déjà

[3] D'après PubMed (http://www.ncbi.nlm.nih.gov/pubmed), consultée le 2 février 2012.

discutées plus haut (comme sUPGMA, TREBLE ou encore *Root-to-tip*). Elle est mise en œuvre sur le site web de la base de données sur le VIH du laboratoire national de Los Alamos[4]. Avec cette méthode, les auteurs montrent que sur la région *env* du génome du VIH-1 le taux de substitution du sous-type C est de $9{,}65 \times 10^{-3}$ [$8{,}88 \times 10^{-3}$; $10{,}4 \times 10^{-3}$] substitutions par site et par année alors que celui du sous-type A est de $16{,}9 \times 10^{-3}$ [$12{,}1 \times 10^{-3}$; $21{,}6 \times 10^{-3}$] substitutions par site et par année (à partir d'échantillons collectées en Afrique uniquement) (Maljkovic Berry *et al*, 2007).

Figure 16. Illustration de la méthode *TreeRate*.
La distance moyenne des feuilles du groupe T1 à la racine est noté $\bar{X}_1$, celle des feuilles du groupe T2 à la racine $\bar{X}_2$. La mesure $\Delta\hat{d} = \bar{X}_2 - \bar{X}_1$ représente la distance qui sépare les feuilles du groupe T1 et T2. Le taux de substitution $\hat{\omega}$ vaut $\Delta\hat{d}/\Delta t$, où Δt représente l'intervalle de temps entre la moyenne arithmétique des dates d'échantillonnage des feuilles de T1 et celle des feuilles de T2.
Source : http://www.hiv.lanl.gov/content/sequence/TREERATE/treerate_explanation.html



2.4.6 Méthode de Langley-Fitch

La méthode de Langley-Fitch (1974), mise en œuvre dans r8s (Sanderson, 2003), permet d'estimer simultanément le taux de substitution et les dates des ancêtres communs d'une phylogénie racinée sous les hypothèses du modèle SRDT. Plus tard, Sanderson (2002) propose la méthode *Penalized Likelihood* qui est une adaptation de la méthode Langley-Fitch au modèle DR, c'est-à-dire avec un taux de substitution pour chaque branche de la phylogénie, et une corrélation entre les taux des branches adjacentes. Ces deux méthodes suivent une approche de distance mais posent un modèle probabiliste sur le bruit et les (éventuelles) corrélations, si bien que ce sont aussi des méthodes probabilistes avec des temps de calcul plus importants que les méthodes précédentes, mais moins que ceux des méthodes pleinement probabilistes basées sur les caractères, comme BEAST par exemple.

[4] http://www.hiv.lanl.gov/content/sequence/TREERATE/combinedBranchlength.html

Supposons que la phylogénie a S nœuds internes représentés par un nombre entier de 0 à S, où 0 représente le nœud racine, et M feuilles représentées par les nombres entiers $S+1$ à $S+M$. L'âge du nœud k est noté t_k et $\mathrm{anc}(k)$ représente le nœud ancestral à k dans la phylogénie. Notons par ω le taux de substitution et par b_i la longueur de la branche $\big(i,\mathrm{anc}(i)\big)$. Les paramètres du modèle (SRDT) à estimer dans la méthode de Langley-Fitch sont donc $\Omega = \{t_0, \cdots, t_S, \omega\}$. On s'intéresse ici au nombre de substitutions par site sur des durées déterminées par les temps de prélèvement et sur la phylogénie. Le modèle Poissonien offre un cadre naturel.

Supposons que le nombre de substitutions par unité de temps et par site suit une loi de Poisson de moyenne ω. Alors le nombre de substitutions par site sur une branche $\big(i,\mathrm{anc}(i)\big)$ suit aussi une loi de Poisson mais de moyenne $\omega\big(t_{\mathrm{anc}(i)} - t_i\big)$. Autrement dit, la probabilité d'avoir b_i substitutions par site sur la branche i est $P(b_i|\,\omega\big[t_{\mathrm{anc}(i)} - t_i\big])$, avec $P(b|a) = a^b \exp(-a)/b!$. Le logarithme de la vraisemblance de l'arbre tout entier est donné par

$$\log L(\Omega|b_1, \cdots, b_{S+M}) = \sum_{k=1}^{S+M} \log P\big(b_k|\omega\big[t_{\mathrm{anc}(k)} - t_k\big]\big),$$

et les valeurs des paramètres Ω qui maximisent ce logarithme sont les estimateurs du maximum de vraisemblance. Maximiser cette expression ne pose pas de problème majeur et peut être réalisé par une approche standard.

2.5 Quelques méthodes pleinement probabilistes

Les méthodes probabilistes présentent un avantage certain en précision d'estimation par rapport aux méthodes de distances. Également, avec des méthodes probabilistes, les paramètres du modèle d'évolution peuvent être auto-estimés (cf. Chapitre 1). Toutefois, ces méthodes ne peuvent être appliquées qu'à des petits jeux de données (quelques centaines de séquences au plus), en raison des temps de calcul considérables qu'elles nécessitent. Dans cette section nous présentons brièvement deux méthodes probabilistes permettant d'estimer le taux de substitution à partir des hypothèses du modèles SRDT. La première méthode, *TipDate*, utilise le principe du maximum de vraisemblance et la seconde, BEAST, utilise une approche bayésienne (Drummond *et al*, 2012; Drummond & Rambaut, 2007; Rambaut, 2000). Cette dernière est actuellement la méthode de référence dans le domaine.

TipDate est une méthode développée par Rambaut (2000) qui permet d'estimer simultanément une phylogénie, les dates associées à chaque nœud interne de celle-ci, ainsi que le taux de substitution et cela sous les hypothèses du modèle SRDT. Plus tard, Drummond *et al*. (2001) l'adaptent au modèle MRDT. Cette méthode estime les dates des ancêtres communs et le taux de substitution en

remplaçant dans la procédure décrite par Felsenstein (1981), les longueurs de branche par le produit du taux de substitution et de l'intervalle de temps correspondant à cette branche (obtenu en soustrayant les dates associées aux nœuds adjacents). Les estimations de ces paramètres sont alors ceux qui maximisent la fonction de vraisemblance

$$L(\omega) = P(D|T, \omega, M),$$

où D représente l'alignement, ω est le taux de substitution, T la phylogénie (supposée suivre ici une horloge moléculaire stricte) et M les paramètres associés au modèle d'évolution.

BEAST (*bayesian evolutionary analysis by sampling trees*) est le logiciel d'estimation de taux de substitution le plus utilisé aujourd'hui (Drummond *et al*, 2012; Drummond & Rambaut, 2007; Drummond *et al*, 2002). Ce qui en fait son succès est sans doute les multiples services qu'il propose. Il est bien sûr possible d'y estimer le taux de substitution sous le modèle SRDT, mais ce logiciel donne aussi la possibilité d'utiliser d'autres modèles d'horloge moléculaire, comme par exemple des horloges moléculaires relâchées où les taux de substitution varient au niveau des nœuds internes (horloge moléculaire relâchée en exponentiel) ou le long des branches auxquelles ils sont associés (horloge moléculaire relâchée en log-normal) (Drummond *et al*, 2006). Il donne aussi la possibilité d'inférer une phylogénie mise à l'échelle temporelle (et donc il estime aussi les dates des ancêtres communs à chaque nœud) sous une large gamme de modèles d'évolution, ou d'obtenir le graphique représentant la taille effective de la population en fonction du temps $N_e(t)$ (cf. section 2.4.2.1) sous plusieurs modèles démographiques. Une option spéciale *BEAST (prononcée « star BEAST ») permet d'utiliser simultanément plusieurs régions d'un génome afin d'obtenir des résultats globaux (Heled & Drummond, 2010). Depuis peu, la reconstruction de caractères ancestraux, comme des régions géographiques est mise à disposition (Lemey *et al*, 2010, 2009a). Ce logiciel offre de multiples autres possibilités et les études l'utilisant sont très nombreuses, notamment en épidémiologie moléculaire. Citons en exemple, Dalai *et al*. (2009) qui estiment à $2{,}19 \times 10^{-3}$ [$1{,}83 \times 10^{-3}$; $2{,}56 \times 10^{-3}$] substitutions par site et par année le taux de substitution du VIH-1 sur le gène *pol*, avec le modèle d'horloge moléculaire stricte, et Bello *et al*. (2008) qui estiment à $1{,}5 \times 10^{-3}$ [$1{,}0 \times 10^{-3}$; $2{,}0 \times 10^{-3}$] et à $5{,}8 \times 10^{-3}$ [$3{,}8 \times 10^{-3}$; $7{,}8 \times 10^{-3}$] substitutions par site et par année le taux de substitution du VIH-1 sous le modèle d'horloge moléculaire stricte pour les gènes *pol* et *env* respectivement. Cependant, le point faible de ce logiciel est le temps de calcul considérable qu'il demande sur un jeu de données d'à peine quelques centaines de séquences. En effet, ce logiciel utilise le principe bayésien des chaînes de Markov par technique de Monte Carlo (MCMC) (cf. Chapitre 1), avec la variante de Metropolis-Hasting (Hastings, 1970; Metropolis *et al*, 1953), qui nécessite une quantité très importante de calculs pour approximer au mieux la distribution *a posteriori* des paramètres

d'intérêt à partir de données et d'une distribution *a priori* ou *prior*. De plus, cette *prior* en fait une méthode assez controversée puisqu'utilisée à tort, elle permet généralement d'obtenir des résultats souhaités.

Tableau 1. Récapitulatif des taux de substitution du VIH estimés par les différentes méthodes.
Les taux de substitution du VIH sont donnés en substitutions par site et par année. Les gènes sur lesquels le taux de substitution est estimé sont précisés et lorsque la méthode utilisée ne porte pas de nom particulier, la référence de l'article est donnée à la place. La liste des taux de substitution, triée par gène, correspond aux estimations citées dans le chapitre et n'est en rien exhaustive par rapport à la littérature.

Méthode	Gène	Taux de substitution ($\times 10^{-3}$)			Référence
			Min	Max	
Hahn *et al*. (1996)	*env*	-	3,17	15,80	Hahn *et al*. (1986)
Li *et al*. (1988)	*env*	5,90	-	-	Li *et al*. (1988)
Li *et al*. (1988)	*env*	35,50[a]	-	-	Gojobori *et al*. (1994)
Li *et al*. (1988)	*env*	3,90[b]	-	-	Gojobori *et al*. (1994)
Pairwise-Distance	*env*	6,70	4,60	8,80	Leitner et Albert (1999)
Root-to-tip	*env*	2,40	1,80	2,80	Korber *et al*. (2000)
sUPGMA	*env*	3,00	-1,34	14,89	Drummond et Rodriguo (2000)
TreeRate	*env*	9,65	8,88	10,40	Maljkovic Berry *et al*. (2007)
TreeRate	*env*	16,90	12,10	21,60	Maljkovic Berry *et al*. (2007)
BEAST	*env*	5,80	3,80	7,80	Bello *et al*. (2008)
Hahn *et al*. (1986)	*gag*	-	0,37	1,85	Hahn *et al*. (1986)
Li *et al*. (1988)	*gag*	26,00[a]	-	-	Gojobori *et al*. (1994)
Li *et al*. (1988)	*gag*	1,00[b]	-	-	Gojobori *et al*. (1994)
Pairwise-Distance	*gag*	2,70	2,20	3,20	Leitner et Albert (1999)
Root-to-tip	*gag*	1,90	0,90	2,70	Korber *et al*. (2000)
BEAST	*pol*	2,19	1,83	2,56	Dalai *et al*. (2009)
BEAST	*pol*	1,50	1,00	2,00	Bello *et al*. (2008)

[a]Taux de substitution synonyme
[b]Taux de substitution non synonyme

2.6 Conclusion

Nous présentons dans ce chapitre différentes méthodes qui permettent d'estimer le taux de substitution, c'est-à-dire la vitesse évolutive, sous les hypothèses du modèle SRDT (horloge moléculaire stricte et séquences hétérochrones), comme les régressions linéaires *Pairwise-Distance* et *Root-to-tip*, les méthodes de distances sUPGMA, TREBLE et *TreeRate*, la méthode probabiliste Langley-Fitch qui utilise une approche de distance et les méthodes pleinement probabilistes *TipDate* (vraisemblance) et BEAST (bayésien). Certaines de ces méthodes sont étendues à des modèles d'horloge moléculaire plus complexes, comme le modèle MRDT (sUPGMA ou *TipDate*) ou le modèle DR (Langely-Fitch ou BEAST) et d'autres nécessitent de l'information supplémentaire, comme un arbre enraciné (Langley-Fitch) ou l'intervention de l'utilisateur afin de considérer deux groupes de séquences à par-

tir desquels le taux sera estimé (*TreeRate*). Les estimations du taux de substitution du VIH données tout au long de ce chapitre permettent de se faire une bonne idée sur l'ordre de grandeur de celui-ci et ne peuvent en aucun cas être utilisées pour comparer la performance des méthodes entre elles, étant donné que les jeux de données sont différents les uns des autres (Tableau 1). Ces estimations suggèrent que la vitesse évolutive du VIH est plus élevée sur le gène *env* que sur les gènes *gag* et *pol*. En effet, le gène *env* code pour un précurseur des glycoprotéines gp120 et gp41 qui sont exposées à la surface du virion et mutent beaucoup afin de chercher à échapper au système immunitaire.

Dans cette thèse nous proposons une méthode de distances qui permet d'estimer rapidement le taux de substitution sous les hypothèses du modèle SRDT, tout en gardant une bonne précision d'estimation. Un des objectifs est de pouvoir analyser de très grands jeux de données afin de stabiliser les estimations du taux de substitution proposées dans la littérature, et cela sous un modèle donné.

Chapitre 3

Diversité génétique, épidémiologie moléculaire et origine du virus de l'immunodéficience humaine (VIH), l'agent responsable du SIDA

Le virus de l'immunodéficience humaine (VIH) présente une grande diversité génétique. Deux types (VIH-1 et VIH-2), quatre groupes pour le VIH-1 (M, N, O et P) et huit pour le VIH-2 (A à H). Chaque groupe est le résultat d'une transmission inter-espèce d'un virus infectant les singes d'Afrique à l'homme. Les virus du groupe M du VIH-1 sont responsables de l'épidémie mondiale et sont sous-divisés en sous-types, sous-sous-types et de nombreux recombinants circulants ou uniques. Déjà neuf sous-types (A à D, F à H, J et K) sont décrits pour le groupe M du VIH-1 et un nombre croissant de variants recombinants. Cependant, tous ces variants génétiques n'ont pas la même implication dans l'épidémie mondiale, certains sont très prévalents, d'autres peu et leur distribution est hétérogène. Les facteurs biologiques (recombinaisons, sélection naturelle, etc.) ayant conduit à cette diversité ainsi que les facteurs sociologiques (guerres, migrations, etc.) liés à l'expansion sont décrits dans ce chapitre, tout comme ses conséquences. Enfin, les origines zoonotiques de ces virus sont aussi discutées.

Sommaire

3.1 Introduction

Au début des années quatre-vingt, des médecins américains s'aperçoivent que certains de leurs patients présentent des infections généralement observées chez les nouveau-nés ou chez les personnes ayant un système immunitaire affaibli (pneumonies dues à *Pneumocystis carinii*, sarcomes de Kaposi, etc.). Ces patients étaient tous des hommes jeunes, préalablement en parfaite santé mais avaient des rapports sexuels avec d'autres hommes. L'examen de leur sang a montré une baisse du nombre de lymphocytes, confirmant un disfonctionnement de leur système immunitaire. Le 5 juin 1981, le *Center for Disease Control* (CDC) d'Atlanta publie dans son bulletin hebdomadaire *Morbidity and Mortality Weekly Report* (MMWR) la description de ces cas qui ont été observés à Los Angeles entre octobre 1980 et mai 1981 (Pneumocystis pneumonia--Los Angeles, 1981). L'année suivante, le terme SIDA pour « syndrome de l'immunodéficience acquise » est pour la première fois employé dans la littérature afin de désigner cette nouvelle maladie (Update on acquired immune deficiency syndrome (AIDS)--United States, 1982). L'agent viral du SIDA, quant à lui, est identifié en 1983 à l'Institut Pasteur de Paris par l'équipe de Luc MONTAGNIER (Barré-Sinoussi *et al*, 1983), mais c'est seulement en 1986 que le terme HIV, acronyme de « *human immunodeficiency virus* », est proposé afin de désigner ce virus (Coffin *et al*, 1986).

Depuis, et selon les estimations du programme commun des Nations Unies sur le VIH/SIDA (ONUSIDA), le virus de l'immunodéficience humaine (VIH) a déjà causé plus de 27 millions de décès (ONUSIDA, 2010). Le nombre annuel de nouvelles infections au VIH a connu un pic en 1996 (3,5 millions de nouvelles infections au VIH) suivi d'une diminution régulière de ce chiffre (estimé à 2,6 millions en 2009) mais qui reste toujours alarmant (Figure 17). En raison de la latence du virus entre le moment de l'infection et celui de la phase symptomatique, le nombre annuel de décès liés au SIDA atteint son paroxysme en 2004 (2,2 millions de décès). Au total, le nombre de personnes vivant avec le VIH en 2009 est estimé à 33,3 millions (ONUSIDA, 2010), mais les régions du globe ne sont pas égales face à l'ampleur de cette épidémie (Tableau 2). Plus des deux-tiers des personnes infectées vivent en

Afrique subsaharienne (22,5 millions de personnes), suivie de loin par l'Asie (6,2 millions de personnes).

Figure 17. Nombre de personnes nouvellement infectées par le VIH.
La courbe en rouge représente l'évolution du nombre de personnes nouvellement infectées par le VIH (en millions) entre 1990 et 2009. Les courbes en pointillés indiquent l'intervalle de confiance de ces estimations.
Adaptation de ONUSIDA (2010).

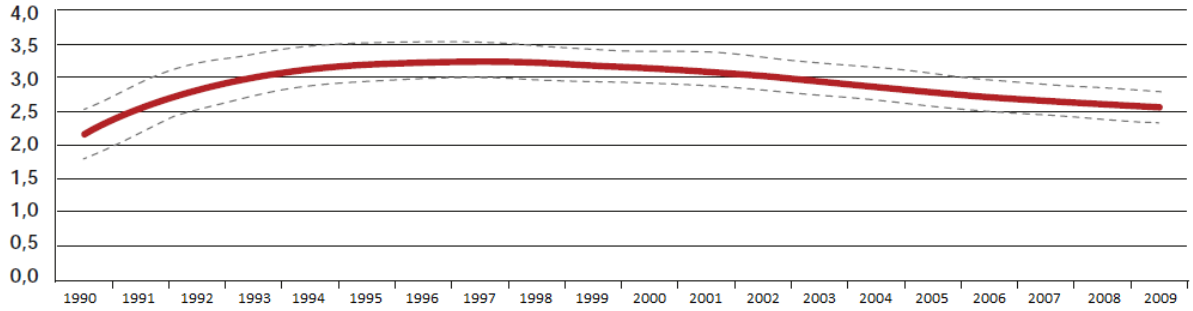


Tableau 2. Estimations de l'ONUSIDA du nombre de personnes vivant avec le VIH en 2009.
Estimations de l'ONUSIDA du nombre de personnes vivant avec le VIH et de la prévalence chez les adultes (15-49 ans) en 2009 dans les différentes régions du globe.
Adaptation de ONUSIDA (2010).

	Nombre de personnes vivant avec le VIH	Prévalence des 15-49 ans (%)
Afrique subsaharienne	22,5 millions [20,9-24,2 millions]	5,0 [4,7-5,2]
Moyen-Orient et Afrique du nord	460 000 [400 000-530 000]	0,2 [0,2-0,3]
Asie du sud et du sud-est	4,1 millions [3,7-4,6 millions]	0,3 [0,3-0,3]
Asie de l'est	770 000 [560 000-1,0 million]	0,1 [0,1-0,1]
Océanie	57 000 [50 000-64 000]	0,3 [0,2-0,3]
Amérique centrale et du sud	1,4 millions [1,2-1,6 millions]	0,5 [0,4-0,6]
Caraïbes	240 000 [220 000-270 000]	1,0 [0,9-1,1]
Europe orientale et Asie centrale	1,4 millions [1,3-1,6 millions]	0,8 [0,7-0,9]
Europe occidentale et centrale	820 000 [720 000-910 000]	0,2 [0,2-0,2]
Amérique du nord	1,5 millions [1,2-2,0 millions]	0,5 [0,4-0,7]
Total	**33,3 millions [31,4-35,3 millions]**	**0,8 [0,7-0,8]**

La prévalence des personnes vivant avec le VIH dans le monde entier ne cesse de croître en raison du succès des thérapies antirétrovirales hautement actives (HAART, *Highly Active Antiretroviral The-*

rapy) introduites en 1996. Elles permettent aux personnes infectées par le VIH de vivre plus longtemps et dans de meilleures conditions, de réduire les transmissions sexuelles et la transmission mère-enfant (ONUSIDA, 2009). Mais l'accès à ces thérapies est inégalement réparti entre les régions du monde. Les populations des pays à revenu faible ou intermédiaire peuvent difficilement accéder à ces traitements en raison de leur coût élevé et de l'absence d'infrastructures spécialisées nécessaires au suivi de l'infection. Dans ce contexte, plusieurs initiatives de la communauté internationale (ONUSIDA, Fondation Clinton, Fonds mondial de lutte contre le SIDA, la tuberculose et le paludisme, etc.) permettent aux gouvernements des pays à revenu faible ou intermédiaire d'assurer au plus grand nombre l'accès gratuit à ces traitements, ainsi que l'apport d'infrastructures pour le suivi des patients. Toutefois, ces infrastructures sont principalement implantées dans des zones urbaines, d'accès difficile pour les populations des zones rurales dont le suivi des patients est souvent irrégulier. Des efforts doivent encore être faits afin de décentraliser ces centres de soin (Bouchaud *et al*, 2011). Rappelons que ces thérapies ne permettent pas d'éradiquer le virus, mais seulement de le contrôler, et qu'en raison de la diversité génétique du VIH, aucun vaccin efficace n'a encore été élaboré.

Nous présentons ici les informations concernant la diversité génétique et la classification du VIH. Puis nous présentons la répartition géographique de ces différents variants, ainsi que l'origine du VIH. Nous présentons ensuite, les causes et les conséquences d'une telle diversité génétique sur les aspects biologiques et médicaux et enfin les facteurs humains ayant contribué à l'expansion mondiale de ce virus.

3.2 Virus de l'immunodéficience humaine (VIH)

3.2.1 La classification taxonomique des VIH

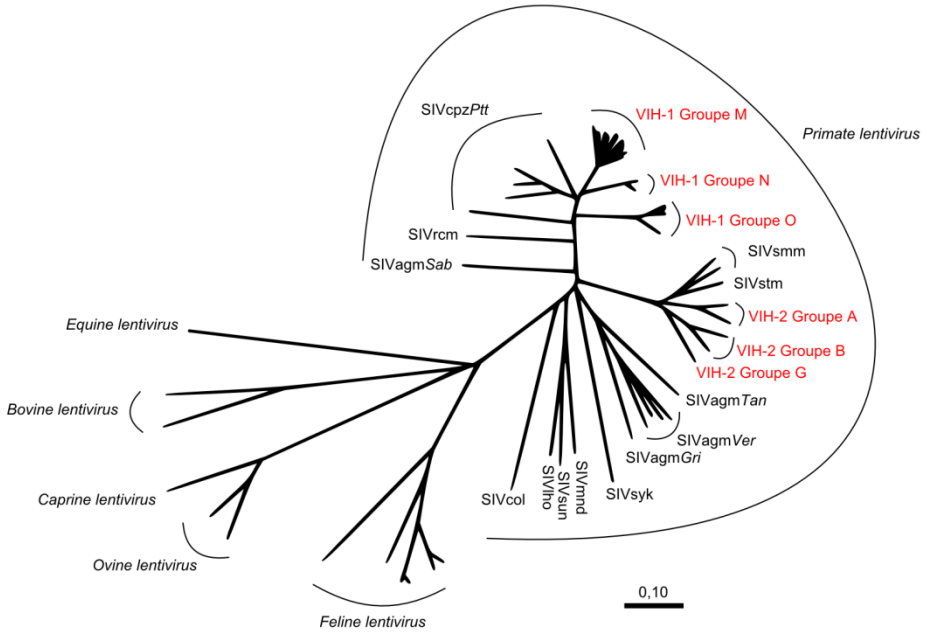
Les VIH appartiennent à la famille *Retroviridae*. Les membres de cette famille s'appellent communément des rétrovirus. Ce sont des virus à acide ribonucléique (ARN) qui ont la particularité de posséder une enzyme, la transcriptase inverse ou rétrotranscriptase (RT du terme anglo-saxon *reverse transcriptase*), qui permet de transcrire leur ARN viral en molécule d'acide désoxyribonucléique (ADN) capable de s'intégrer à l'ADN de la cellule hôte. Les rétrovirus sont subdivisés en deux sous-familles et sept genres suivant leur pathogénicité et leur morphologie : *Alpharetrovirus*, *Betaretrovirus*, *Deltaretrovirus*, *Epsilonretrovirus*, G*ammaretrovirus* et *Lentivirus* dans la sous-famille des O*rthoretrovirinae* et *Spumavirus* dans la sous-famille des *Spumaretrovirinae*. À l'exception des *Lentivirus*, les rétrovirus de la sous-famille des *Orthoretrovirinae* induisent des leucémies et des tumeurs chez leur hôte. Ces rétrovirus sont aussi appelés des oncovirus. Les virus T-lymphotropiques humains

(HTLV) et leurs homologues chez les primates non humains (PNH), les virus T-lymphotropiques simiens (STLV) appartiennent au genre *Deltaretrovirus*. Quant aux *Spumavirus*, ils sont considérés comme non pathogènes pour leur hôte. Les *foamy virus* appartiennent à ce genre. Les *Lentivirus* sont responsables de maladies à évolution lente, caractérisées par une longue période de latence aboutissant à la mort. Ils ont la particularité d'être cytopathogènes, c'est-à-dire qu'ils tuent les cellules qu'ils infectent. Les virus de l'immunodéficience humaine mais également ceux d'autres espèces (féline, bovine, simienne, etc.) appartiennent à ce genre. Actuellement, il existe deux types de VIH : le VIH-1 et le VIH-2. Les différences entre ces deux virus se font principalement au niveau génétique : plus de 50% de leur génome est différent. Au niveau morphologique, seuls les poids moléculaires des protéines et des enzymes constitutives du virus changent. La réplication virale (Marlink *et al*, 1994) et la transmission (Kanki *et al.*, 1994; De Cock *et al.*, 1993), aussi bien sexuelle que mère-enfant, sont moindres pour le VIH-2. Néanmoins, au stade final, le VIH-2 induit les mêmes symptômes que le VIH-1, malgré une phase asymptomatique plus longue (Ancelle *et al*, 1987).

3.2.2 Phylogénie et diversité génétique des VIH

Dans l'arbre phylogénétique des lentivirus (Figure 18), les VIH se placent à proximité des SIV (*simian immunodeficiency virus*), virus infectant les primates non humains. Chaque espèce de primate est infectée avec une lignée monophylétique spécifique. Par exemple, les SIV infectant naturellement les mangabeys à collier blanc (*Cercocebus torquatus*), aussi dénommé mangabeys couronnés, SIVrcm, forment une lignée distincte des SIVcol, infectant les colobes guéréza (*Colobus guereza*). De ces observations, les virus SIV sont nommés en fonction de l'espèce dans laquelle ils sont observés. Pour cela, le sigle SIV est suivi par trois lettres minuscules qui réfèrent au nom commun anglais de l'espèce hôte considérée. Par exemple, SIVsyk réfère à l'espèce *Cercopithecus albogularis* (cercopithèque à diadème) dont le nom commun anglais est : « *Sykes' monkey* ». Si nécessaire, les initiales du nom latin de la sous-espèce peuvent être ajoutées (ex. SIVcpz*Ptt* réfère aux SIV qui infectent naturellement les chimpanzés *Pan troglodytes troglodytes* et SIVcpz*Pts* les *Pan troglodytes schweinfurthii*). À l'intérieur des clades correspondant aux VIH-1 et VIH-2, des lignées monophylétiques, que l'on nomme des groupes, sont observées. Chaque groupe correspond à une transmission inter-espèce d'un SIV à l'homme. À ce jour, le VIH-1 dénombre quatre groupes (M, N, O et P) et le VIH-2 huit (A à H). Parmi ces groupes, seuls les virus du groupe M sont responsables de la pandémie.

Figure 18. Phylogénie des lentivirus.
Phylogénie obtenue d'après un alignement du gène *pol* comprenant les régions codantes de la protéase, de la transcriptase inverse, de la RNAse H et de l'intégrase. La phylogénie est calculée à l'aide des logiciels DNAdist et NEIGHBOR du package PHYLIP (Felsenstein, 1993), sous le modèle d'évolution F84.
Adaptation de Foley (2000).

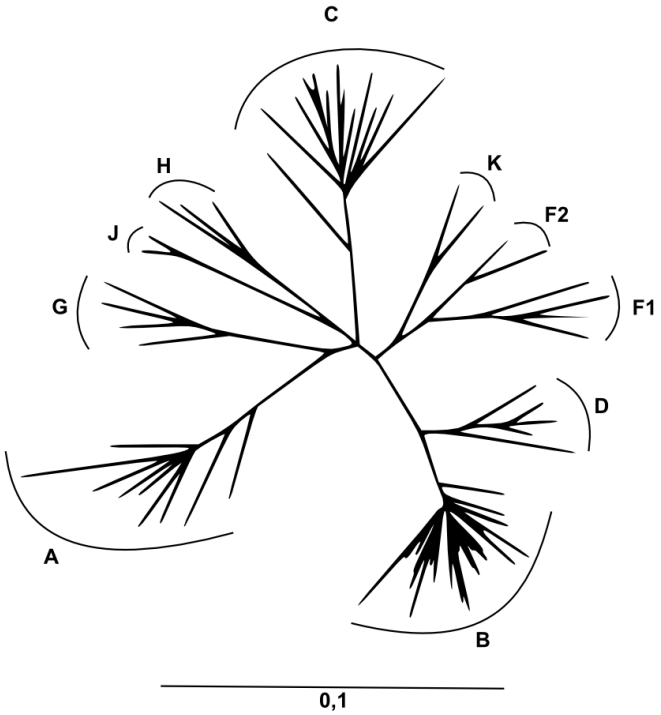


La phylogénie des virus appartenant au groupe M du VIH-1 (Figure 19) montre que certains variants génétiques forment des lignées monophylétiques que l'on appelle des sous-types. Actuellement 9 sous-types sont dénombrés (A à D, F à H, J et K). Les sous-types E et I ne sont plus considérés comme des sous-types « purs ». D'abord identifiés sur l'enveloppe, ils étaient reconnus comme « nouveaux », mais l'analyse de leur génome complet a révélé des virus recombinants, c'est-à-dire qu'ils sont formés de fragments appartenant à des sous-types différents, voire pas référencés. Les recombinants peuvent jouer un rôle important dans l'épidémie des VIH, ils sont alors appelés « formes recombinantes circulantes » (CRF, *Circulating Recombinant Forms*), dans le cas contraire le terme « formes recombinantes uniques » (URF, *Unique Recombinant Forms*) est utilisé pour les désigner. Des fragments du sous-type E peuvent être retrouvés sur les CRF01_AE et CRF27_cpx et des fragments du sous-type I sur le CRF04_cpx. La nomenclature impose l'identification d'au moins trois souches virales séquencées sur la totalité du génome, sans lien épidémiologique proche (c.-à-d. hors couple, mère-enfant, etc.), afin de proposer un nouveau sous-type ou un nouveau CRF (Robertson *et al*, 2000).

En général, deux souches virales appartenant à deux sous-types différents diffèrent d'environ 25% à 30% au niveau de l'enveloppe (*env*), d'environ 15% pour le gène *gag* et d'environ 10% pour *pol* (Gao *et al*, 1998). La variabilité génétique de souches appartenant au même sous-type est inférieure à 20% au niveau de l'enveloppe (Robertson *et al*, 2000). Toutefois, la diversité génétique à l'intérieur d'un sous-type n'est pas identique pour tous les sous-types. Par exemple, on observe une plus

grande diversité pour le sous-type A (Gao *et al*, 2001), tandis que les sous-types B ou C sont plus homogènes. Cela suggère que l'épidémie du sous-type A est plus ancienne que celle des sous-types B et C. La nomenclature propose de nommer les lignées distinctes à l'intérieur d'un sous-type, des sous-sous-types. Actuellement, le sous-type A possède quatre sous-sous-types, identifiés A1 à A4 (Vidal *et al*, 2006), et le sous-type F deux, identifiés F1 et F2 (Triques *et al*, 1999). En considérant cette définition, le sous-type D peut alors être considéré comme un sous-sous-type du sous-type B, mais pour des raisons historiques la désignation D a été conservée.

Figure 19. Phylogénie des virus du groupe M du VIH-1.
Phylogénie de maximum de vraisemblance des virus appartenant au groupe M du VIH-1 obtenue sur le gène *pol* presque complet. La phylogénie est calculée avec fastDNAml (Olsen *et al*, 1994) sous le modèle d'évolution F84.
Adaptation de Robertson *et al.* (2000).



3.3 Distribution géographique des différents variants génétiques du VIH

3.3.1 Les VIH de type 1

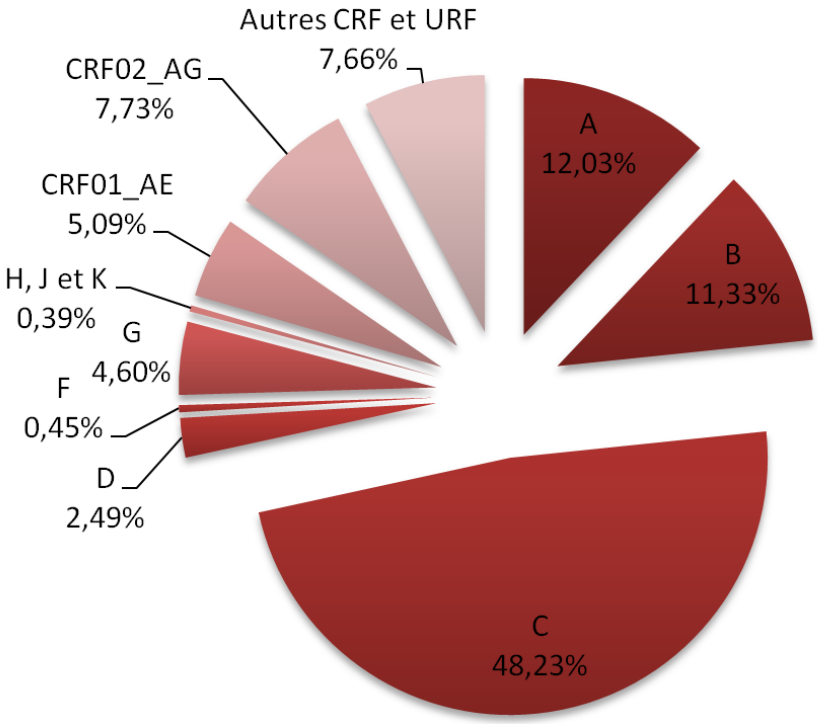
Les souches du VIH-1 sont phylogénétiquement classées en quatre groupes : M, N, O et P, résultat de quatre anthropozoonoses indépendantes. Seul le groupe M est pandémique, les autres sont surtout responsables d'infections en Afrique centrale (particulièrement au Cameroun). Les raisons pour lesquelles les virus du groupe M sont pandémiques sont inconnues. Peut-être ont-ils une propriété intrinsèque qui leur permettent de se transmettre (et donc de se répandre) plus facilement que ceux

des autres groupes ou peut-être ont-ils eu plus de chance à se retrouver dans une population présentant des conditions épidémiologiques idéales pour se diffuser.

3.3.1.1 Le groupe M

La plus grande diversité génétique intragroupe est observée avec les souches virales du groupe M. Pour cette raison et sur l'appui d'arbres phylogénétiques, la nomenclature les subdivise en neuf sous-types (A à D, F à H, J et K) représentant les différentes lignées du groupe M du VIH-1, et en 51 CRF (http://www.hiv.lanl.gov) ; les URF ne sont pas répertoriées. Le sous-type C est responsable de presque 50% des infections mondiales au VIH-1 (Figure 20). Vient ensuite le sous-type A responsable d'environ 12% des infections, puis le sous-type B avec 11,33%. Face à ces trois sous-types, la proportion d'individus infectés par les sous-types restants (D, F, G, H, J et K) semble négligeable, elle vaut moins de 8%. Quant aux CRF et aux URF, ils sont responsables d'environ 20% des infections mondiales au VIH-1, mais seuls les CRF01_AE (environ 5%) et CRF02_AG (environ 8%) sont réellement pandémiques. Les autres sont responsables d'épidémies localisées (Hemelaar *et al*, 2011).

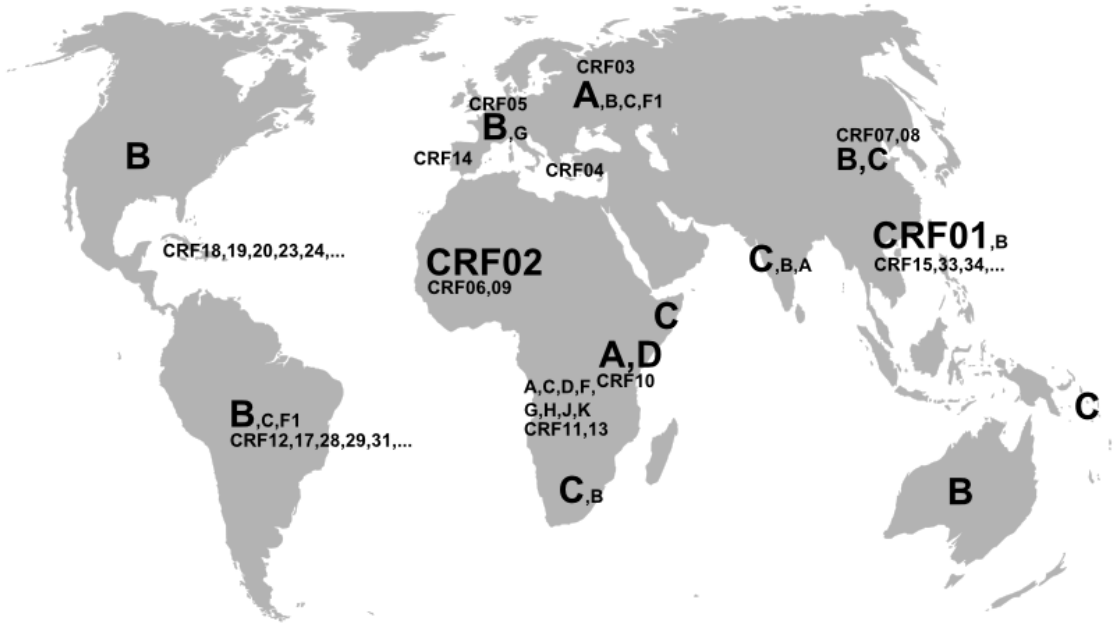
Figure 20. Distribution globale des variants génétiques du groupe M du VIH-1 sur la période 2004-2007.
Pourcentage des infections causées par les différents variants génétiques du groupe M du VIH-1 par rapport au nombre total d'individus infectés par ces variants (plus de 35 millions) (Hemelaar *et al*, 2011).



La distribution géographique des sous-types et des CRF est hétérogène et en perpétuelle évolution (Figure 21). En Amérique, en Europe de l'ouest et en Australie le sous-type B est prédominant. Toutefois, en Amérique Latine les sous-types B, C et F sont aussi observés, ainsi que de nombreux recombinants B/F ou B/C qui y circulent. En Europe de l'est ce sont les sous-types A et B, ainsi que des virus recombinants A/B qui prédominent. L'essentiel des variants qui circulent en Asie sont les

sous-types B, C et le CRF01_AE (Lau *et al*, 2007). Néanmoins, le sous-type quasi-omniprésent en Inde est le C, il est presque responsable de 98% des infections (Hemelaar *et al*, 2011). Le CRF01_AE est surtout épidémique en Asie du sud-est tandis que les sous-types B et C sont prédominants en Chine où les recombinants CRF07_BC et CRF08_BC ont émergé. Les souches B qui circulent en Asie sont génétiquement différentes de celles qui circulent en Europe ou en Amérique. De ce fait, elles sont parfois nommées B' ou Thai B afin de les distinguer du variant occidental.

Figure 21. Distribution géographique des principaux variants génétiques du groupe M du VIH-1.
Ce planisphère indique l'emplacement géographique des principaux variants génétiques du VIH-1. Les variants prédominants dans une zone géographique donnée sont représentés en caractère plus grand.



C'est en Afrique, et particulièrement au centre, qu'est observée la plus grande diversité génétique du VIH-1 en terme de sous-types et de recombinants (Peeters *et al*, 2003; Toure-Kane *et al*, 2000; Vidal *et al*, 2000). Le sous-type C est responsable de la quasi-totalité des infections en Afrique australe et dans la Corne de l'Afrique, ainsi qu'au Burundi. L'Afrique du Sud est le seul pays de la région où le sous-type B est observé, mais uniquement chez les hommes ayant des rapports sexuels avec des hommes (van Harmelen *et al*, 1997). Les pays de l'Afrique de l'est sont touchés par les sous-types A et D, tandis que les pays de l'Afrique de l'ouest par le CRF02_AG. La plus grande diversité génétique des souches du groupe M est observée dans les pays du centre de l'Afrique (Cameroun, Centrafrique, Gabon, République Démocratique du Congo et le Congo) (Marechal *et al*, 2006; Niama *et al*, 2006; Vidal *et al*, 2005, 2000; Fonjungo *et al*, 2000; Delaporte *et al*, 1996). Tous les sous-types y ont été identifiés, de nombreuses CRF, URF, ainsi que des souches virales non encore classifiées. Ces résultats suggèrent que l'épidémie en Afrique centrale est ancienne et que l'Afrique centrale serait probablement l'épicentre des virus du groupe M, en particulier la République Démocratique du Congo qui montre un degré nettement supérieur de diversité génétique (Vidal *et al*, 2000).

3.3.1.2 Le groupe O

En 1990, De Leys *et al*. (1990) observent deux cas d'infection au VIH chez un couple d'origine camerounaise mais installé en Belgique. Les comparaisons génétiques de l'isolat ANT70, isolé chez la femme et séquencé sur la région LTR (*long terminal repeat*), montre des différences significatives avec les autres isolats connus de l'époque. Quatre ans plus tard, le génome complet de cet isolat est rendu disponible (Vanden Haesevelde *et al*, 1994) et simultanément un nouvel isolat MVP-5180, similaire à ANT70 sur LTR, est isolé chez une patiente camerounaise atteinte du SIDA (Gürtler *et al*, 1994). Dès lors, l'appellation sous-type O est proposée par les auteurs pour désigner les virus génétiquement proches de ces deux variants. Ce n'est qu'avec l'identification d'un troisième isolat (VAU), chez une patiente séropositive française, et sur la base d'analyses phylogénétiques, que ces variants génétiques sont désormais classés dans un nouveau groupe, le groupe O (*outgroup* ou *outlier*) (Charneau *et al*, 1994).

Depuis, d'autres cas d'infections par le groupe O du VIH-1, principalement au Cameroun ou chez des patients Camerounais vivant en Europe sont observés. Plusieurs études en Afrique montrent que l'épicentre de cette infection se situe dans la partie ouest de l'Afrique centrale, plus particulièrement au Cameroun et dans les pays voisins comme la Guinée Équatoriale où ce variant représente 1% des infections au VIH-1 (Ayouba *et al*, 2000). Des infections VIH-1 groupe O sont aussi documentées dans plusieurs pays de Afrique de l'ouest (Figure 22), comme le Tchad, le Nigéria (Peeters *et al*, 1997), le Bénin (Heyndrickx *et al*, 1996), la Côte d'Ivoire (Nkengasong *et al*, 1998), le Togo, le Sénégal et le Niger (Peeters *et al*, 1996), mais aussi au Kenya (Songok *et al*, 1996), un pays de l'Afrique de l'est, et en Zambie, un pays de l'Afrique australe (Peeters *et al*, 1997). Des cas sporadiques sont retrouvés en Europe (France (Loussert-Ajaka *et al*, 1995), Allemagne (Hampl *et al*, 1995), Belgique (Peeters *et al*, 1995), Espagne (Quiñones-Mateu *et al*, 1998; Soriano *et al*, 1996) et Norvège (Jonassen *et al*, 1997)) et aux États-Unis (Sullivan et al., 2000; « Identification of HIV-1 group O infection--Los Angeles county, California, 1996 », 1996). Néanmoins, les investigateurs ont à chaque fois démontré l'existence d'un lien épidémiologique fort avec les pays de l'Afrique centrale, surtout avec le Cameroun et la Guinée Équatoriale. L'isolat VAU est le seul qui y fait exception. Des virus recombinants intergroupes O/M sont également identifiés au Cameroun (Yamaguchi *et al*, 2004; Peeters *et al*, 1999; Takehisa *et al*, 1999).

Du fait du peu de cas observés, la prévalence des infections au VIH-1 groupe O reste faible. La plus forte prévalence est documentée au Cameroun (2,1%), suivi du Nigéria (1,1%) et du Gabon (0,9%) (Peeters *et al*, 1997), deux pays limitrophes au Cameroun. Ces résultats suggèrent que le foyer épidémique de ce variant génétique semble être le Cameroun où la prévalence reste stable et très faible (1,1%) (Vessière *et al*, 2010; Ayouba *et al*, 2001).

Figure 22. Répartition géographique des différents cas d'infection au VIH-1 groupe O.
Cette figure montre la dispersion géographique des différents cas d'infection reportés du VIH-1 groupe O. L'intensité du gris est proportionnelle au nombre d'infections au groupe O de chaque pays (quelques cas à plusieurs dizaines). Aucun cas n'est documenté pour les pays coloriés en blanc. L'Afrique centrale, plus particulièrement le Cameroun et ses proches voisins, concentre l'essentiel de l'épidémie du VIH-1 groupe O. Des cas sont tout de même observés à l'ouest, à l'est et au sud de l'Afrique. En dehors de l'Afrique, seuls les États-Unis et quelques pays d'Europe ont observé des infections au groupe O, mais chez des patients ayant des liens épidémiologiques forts avec l'Afrique centrale.
Adaptation de Quiñones-Mateu *et al.* (2000).

3.3.1.3 Le groupe N

Le premier cas d'infection au VIH-1 groupe N est identifié chez une patiente d'une trentaine d'années qui s'est présentée à l'hôpital de Yaoundé (Cameroun) en mai 1995 avec des signes cliniques de SIDA (Simon *et al*, 1998). La souche virale isolée à cette occasion est désignée sous le terme YBF30. Les études phylogénétiques montrent qu'elle n'est semblable à aucune autre souche de référence du VIH-1, mais qu'elle est proche des SIV de chimpanzés. Un autre cas de transmission inter-espèce est à l'origine de cette souche virale et un nouveau groupe est proposé afin de la classer : N pour « *new* » ou « non-M, non-O ». Afin de confirmer la circulation de ce variant génétique au sein de la population, Simon *et al*. (1998) ont caractérisé 700 échantillons VIH-1 entre 1988 et 1997 parmi lesquels seulement un autre échantillon VIH-1 groupe N est identifié. Par la suite, plusieurs autres infections anecdotiques furent décrites et uniquement au Cameroun (Vallari *et al*, 2010; Yamaguchi *et al*, 2006b, 2006a; Bodelle *et al*, 2004; Roques *et al*, 2004; Vergne *et al*, 2003; Ayouba *et al*, 2000). Au total, moins d'une vingtaine de cas sont reportés dans la littérature et la prévalence du groupe N au Cameroun est donc de 0,1% (Vallari *et al*, 2010).

Bien que l'observation de souches appartenant au groupe N est rare, les modes de transmissions et la pathogénicité de ce variant sont comparables à ceux du groupe M. Par exemple, l'identification d'un enfant âgé de 7 ans porteur d'un virus du groupe N suggère qu'une transmission verticale (mère-enfant) peut se produire (Ayouba *et al*, 2000). La transmission horizontale (par contacts sexuels) est attestée lors d'une étude sur un couple (Yamaguchi *et al*, 2006a) et la mort par le SIDA de la patiente porteuse de la souche YBF30 confirme la pathogénicité de ce variant. Récemment, un cas identifié en France chez un patient revenant d'un voyage au Togo (pays de contamination) confirme la circulation de ce virus en-dehors du Cameroun (Delaugerre *et al*, 2011).

3.3.1.4 Le groupe P

A ce jour, seulement deux individus porteurs du VIH-1 groupe P sont identifiés. Le premier est une femme originaire du Cameroun mais résidant en France depuis 2004 (Plantier *et al*, 2009). Avant son arrivée, elle a vécu dans plusieurs villes semi-urbaines situées aux alentours de Yaoundé, la capitale du Cameroun et lieu probable de contamination. En ayant connaissance de cette information, Vallari *et al*. (2011) ont effectué une enquête épidémiologique afin d'estimer la prévalence du groupe P au Cameroun. Sur 1 736 échantillons VIH-1 positifs examinés (collectés entre 2006 et 2009), seule une autre souche virale appartenant au groupe P est identifiée. Ainsi, la prévalence du groupe P au Cameroun est de 0,06%. L'identification de cette dernière souche virale valide la circulation de ce variant au sein de la population humaine.

3.3.2 Les VIH de type 2

Contrairement à son homologue le VIH-1, le VIH-2 n'a pas connu une expansion épidémique mondiale. Il est aujourd'hui caractérisé par huit groupes (A à H) (de Silva *et al*, 2008; Damond *et al*, 2004), mais seuls les groupes A et B sont épidémiques en Afrique de l'ouest. De rares cas sont aussi retrouvés dans les autres pays africains et dans le reste du monde. Le groupe A circule principalement en Guinée-Bissau et au Sénégal, tandis que le groupe B est particulièrement présent au Mali et en Côte d'Ivoire. Néanmoins, les prévalences VIH-2 sont peu élevées et vont en décroissance. Par exemple, en Guinée-Bissau, la prévalence du VIH-2 est de 10,1% en 1989 alors qu'en 2008 elle est de 4,4% (da Silva *et al*, 2008; Poulsen *et al*, 1993). Parallèlement une augmentation de la prévalence du VIH-1 est observée. Un phénomène similaire est remarqué dans les autres pays où le VIH-2 est épidémique (Hamel *et al*, 2007). Ces informations suggèrent que l'épidémie du VIH-2 tend à diminuer, voire à disparaître.

Les autres groupes du VIH-2 ne sont pas du tout épidémiques. Chacun des groupes restant est seulement observé chez un seul patient. Les groupes C et D sont identifiés en 1989 chez deux patients vivant au Libéria (Gao *et al*, 1994) et les groupes E et F en 1992 et 1991 respectivement chez

deux patients originaires de Sierra Leone (Chen et al., 1997; Gao et al., 1994). Le groupe G est identifié en 1992 chez un patient vivant en Côte d'Ivoire (Brennan *et al*, 1997) et le groupe H en 1996 chez un patient originaire de l'Afrique de l'ouest, mais vivant en France (Damond *et al*, 2004). Récemment, un autre cas d'infection par un virus VIH-2 appartenant au groupe F est identifié aux États-Unis chez un patient aussi originaire de Sierra Leone (Smith *et al*, 2008).

À l'instar du VIH-1, des phénomènes de recombinaisons intergroupes (A/B) sont aussi observés pour le VIH-2 (Yamaguchi *et al*, 2008; Gao *et al*, 1994). Mais ce n'est qu'avec l'identification récente de trois isolats de patients vivant au Japon (dont deux originaires du Nigéria) et leur correspondance génomique avec un isolat de 1990 collecté en Côte d'Ivoire (Gao *et al*, 1994) que la première CRF du VIH-2 est proposée par la nomenclature : CRF01_AB (Ibe *et al*, 2010). Même si des cas d'infections double aux VIH-1 et VIH-2 sont reportés (Gottlieb *et al*, 2003), aucune souche recombinante entre ces deux types n'est observée.

3.4 L'origine africaine des VIH

C'est avec l'observation de symptômes similaires au SIDA chez l'homme dans une colonie de 64 macaques rhésus (*Macaca mulatta*) du centre de primatologie de Californie (*California Primate Research Center*) (Henrickson *et al*, 1983), et l'isolation de trois SIV sur ces animaux malades au centre de primatologie *New England Regional Primate Research Center* (Daniel *et al*, 1985) que l'hypothèse d'une origine simienne au VIH fut suspectée dès 1985. L'isolation de souches SIV phylogénétiquement proches du VIH-1 sur deux chimpanzés (*Pan troglodytes troglodytes*) du Gabon vers la fin des années quatre-vingt a considérablement favorisé l'hypothèse de chimpanzés comme réservoir direct du VIH-1 (Huet *et al*, 1990; Peeters *et al*, 1989). Mais des doutes subsistaient encore à cause du peu d'animaux infectés retrouvés. L'hypothèse d'une troisième espèce de primate non humain comme réservoir était toujours ouverte. Mais où, quand et comment cette transmission inter-espèce a eu lieu étaient encore des questions ouvertes.

Les chimpanzés sont classés en quatre sous-espèces : *Pan troglodytes verus* vivant en Afrique de l'ouest, *Pan troglodytes ellioti* à l'ouest du Cameroun, *Pan troglodytes troglodytes* au sud du Cameroun, en Guinée Équatoriale, au Gabon et au nord du Congo, et *Pan troglodytes schweinfurthii* au nord de la République Démocratique du Congo, ainsi qu'à l'est de la République Centrafricaine, en Ouganda et Tanzanie (Figure 23). Toutefois, seules les sous-espèces *P. t. troglodytes* et *P. t. schweinfurthii* sont naturellement infectées par des SIV, respectivement dénommés SIVcpz*Ptt* et SIVcpz*Pts* (Leendertz *et al*, 2011; Van Heuverswyn *et al*, 2007; Switzer *et al*, 2005). Les premières études phylogénétiques, incluant des souches virales VIH-1, SIVcpz*Ptt* et SIVcpz*Pts* ne permettaient pas

l'identification certaine du réservoir du VIH-1 car le nombre de souches SIV était limité et provenaient d'animaux captifs (Gao *et al*, 1999). Néanmoins, elles montraient que les virus SIVcpz*Ptt* sont phylogénétiquement plus proches du VIH-1 que SIVcpz*Pts*. Le développement en 2002 d'une méthode non invasive (utilisant des échantillons fécaux) pour les espèces protégées a permis la caractérisation de nouvelles souches de SIV provenant de chimpanzés sauvages (Santiago *et al.*, 2003, 2002). À partir de ces nouvelles données, des études phylogénétiques confirment que SIVcpz*Pts* n'est pas l'ancêtre du VIH-1 (Worobey *et al*, 2004) et en 2006, les réservoirs exacts du VIH-1 groupe M, virus pandémique, et du VIH-1 groupe N, non pandémique, sont identifiés (Keele *et al*, 2006). Les ancêtres des virus du groupe M prennent source dans une communauté de chimpanzés sauvages vivant à l'extrême sud-est du Cameroun et ceux du groupe N dans une autre communauté de chimpanzés sauvages située aux environs de la forêt du Dja au centre-sud du Cameroun (Figure 24). Les ancêtres des groupes O et P ne sont pas encore identifiés. Néanmoins, ces virus sont proches des SIV infectant les gorilles (*Gorilla gorilla gorilla*) vivant aussi au Cameroun. En ce qui concerne les virus du groupe M, il a été montré que l'épicentre de l'épidémie se situe en République Démocratique du Congo à des centaines de kilomètres du lieu de contamination initiale (Vidal *et al*, 2000). Les raisons exactes entre ces différences de localisation ne sont pas connues, mais plusieurs hypothèses sont formulées (démographiques, sociologiques, économiques, etc.).

Figure 23. Aires de répartition des différentes sous-espèces de chimpanzés en Afrique.
Les chimpanzés (*Pan troglodytes*) sont divisés en quatre sous-espèces. Les chimpanzés *Pan troglodytes versus* (en jaune) vivent à l'ouest de l'Afrique, les *Pan troglodytes ellioti* (en vert) à l'ouest du Cameroun, les *Pan troglodytes troglodytes* (en mauve) au sud du Cameroun, en Guinée Équatoriale, au Gabon et au nord du Congo et les *Pan troglodytes schweinfurthii* (en bleu) au nord de la République Démocratique du Congo, au sud-est de la Répulique Centrafricaine, ainsi qu'à l'ouest de l'Ouganda et de la Tanzanie.
Adaptation d'une image de Wikipédia.

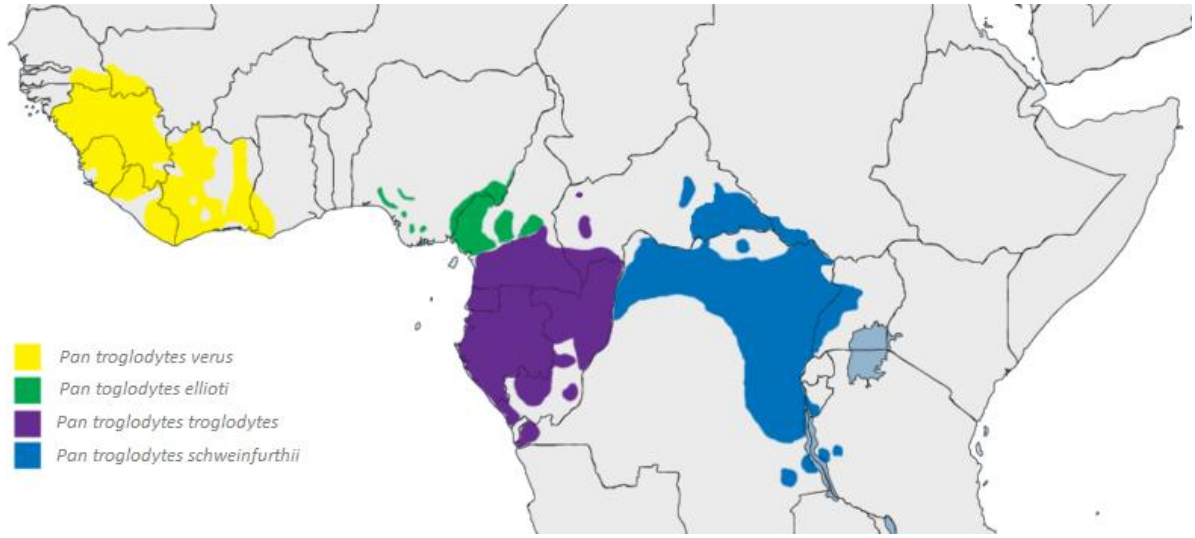
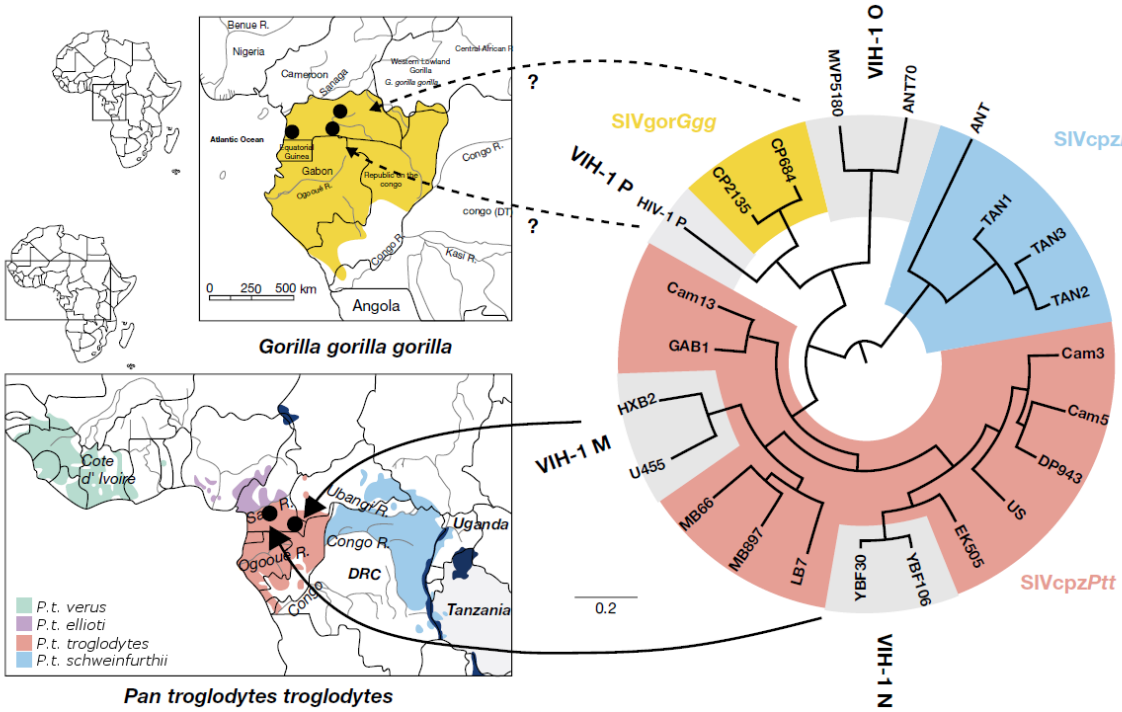


Figure 24. Liens de parenté entre les virus VIH-1 et SIV.
Phylogénie (inférée à partir d'une méthode de maximum de vraisemblance sur la région *env)* montrant que les ancêtres directs des groupes M et N du VIH-1 sont des chimpanzés (*Pan troglodytes*) de la sous-espèce *troglodytes* vivant respectivement au sud-est et au centre du Cameroun. Les ancêtres des groupes O et P ne sont pas encore identifiés, mais les virus SIV infectant les gorilles (*Gorilla gorilla gorilla*) du Cameroun sont de proches parents et peut être leur ancêtre.
Adaptation de Etienne & Peeters (2010).



Dans la République Démocratique du Congo, particulièrement à Kinshasa (la capitale), une diversité génétique impressionnante des souches virales du groupe M du VIH-1 est actuellement observée (Vidal *et al*, 2005, 2000), suggérant que l'épidémie dans ce pays est ancienne et qu'il est l'épicentre de la pandémie actuelle. Deux souches virales, une provenant d'un sérum de 1959 (Zhu *et al*, 1998) et l'autre d'une biopsie de 1960 (Worobey *et al*, 2008), confirment l'ancienneté de l'épidémie dans ce pays car elles présentent une grande diversité génétique. Elles appartiennent respectivement aux sous-types D et A. Des études de datation moléculaire sont utilisées afin d'estimer la date de l'ancêtre commun aux souches des VIH-1. La date de l'ancêtre commun aux souches appartenant au groupe M est estimée au début du xx^{e} siècle (1908 [1884-1924]), celle des souches appartenant au groupe N au début des années soixante (1963 [1948-1977]) et celle des souches appartenant au groupe O au début des années vingt (1920 [1890-1940]) (Wertheim & Worobey, 2009). Au vu du nombre insuffisant de souches appartenant au groupe P, aucune publication ne relate les estimations de la date de son ancêtre commun, car elle manquerait de crédibilité.

En Afrique, les chimpanzés (aussi bien que d'autres espèces de primates non humains) sont chassés et domestiqués. Ils sont une source de revenu et d'alimentation pour les populations locales. L'exposition directe avec le sang ou d'autres sécrétions corporelles lors de la chasse, lors de la préparation de la viande de brousse ou même lors de blessures infligées par des singes domestiqués (p. ex.

morsures) sont les explications les plus plausibles pour expliquer le franchissement de la barrière d'espèce (Peeters *et al*, 2002) (Figure 25). Mais le mode exact de contamination reste indéterminé.

Figure 25. Illustrations de situation à risque.
Illustrations de la proximité qu'ont les habitants de l'Afrique centrale avec des singes possiblement contaminés. Les photos du bas montrent un habitant préparant, à mains nues, de la viande de brousse. Une blessure lors du maniement de son ustensile l'exposerait directement au sang de l'animal.
Photos de Steve AHUKA MUNDEKE et Bernadette ABELA.

Contrairement au VIH-1, le VIH-2 a pour origine le SIVsm infectant le mangabey enfumé (*Cercocebus atys*) de l'Afrique de l'ouest qui est aussi chassé et consommé par les habitants locaux. Le réservoir des groupes épidémiques (A et B), ainsi que ceux des groupes C, G et H, infectant peu d'individus, est identifié. Il s'agit de mangabeys enfumés sauvages vivant dans la forêt de Taï en Côte d'Ivoire, limitrophe au Libéria (Santiago *et al*, 2005). Les souches des groupes E et F sont phylogénétiquement proches de souches SIVsmm de Sierra Leone et celle du groupe D de souches SIVsmm du Libéria (Chen *et al*, 1997, 1996). Les études de datation estiment la date de l'ancêtre commun des groupes A et B à 1932 [1906-1955] et 1935 [1907-1961] respectivement (Wertheim & Worobey, 2009).

3.5 Causes de la diversité génétique

Le VIH présente une grande diversité génétique. Plusieurs facteurs biologiques sont responsables de cette diversité.

Une caractéristique des virus appartenant à la famille des rétrovirus est de posséder la transcriptase inverse (ou rétrotranscriptase), une enzyme qui permet de synthétiser le matériel génétique viral (initialement en ARN) en ADN afin de l'intégrer dans le génome de la cellule hôte pour devenir un provirus. C'est une étape clef du cycle réplicatif viral. Cependant, lors de la transcription de l'ARN viral en ADN, la transcriptase inverse commet un nombre important d'erreurs. Ce nombre est estimé *in vivo* à environ une substitution sur 10 000 bases par cycle réplicatif viral (Preston & Dougherty, 1996). De plus, la transcriptase inverse ne possède pas d'activité exonucléasique permettant la correction des erreurs d'appariement (Roberts *et al*, 1988). Associer ce facteur avec une forte réplication virale, environ 10^{10} à 10^{12} nouveaux virions chaque jour (Perelson *et al*, 1996), implique qu'un individu est infecté par une pléthore de virus génétiquement différents et dont la population est appelée une « quasi-espèce ».

Outre le fait de commettre des erreurs, la transcriptase inverse est aussi connue pour sauter (« *switch* ») d'un brin d'ARN à un autre pendant la transcription (An & Telesnitsky, 2002). Lorsque les brins d'ARN sont identiques, ce phénomène peut créer des insertions ou des délétions (Zhang *et al*, 2000). Mais lorsque la cellule hôte est infectée par plusieurs variants génétiques différents (uniquement possible lors d'une surinfection ou co-infection), des virus recombinants peuvent émerger et donc participer à la diversification génétique du VIH. Le taux de recombinaison du VIH est estimé à trois évènements de recombinaison par génome et par cycle de réplication virale (Zhuang *et al*, 2002). Ce phénomène pose problème dans le développement de vaccins à virus atténués, en effet ces virus peuvent recombiner avec les virus naturels et devenir infectieux. Les recombinaisons les plus souvent reportées sont ceux entre des virus de sous-type différent, mais des recombinaisons intra-sous-types (Rousseau *et al*, 2007) ou intergroupes (Peeters *et al.*, 1999; Takehisa *et al.*, 1999) sont aussi documentées. Malgré l'observation de nombreuses infections doubles VIH-1/VIH-2 (Gottlieb *et al*, 2003), aucun recombinant VIH-1/VIH-2 n'est référencé (Curlin *et al*, 2004) et peu d'études s'y attachent.

La sélection naturelle est aussi un acteur important de la diversité génétique des VIH (Kils-Hütten *et al*, 2001). Elle se joue sur deux fronts : les pressions de sélection du système immunitaire (commun à tout individu) et les pressions de sélection dues aux traitements antirétroviraux (uniquement pour les patients sous traitement). Ces pressions peuvent être de deux sortes soit positives, soit né-

gatives. Lorsque la pression est négative, les mutations se produisent essentiellement sur les parties non codantes du génome. Dans le cas contraire, elles ne doivent pas changer la structure et la fonctionnalité des protéines (en particulier celles nécessaires au cycle réplicatif virale) sous peine de rendre inaptes les nouveaux virions. Ces mutations sont dites synonymes ou silencieuses. Les changements d'acides aminés sur les protéines sont la preuve d'une pression positive. Dans ce cas, les mutations sont dites non synonymes. En ce qui concerne le VIH, les mutations positives sur les protéines de l'enveloppe permettent au virus d'échapper au système immunitaire : des changements dans les déterminants antigéniques permettent à ces nouveaux variants de ne plus être reconnus par les anticorps et les cellules immunitaires. De même, un variant résistant à un traitement spécifique aura un avantage par rapport aux autres variants et sera sélectionné. De ce fait, la région gp120 de l'enveloppe possède une grande variabilité génétique (région en contact avec l'environnement extérieur). Les mutations de résistance aux traitements antirétroviraux peuvent apparaître dans les régions qui sont ciblées par les antirétroviraux, essentiellement *pol* (transcriptase inverse et protéase), mais aussi les gènes *gag* (site de clivage) et *env* (sur la gp41). Par ailleurs, des variants naturellement résistants aux traitements antirétroviraux sont aussi observés (Shafer *et al*, 1999).

3.6 Conséquences de cette diversité génétique

Les conséquences de cette diversité génétique sont nombreuses. Elles concernent la virulence, la transmission, les tests de dépistages, le suivi de l'infection par la quantification de la charge virale, les traitements antirétroviraux et l'élaboration d'un vaccin.

Plusieurs études indépendantes ont en effet montré que la progression vers le stade SIDA peut dépendre du variant génétique (Kaleebu *et al*, 2002; Kanki *et al*, 1999; Neilson *et al*, 1999). Par exemple, une étude sur 1 045 adultes vivant en Ouganda montre que les individus infectés par le sous-type D ont une progression vers la maladie plus rapide que ceux infectés par le sous-type A. Mais il ne semble pas y avoir de différence dans le taux de transmission de la mère à l'enfant entre ces deux variants (Eshleman *et al*, 2001). Une autre étude sur une cohorte de 320 femmes vivant à Nairobi (Kenya) montre que les femmes infectées par le sous-type C se retrouve plus tôt dans un état avancé d'immunodépression par rapport aux femmes infectées par les sous-types A ou D. Toutefois aucune règle générale ne peut être établie puisque selon les études les résultats peuvent différer (Alaeus *et al*, 1999) et les raisons biologiques ne sont pas encore connues (Wright *et al*, 2011).

Les tests de dépistage sont primordiaux dans le suivi de l'épidémie et de l'infection au VIH, ils doivent donc être d'une sensibilité extrême. Même si actuellement ces tests détectent tous les variants génétiques connus (type, sous-type ou groupe), ce n'était pas le cas au début de la pandémie. En effet, les premiers tests commercialisés étaient développés sur la base du sous-type B circulant dans

les pays à revenu élevé et ils ne permettaient pas de détecter les variants génétiques appartenant au groupe O (Loussert-Ajaka *et al*, 1994). Cela montre qu'une surveillance des variants génétiques est nécessaire, surtout dans les pays d'Afrique centrale ou la diversité génétique est très importante et où de nouveaux variants peuvent facilement émerger suite à d'éventuelles autres transmissions zoonotiques. L'échec de la détection de l'émergence d'un nouveau variant aurait des conséquences dramatiques.

La quantification de la charge virale est un outil indispensable au suivi de l'infection, mais aussi pour suivre l'impact des traitements antirétroviraux. Hélas, la diversité génétique influe aussi sur ce procédé qui est basé sur une détection moléculaire d'un fragment génomique des virus. À l'instar des tests de dépistage, les premiers tests de quantification de la charge virale étaient inadaptés à détecter des sous-types non-B car leur développement était basé sur le sous-type B (Gueudin *et al*, 2003). Depuis, la performance de ces tests a été accrue. Il en existe capables d'identifier de nombreux variants génétiques, mais aucun ne peut les identifier tous (Peeters *et al.*, 2010; Rouet & Rouzioux, 2007).

Les traitements antirétroviraux ont pour but de réduire le taux de réplication virale et de le maintenir à un niveau très bas. Cela permet au système immunitaire de se restaurer, réduisant ainsi la mortalité, la morbidité et l'impact du virus sur le patient pour améliorer sa qualité de vie (Palella *et al*, 1998). Néanmoins, la diversité génétique a une influence sur l'efficacité de ces traitements thérapeutiques. Par exemple, il existe des souches (VIH-2 et VIH-1 groupe O) naturellement résistantes aux inhibiteurs non nucléosidiques de la transcriptase inverse (INNTI) (Lal *et al*, 2005; Quiñones-Mateu *et al*, 1998) et les souches virales appartenant au sous-type G semblent moins sensibles à certains inhibiteurs de protéase (IP) (Descamps *et al*, 1998). De plus, sous l'action de ces thérapies, des mutations de résistance peuvent apparaître, permettant au virus d'échapper à la pression médicamenteuse. La rapidité avec laquelle ces souches résistantes émergent peut varier pour certains sous-types. Par exemple, les virus du sous-type C développent plus rapidement des souches résistantes aux NNRTI que celles du sous-type B (Loemba *et al*, 2002).

L'extraordinaire diversité génétique du VIH, résultat de l'adaptation du virus à son environnement, est aussi un obstacle majeur à l'élaboration d'un vaccin (Gaschen *et al*, 2002). En effet, l'hypervariabilité antigénique produite par le virus ne permet pas d'élaborer un vaccin préventif contre tous les variants génétiques simultanément (Berman *et al*, 1999). Les premiers essais vaccinaux n'ont pas donné de résultats encourageants (Bolognesi & Matthews, 1998), mais les tests continuent toujours (Rerks-Ngarm *et al*, 2009).

3.7 Facteurs sociologiques de la diffusion mondiale du VIH

Les raisons pour lesquelles le VIH est devenu pandémique ne sont pas uniquement biologiques, elles sont aussi sociologiques. Une grande partie de la dissémination mondiale du VIH dépend du comportement humain, aidé par la latence de la maladie, sa capacité à échapper au système immunitaire et ses modes de transmissions.

Le VIH peut se transmettre d'un individu à un autre suivant deux voies horizontales et une voie verticale. Les deux transmissions horizontales sont les voies sexuelle et sanguine. La transmission par voie sexuelle s'effectue essentiellement par les muqueuses vaginales, vulvaires, péniennes ou rectales. Ces membranes sont des portes d'entrée efficaces pour le virus, en particulier pour le partenaire receveur. Toutefois, la probabilité de contamination n'est pas de 100%, elle est d'environ 1 cas sur 1 000 épisodes de rapports sexuels (Galvin & Cohen, 2004). Cela dépend notamment du type de rapport sexuel (vaginal/anal), de la durée et de la fréquence des expositions, de la charge virale présente dans les sécrétions, du stade de l'infection, de la présence de microlésions sur les muqueuses ou de la présence d'autres infections sexuellement transmissibles (Galvin & Cohen, 2004). Il semblerait que la circoncision diminuerait considérablement les probabilités de contamination des hommes lors de rapports hétérosexuels (Auvert *et al*, 2005) ; c'est pour cela que des campagnes de circoncision sont effectuées en Afrique comme acte de prévention (Bailey *et al*, 2007; Gray *et al*, 2007). La transmission par voie sanguine est maintenant éradiquée dans le milieu hospitalier (lors de transfusions), excepté dans les pays du Sud où le dépistage des dons de sang n'est pas toujours et partout disponible. Le dernier mode de transmission du VIH est celui de la mère à l'enfant. La transmission peut se produire pendant la grossesse, l'accouchement mais aussi durant l'allaitement. Le taux de transmissions de la mère à l'enfant est en nette diminution et, suite à l'avancée des traitements préventifs à ce mode de transmission, l'ONUSIDA envisage l'élimination quasi-totale des transmissions mères-enfants à l'horizon 2015 (ONUSIDA, 2010).

Selon les modes de transmission du VIH, il existe certaines personnes qui sont « statistiquement plus exposées à l'infection que le reste de la population » (Grmek, 1990). Pour les désigner, la littérature utilise le terme de « groupes à risque » ou « populations clés » (ONUSIDA, 2009). Ces groupes à risque ont joué un rôle dans la diffusion de l'épidémie de SIDA en Afrique et en-dehors de l'Afrique par le biais d'évènements fondateurs qui correspondent à la diffusion rapide d'un variant au sein d'un groupe, lequel peut après se répandre dans la population générale. Par exemple, en Russie le sous-type A s'est répandu parmi les consommateurs de drogues injectables (Bobkov *et al*, 1997). Le terme de consommateurs de drogues injectables (IDU pour *intraveinous drug users*) désigne des individus qui se droguent à l'aide de seringues. Ces seringues peuvent être contaminées par le virus et

l'échange de seringues contaminées diffuse le virus au sein de cette population. Certains toxicomanes s'adonnent à la prostitution afin d'avoir des revenus supplémentaires qui leur permettent de se procurer de la drogue. Ils peuvent ainsi servir de pont envers la population générale. Les hommes ayant des rapports sexuels avec des hommes (désignés dans la littérature internationale sous le terme MSM pour « *men having sex with men* ») est un autre groupe à risque. Certains individus de cette population ont une activité sexuelle impliquant de nombreux partenaires (Grmek, 1990) favorisant la dispersion du virus lors de rapports non protégés. Ce dernier groupe a souvent été stigmatisé au début de l'épidémie du SIDA, parce que la maladie a été pour la première fois identifiée sur des individus appartenant à cette communauté. Le dernier groupe à risque couramment retrouvé dans la littérature, sont les professionnel(le)s du sexe (CSW pour *commercial sex workers* ou parfois FSW pour *female sex workers*). En Afrique, les premières études sur le VIH ont été faites sur des cohortes de professionnel(le)s du sexe où la prévalence observée était très élevée (Van de Perre *et al*, 1985). Dans de nombreux pays, l'utilisation de préservatifs pendant des rapports sexuels tarifés est peu fréquente ce qui participe ainsi à la diffusion du VIH. Depuis le début de la pandémie, des campagnes de prévention ont été mises en place et ont considérablement réduit les risques de transmission dans ces différents groupes.

L'état actuel de l'épidémie du VIH et la dispersion de ses variants génétiques ne peuvent s'expliquer uniquement sur la base de ces groupes à risque. D'autres acteurs ont joué un rôle plus ou moins important dans la diffusion de cette épidémie (Perrin *et al*, 2003). Les mouvements de populations, qu'ils soient liés au tourisme (notamment sexuel), à l'immigration ou à l'éloignement familial pour raisons économiques, mais aussi aux conséquences des conflits ou des guerres, participent à la diffusion de certains variants génétiques (Belda *et al*, 1998; Lasky *et al*, 1997; Kane *et al*, 1993; Bwayo *et al*, 1994). Par exemple, le déploiement de troupes militaires (éloignement du domicile, situations de stress dû aux conflits, etc.) peut favoriser l'exposition des soldats aux maladies sexuellement transmissibles, incluant le VIH, puis leur diffusion après rapatriement (Azuonwu *et al*, 2012; Djoko *et al*, 2011). Le développement économique massif de certains secteurs reculés (exploitations forestières et minières) induit de manière compréhensible la formation de réseaux sociaux et économiques (incluant la prostitution) autour de ces secteurs, pouvant ainsi participer à la diffusion de l'épidémie du VIH dans les zones rurales (Laurent *et al*, 2004).

Chapitre 4

Ultrametric Least Squares : une méthode de distances rapide et précise pour estimer le taux de substitution à partir d'un ensemble de séquences hétérochrones

Nous présentons ici une méthode de distances, Ultrametric Least Squares *(ULS), qui estime le taux de substitution d'un ensemble de séquences hétérochrones, en faisant l'hypothèse d'une horloge moléculaire stricte. Cette méthode corrige la distance évolutive entre séquences, par l'adjonction d'un facteur correctif aux séquences non contemporaines. Ce facteur est proportionnel au taux de substitution à estimer, ainsi qu'à l'ancienneté de la séquence en question. Le taux de substitution est alors estimé par la minimisation d'un critère quadratique, qui mesure l'ultramétricité de la distance corrigée. Nous montrons que ce critère est parabolique par morceaux, et proposons un algorithme efficace, en* $O(n^3 \log n)$ *où* n *est le nombre de séquences, pour minimiser ce critère. Nous montrons aussi qu'il est possible de borner cette complexité et sans perte de précision par un tirage aléatoire de triplets. Notre méthode peut être étendue à l'estimation de plusieurs taux de substitution variant au cours du temps, par exemple pour prendre en compte la prise d'un traitement et sa date de début, ou par lignage (horloges moléculaires locales). ULS est confrontée sur données simulées à d'autres méthodes de distances, comme sUPGMA ou TREBLE, aux régressions linéaires* Root-to-Tip *et* Pairwise-Distance*, ainsi qu'à l'approche probabiliste développée dans le logiciel BEAST, qui est à l'heure actuelle considérée comme l'une des plus précises mais est handicapée par un temps de calcul très important. Les expériences montrent qu'ULS est plus précise ou aussi précise que les autres méthodes de distances et que BEAST, tout en étant extrêmement rapide. Nous présentons ensuite une application d'ULS sur deux jeux de données du VIH.*

Sommaire

4.1 Introduction

Les méthodes de distances définissent avec les méthodes probabilistes et les méthodes de parcimonie les trois approches principales permettant l'inférence de phylogénies moléculaires (cf. Chapitre 1). Un des principes souvent utilisé avec les méthodes de distances est celui des moindres carrés (en anglais *Least Squares*) qui compare les distances évolutives estimées entre paires de séquences, contenant des erreurs dues à l'échantillonnage et inhérentes au modèle d'évolution, aux distances patristiques (ou distances de chemin) calculées dans l'arbre estimé (Felsenstein, 1997; Bulmer, 1991; Fitch & Margoliash, 1967). Ce principe est non seulement rapide en temps de calcul, mais augmente en précision au fur et à mesure que les erreurs d'estimation dans les distances tendent à disparaître. En pratique, il est impossible d'estimer les vraies distances évolutives puisque les modèles d'évolution font des hypothèses simplificatrices, comme, par exemple, l'indépendance des sites. Pour contrer cet effet, plus marqué sur les grandes distances que sur les petites, les méthodes de moindres carrés utilisent généralement une valeur de pondération devant chaque terme de la somme, qui est inverse à la variance de l'estimateur et donc plus faible pour les grandes distances. Ainsi, les méthodes de moindres carrés pondérées *Weighted Least Squares*, WLS, généralisent la

méthode des moindres carrés ordinaires *Ordinary Least Squares*, OLS, qui n'utilise pas de pondérations, ou, ce qui revient au même, des pondérations toutes identiques.

Cette approche de pondération est peu exploitée (hormis TREBLE) par les méthodes de distances qui permettent d'estimer le taux de substitution, alors qu'elle est presque universelle pour les méthodes d'inférence phylogénétique. À notre connaissance, la méthode sUPGMA est la seule qui utilise le principe des moindres carrés, mais sans valeurs de pondération (OLS) (Rodrigo *et al*, 2007; Drummond & Rodrigo, 2000).

Nous présentons dans ce chapitre une méthode de distances, *Ultrametric Least Squares* (ULS), qui estime le taux de substitution d'un ensemble de séquences hétérochrones sous l'hypothèse d'une horloge moléculaire stricte, c'est-à-dire sous les hypothèses du modèle SRDT. Cette méthode utilise des triplets de séquences et des pondérations, comme TREBLE, mais propose un algorithme radicalement différent où on optimise un critère global dont nous montrons qu'il est parabolique par morceaux. Cette méthode est ensuite étendue aux modèles MRDT et DR (mais avec des horloges moléculaires locales). Les performances de cette méthode sont simultanément comparées avec celles des autres méthodes de distances et celles de la méthode probabiliste de référence, BEAST.

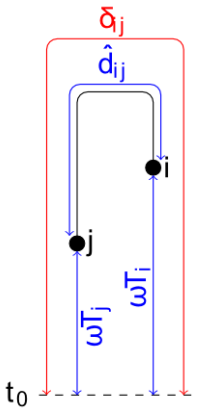
4.2 Description de la méthode

Considérons un ensemble de n séquences homologues alignées $\mathcal{S} = \{S_i, i = 1, \cdots, n\}$, associées aux souches $\mathcal{E} = \{i, i = 1, \cdots, n\}$ échantillonnées aux temps $\mathcal{T} = \{t_i, i = 1, \cdots, n\}$. Soit $d: \mathcal{S} \times \mathcal{S} \to \mathbb{R}_+$ la distance arborée entre ces souches correspondant à la vraie phylogénie (inconnue). Pour simplifier, $d(S_i, S_j)$ est noté d_{ij}. Soit $\hat{d}$ une estimation de d satisfaisant les conditions de symétrie ($\hat{d}_{ij} = \hat{d}_{ji}$), de positivité ($\hat{d}_{ij} \geq 0$ et $\hat{d}_{ii} = 0$) et de réflexivité ($\hat{d}_{ij} = 0 \Leftrightarrow i = j$). On dit que $\hat{d}$ est un indice de distance ou une dissimilarité sur $\mathcal{S}$. L'inégalité triangulaire est une condition non forcément respectée par $\hat{d}$ et non nécessaire ici. Pour tout t, $t' \in \mathcal{T}$, $t < t'$ signifie que t est plus ancien que t' ou que t' est plus récent que t. Soit $t_0 \in \mathcal{T}$ la date d'échantillonnage la plus récente, c'est-à-dire que $t_0 = \max_{i \in \mathcal{E}} t_i$. L'intervalle de temps qui sépare la date d'échantillonnage $t_i \in \mathcal{T}$ de t_0 se note $T_i = t_0 - t_i$ et s'exprime en unité de temps (généralement en jours, années ou générations). Comme $t_0 \geq t_k$, pour tout $k \in \mathcal{E}$, alors $t_0 - t_k = T_k \geq 0$ pour tout $k \in \mathcal{E}$.

Pour estimer le taux de substitution ω, relatif à l'ensemble $\mathcal{S}$, nous « corrigeons » $\hat{d}$, à l'aide de ω, en une mesure δ, fonction du taux de substitution, qui représente une distance ou dissimilarité où chaque souche est vue comme contemporaine (Figure 26). Soient deux souches i et j de $\mathcal{E}$, $i \neq j$, alors

(1) $$\delta_{ij}(\omega) = \hat{d}_{ij} + \omega \times (T_i + T_j).$$

Figure 26. Schéma représentant la définition de l'équation (1).
Soient deux taxa i et j échantillonnés aux temps t_i et t_j. La distance estimée séparant i et j est $\hat{d}_{ij}$, la distance restante pour ramener j (respectivement i) au temps contemporain t_0 est $\omega \times (t_0 - t_j) = \omega T_j$ (respectivement ωT_i). Donc la distance δ_{ij} qui voit i et j comme contemporains vaut $\hat{d}_{ij} + \omega \times (T_i + T_j)$.



Remarque 1. La mesure δ_{ij} est une fonction affine de ω ayant pour coefficient directeur $T_i + T_j$ et pour ordonnée à l'origine $\hat{d}_{ij}$. Comme $T_i \geq 0$ et $T_j \geq 0$, il en va de même pour leur somme $T_i + T_j \geq 0$. Ainsi, δ_{ij} est strictement croissante lorsque $T_i + T_j \neq 0$, c'est-à-dire lorsqu'au moins une des deux souches est échantillonnée à un temps différent de t_0 et est constante lorsque $T_i + T_j = 0$, c'est-à-dire lorsque les deux souches sont échantillonnées au temps t_0.

L'hypothèse de l'horloge moléculaire stricte stipule que l'évolution est un processus constant (à travers le temps) et uniforme (à travers les lignées). Donc, les souches identifiées comme contemporaines dans la vraie phylogénie sont à égale distance de leur ancêtre commun. Cette phylogénie peut se représenter par un dendrogramme et la distance additive associée est dite sphérique ou encore ultramétrique. Nous allons préciser cette notion, qui va nous servir à établir le critère sur lequel est basé ULS.

Définition 1. *Soit E un ensemble. Une distance $d: E \times E \to \mathbb{R}_+$ est ultramétrique si pour tout triplet i, j et k de E, la condition*

$$d_{ij} \leq max\{d_{ik}, d_{jk}\}$$

est vérifiée. Cette condition se nomme aussi la condition des trois points.

Soient i, j et k de $\mathcal{E}$, la définition de l'ultramétricité implique que deux des trois nombres d_{ij}, d_{ik} et d_{jk} sont égaux et maximaux. Il est évident que cette condition implique la condition des quatre points (pour tout x, y, z et t de $\mathcal{E}$: $d_{xy} + d_{zt} \leq max\{d_{xz} + d_{yt}, d_{xt} + d_{yz}\}$) et, par conséquent,

l'inégalité triangulaire (pour tout x, y et z de $\mathcal{E}$: $d_{xy} \leq d_{xz} + d_{zy}$). En notant respectivement par $d_{ijk}^{(1)}$ et $d_{ijk}^{(2)}$ le plus grand nombre et le deuxième plus grand nombre parmi d_{ij}, d_{ik} et d_{jk}, nous avons

$$d_{ijk}^{(1)} - d_{ijk}^{(2)} = 0 = d_{ijk}^{(2)} - d_{ijk}^{(1)}.$$

Proposition 1. *Soient un espace E et d une distance sur E. La distance d est ultramétrique si, et seulement si, le critère*

(2) $$Q(d) = \sum_{i<j<k} \left(d_{ijk}^{(1)} - d_{ijk}^{(2)}\right)^2$$

est nul.

Le critère Q utilise la méthode des moindres carrés pour tester l'ultramétricité d'une distance. L'équation (2) montre le critère sous sa forme la plus simple (OLS). Sachant que δ représente les distances entre les feuilles vues comme contemporaines et que nous faisons l'hypothèse d'un taux de substitution ω unique, alors une méthode naturelle d'estimation du taux de substitution ω consiste à minimiser le critère $Q\big(\delta(\omega)\big)$, ce qui revient à rendre δ le plus « ultramétrique possible ».

La formule (2) sous-entend que les estimations $\hat{d}_{ij}$ ont le même taux d'erreur quels que soient i et j. Une supposition fausse car plus la distance entre i et j est grande, plus l'erreur sur $\hat{d}_{ij}$ est importante. Pour prendre en compte cet effet Fitch et Margoliash (1967), Felsenstein (1997) et d'autres proposent d'ajouter une valeur de pondération à chaque terme de la somme, en suivant une approche de type WLS. Cette valeur de pondération augmente l'importance donnée aux estimations ayant un faible taux d'erreur, c'est-à-dire associées à une petite variance, et une faible importance aux estimations ayant un fort taux d'erreur, c'est-à-dire associées à une grande variance. Par ce procédé, la formule (2) devient :

(3) $$Q\big(\delta(\omega)\big) = \sum_{i<j<k} w_{ijk} \left(\delta_{ijk}^{(1)}(\omega) - \delta_{ijk}^{(2)}(\omega)\right)^2$$

où w_{ijk} est la valeur de pondération, dépendante de i, j et k, attribuée à chaque terme de la somme. Le choix de cette valeur de pondération, correspondant à l'inverse d'une variance, sera discuté à la section 4.2.3.

4.2.1 Minimisation du critère d'ultramétricité sur un triplet

Dans un premier temps, nous allons étudier le comportement du critère en ne considérant qu'un seul triplet, puis nous le généraliserons sur plusieurs triplets. De plus, avec un seul triplet la pondéra-

tion WLS n'intervient pas, simplifiant alors l'analyse. Notons par $\delta_{|ijk}(\omega)$ la distance corrigée $\delta(\omega)$ restreinte au seul triplet i, j et k de $\mathcal{E}$.

Sur un triplet, l'équation (2) se réduit à un terme

$$Q\left(\delta_{|ijk}(\omega)\right) = \left(\delta_{ijk}^{(1)}(\omega) - \delta_{ijk}^{(2)}(\omega)\right)^2. \tag{4}$$

Ce terme est dépendant des trois droites $\delta_{ij}(\omega)$, $\delta_{ik}(\omega)$ et $\delta_{jk}(\omega)$. Leurs équations sont définies en (1). Dans la suite et pour simplifier, on omettra d'indiquer ω lorsque ce ne sera pas nécessaire.

Remarque 3. Soit un intervalle $I = [a, b]$ où les droites δ_{ij}, δ_{ik} et δ_{jk} n'ont aucun point d'intersection. Alors les deux plus hautes droites correspondent aux termes $\delta_{ijk}^{(1)}(\omega)$ et $\delta_{ijk}^{(2)}(\omega)$, qui restent identiques sur I. Posons $\delta_{ijk}^{(1)}(\omega) = a_1\omega + b_1$ et $\delta_{ijk}^{(2)}(\omega) = a_2\omega + b_2$. Alors

$$Q\left(\delta_{|ijk}(\omega)\right) = \left((a_1 - a_2)\omega + b_1 - b_2\right)^2 = A^2\omega^2 + 2AB\omega + B^2,$$

avec $A = a_1 - a_2$ et $B = b_1 - b_2$. Ainsi, $Q\left(\delta_{|ijk}(\omega)\right)$ est une parabole de variable ω. Elle est positive (c'est une différence au carré) et convexe (le coefficient de son monôme de plus haut degré est positif). De plus, lorsque les droites $\delta_{ijk}^{(1)}$ et $\delta_{ijk}^{(2)}$ sont parallèles ($a_1 = a_2$), le critère est constant sur I ($A = 0$).

Regardons maintenant ce qu'il se passe autour des points d'intersection des trois droites $\delta_{ij}(\omega)$, $\delta_{ik}(\omega)$ et $\delta_{jk}(\omega)$, c'est-à-dire là où les termes $\delta_{ijk}^{(1)}$ et $\delta_{ijk}^{(2)}$ sont modifiés, correspondant alors à des droites différentes. Il convient de distinguer deux types de points d'intersection. Le premier regroupe les points d'intersection qui ne modifient pas l'expression du critère. Ce sont les points d'intersection entre les deux plus hautes droites ; ils permutent les droites représentées par les termes $\delta_{ijk}^{(1)}$ et $\delta_{ijk}^{(2)}$ entre elles, mais leur différence au carré reste identique avant et après le point d'intersection qui a la particularité d'annuler le critère. Le second regroupe les points d'intersection qui modifient l'expression du critère. Ce sont les points d'intersection entre la deuxième et la troisième plus haute droite. Ils modifient uniquement la droite représentée par le terme $\delta_{ijk}^{(2)}$, et, donc, le comportement du critère avant et après ce point change, au sens où on a toujours une parabole, mais d'équation différente.

Définition 2. *Soient trois droites quelconques A, B et C définies sur $\mathbb{R}$, et soit x_{ab} le point d'intersection entre les droites A et B. Le point x_{ab} est dit « frontière » si l'inégalité*

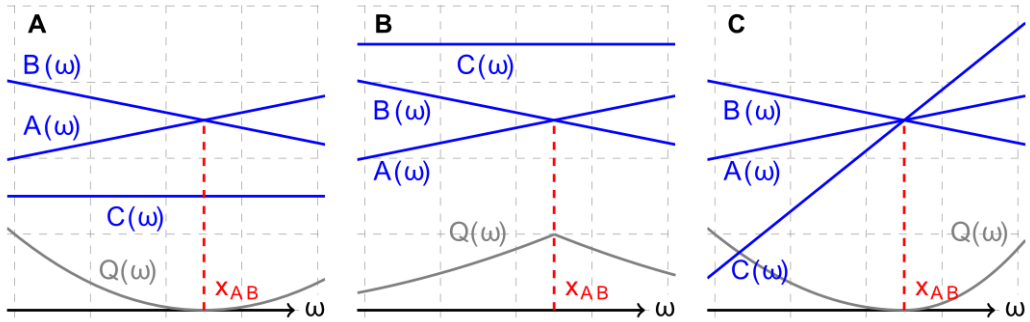
$$A(x_{AB}) = B(x_{AB}) \leq C(x_{AB})$$

est vérifiée. Il est dit « solution » si l'inégalité

$$A(x_{AB}) = B(x_{AB}) \geq C(x_{AB})$$

*est vérifiée (*Figure 27*).*

Figure 27. Différence entre point solution et point frontière.
Le graphique A montre un point solution (une seule parabole dont on atteint le minimum), le graphique B un point frontière (on change de parabole) et le graphique C montre qu'il est possible pour un point d'être à la fois frontière et solution (deux paraboles se rencontrent sur leur minimum).

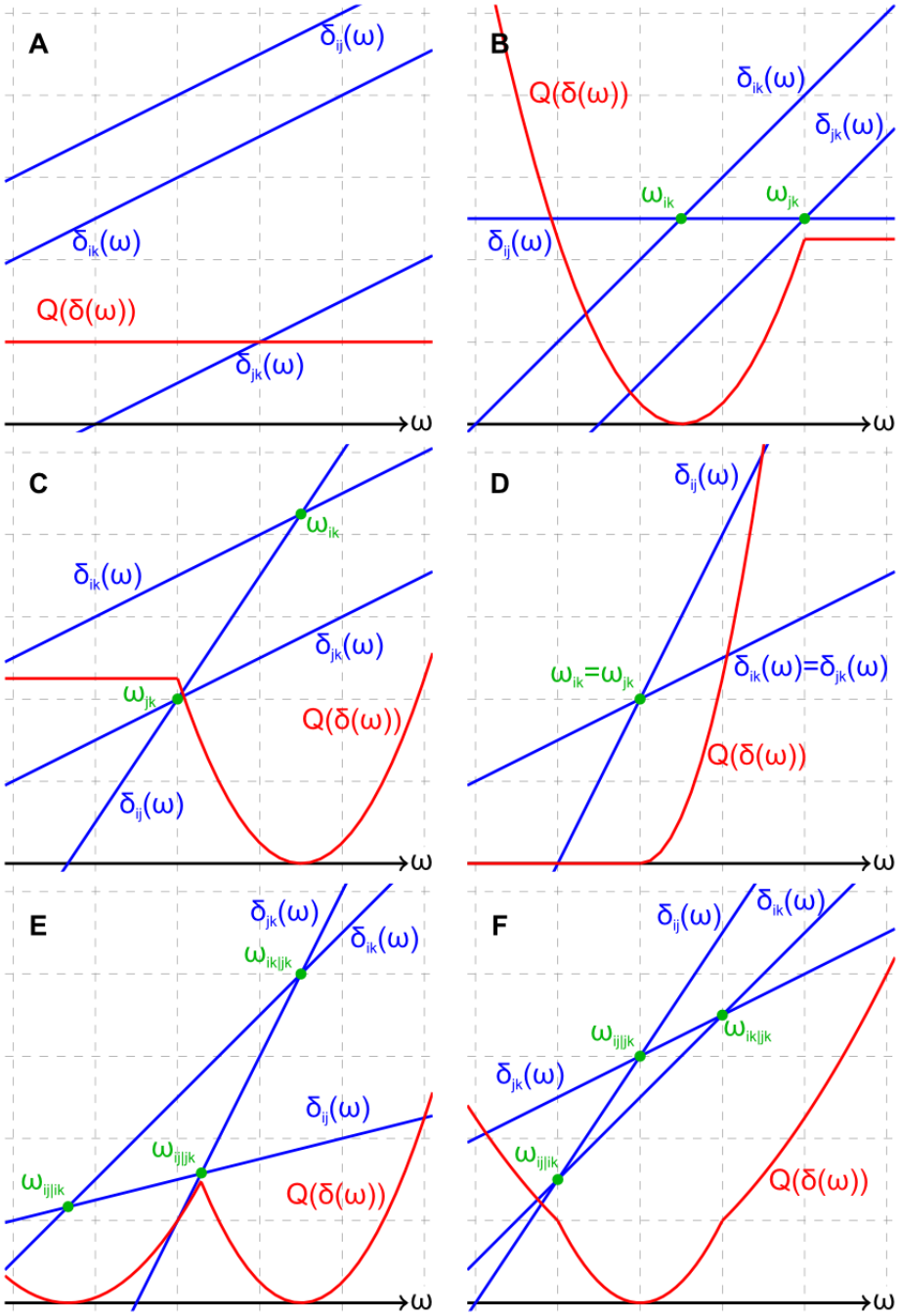


Les points d'intersection solutions entre les droites δ_{ij}, δ_{ik}et δ_{jk} sont ceux qui annulent le critère $Q(\delta_{|ijk})$. En effet, en ces points, les deux termes $\delta_{ijk}^{(1)}$ et $\delta_{ijk}^{(2)}$ sont égaux et leur différence est donc nulle. Ils correspondent à une solution optimale pour l'estimation du taux de substitution avec le critère $Q(\delta_{|ijk})$. Les autres points d'intersection, les points frontières, n'annulent pas le critère, mais changent la fonction parabolique représentant sa valeur. Ainsi, le critère est une fonction parabolique par morceaux. La définition 2 n'exclut pas le fait qu'un point d'intersection peut être à la fois frontière et solution (Figure 27C). Ce dernier cas se produit lorsque les trois droites sont concourantes. La Figure 28 montre le comportement du critère sur un triplet, ainsi que les droites δ_{ij}, δ_{ik} et δ_{jk} ayant permis d'obtenir son allure.

Remarque 4. *L'expression $Q(\delta_{|ijk})$, définie en* (4)*, est continue sur* $\mathbb{R}$.

Les différents cas de figures montrés à la Figure 28 suggèrent qu'il y a toujours au moins une solution pour le taux de substitution ω, qui rende $\delta_{|ijk}$ ultramétrique, sauf lorsque les droites δ_{ij}, δ_{ik} et δ_{jk} sont parallèles. Cependant, les minimums de la fonction ne sont pas toujours acceptables (Figure 29). En effet, comme le taux de substitution est une vitesse, les valeurs négatives ne peuvent lui être assignées ; dans ce cas, la solution optimale est parfois zéro, comme montré dans la Figure 29.

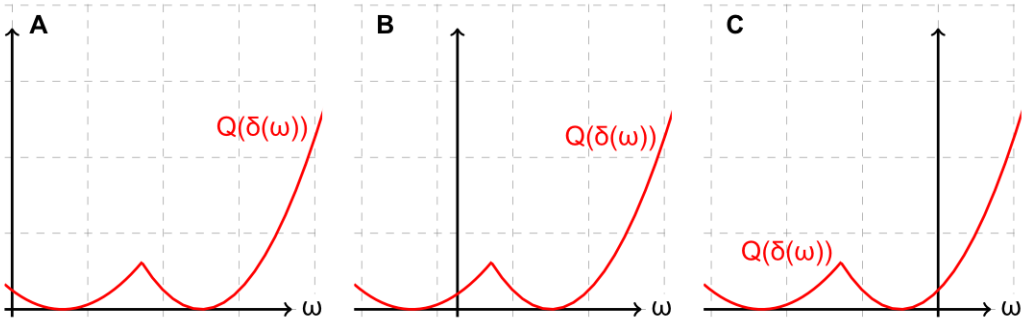
Figure 28. Quelques exemples (non exhaustifs) de l'allure du critère restreint à un triplet.
Le comportement du critère restreint à un triplet sur $\mathbb{R}$ est montré en rouge, les trois droites δ_{ij}, δ_{ik} et δ_{jk} en bleu et les points d'intersection en vert. Pour l'exemple A, les trois droites sont parallèles et le critère résultant est une droite. Les exemples B et C montrent les cas où seulement deux droites sont parallèles et l'exemple D le cas où deux droites sont confondues. Les exemples E et F se produisent dans le cas où aucune des droites n'est parallèle. L'exemple E a deux solutions tandis que l'exemple F a deux frontières.



Dans le cas d'un seul triplet, il est généralement possible (sauf Figure 29C) de trouver au moins une valeur pour le taux de substitution rendant la distance $\delta_{|ijk}(\omega)$ ultramétrique. Cette valeur est déterminée par le(s) point(s) d'intersection solution(s), qui coïncide(nt) avec un minimum du critère. La valeur de ces points peut aussi être obtenue en annulant la dérivée du critère. Lorsque nous considérerons plusieurs triplets (cf. ci-dessous), les points d'intersection solutions des différents triplets

ne correspondront plus avec un minimum du critère (sauf cas très particulier), et pour en déterminer le minimum global, il faudra dériver chaque morceau du critère, lui aussi parabolique par morceaux. Par ailleurs, dans le cas d'un seul triplet, le minimum rend toujours la distance corrigée parfaitement ultramétrique. Mais cette observation n'est plus vraie lorsque plusieurs triplets sont considérés et le résultat rendra la distance corrigée « aussi ultramétrique que possible ».

Figure 29. Solutions considérées par l'algorithme ULS suivant la positivité ou la négativité des points solutions.
Le critère restreint à un triplet a toujours au plus deux solutions qui l'annulent. Cependant, ces solutions ne sont pas toujours convenables suivant qu'elles soient positives ou négatives (graphiques A et B). Dans le cas du graphique C, aucune des solutions proposées par le triplet n'est considérée, et zéro sera présenté comme solution optimale.



4.2.2 Minimisation du critère d'ultramétricité sur plusieurs triplets

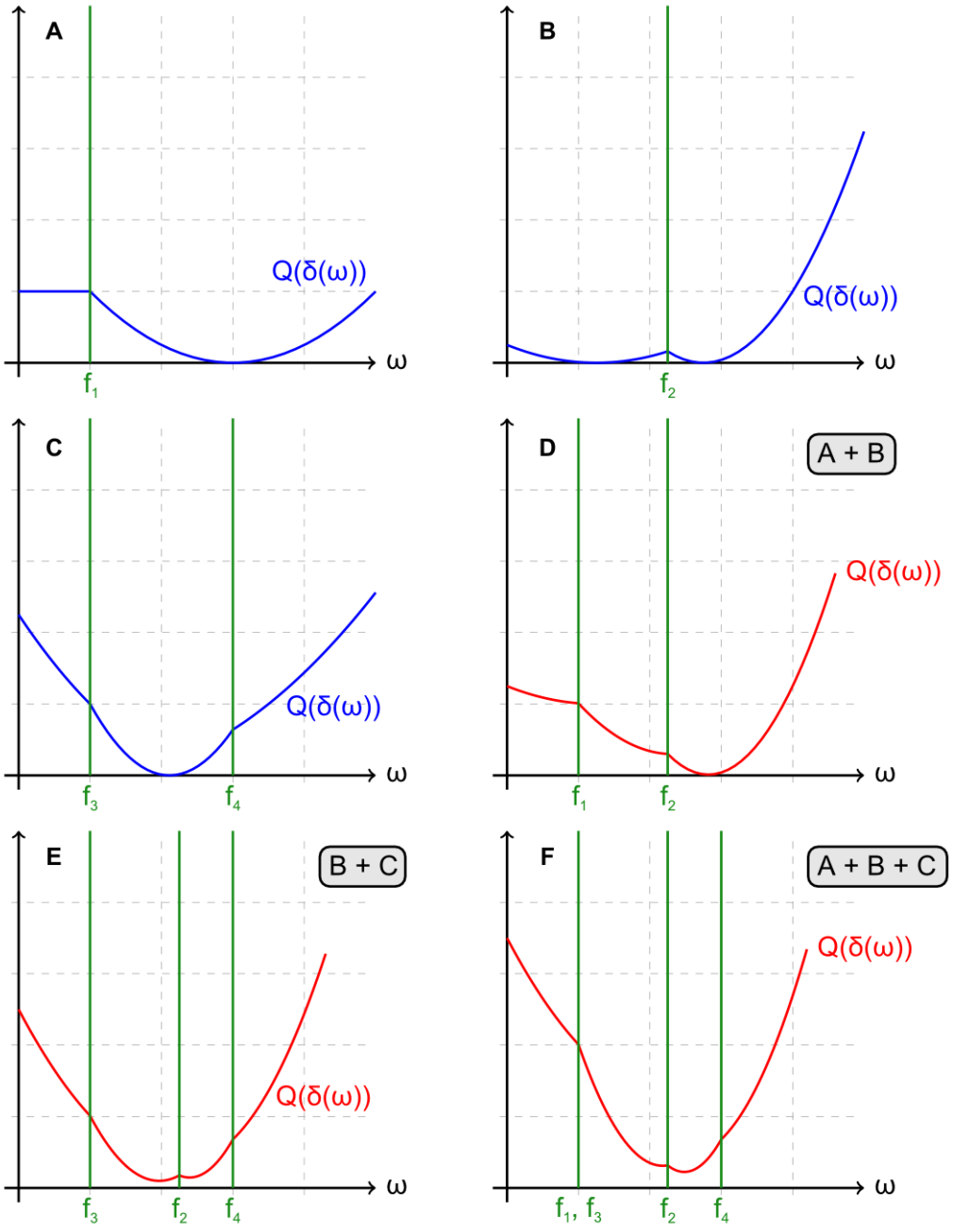
Le critère d'ultramétricité restreint à un triplet se comporte sur $\mathbb{R}$ comme une fonction parabolique par morceaux, où chaque morceau est soit une constante ou soit une parabole convexe positive. Quand il est étudié sur plusieurs triplets, il devient une somme de fonctions paraboliques par morceaux (Figure 30), c'est donc toujours une fonction parabolique par morceaux, mais elle est plus complexe. Néanmoins, l'objectif reste d'en déterminer le minimum, considéré comme la meilleure estimation possible pour le taux de substitution ω, et qui rend la distance corrigée δ aussi ultramétrique que possible.

L'étude du critère sur un triplet a montré que le minimum du critère correspond au(x) point(s) solution(s)**.** Les exemples E et F de la Figure 30 montrent que cela n'est plus systématiquement vrai. Donc, le minimum du critère ne peut être obtenu que par dérivation des morceaux de parabole du critère. Il convient ensuite de vérifier si ce minimum est inclus dans le domaine de définition du morceau considéré. Il devient alors une solution potentielle, et le minimum global correspond alors au minimum de toutes les solutions potentielles (le plus souvent uniques).

Le critère est continu en tout point de $\mathbb{R}$, puisqu'il est la somme de critères continus définis sur chaque triplet, mais il n'en est pas dérivable aux points frontières. Cependant, en ces points la propriété de dérivabilité n'est pas nécessaire. En effet, sauf dans un cas particulier (Figures 28A et 28D), extrêmement rare avec des données réelles et correspondant à un plateau, on peut montrer que les

morceaux de parabole ont une pente croissante après le passage de chaque point frontière (Figures 30D, 30E et 30F). De ce fait, il est inutile de rechercher une solution aux points frontières puisqu'elle se trouvera forcément entre ces points (à moins d'un plateau, ou lorsque le changement de paraboles correspond aussi à leur minimum, cf. Figure 27C, ces deux cas nécessitant un soin particulier, bien que très peu probables).

Figure 30. Comportement du critère sur plusieurs triplets.
Les graphiques A, B et C montrent le critère obtenu à partir d'un triplet, tandis que les graphiques D, E et F le montrent comme étant la somme de plusieurs critères définis sur un triplet. Le critère du graphique D (respectivement E) correspond à la somme des critères A et B (respectivement B et C). Le critère du graphique F correspond à la somme des trois critères A, B et C. Les graphiques E et F montrent qu'il n'est pas possible d'obtenir une valeur qui annule le critère, mais un minimum existe et il est alors choisi comme solution optimale. Le graphique F montre qu'il est possible d'avoir plusieurs points frontières identiques et E qu'il est possible d'avoir plusieurs minima locaux.



Remarque 5. Soit un intervalle $I \subset \mathbb{R}$ sur lequel il n'y a aucun point frontière. Alors les variables $\delta_{ijk}^{(1)}$ et $\delta_{ijk}^{(2)}$, de chaque terme de la somme (2), représentent toujours les mêmes droites. Pour un triplet i, j et k de $\mathcal{E}$, posons $\delta_{ijk}^{(1)}(\omega) = a_{ijk}^{(1)}\omega + b_{ijk}^{(1)}$ et $\delta_{ijk}^{(2)}(\omega) = a_{ijk}^{(2)}\omega + b_{ijk}^{(2)}$, alors

$$Q\big(\delta(\omega)\big) = \sum_{i<j<k} \left(\delta_{ijk}^{(1)}(\omega) - \delta_{ijk}^{(2)}(\omega)\right)^2 = \sum_{i<j<k} \left(\left(a_{ijk}^{(1)} - a_{ijk}^{(2)}\right)\omega + b_{ijk}^{(1)} - b_{ijk}^{(2)}\right)^2$$

d'où

$$Q\big(\delta(\omega)\big) = \omega^2 \sum_{i<j<k} \left(a_{ijk}^{(1)} - a_{ijk}^{(2)}\right)^2 + 2\omega \sum_{i<j<k} \left(a_{ijk}^{(1)} - a_{ijk}^{(2)}\right)\left(b_{ijk}^{(1)} - b_{ijk}^{(2)}\right) + \sum_{i<j<k} \left(b_{ijk}^{(1)} - b_{ijk}^{(2)}\right)^2.$$

Le minimum ω^* de cette parabole est déductible par dérivation et vaut

$$\omega^* = -\frac{\sum_{i<j<k}\left(a_{ijk}^{(1)} - a_{ijk}^{(2)}\right)\left(b_{ijk}^{(1)} - b_{ijk}^{(2)}\right)}{\sum_{i<j<k}\left(a_{ijk}^{(1)} - a_{ijk}^{(2)}\right)^2}.$$

Si $\omega^* \in I$, alors ω^* est un minimum local au critère $Q(\delta)$. Dans le cas contraire, $Q(\delta)$ n'a aucun minimum sur I, du fait de la propriété précédente.

Pour obtenir le minimum global du critère, il faut donc rechercher le minimum local de chaque morceau. La remarque 5 fournit, pour un morceau de parabole, la formule donnant son minimum local. L'algorithme va donc consister à parcourir les points frontières et à vérifier sur chaque intervalle si celui-ci contient ou non une solution potentielle.

Comme le taux de substitution est nécessairement positif ou nul, seule la définition du critère sur $\mathbb{R}_+$ a besoin d'être considérée. Mais il se peut que certains triplets aient leurs points frontières et solutions tous négatifs (Figure 29C). Ce type de triplets biaise l'estimation du taux de substitution et « pousse » celle-ci vers zéro. Pour éviter un tel biais, les triplets ayant cette caractéristique sont omis de l'analyse.

Le comportement du critère reste identique avec la version pondérée du critère. En effet, multiplier chaque terme de la somme par une constante positive ne modifie en rien ses propriétés. Cependant, cette constante influe sur l'importance à donner à tels ou tels termes. En effet, plus cette constante est proche de zéro, plus le terme correspondant, jugé non informatif, tend à ressembler à une droite horizontale passant par zéro (et a donc peu d'influence sur le critère), tandis que plus sa valeur est élevée, plus elle donne de l'impact à ce terme (jugé informatif). Ainsi, considérer des pondérations sur chaque terme de la somme n'a pas de conséquence sur les propriétés énoncées précédemment, mais uniquement sur le résultat final.

4.2.3 Détermination de la valeur de pondération optimale

Pour les méthodes standards de reconstruction phylogénétique, la valeur de pondération associée à chaque terme de la somme des moindres carrés pondérés (WLS) est $1/\hat{d}_{ij}$ (Swofford *et al*, 1996), car $\hat{d}_{ij}$ est une bonne approximation (à un facteur constant dépendant du nombre de sites) de la variance associée à $\hat{d}_{ij}$ (Felsenstein, 1984), signifiant que la confiance dans l'estimation des distances évolutives devient de plus en plus faible au fur et à mesure que $\hat{d}_{ij}$ est grand. Une variante de pondération largement utilisée, proposée par Fitch et Margoliash (1967), est $1/\hat{d}_{ij}^2$. Sur un jeu de données comprenant 43 mammifères, Sanjuán et Wróbel (2005) montrent que 1,823 est une valeur d'exposant optimal, plutôt que 2. Sur un autre jeu de données contenant des souches *env* du VIH-1 provenant de différents organes (moelle osseuse, cerveau, liquide cérébro-spinal, rein, foie, poumon, ganglion lymphatique et rate), ils montrent, cette fois-ci, que la valeur d'exposant 1,766 est la plus appropriée. Ce dernier jeu de données avait pour but de montrer le phénomène de compartimentalisation du VIH (McGrath *et al*, 2001). Dans tous les cas, ils indiquent que $1/\hat{d}_{ij}^2$ fournit des résultats similaires. Cette valeur d'exposant (2) semble être *a priori* optimale (ou n'en est pas loin) et nous la conserverons dans les simulations ci-après.

Les formules de pondération précédentes ne peuvent pas être utilisées directement pour w_{ijk} puisque chaque terme de la somme du critère dépend sur $\mathbb{R}$ des trois distances $\hat{d}_{ij}$, $\hat{d}_{ik}$ et $\hat{d}_{jk}$ (sauf cas particulier). De plus, nous avons montré que le critère est une fonction continue et, afin de conserver cette propriété remarquable, il est important que la valeur de pondération associée à chaque terme de la somme soit constante sur $\mathbb{R}$. Une mesure naturelle est alors d'utiliser l'inverse de la somme des trois distances, soit

$$w_{ijk} = \frac{1}{\hat{d}_{ij} + \hat{d}_{jk} + \hat{d}_{ik}}.$$

Cette valeur de pondération a le même comportement que celle utilisée classiquement. Néanmoins, si la somme des trois distances est très proche de zéro, une confiance presque absolue est donnée au terme correspondant et faussera inévitablement les estimations du taux de substitution. Il est d'usage de supposer que chaque mesure $\hat{d}_{ij}$ ne peut pas être plus petite que la moitié d'une substitution observée, c'est-à-dire $\hat{d}_{ij} \geq 1/(2N)$, où N est la longueur de l'alignement (Swofford *et al*, 1996). Ainsi, nous pouvons rajouter un pseudo-compte au dénominateur afin d'éviter les valeurs trop petites ou nulles, d'où

$$w_{ijk} = \frac{1}{\hat{d}_{ij} + \hat{d}_{ik} + \hat{d}_{jk} + k/N},$$

où k est choisi sur la base de simulations de manière à optimiser l'impact de w_{ijk} sur chaque terme de la somme. En suivant l'approche classique de Fitch et Margoliash (1967), on peut employer le carré de cette pondération, soit

(5) $$w_{ijk} = \frac{1}{\left(\hat{d}_{ij} + \hat{d}_{ik} + \hat{d}_{jk} + k/N\right)^2}.$$

Le produit des trois distances (au lieu de leur somme) peut aussi être considéré comme valeur de pondération, bien que l'interprétation en soit moins naturelle, soit

$$w_{ijk} = \frac{1}{\hat{d}_{ij} \times \hat{d}_{ik} \times \hat{d}_{jk} + k/N},$$

et à nouveau on peut passer au carré

(6) $$w_{ijk} = \frac{1}{\left(\hat{d}_{ij} \times \hat{d}_{ik} \times \hat{d}_{jk} + k/N\right)^2}.$$

Dessimoz *et al*. (2006) ont proposé des formules moins empiriques, bien adaptées à notre cas de figure. Ils ont calculé la variance de la différence $\left(\hat{d}_{ik} - \hat{d}_{jk}\right)$, et notre critère est justement calculé à partir de différences de distances, qui est approximée par la formule

(7) $$\sigma^2\left(\hat{d}_{ik} - \hat{d}_{jk}\right) = \frac{\hat{d}_{ij}^{1,3182}}{v_{ij}^{0,3026}} \times \frac{\left(v_{ik} + v_{jk}\right)^{1,0933}\left(v_{ik}v_{jk}\right)^{0,1181}}{\left(\hat{d}_{ik} + \hat{d}_{jk}\right)^{1,2449}},$$

où les termes v_{ij}, v_{ik} et v_{jk} sont les variances respectivement associées aux distances $\hat{d}_{ij}$, $\hat{d}_{ik}$ et $\hat{d}_{jk}$. Ces variances peuvent être calculées pour la plupart des modèles d'évolution, mais elles sont rarement données par les logiciels les plus couramment utilisés. Elles peuvent alors être approximées par $\hat{d}_{ij}/N$, $\hat{d}_{ik}/N$ et $\hat{d}_{jk}/N$ respectivement, où N est la longueur de l'alignement. Le problème avec cette variance est qu'elle considère une paire de distances, issues de $\delta_{ijk}^{(1)}$ et $\delta_{ijk}^{(2)}$, et que cette paire peut varier aux points frontières. Afin de conserver une variance constante sur $\mathbb{R}$ (cf. ci-dessus), nous considérons, pour chaque terme de la somme, la variance associée aux droites avec un point solution, et si plusieurs variances sont possibles, la plus grande est choisie. Ainsi, la valeur de pondération correspond à l'inverse de cette variance maximum plus un pseudo-compte de la forme k/N^2.

La pertinence des différentes valeurs de pondérations proposées est testée sur un jeu de données contenant 800 simulations (Figure 31) et pour lequel les séquences sont de longueur 300 paires de bases pour la figure A et de 1 000 paires de bases pour la figure B. La description complète du proto-

cole de simulation est donnée à la section 4.3.1.1. Les valeurs de pondération testées sont : sans pondération en rouge, pondération tenant compte de la variance de Dessimoz (7) pour k valant 0 et 1 en jaune et vert respectivement, pondération (5) pour k valant 0 et 1 respectivement en cyan et bleu et pondération (6) pour k valant 1 en violet. L'ordonnée indique la précision d'estimation des différentes valeurs de pondération (fonction déviation relative, cf. section 4.3.1.2). Pour l'essentiel, plus cette valeur est petite, plus les estimations sont précises. Différentes entrées sont aussi étudiées, soit avec des matrices de distances (calculées avec DNAdist), soit avec des arbres FastME (calculés à partir des matrices de distances) ou soit avec des arbres PhyML. Dans ces deux derniers cas, on commence par calculer la distance patristique avant d'appliquer ULS. Les précisions d'estimation des différentes valeurs de pondération proposées sont quasi-identiques pour des valeurs de k oscillant entre 0,5 et 2, et ne sont donc pas toutes montrées. De même, les précisions d'estimation de la valeur de pondération (6) pour $k = 0$ ne sont pas montrées car elles sont systématiquement supérieures à 0,2 pour des séquences de 300 paires de bases et supérieures à 0,1 pour des séquences de 1 000 paires de bases. Les différences en précision d'estimation sont flagrantes en passant d'alignements de 300 sites à 1 000 sites, comme on peut s'y attendre. Dans ce dernier cas, il n'y a presque plus de différence en précision d'estimation entre les différentes valeurs de pondération puisqu'au fur et à mesure que la longueur des séquences augmente les estimations des distances évolutives deviennent de plus en plus justes et la pondération devient alors plus ou moins superflue, en particulier avec des arbres FastME et PhyML. Les alignements de 300 sites semblent dont être le cas de figure idéal pour trancher entre les différentes valeurs de pondération où l'on observe, qu'en moyenne, la pondération (6) pour $k = 1$ est la plus performante, en particulier avec des matrices de distances. Donc, la valeur de pondération (6) pour $k = 1$ est conservée pour notre méthode. Remarquons que la valeur de pondération (5) et celle avec la variance de Dessimoz (7), pour $k = 0$, semblent inadaptées avec des arbres PhyML et des séquences de 300 paires de bases, mais ce problème est corrigé en utilisant un pseudo-compte ou en considérant un alignement plus grand.

4.2.4 Limites algorithmiques et solutions proposées

La mise en œuvre d'ULS nécessite l'utilisation de techniques algorithmiques avancées pour résoudre certains problèmes pratiques (essentiellement place mémoire) et d'observer les propriétés du critère (par exemple, la possibilité d'échantillonner les triplets) afin d'obtenir un programme rapide et précis.

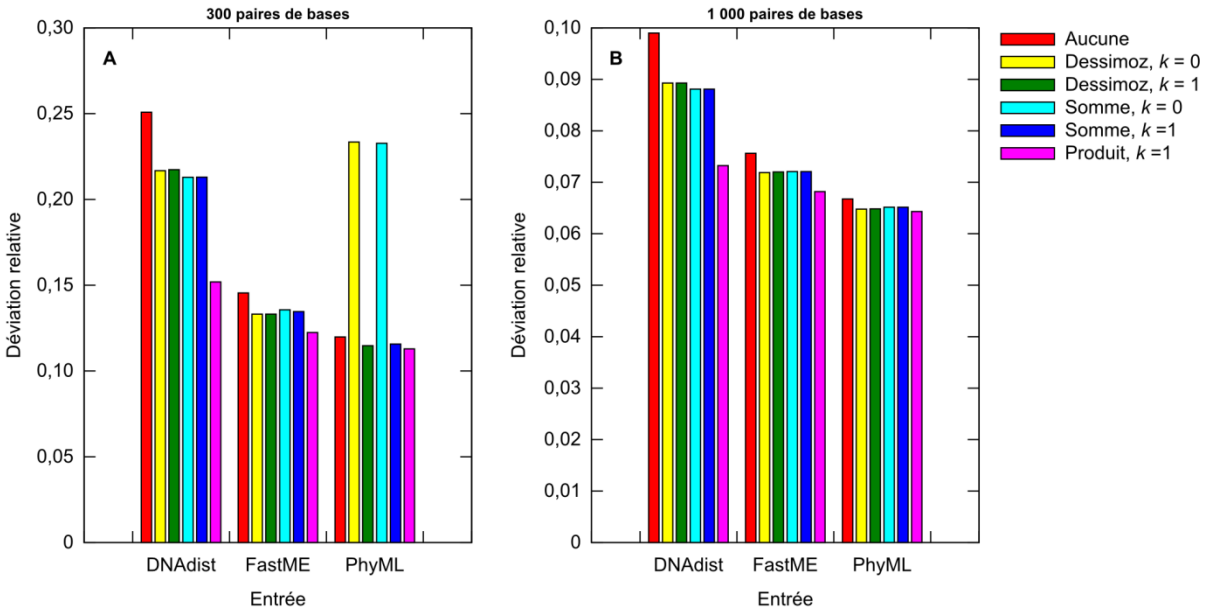
4.2.4.1 Conservation des coefficients de chaque morceau de parabole

ULS minimise un critère qui est une fonction parabolique par morceaux pour déterminer la meilleure estimation du taux de substitution. Chaque morceau du critère est un polynôme du second

degré (ou une constante qui en est un cas particulier) et peut donc être représenté par les coefficients de ses trois monômes (dans le cas d'une constante, les coefficients des monômes de degré un et deux sont nuls). Pour connaître le minimum global du critère, il faut donc connaître avec exactitude la définition numérique de chaque morceau de parabole, ainsi que leur intervalle de définition.

Initialement, les frontières sont indéterminées. Elles apparaissent progressivement et de façon aléatoire lors du parcours des triplets. Ces contraintes obligent à devoir construire les coefficients associés à chaque morceau de façon progressive mais surtout non dépendante de l'intervalle auquel il appartient (puisque cet intervalle est uniquement connu à la fin du parcours des triplets). Autrement, il serait nécessaire de parcourir l'ensemble des triplets deux fois de suite : une première fois pour connaître toutes les frontières et la seconde pour connaitre les coefficients de chaque morceau de parabole.

Figure 31. Performance en précision d'estimation des différentes valeurs de pondération étudiées.
Ces graphiques montrent en ordonnée les performances en précision d'estimation (plus les valeurs sont petites, meilleures sont les estimations ; fonction déviation relative, cf. section 4.3.1.2) pour les différentes valeurs de pondération proposées (de gauche à droite : aucune pondération en rouge, pondération avec variance de Dessimoz (7) pour $k = \{0; 1\}$ en jaune et vert respectivement, pondération (5) pour $k = \{0; 1\}$ en cyan et bleu respectivement, pondération (6) pour $k = 1$ en violet) et pour différentes entrées (matrices de distances (DNAdist), arbres FastME et PhyML). Les figures A et B correspondent aux mêmes jeux de données contenant chacun 800 simulations (cf. section 4.3.1.1), mais avec des séquences de 300 paires de bases pour la figure A et de 1 000 paires de bases pour la figure B. Les précisions d'estimation de la pondération (6) pour $k = 0$ sont toutes supérieures à 0,2 avec des séquences de 300 paires de bases et à 0,1 avec des séquences de 1 000 paires de bases. Enfin, aucune différence notable n'existe pour différentes valeurs de k oscillant entre 0,5 et 2.



Une méthode simple et efficace qui permet de construire les coefficients de chaque morceau de parabole, en s'abstenant de la connaissance de leur intervalle de validité, est d'utiliser des coefficients temporaires associés à chaque frontière. Notons par (a_u, b_u, c_u) les coefficients du morceau de parabole $m_u(\omega) = a_u\omega^2 + b_u\omega + c_u$ défini entre les frontières f_u et f_{u+1} ($f_u < f_{u+1}$). Chaque frontière f_u correspond à la borne inférieure de l'intervalle de définition du morceau m_u et à la

borne supérieure de l'intervalle de définition du morceau m_{u-1}. Elle contient les coefficients temporaires (a'_u, b'_u, c'_u) qui permettent de connaître les coefficients (a_u, b_u, c_u) du morceau de parabole m_u en les sommant aux coefficients $(a_{u-1}, b_{u-1}, c_{u-1})$ du morceau de parabole m_{u-1}, c'est-à-dire

$$(a_u, b_u, c_u) = (a_{u-1}, b_{u-1}, c_{u-1}) + (a'_u, b'_u, c'_u). \quad (8)$$

Concevoir de tels coefficients temporaires est chose aisée. En effet, les frontières associées à un triplet, ainsi que les coefficients des morceaux de parabole entre ces frontières, sont facilement calculables et constituent la base pour concevoir les coefficients temporaires. Imaginons que pour un triplet i, j et k de $\mathcal{E}$ nous avons deux frontières positives f^1_{ijk} et f^2_{ijk}, $f_0 < f^1_{ijk} < f^2_{ijk}$, avec $f_0 = 0$. Ajoutons aux coefficients temporaires de f_0 les coefficients du morceau propre à i, j et k passant par f_0 (ici défini sur $]-\infty; f^1_{ijk}]$) et retranchons-les aux coefficients temporaires de f^1_{ijk}. Ensuite, ajoutons aux coefficients temporaires de f^1_{ijk} les coefficients du morceau propre à i, j et k défini entre $[f^1_{ijk}; f^2_{ijk}]$, puis après avoir retranché ces derniers aux coefficients temporaires de f^2_{ijk}, ajoutons-leur les coefficients du morceau de parabole défini entre $[f^2_{ijk}; +\infty[$. En procédant de même pour tous les triplets, les coefficients associés à chaque frontière ont la propriété de l'équation (8) et sont calculés itérativement en parcourant l'ensemble des triplets.

Comme le taux de substitution est supposé être positif ou nul (hypothèse biologique évidente), la définition du critère sur $\mathbb{R}_-$ est inutile. Aussi, les frontières négatives ne sont pas considérées. Cependant, chaque triplet a une influence non négligeable sur la partie positive ($\mathbb{R}_+$) du critère, même si toutes ses frontières sont négatives (sauf, bien sûr, dans le cas d'une constante ; l'influence est alors identique en tout point de $\mathbb{R}$ et peut être négligée). Dans ce dernier cas, le triplet est uniquement considéré s'il a au moins une solution positive, et cela même si toutes ses frontières sont négatives. Les triplets n'ayant pas de solution(s) positive(s) ne sont jamais considérés (cf. section 4.2.1). Ainsi, $f_0 = 0$ est considéré comme la première frontière de chaque triplet. Elle contient donc l'information de tous les morceaux de parabole considérés. De ce fait, les coefficients temporaires (a'_0, b'_0, c'_0) associés à cette frontière correspondent aux vrais coefficients (a_0, b_0, c_0) associés au premier morceau de parabole m_0. Donc, les coefficients du morceau m_u, défini sur l'intervalle $[f_u; f_{u+1}]$, s'expriment par la relation

$$(a_u, b_u, c_u) = \sum_{k=0}^{u} (a'_k, b'_k, c'_k),$$

où (a'_k, b'_k, c'_k) représentent les coefficients temporaires associés à la frontière f_k, $k = \{0, \cdots, u\}$.

Il convient de noter que, lors d'un cas particulier où le critère associé à un triplet a deux solutions, une positive et l'autre négative (Figure 29B), le morceau de parabole défini sur $\mathbb{R}_+$ qui correspond à la solution négative (s'il y en a un) n'est pas considéré (puisqu'il biaisera l'estimation du taux de substitution vers zéro) et que, dans ce cas, le morceau de parabole associé à la solution positive débutera à zéro et non au point frontière.

4.2.4.2 Parcours de chaque morceau du critère et estimation des minima locaux

Pour calculer le minimum global du critère, nous devons estimer le minimum de chacun de ses morceaux et vérifier s'il correspond à un minimum à considérer, c'est-à-dire si le minimum du morceau est compris dans l'intervalle de définition de celui-ci. Dans la section précédente, nous décrivons la manière de calculer les coefficients temporaires stockés dans chaque frontière. Imaginons que nous disposons d'un tableau contenant ces frontières de façon unique et par ordre croissant (cela fera l'objet de la section suivante). Pour déterminer le minimum global du critère nous devons parcourir chaque morceau de celui-ci. Comme les frontières négatives ne sont pas considérées, la première dans ce tableau est donc $f_0 = 0$ et c'est celle dont les coefficients temporaires (a'_0, b'_0, c'_0) représentent les vrais coefficients (a_0, b_0, c_0) associés au premier morceau de parabole m_0. L'abscisse du minium de m_0 est donc $\omega_0 = -b_0/(2a_0)$. Si $f_0 \leq \omega_0 \leq f_1$, ce minimum est à considérer comme un minimum local, sinon il est ignoré. Les coefficients du morceau m_1 s'obtiennent en $O(1)$ en ajoutant aux coefficients (a_0, b_0, c_0) les coefficients temporaires associés à la frontière f_1. Puis, on calcule le minimum, vérifie s'il constitue un minimum local, et on passe à m_2. Et ainsi de suite. En parcourant de cette manière le tableau, l'équation numérique de chaque morceau m_u du critère est déductible facilement et son minimum $\omega_u = -b_u/2a_u$ est local si, et seulement si, $f_u \leq \omega_u \leq f_{u+1}$. Chaque frontière nécessite un calcul en $O(1)$, si bien que la complexité en temps (il faut parcourir l'ensemble des triplets) et en espace (il faut stocker les frontières ; cf. ci-après) est en $O(n^3)$.

4.2.4.3 Structure de données associée aux frontières

Pour n souches, il y a au plus C_n^3 combinaisons sans répétitions de trois souches parmi n et comme chaque triplet peut avoir au plus deux frontières (impossible d'obtenir trois points frontières, au moins un des points est solution), il y a au plus

$$2 \times C_n^3 = \frac{n(n-1)(n-2)}{3}$$

frontières à conserver en mémoire. Avec $n = 200$, nous avons 7 880 400 frontières et en considérant le fait que nous devons au moins avoir quatre nombres réels associés à chaque frontière (un pour représenter la frontière et trois pour les coefficients temporaires), nous utilisons au plus environ 640

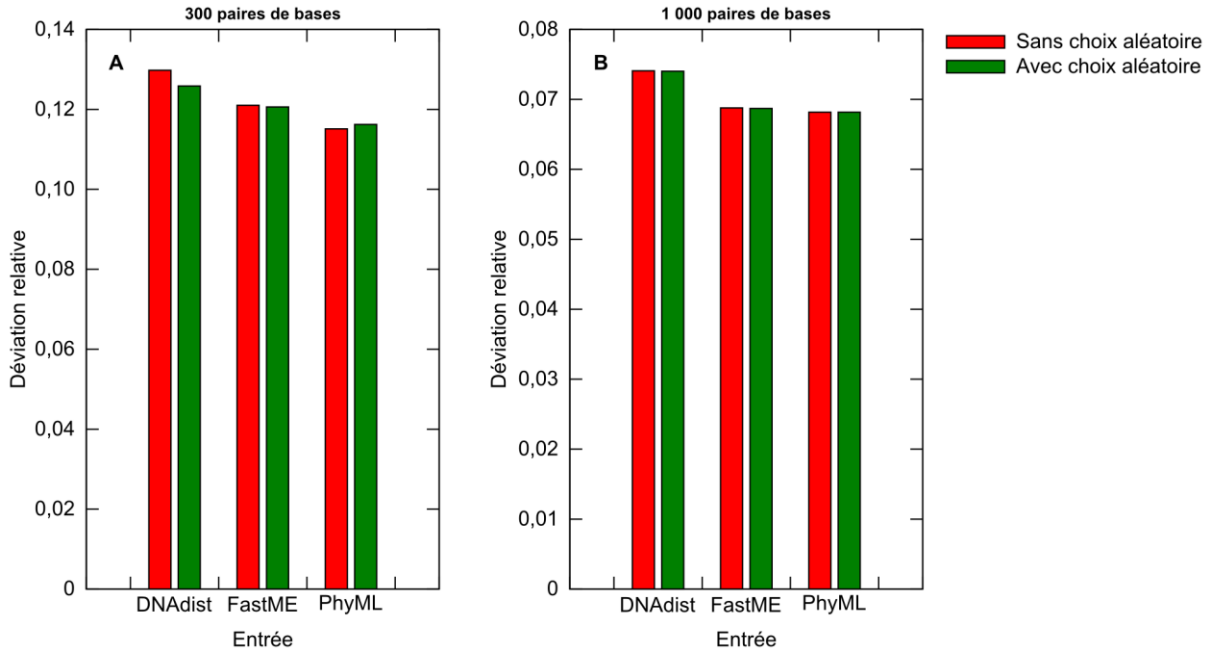
mégabits de mémoire (sachant qu'un nombre réel est codé sur 64 bits) et cela dépasse les gigabits pour $n = 300$. Bien entendu, cette estimation du nombre de frontières est beaucoup plus importante qu'en pratique. En effet, une même frontière peut appartenir à plusieurs triplets et les duplicata ne sont pas conservés. De plus, nous estimons beaucoup plus de combinaisons de triplets qu'il y en a en réalité, car considérer trois souches échantillonnées à une même date est inutile (il ne donne aucune information sur le taux de substitution puisque le critère résultant de ce triplet est constant ; en outre il n'a aucun point frontière), et tous les triplets n'ont pas forcément deux points frontières positifs. Malgré cela, le nombre de frontières reste tout de même important.

Pour stocker les frontières en mémoire, nous devons donc choisir une structure de données dynamique. Cette structure doit permettre de rechercher un élément (pour ne pas avoir deux fois le même) le plus rapidement possible et d'en insérer un nouveau (lorsqu'il n'est pas encore présent dans l'arbre) tout en conservant l'ordre établi sur la structure. La structure de données des arbres AVL (Adelson-Velskii & Landis, 1962) a l'avantage d'être rapide en termes de recherche et d'ajout d'un élément et conserve la propriété d'ordonnancement établie sur les éléments déjà lus (lors de l'ajout). Ce sont des arbres binaires de recherche dont la différence de hauteur entre le sous-arbre droit et le sous-arbre gauche d'un nœud n'excède pas un. L'ajout et la suppression d'un nœud de l'arbre nécessitent éventuellement une étape de rééquilibrage afin de conserver les propriétés spécifiques de cette structure. Le temps de calcul nécessaire pour accomplir ces deux tâches (ajout et recherche d'un élément) est en $O(\log k)$, où k est le nombre de nœuds existant dans l'arbre. Ensuite, un simple parcours infixe permet d'obtenir une liste triée des frontières. La complexité en temps est donc en $O(n^3 \log n)$, où n est le nombre de séquences.

La seule utilisation de cette structure de données n'est pas suffisante à l'obtention d'une bonne vitesse d'exécution, mais surtout elle n'influe pas sur la taille de la mémoire nécessaire au programme. En utilisant un algorithme simple d'épaississement des frontières (c.-à-d. que deux frontières f_i et f_j sont considérées comme identiques si $|f_i - f_j| < k$, avec k variable), le nombre de frontières retenues peut considérablement diminuer. Toutefois, le seuil d'acceptabilité de k dépend de la grandeur du taux de substitution à estimer et lorsqu'il est inconnu, une estimation au préalable de celui-ci est alors nécessaire. Nous préférons donc une autre alternative qui permet de gagner en plus du temps de calcul : le tirage aléatoire de triplets. Pour cela, nous fixons un seuil $s = 10^5$ au-delà duquel les triplets considérés sont obtenus par tirage aléatoire sur l'ensemble des triplets. Si le nombre « théorique » de triplets (obtenue par une fonction qui prend en considération le nombre de souches et le nombre de souches par date d'échantillonnage) est inférieur à ce seuil, alors tous les triplets sont utilisés. Dans le cas contraire seuls 10^5 triplets, choisis aléatoirement, sont considérés.

Non seulement ce principe permet de gagner considérablement de l'espace mémoire, mais il permet aussi d'avoir un gain, non négligeable, en temps de calcul puisque la méthode ne dépend plus du nombre de triplets (à partir d'un certain seuil), elle est bornée. La Figure 32 montre, sur un jeu de données de 200 simulations comprenant chacun 550 taxa, qu'il n'y a aucune différence (ou une différence négligeable) entre la précision d'estimation de la version considérant tous les triplets (sans choix aléatoire), soit environ 5×10^6 triplets, et celle utilisant le tirage aléatoire, donc 10^5 triplets, c'est-à-dire seulement 2% de la totalité des triplets possibles.

Figure 32. Performance en précision d'estimation avec ou sans choix aléatoire de triplets.
Ces graphiques indiquent en ordonnée la précision d'estimation (plus les valeurs sont petites, meilleures sont les estimations ; fonction déviation relative, cf. section 4.3.1.2) d'ULS sans ou avec choix aléatoire de triplets (en rouge et vert respectivement) pour différentes entrées (matrices de distances (DNAdist), arbres FastME et PhyML) et différentes longueurs d'alignement (300 et 1 000 sites). Les 200 simulations de ce jeu de données contiennent chacun 550 taxa (cf. section 4.3.1.1).



4.2.5 Description de l'algorithme

L'algorithme ULS permet d'estimer le taux de substitution relatif aux séquences hétérochrones dont la matrice de distances D est donnée en entrée (Figure 33). Le nombre n de séquences ainsi qu'un vecteur T contenant les intervalles de temps, exprimés en unité de temps, de chaque séquence entre leur date d'échantillonnage et la date d'échantillonnage la plus récente, sont aussi donnés en entrée. Cet algorithme renvoie un nombre qui correspond au taux de substitution ω, exprimé en substitutions par site et par unité de temps, à estimer.

Cet algorithme fonctionne en deux étapes. La première (lignes 2 à 11) parcourt l'ensemble des triplets et construit progressivement l'arbre AVL contenant les frontières et leurs coefficients temporaires (en $O(n^3 \log n)$). Il existe deux manières différentes de parcourir les triplets, soit en utilisant le

principe du tirage aléatoire (lignes 3 à 6), soit en parcourant la totalité des triplets (lignes 8 à 10). Le choix de considérer l'une ou l'autre manière est déterminé à la ligne 2. La deuxième étape (lignes 15 à 21) balaie l'ensemble ordonné des frontières pour rechercher le minimum global du critère (en $O(n^3)$). Lorsque tous les morceaux ont été parcourus et que le minimum global, correspondant alors à l'estimation du taux de substitution, est trouvé, l'algorithme le renvoie (ligne 22). La complexité algorithme est donc en $O(n^3 \log n)$, mais elle est bornée à partir d'un certain seuil de n.

Figure 33. Description de l'algorithme *Ultrametric Least Squares*.
Cet algorithme estime le taux de substitution ω à partir d'une matrice de distances D de taille n et de T un vecteur contenant pour chaque taxon i l'intervalle $T_i = t_0 - t_i$.

Entrée : D une matrice de distances, T un vecteur temps, n le nombre de taxa
1. $r \longleftarrow$ Créer un nœud AVL et l'initialiser avec la frontière $f_0 = 0$ et les coefficients temporaires $(a, b, c) = (0,0,0)$;
2. **si** $\mathrm{nombreTriplet}(n, T) \geq 10^5$ **alors**
3. **répéter**
4. Choisir un triplet au hasard ;
5. Ajouter ou rechercher dans r les frontières du triplet et actualiser leurs coefficients temporaires ;
6. **jusqu'à** 10^5 fois
7. **sinon**
8. **pour** chaque triplet **faire**
9. Ajouter ou rechercher dans r les frontières du triplet et actualiser leurs coefficients temporaires ;
10. **fin pour**
11. **fin si**
12. $t \longleftarrow$ Lister les nœuds de r à l'aide d'un parcours infixe ;
13. $(\omega, q) \longleftarrow (0, t[0].c)$;
14. $(a, b, c) \longleftarrow (0,0,0)$;
15. **pour** chaque élément de t **faire**
16. Mettre à jour les coefficients (a, b, c) ;
17. Calculer les coordonnées(ω_0, q_0) du minimum ;
18. **si** il est à considérer **et** $q_0 \leq q$ **alors**
19. $(\omega, q) \longleftarrow (\omega_0, q_0)$;
20. **fin si**
21. **fin pour**
22. **retourner** ω ;

4.2.6 Utilisation de la méthode dans le cas de taux variant par intervalle de temps

ULS peut facilement s'adapter à l'estimation de k taux de substitution $\omega_1, \cdots, \omega_k$ dans le cadre du modèle MRDT (Drummond *et al*, 2001), c'est-à-dire avec un taux de substitution par intervalle de temps entre deux dates d'échantillonnage consécutives. Soient deux souches i et j, échantillonnées

aux temps t_i et t_j, alors $\delta_{ij}(\omega_1, \cdots, \omega_k)$ (qui voit i et j comme contemporains) s'exprime par la relation

(9) $$\delta_{ij}(\omega_1, \cdots, \omega_k) = \hat{d}_{ij} + \sum_{m=1}^{i} \omega_m(t_m - t_{m-1}) + \sum_{m=1}^{j} \omega_m(t_m - t_{m-1}).$$

Il est impossible d'estimer simultanément ces k taux de substitution avec l'algorithme ULS. Pour estimer ces k taux de substitution, nous en supposons $k-1$ fixes et estimons le dernier taux, puis itérons le processus jusqu'à convergence. Dans ce cas, l'équation (9) est uniquement dépendante d'un seul taux de substitution et peut être exprimée comme à l'équation (1). L'algorithme ULS peut alors être utilisé pour estimer le taux de substitution non fixé.

Ce procédé nécessite d'abord d'initialiser les k taux de substitution, par exemple avec la valeur obtenue par ULS dans le cadre du modèle SRDT (où un seul taux de substitution est supposé). Ensuite, les taux de substitution sont fixés, hormis le $u^{\text{ème}}$, et la matrice de distances et le vecteur temps sont modifiés en conséquence, c'est-à-dire que pour chaque distance $\hat{d}_{ij}$ la mesure

$$\sum_{m=1, m\neq u}^{i} \omega_m(t_m - t_{m-1}) + \sum_{m=1, m\neq u}^{j} \omega_m(t_m - t_{m-1})$$

y est ajoutée et les intervalles de temps sont modifiés en retranchant les mesures

$$\sum_{m=1, m\neq u}^{i} t_m - t_{m-1}$$

et

$$\sum_{m=1, m\neq u}^{j} t_m - t_{m-1}$$

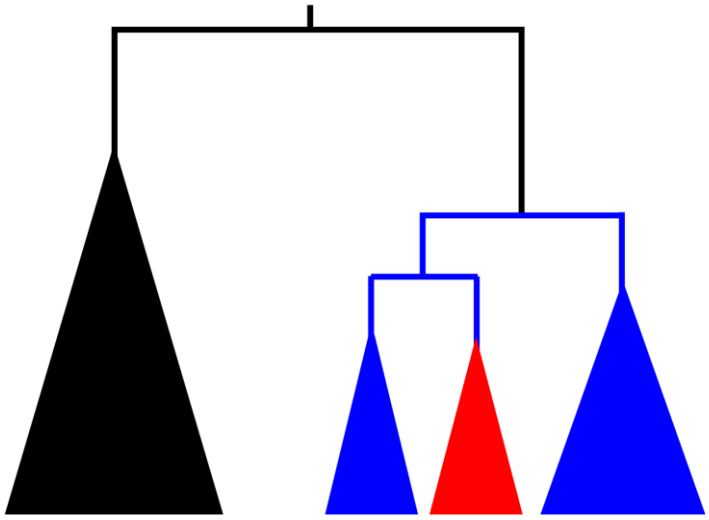
à T_i et à T_j respectivement. Dès lors, nous sommes dans le contexte du modèle SRDT et pouvons estimer le $u^{\text{ème}}$ taux de substitution. Cette procédure est ensuite itérée pour le $u+1^{\text{ème}}$ taux de substitution et, ainsi de suite, jusqu'au $k^{\text{ème}}$. Lorsque les k taux de substitution sont estimés et si au moins un est modifié significativement, alors la procédure est de nouveau itérée, mais en conservant cette fois-ci les dernières estimations des k taux de substitution comme valeurs initiales, et cela jusqu'à stabilisation des valeurs.

Ce procédé est facilement adaptable au modèle MRDTa, c'est-à-dire le modèle où l'utilisateur choisit lui-même les intervalles de temps pour lesquels il souhaite connaître un taux de substitution. Ces intervalles ne sont pas forcément en adéquation avec les dates de collecte, mais compris entre la date d'échantillonnage la plus ancienne et la plus récente. Toutefois, il ne peut y avoir plus d'intervalles de temps qu'en a le modèle MRDT sur le même jeu de donnée. En effet, il faut au moins avoir une souche par intervalle de temps pour que le taux de substitution correspondant puisse être estimé. Donc, le nombre maximum d'intervalles de temps est donné par le nombre de dates d'échantillonnage différentes moins une (une date est utilisée comme référence) (Drummond *et al*, 2001). Dans le cas contraire, il y a forcément un intervalle de temps qui ne contient pas de souches et le taux de substitution correspondant ne peut être estimé.

4.2.7 Utilisation de la méthode dans le cas de taux variant par lignage

Le modèle par lignage (horloges moléculaires locales) permet d'estimer un taux de substitution par lignage (ou sous-arbre). La donnée comprend un arbre (non nécessairement enraciné) et les sous-arbres où le taux de substitution est différent du taux global affecté au reste de l'arbre (Figure 34). Ce modèle peut, par exemple, être utilisé lorsque l'on considère plusieurs sous-types du VIH-1 dans une même phylogénie, on peut alors affecter à chaque sous-type un taux de substitution différent.

Figure 34. Schéma représentant le modèle par lignage.
Ce modèle prend en considération le fait qu'un ou des sous-arbres peuvent évoluer à des vitesses différentes de celle de la phylogénie globale, c'est-à-dire en considérant des horloges moléculaires locales. Dans cet exemple, il y a trois taux de substitution, un global (en noir) et deux locaux (en bleu et rouge).



Soit un arbre binaire A. L'ensemble de ses nœuds est noté V et l'ensemble de ses feuilles $F \subset V$. On pose $|F| = n$. Soit maintenant $\hat{d}: V \times V \to \mathbb{R}_+$ une mesure de distance telle que, pour tout i, j de V, $\hat{d}(i,j) = \hat{d}_{ij}$ renvoie la longueur du chemin reliant i avec j et soit P_{ij} l'ensemble contenant la

suite de nœuds consécutifs de ce chemin. Le premier nœud dans la suite P_{ij} est i et le dernier j. Supposons maintenant que chaque arête est associée à un taux de substitution. Le taux de substitution correspondant à l'arête ayant u et v comme sommets est noté ω_{uv} et $m = 2n - 3$ correspond au nombre d'arêtes.

La distance corrigée δ_{ij}, exprimée ici en unité de temps, entre deux feuilles i et j vues comme contemporaines, vaut

$$\delta_{ij} = T_i + T_j + \sum_{k \in P_{ij}} \frac{\hat{d}_{k(k+1)}}{\omega_{k(k+1)}}.$$

Procédons maintenant d'une manière analogue à celle du modèle MRDT. Soit c le nombre de taux de substitution à estimer. Au préalable, on initialise les c taux de substitution avec la valeur retournée par l'algorithme ULS, en considérant le modèle SRDT. Puis, fixons $c - 1$ taux de substitution et imagions que l'on souhaite estimer le taux de substitution ω_c. Soit maintenant, $V_c = \{(u, v) \in V^2 | \omega_{uv} = \omega_c\}$ l'ensemble d'arêtes correspondant au taux de substitution ω_c à estimer. Ainsi,

$$\delta'_{ij} = \omega_c \delta_{ij} = \omega_c \lambda_{ij} + \mu_{ij}$$

avec

$$\lambda_{ij} = T_i + T_j + \sum_{k \in P_{ij}, (k,(k+1)) \notin V_c} \frac{\hat{d}_{k(k+1)}}{\omega_{k(k+1)}}$$

et

$$\mu_{ij} = \sum_{k \in P_{ij}, (k,(k+1)) \in V_c} \hat{d}_{k(k+1)}.$$

Les termes λ_{ij} et μ_{ij} sont indépendants de ω_c. Ainsi, δ'_{ij} a la même forme que l'équation (1) et nous pouvons y appliquer le critère Q afin d'estimer le taux de substitution ω_c. L'algorithme ULS peut donc être appliqué, et on itère ce processus jusqu'à convergence des c taux de substitution.

4.2.8 Mise en œuvre

L'algorithme ULS est implémenté en langage C et présente une interface similaire à celle des logiciels de la suite PHYLIP (Felsenstein, 1993). Ce logiciel permet d'estimer le taux de substitution à partir d'une matrice de distances ou d'un arbre (raciné ou non) dont on extrait les distances patristiques. Il permet aussi d'estimer la date de l'ancêtre commun aux taxa (à l'aide d'un arbre UPGMA calculé d'après la matrice de distances corrigées) et propose l'enracinement d'un arbre avec la méthode de

minimisation de la variance spécifiquement adaptée aux arbres avec feuilles hétérochrones. Les adaptations d'ULS aux modèles MRDT, MRDTa et lignage sont aussi disponibles mais nécessiteraient d'être testées avec soin.

4.3 Confrontation aux autres méthodes de distances et à celle de référence (BEAST)

Nous avons confronté, sur données simulées, ULS aux autres méthodes de distances qui permettent d'estimer le taux de substitution sous les hypothèses du modèle SRDT (horloge moléculaire stricte et feuilles hétérochrones), à savoir *Pairwise-Distance*, *Root-to-Tip*, sUPGMA et TREBLE, ainsi qu'à la méthode probabiliste de référence BEAST (cf. Chapitre 2). Dans un second temps, ULS est appliquée sur deux jeux de données du sous-type C du virus de l'immunodéficience humaine de type 1 (VIH-1C).

4.3.1 Confrontation sur jeux de données simulées

À notre connaissance, il n'existe qu'un seul générateur d'arbres qui émette les hypothèses du modèle SRDT : *Serial SimCoal* (Anderson *et al*, 2005). Cependant, il génère des arbres sous le modèle du coalescent (Kingman, 1982), basé sur la génétique des populations, et est une approche différente au modèle phylogénétique classique qui est celui supposé par ULS. Nous avons donc généré nos propres jeux de données simulées, en adaptant le modèle de Yule (Yule, 1925), étendu par Raup et al. (1973) en y incluant un taux de mort constant, aux hypothèses du modèle SRDT.

4.3.1.1 Construction des jeux de données simulées

Le processus stochastique utilisé pour générer les jeux de données simulées démarre d'un individu. Puis, chaque individu vivant a autant de chance de donner naissance à un nouvel individu, jusqu'à en obtenir n. Dès que les n individus sont obtenus, $m < n$ individus sont choisis aléatoirement et disparaissent du processus (ils sont morts). Parmi ces individus morts, $N < m$ sont sélectionnés aléatoirement et sont considérés comme échantillonnés au temps t_{k-1}. Puis ce processus est recommencé avec les $n - m$ individus vivants jusqu'à en obtenir de nouveau n. Et ceci encore $k - 1$ fois, jusqu'au temps d'échantillonnage t_0. Ainsi, le sous-arbre contenant toutes les feuilles échantillonnées correspond à l'arbre souhaité. Ce processus est à l'opposé de celui du coalescent où l'on démarre avec n individus pour ne terminer qu'avec un seul individu (Steel & McKenzie, 2001), aussi appelé modèle de Hey (Hey, 1992). L'algorithme *GenTree* permet de générer un arbre avec le principe décrit ci-dessus et de telle sorte qu'il se passe a années entre deux dates d'échantillonnage (Figure 35).

Classiquement, sous le modèle de Yule, la probabilité qu'un évènement de spéciation survienne dans une lignée à l'instant t suit une loi exponentielle de paramètre λ, où λ représente le nombre moyen d'évènements de spéciation qui se produisent dans une lignée par unité de temps (Mooers *et al*, 2007), et de moyenne $1/\lambda$. Si k lignées sont présentes à un moment donné, alors l'instant t jusqu'au prochain évènement de spéciation suit aussi une loi exponentielle mais d'espérance $1/(k\lambda)$ (Steel & Mooers, 2009). Afin de simplifier cette hypothèse, nous supposons que l'espérance de la loi exponentielle fournit le temps jusqu'au prochain évènement de spéciation, soit $1/k$ lorsque k lignées sont présentes, en posant $\lambda = 1$. Ceci permet d'avoir des arbres dont les intervalles de temps sont identiques et non stochastiques.

Figure 35. Description de l'algorithme *GenTree*.
Cet algorithme génère un arbre sous les hypothèses du modèle SRDT.

Entrée : m le nombre d'individus morts, n le nombre d'individus à chaque temps d'échantillonnage, N le nombre d'individus échantillonnés à chaque temps d'échantillonnage, k le nombre de temps d'échantillonnage, a le nombre d'années entre deux dates d'échantillonnage et ω la valeur du taux de substitution souhaité.

Créer une feuille et la stocker dans un tableau T ;
$x \longleftarrow 1$;
$C \longleftarrow \sum_{i=n-m+1}^{n} \frac{1}{i}$;
répéter
    répéter
        Choisir aléatoirement une feuille f dans T ;
        Créer deux nouvelles feuilles qui sont les fils gauche (g) et droit (d) de f ;
        Supprimer f du tableau et y ajouter g et d ;
        $x \longleftarrow x + 1$;
        pour toutes les longueurs de branche l des feuilles de T **faire**
            $l \longleftarrow l + a\omega/(Cx)$;
        fin pour
    jusqu'à $x = n$
    Choisir aléatoirement N feuilles que l'on marque et que l'on supprime du tableau T ;
    Choisir aléatoirement $m - N$ feuilles que l'on supprime du tableau ;
    $x \longleftarrow x - m$;
jusqu'à k fois
pour chaque feuille i marquée **faire**
    Affecter le temps d'échantillonnage d_i/ω, où d_i est la distance séparant la feuille i de la racine ;
fin pour
$R \longleftarrow$ Extraire le sous-arbre contenant toutes les feuilles marquées ;
retourner R ;

Nous voulions que les paramètres utilisés pour générer les jeux de données reflètent au mieux la topologie des phylogénies intra- et inter-hôtes du VIH. Pour cela, nous avons respectivement utilisé un taux de mort à 995 et 750 sur 1 000 taxa à chaque temps d'échantillonnage ($m = \{995,750\}$,

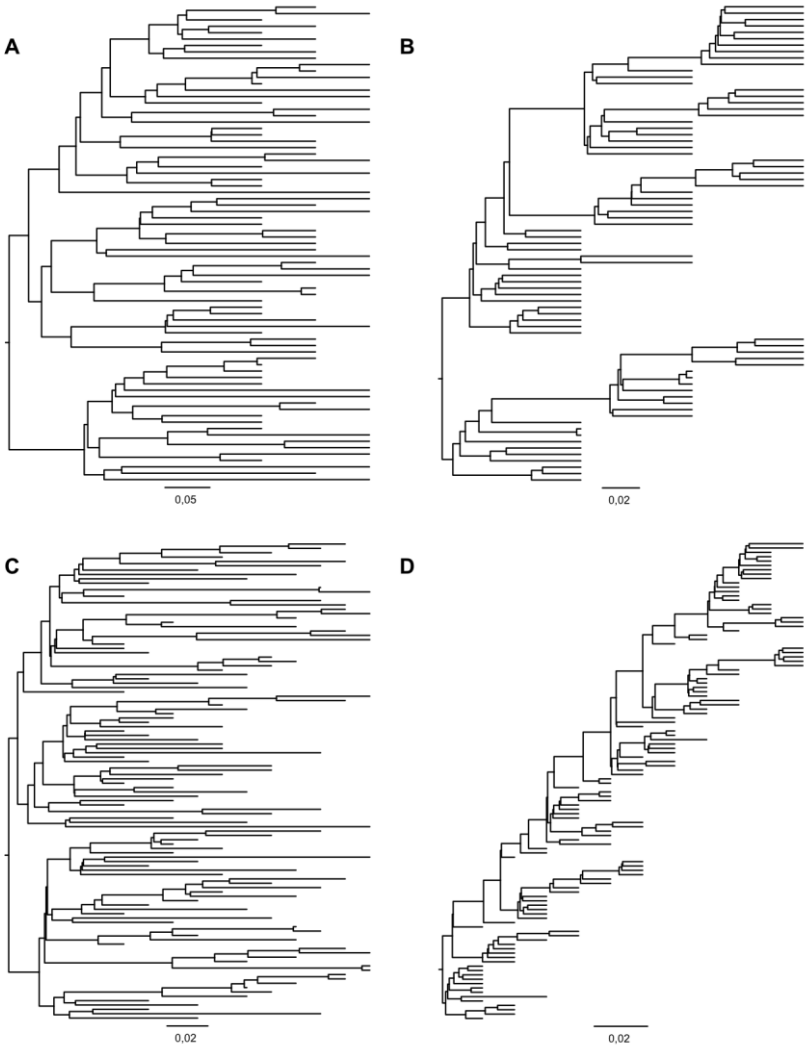
$n = 1000$). Pour chaque taux de mort, quatre jeux de données, contenant chacun 100 arbres, sont générés. Les deux premiers contiennent 3 temps d'échantillonnage chacun séparé de 10 ans ($k = 3$ et $a = 10$), avec respectivement 25 et 100 feuilles collectées à chaque date ($N = \{25,100\}$). Les deux derniers contiennent 11 temps d'échantillonnage chacun séparé de 2 ans ($k = 11$ et $a = 2$), avec respectivement 10 et 50 feuilles collectées à chaque date ($N = \{10,50\}$). Ces quatre jeux de données veulent représenter le suivi de l'infection au VIH pour un individu ($m = 995$) ou une population ($m = 750$) sur 20 ans, avec un échantillonnage de la population virale tous les 10 ans ou tous les 2 ans. Le taux de substitution attribué à chaque jeu de données est de 6×10^{-3} substitutions par site et par année. Il correspond approximativement à celui obtenu par Bello *et al*. (2008) sur la région *env* du génome ($5{,}8 \times 10^{-3}$ dans leur étude pour une estimation avec une horloge moléculaire stricte). Ainsi, nous avons huit collections d'une centaine d'arbres générés sous le modèle SRDT et reflétant l'évolution intra- et inter-hôte du VIH, avec un taux de substitution de 6×10^{-3} substitutions par site et par année. La Figure 36 montre quatre exemples de topologies extraites de ces jeux de données simulées.

Ces arbres servent ensuite de guide à SeqGen (Rambaut & Grassly, 1997) pour générer les alignements correspondants (un de 1 000 sites et l'autre de 300 sites) sous le modèle d'évolution F84 (Felsenstein, 1984; Kishino & Hasegawa, 1989), similaire au modèle d'évolution HKY85 (Hasegawa *et al*, 1985), avec une loi gamma de paramètre 1 à 8 catégories de taux et un taux de transition/transversion (Ts/Tv) à 2,5. Les fréquences des nucléotides sont respectivement de 0,35, 0,2, 0,2 et 0,25 pour les bases A, C, G et T. Ces paramètres correspondent approximativement à ceux obtenus sur *env* pour des virus appartenant au groupe M du VIH-1 (Posada & Crandall, 2001), groupe responsable de la pandémie mondiale. Ces 16 jeux de données (8 de longueurs 300 sites et 8 de longueurs 1 000 sites) servent de base pour nos analyses comparatives. Comme les formats d'entrée des différentes méthodes varient (alignements, matrices de distances, arbres), nous avons généré les matrices de distances correspondantes avec DNAdist v3.69 du package PHYLIP (Felsenstein, 1989). Le modèle d'évolution utilisé et les paramètres sont choisis en concordance avec ceux utilisés pour générer les alignements. À partir des matrices de distances, des arbres FastME v2.07 (Desper & Gascuel, 2002) sont calculés en utilisant l'option SPR pour parcourir l'espace des arbres. Thu Hien TO a généré les arbres calculés avec PhyML v3.0 (Guindon *et al*, 2010; Guindon & Gascuel, 2003) sous le modèle F84 et en laissant PhyML optimiser les paramètres. À nouveau l'option SPR est choisie pour parcourir l'espace des arbres. Les arbres ainsi obtenus sont aussi convertis en matrices de distances patristiques pour les méthodes prenant celles-ci en entrée.

Les méthodes de distances utilisées pour ces tests sont de notre propre implémentation. Toutefois, quand cela a été possible, nous avons vérifié les résultats obtenus par nos implémentations

contre celles des auteurs sur quelques jeux de données. L'algorithme *Root-to-tip* utilisé ici est celui implémenté dans la version 1.3 de Path-O-Gen, c'est-à-dire celui de la minimisation des résidus (cf. Chapitre 2). L'implémentation de TREBLE est une adaptation en C de celle disponible à l'adresse http://jacobian.wikidot.com/software (programmée pour R). Cette version correspond à celle publiée mais où le paramètre α est mis à 0 (selon l'auteur, lors d'une conversation par courriel ; cf. Chapitre 2). La valeur de pondération d'ULS correspond à celle de l'équation (6) pour $k = 1$. Pour BEAST v1.6.2, le modèle d'évolution choisi est HKY85 et le modèle démographique est *Constant Size* avec une horloge moléculaire stricte. La prior pour le paramètre *clock.rate* est une loi uniforme entre 0 et 1. La longueur de la chaîne de Markov par technique de Monte Carlo (MCMC) est de 5×10^6 générations avec un échantillonnage toutes les 5×10^3 générations.

Figure 36. Exemples de topologies d'arbre simulé.
Quatre exemples de topologies d'arbre extrait de nos jeux de données. Les arbres A et C (topologie approximant une phylogénie du VIH inter-hôte) proviennent des jeux de données ayant 750 morts (sur 1 000 taxa) par date d'échantillonnage et les arbres B et D (topologie approximant une phylogénie du VIH intra-hôte) des jeux de données ayant 995 morts par date d'échantillonnage. Les arbres A et B ont chacun 3 temps d'échantillonnage avec 25 feuilles par date d'échantillonnage et les arbres C et D ont chacun 11 dates d'échantillonnage avec 10 feuilles par date d'échantillonnage.



4.3.1.2 Performance en précision d'estimation

Connaissant, pour chaque jeu de données, le taux de substitution théorique ω (6×10^{-3} substitutions par site et par année), il est alors facile de mesurer la performance des méthodes en comparant les taux estimés $\Omega = \{\widehat{\omega}_k, k = 1, \cdots, 100\}$ à ω. Ainsi, nous définissons les fonctions déviation relative

(10) $$D(\omega, \Omega) = \frac{1}{\omega}\sqrt{\frac{1}{100}\sum_{k=1}^{100}(\widehat{\omega}_k - \omega)^2}$$

et biais relatif

(11) $$B(\omega, \Omega) = \frac{1}{\omega}\left(\frac{1}{100}\left(\sum_{k=1}^{100}\widehat{\omega}_k\right) - \omega\right).$$

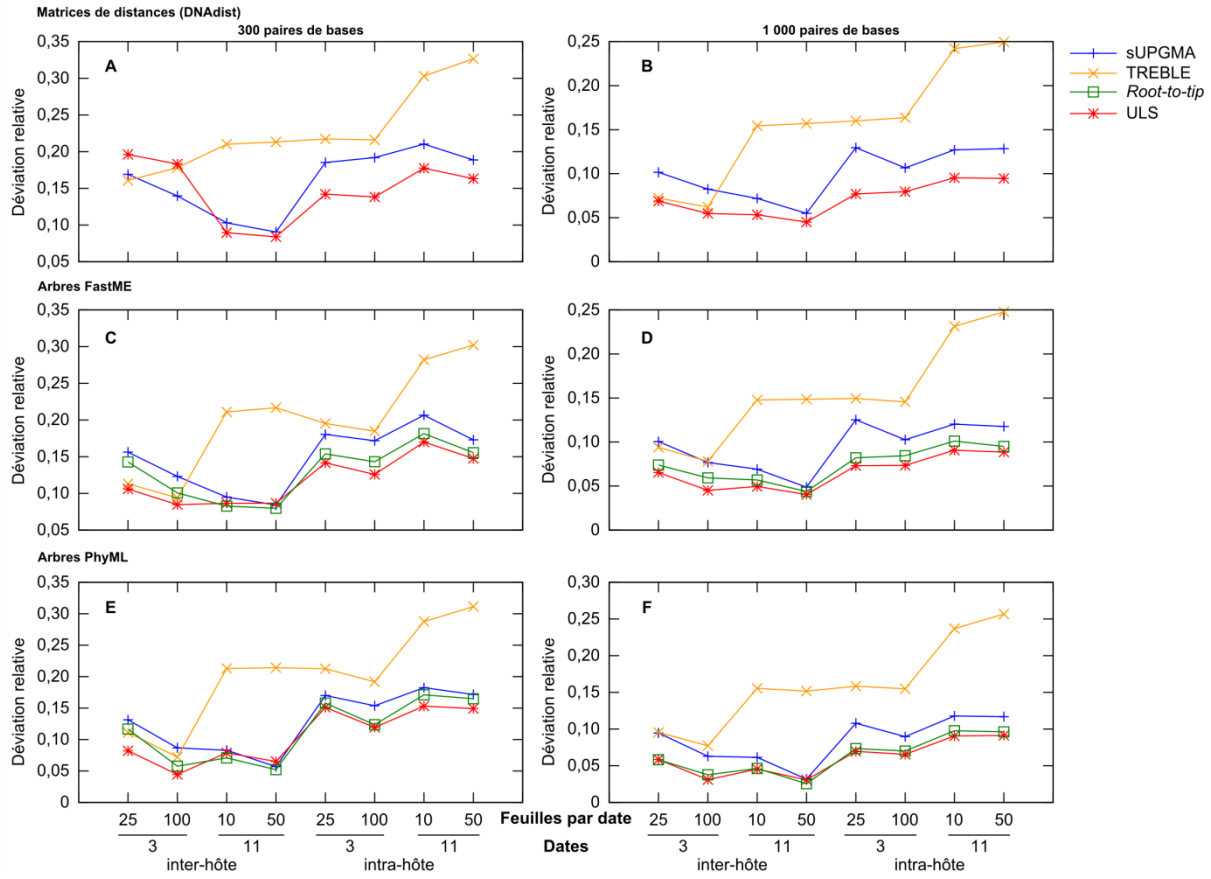
La fonction (10) mesure la performance moyenne des méthodes. Plus les valeurs de cette fonction sont petites, plus les estimations des taux de substitution sont proches de la valeur théorique. Ainsi, la méthode parfaite, celle qui estime à chaque fois le bon taux de substitution, a zéro comme valeur de déviation relative. La fonction (11) mesure une autre information, la tendance moyenne à sur- ou sous-estimer le taux de substitution réel. Si sa valeur est négative alors les estimations du taux de substitution sont majoritairement sous-estimées, tandis que si elle est positive, les estimations sont majoritairement surestimées.

La Figure 37 montre pour chaque jeu de données et pour chaque entrée possible (matrices de distances (DNAdist), arbres FastME et PhyML) la performance en précision d'estimation (déviation relative) des méthodes de distances testées (sUPGMA en bleu, TREBLE en orange, *Root-to-tip* en vert et ULS en rouge). Les performances de la méthode *Pairwise-Distance* sont aussi mesurées, mais elles sont trop faibles pour être représentées. En effet, pour les jeux de données avec 750 morts (inter-hôte) les valeurs de déviation relative sont toujours supérieures à 0,9 et pour les jeux de données avec 995 morts (intra-hôte) elles sont toujours supérieures à 0,24.

Comme attendu, les performances en précision d'estimation augmentent lorsque les séquences contiennent plus d'information, c'est-à-dire plus de nucléotides. Pour les jeux de données intra-hôtes, la méthode ULS est la plus performante, parfois égalée par la régression linéaire *Root-to-tip*, et cela quelle que soit la longueur des séquences ou le format d'entrée (arbres ou matrices de distances). La question est autre sur les jeux de données inter-hôte. En considérant 11 temps d'échantillonnage, la performance en précision d'estimation d'ULS semble à chaque fois égaler la meilleure des autres méthodes, hormis sur le graphique B avec 10 feuilles par temps

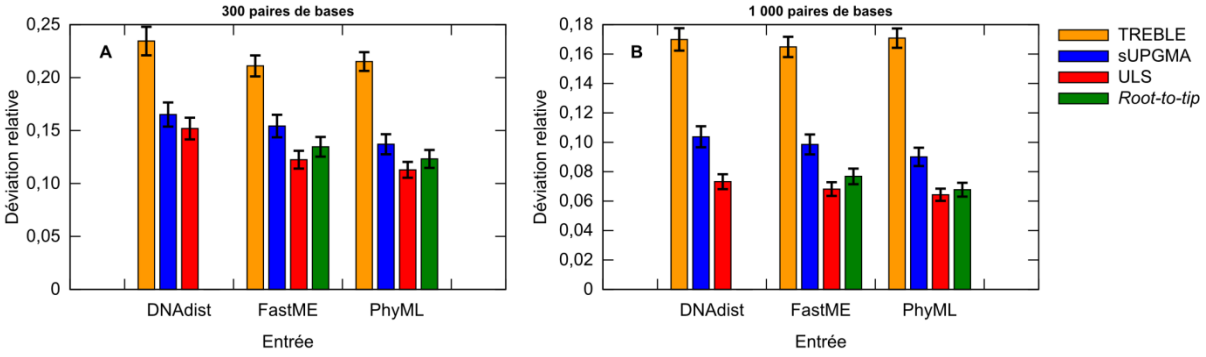
d'échantillonnage où elle est la plus performante. En revanche, avec les jeux de données contenant 3 dates d'échantillonnage, la performance d'ULS est souvent égalée, voire dépassée par d'autres méthodes, comme sur le graphique A. Or, dans ce dernier cas, la distance paire à paire moyenne avoisine les 0,6 substitutions par site, tandis que sur les autres jeux de données elle avoisine les 0,2 substitutions par site, sauf sur les jeux de données intra-hôte avec 11 temps d'échantillonnage où elle avoisine les 0,1 substitutions par site. Cela suggère que la méthode ULS reste dépendante de la précision d'estimation des distances (elles deviennent meilleures au fur et à mesure que la longueur des séquences augmentes où que les distances sont petites). D'ailleurs, les précisions d'estimation de la méthode TREBLE semblent aussi fluctuer en fonction de cette observation (elles deviennent de plus en plus précises au fur et à mesure que la distance paire à paire moyenne augmente, à l'inverse de ce que l'on attend). Remarquons que ce défaut est corrigé avec les arbres FastME et semble inexistant avec les arbres PhyML.

Figure 37. Performance en précision d'estimation des différentes méthodes de distances (fonction déviation relative). Graphiques montrant les valeurs de la fonction déviation relative en ordonnée pour chaque méthode de distances (sUPGMA en bleu, TREBLE en orange, *Root-to-tip* en vert et ULS en rouge) et pour chacun des 8 jeux de données simulées. Les graphiques A, C et E sont obtenus sur l'alignement de 300 sites et les graphiques B, D et F sur celui de 1 000 sites. Les résultats des graphiques A et B sont obtenus à partir de matrices de distances (DNAdist), les graphiques C et D à partir des arbres FastME et les graphiques E et F à partir des arbres PhyML. La méthode *Pairwise-Distance* n'est pas reportée dans les graphiques à cause de valeurs très différentes. En effet, quelle que soit l'entrée, les valeurs de la déviation relative en inter-hôte (750 morts) sont toujours supérieures à 0,9 et en intra-hôte (995 morts) toujours supérieures à 0,24.



Afin de déterminer quelle méthode est, dans l'absolu, la plus performante, leur précision d'estimation (déviation relative) est calculée sur les 800 simulations pour chaque longueur de séquences (300 et 1 000 paires de bases) et pour chaque entrée (matrices de distances (DNAdist), arbres FastME et PhyML) (Figure 38). Les intervalles de confiance à 95% sont précisés au sommet de chaque barre. Contre sUPGMA et TREBLE, la méthode ULS est toujours la plus performante, sauf pour des séquences de 300 paires de bases et avec des matrices de distances où l'intervalle de confiance recouvre celui de sUPGMA. Dans ce cas, la perte en précision d'estimation provient des deux jeux de données inter-hôtes avec 3 dates d'échantillonnage (cf. paragraphe précédent). Autrement, la méthode ULS est toujours, en moyenne, plus performante que la méthode *Root-to-tip*, mais nous ne pouvons affirmer que la précision d'estimation d'ULS est significativement meilleure que celle de *Root-to-tip*, étant donné que les intervalles de confiances à 95% sont systématiquement recouvrants. En revanche, un test du signe qui prend en considération le fait que les échantillons comparés proviennent de la même population (ce qui n'est pas pris en compte en comparant les intervalles de confiance), indique qu'ULS est significativement meilleure que la régression linéaire *Root-to-tip* avec les arbres FastME ($p < 0{,}01$; 445 contre 355 [resp. 485 contre 315] avec des séquences de 300 [1 000] paires de bases), mais rien ne peut être affirmé avec les arbres PhyML ($p = 0{,}15$ [421 contre 379] et $p = 0{,}07$ [426 contre 374] pour 300 et 1 000 paires de bases respectivement). Sur ce graphique, nous observons aussi que la précision d'estimation des méthodes (hormis TREBLE) est, en moyenne, plus performante avec des arbres PhyML qu'avec des arbres FastME, bien que généralement les intervalles de confiance soient recouvrants.

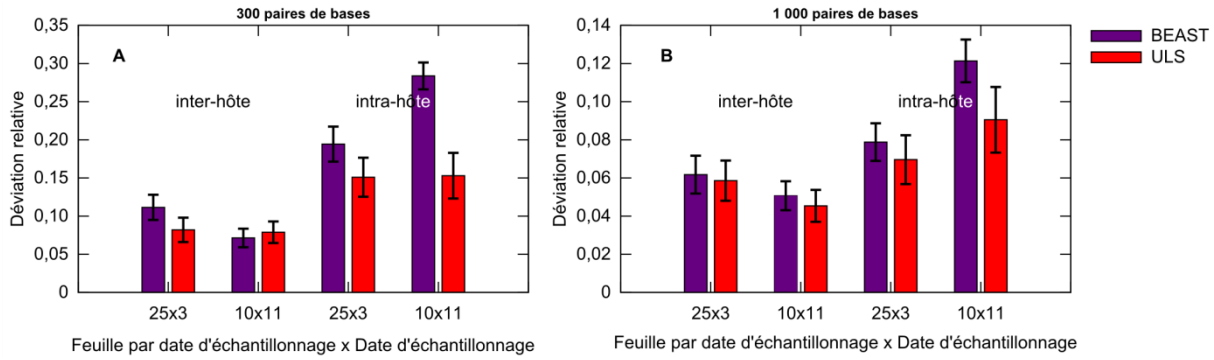
Figure 38. Performance en précision d'estimation (déviation relative) pour toutes simulations confondues.
Ces graphiques représentent la précision d'estimation (déviation relative) des différentes méthodes testées (de gauche à droite : TREBLE en orange, sUPGMA en bleu, ULS en rouge et *Root-to-tip* en vert) sur l'ensemble des 800 simulations et cela pour chaque entrée (matrices de distances (DNAdist), arbres FastME et PhyML) et chaque longueur d'alignement (300 sites pour le graphique A et 1 000 sites pour le graphique B). Les intervalles de confiance à 95% sont indiqués au sommet de chaque barre. Les performances de la méthode *Pairwise-Distance* sont toujours supérieures à 0,70 et ne sont donc pas représentées.



Les précisions d'estimation de la méthode de distances ULS (en rouge) sont comparées avec celles de la méthode probabiliste de référence BEAST (en mauve) (Figure 39). Ces comparaisons sont seulement faites sur les petits jeux de données inter- et intra-hôtes (3 et 11 dates d'échantillonnage avec

respectivement 25 et 10 feuilles par date d'échantillonnage), en raison du temps de calcul prohibitif nécessaire à BEAST sur les grands jeux de données. Les précisions d'estimation d'ULS sont toujours, en moyenne, plus performantes que celles de BEAST (en particulier sur les jeux de données intra-hôtes), exception faite sur un jeu de donnée (graphique A, 10x11). Mais dans ce dernier cas, et dans d'autres (surtout en inter-hôte), la différence de précision n'est pas significative puisque les intervalles de confiance sont recouvrants. Donc ULS est au pire équivalent à BEAST. La perte en précision d'estimation de BEAST sur les jeux de données intra-hôte provient sans soute du fait que BEAST est basé sur le modèle du coalescent, en opposition avec ces jeux de données qui sont générés avec un modèle de spéciation, or il est impossible de choisir un tel modèle avec la version de BEAST utilisée.

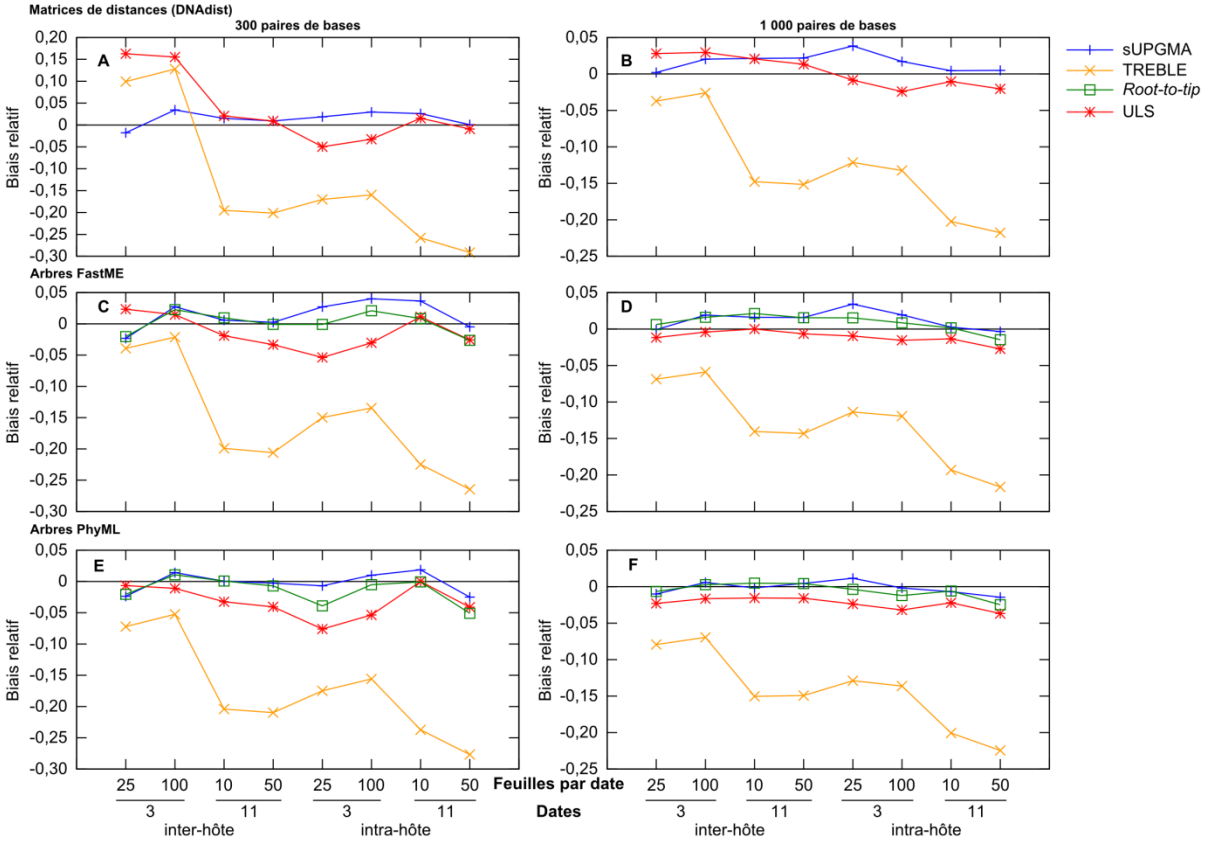
Figure 39. Comparaison de la précision d'estimation entre BEAST et ULS.
Ces graphiques montrent les précisions d'estimation en ordonnée de la méthode de distances ULS (sur arbres PhyML ; en rouge) et celles de la méthode probabiliste de référence BEAST (en mauve). Seuls les petits jeux de données inter-hôte (750 morts) et intra-hôte (995 morts) sont utilisés, à savoir ceux avec 25 feuilles par date d'échantillonnage et 3 dates (25x3) et ceux avec 10 feuilles par date d'échantillonnage et 11 dates (10x11), avec des séquences de 300 (graphique A) et 1 000 (graphique B) paires de bases. Les intervalles de confiance à 95% sont indiqués au sommet de chaque barre.



Les tendances des méthodes de distances (sUPGMA en bleu, TREBLE en orange, *Root-to-tip* en vert et ULS en rouge) à sur- (biais relatif positif) ou sous-estimer (biais relatif négatif) le taux de substitution sur les différents jeux de données sont présentées à la Figure 40. Le biais de la méthode *Pairwise-Distance* n'est pas présenté puisqu'il est toujours inférieur à –0,9 sur les jeux de données inter-hôte (750 morts) et toujours inférieur à –0,17 sur les jeux de données intra-hôte (995 morts). Comme on s'y attend, le biais est moins important lorsque les séquences ont 1 000 paires de bases. La méthode TREBLE a tendance à sous-estimer le taux de substitution (excepté pour les deux premiers cas du graphique A). Cela provient du fait qu'elle suppose initialement que le taux de substitution est zéro. Une correction à ce problème est apportée par les auteurs par l'instauration d'un critère qui rejette successivement les *outgroups* invalides à l'aide d'un processus itératif, mais ce critère n'est pas mis en œuvre dans la dernière version (pour R) proposée par les auteurs. En ce qui concerne les autres méthodes, il n'y a pas de tendance particulière qui ressort (tantôt positif, tantôt négatif). Remarquons, tout de même, qu'ULS semble sous-estimer le taux de substitution lorsque cette méthode utilise des arbres en entrée et à le surestimer avec des matrices de distances. De plus, le

biais d'ULS est plus important que celui des méthodes sUPGMA et *Root-to-tip* avec des arbres PhyML. Ces dernières méthodes, quant à elles, ont tendance à surestimer le taux de substitution sur les arbres FastME, et sUPGMA semble surestimer le taux de substitution avec des matrices de distances.

Figure 40. Biais dans les estimations des différentes méthodes de distances (fonction biais relatif).
Graphiques montrant les valeurs de la fonction biais relatif en ordonnée pour chaque méthode de distances (sUPGMA en bleu, TREBLE en orange, *Root-to-tip* en vert et ULS en rouge) et pour chacun des 8 jeux de données simulées. Les graphiques A, C et E sont obtenus de l'alignement de 300 sites et les graphiques B, D et F de celui de 1 000 sites. Les résultats des graphiques A et B sont obtenus à partir des matrices de distances (DNAdist), les graphiques C et D à partir des arbres FastME et les graphiques E et F à partir des arbres PhyML. La méthode *Pairwise-Distance* n'est pas reportée dans les graphiques à cause de valeurs très différentes. En effet, quelle que soit l'entrée, les valeurs du biais relatif en inter-hôte (750 morts) sont toujours inférieures à –0,9 et en intra-hôte (995 morts) toujours inférieures à –0,17.



En résumé, ces résultats, sur données simulées, suggèrent qu'ULS est plus précise que les méthodes de distances *Pairwise-Distance*, sUPGMA et TREBLE. Elle est aussi plus précise que la régression linéaire *Root-to-tip* avec en entrée des arbres FastME ou sur des jeux de données inter-hôtes, tandis qu'elle est équivalente à cette dernière sur des jeux de données intra-hôtes ou sur des arbres PhyML. Elle est aussi plus précise que la méthode probabiliste BEAST sur des jeux de données intra-hôtes et au pire équivalente à cette dernière sur des jeux de données inter-hôtes.

4.3.1.3 Performance en temps de calcul

Après avoir présenté la performance en précision d'estimation des différentes méthodes d'estimation de taux de substitution, nous montrons dans le Tableau 3 la performance de ces mé-

thodes (*Pairwise-Distance*, sUPGMA, TREBLE, *Root-to-tip*, ULS avec ou sans choix aléatoire et BEAST) en temps de calcul sur l'ensemble des jeux de données simulées (soit 1 600 jeux de données). À titre indicatif, nous présentons aussi les temps de calcul nécessaire aux outils d'inférence phylogénétique (DNAdist, FastME et PhyML) dont dépendent les méthodes de distances. Les temps de calcul sont donnés en minutes et lorsqu'ils dépassent le jour de calcul, ils sont donnés approximativement en jours. Notons que le temps de calcul de BEAST est seulement donné pour une partie des jeux de données simulées (ceux dont le nombre de feuilles dans l'arbre est inférieur ou égal à 110 alors que certains jeux de données vont jusqu'à 550 feuilles).

Tableau 3. Performance en temps de calcul des différentes méthodes d'estimation de taux de substitution.
Ce tableau présente les temps de calcul (en minutes) nécessaires à chaque méthode d'estimation de taux de substitution pour estimer les dits taux sur les 1 600 jeux de données simulées et pour les différentes entrées possibles (alignements, matrices de distances ou arbres). À titre d'information, nous indiquons aussi le temps de calcul de chaque méthode d'inférence phylogénétique utilisée pour générer l'ensemble des jeux de données simulées. Notons que le temps de calcul de la méthode BEAST est uniquement estimé sur les jeux de données comptant au plus 110 feuilles, alors que certaines simulations en comptabilisent 550.

	Entrées		
	Alignements	Matrices de distances	Arbres
Méthodes d'inférence phylogénétique			
DNAdist	≈ 2 jours 200[a]	-	-
FastME	-	109	-
PhyML	≈ 30 jours ≈ 2 jours[a]	-	-
Méthodes d'estimation de taux de substitution			
Pairwise-Distance	-	5	-
sUPGMA	-	14	-
TREBLE	-	60	-
Root-to-tip	-	-	19
ULS sans choix aléatoire	-	367	293
ULS avec choix aléatoire	-	38	30
BEAST	≈ 129 jours[a]	-	-

[a] uniquement les petits jeux de données (<110 feuilles par phylogénie, alors que certains vont jusqu'à 550).

Ces résultats montrent qu'ULS n'est pas la méthode d'estimation la plus rapide, les méthodes *Pairwise-Distance* et sUPGMA sont plus rapides qu'ULS, mais elles ne sont pas très performantes en précision d'estimation (cf. section précédente). Le temps de calcul de la méthode ULS est quasiment multiplié par 10 entre la version avec et sans le choix aléatoire. Dans le cas où l'on considère le choix aléatoire, il faut approximativement 30 minutes de calcul pour obtenir l'ensemble des estimations. Rappelons que les précisions d'estimation entre ces deux versions sont similaires (cf. section 4.2.4.3). Cette amélioration en fait une méthode très rapide (moins de 5 secondes sur un arbre avec 550 feuilles). Le temps de calcul de la méthode *Root-to-tip* est assez semblable au notre (environ 10 minutes d'écart en faveur de *Root-to-tip*), mais rappelons que cette méthode reste dépendante du nombre de feuilles dans la phylogénie, ce qui n'est plus le cas avec ULS en considérant le choix aléatoire. Notons que le temps de calcul d'ULS est plus rapide avec un arbre en entrée qu'avec une ma-

trice de distances, cela provient du format d'encodage des arbres et des matrices où la quantité de données à lire (nombre de caractères) est beaucoup moins importante avec des arbres au format NEWICK, qu'avec des matrices de distances. Le temps de calcul de la méthode BEAST, avoisinant les 129 jours de calcul mais uniquement sur les petits jeux de données (au plus 110 feuilles), en fait la méthode d'estimation de taux de substitution la plus lente, et cela même en considérant des arbres PhyML en entrée puisque, sur l'ensemble des petits jeux de données, il faut environ 2,3 jours de calcul à PhyML pour inférer les arbres. D'autant plus que sa précision d'estimation est équivalente à (ou moins bonne que) celle d'ULS sur des arbres PhyML. Enfin, notons que le temps de calcul des arbres FastME (en considérant bien sûr le temps nécessaire au calcul des matrices de distances) est beaucoup plus rapide que celui des arbres PhyML et pour une précision d'estimation quasi-équivalente (cf. section précédente). En résumé, sur un jeu de données contenant 550 feuilles, les temps de calcul des méthodes de distances sont approximativement de 5 secondes pour ULS, 1 seconde pour *Pairwise-Distance*, 2 secondes pour sUPGMA, 3 secondes pour *Root-to-tip* et 10 secondes pour TREBLE. Le temps de calcul, avec des séquences de 1 000 paires de bases, pour inférer un arbre PhyML est approximativement de 2 heures, celui d'un arbre FastME de 6 secondes et celui d'une matrice de distances DNAdist de 8 minutes. Quant à BEAST, il met environ 30 minutes sur un jeu de données de 1 000 paires de bases contenant 110 feuilles.

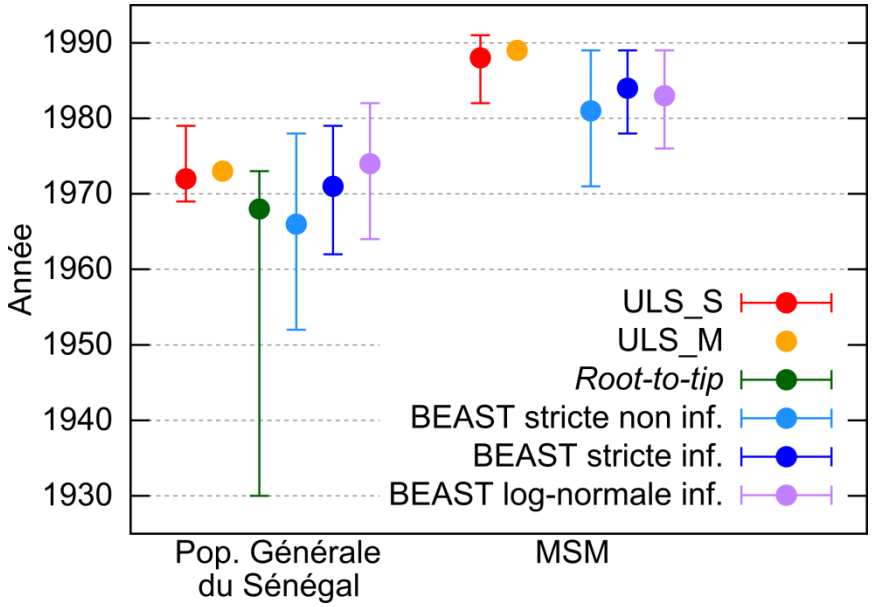
4.3.2 Application au sous-type C du VIH-1

Notre méthode est testée sur deux jeux de données contenant chacun des séquences du sous-type C du VIH-1 (VIH-1C). Le premier est extrait de l'étude sur l'origine géographique et temporelle de l'épidémie du VIH-1C au Sénégal (Chapitre 5) et l'autre de l'étude épidémiologique mondiale du VIH-1C (Chapitre 6).

Le premier jeu de données contient 56 séquences *pol* du VIH-1C collectées au Sénégal (Jung *et al*, 2012). La conception de l'alignement de 1 011 sites est décrite au Chapitre 5. À partir de cet alignement, un arbre de maximum de vraisemblance est calculé avec PhyML v3.0 (Guindon *et al*, 2010; Guindon & Gascuel, 2003) sous le modèle GTR+I+Γ4 (*general time reversible* avec une loi gamma à 4 catégories de taux et des sites invariants), en accord avec Jung *et al*. (2012). Les paramètres du modèle d'évolution sont estimés par PhyML. L'option SPR (*Subtree Pruning and Regrafting*) est choisie afin d'explorer l'espace des arbres. Les intervalles de confiance à 95% sont obtenus à partir de 100 arbres calculés de la même manière, mais sur la base d'alignements obtenus avec la technique du *bootstrap* par le logiciel *seqboot* v3.69 du package PHYLIP (Felsenstein, 1989). Le taux de substitution estimé par ULS sur ce jeu de données est de $2{,}01 \times 10^{-3}$ [$1{,}78 \times 10^{-3}$; $2{,}63 \times 10^{-3}$] substitutions par site et par année, tandis que celui estimé par BEAST dans Jung *et al*. (2012) est de $1{,}59 \times 10^{-3}$

[$1{,}02 \times 10^{-3}$; $2{,}19 \times 10^{-3}$] substitutions par site et par année, sous le modèle d'horloge moléculaire stricte et avec une *prior* non informative (c.-à-d. les mêmes conditions que dans ULS). Ces résultats sont largement recouvrants et la différence n'est pas statistiquement significative. Cependant, le taux de substitution d'ULS semble plus élevé que celui de BEAST et notre intervalle de confiance à 95% plus étroit. La régression linéaire *Root-to-tip*, estime le taux de substitution à $1{,}47 \times 10^{-3}$ [$7{,}06 \times 10^{-4}$; $2{,}11 \times 10^{-3}$] substitutions par site et par année, très proche de celui de BEAST. La date de l'ancêtre commun aux souches du Sénégal est estimée par ULS à 1972 [1969 ; 1979] et par BEAST à 1966 [1952 ; 1978] (Figure 41). L'approche pour ULS consiste à corriger les distances évolutives entre les feuilles non contemporaines et à employer la méthode UPGMA (cf. Chapitre 1) pour reconstruire l'arbre enraciné, dans lequel la distance évolutive qui sépare les feuilles contemporaines de la racine est déduite. Ici aussi ULS propose une estimation plus élevée, mais toujours avec des intervalles de confiance recouvrants et pas de différence significative. De plus, l'intervalle d'ULS est nettement plus serré (environ 10 ans) que celui de BEAST (environ 30 ans). La régression linéaire *Root-to-tip* estime l'ancêtre commun à 1968 [1930 ; 1973], à nouveau proche de l'estimation de BEAST mais avec un intervalle de confiance très large (plus de 40 ans). Cependant, en considérant pour BEAST une *prior* informative (d'après des estimations publiées), ses estimations deviennent assez similaires à celles d'ULS : taux de substitution à $1{,}85 \times 10^{-3}$ [$1{,}36 \times 10^{-3}$; $2{,}37 \times 10^{-3}$] substitutions par site et par année et date de l'ancêtre commun à 1971 [1962 ; 1979]. En revanche, le choix de la *prior* (informative ou non) influe très peu sur les estimations de la date de l'ancêtre commun aux souches des hommes ayant des rapports sexuels avec des hommes (MSM), et est estimée par BEAST au début des années quatre-vingt. Par exemple, pour la *prior* non informative, BEAST date l'ancêtre commun aux souches isolées chez les MSM à 1981 [1971 ; 1989] et pour une *prior* informative à 1984 [1978 ; 1989]. Quant à ULS, il l'estime à 1988 [1982 ; 1991], environ 5 ans plus tard que BEAST (mais avec des intervalles compatibles et donc des différences non significatives). La stabilité de BEAST dans les estimations de la date de l'ancêtre commun aux souches des MSM s'étend aussi au-delà du modèle d'horloge moléculaire. Par exemple, avec une horloge moléculaire relâchée en log-normal et une prior informative, BEAST l'estime à 1983 [1976 ; 1989]. En revanche, sous ce même modèle, l'estimation de la date de l'ancêtre commun aux souches collectées au Sénégal (1974 [1964 ; 1982]) s'accorde mieux avec celle d'ULS. Le temps de calcul nécessaire à BEAST sur ce jeu de données avoisine les 12 heures de temps de calcul, tandis que le temps de calcul de la méthode ULS est à peine de 4 secondes (en considérant, en plus, le calcul de l'intervalle de confiance). Notons cependant que PhyML met environ 4 minutes pour estimer une phylogénie (donc environ 6 heures et demi sont nécessaires à PhyML pour calculer les 100 phylogénies).

Figure 41. Estimations temporelles d'ULS sur deux jeux de données du sous-type C du VIH-1.
Ce graphique montre les estimations par ULS de la date de l'ancêtre commun aux souches collectées au Sénégal et celle de l'ancêtre commun aux souches isolées chez les MSM à partir d'une phylogénie contenant uniquement les séquences collectées au Sénégal (ULS_S, en rouge ; cf. Chapitre 5) et d'une autre phylogénie contenant l'ensemble des séquences du VIH-1C (ULS_M, en orange ; cf. Chapitre 6). L'estimation de la date de l'ancêtre commun des souches collectées au Sénégal par *Root-to-tip* (en vert) et les estimations de BEAST du Chapitre 5 (horloge moléculaire stricte avec une *prior* non informative, en bleu clair, et informative, en bleu foncé, et horloge moléculaire relâchée en log-normal avec une *prior* informative, en mauve) sont aussi présentées. Les estimations de ces deux dernières méthodes sont réalisées à partir d'un jeu de données contenant uniquement les séquences collectées au Sénégal. La date de l'ancêtre commun aux souches du VIH-1C est estimée par ULS à 1964 et par *Root-to-tip* à 1782.



Le second jeu de données considéré contient 3 609 taxa. Les détails de la conception de l'alignement de 1 011 sites et les paramètres utilisés par PhyML pour inférer la phylogénie sont donnés au Chapitre 6. Sur ce jeu de données (*ingroup* uniquement) qui contient l'ensemble des souches *pol* disponibles du VIH-1C, ULS estime le taux de substitution à $4{,}70 \times 10^{-3}$ substitutions par site et par année et la date de l'ancêtre commun à 1964. L'estimation de la date de l'ancêtre commun d'ULS semble être du même ordre de grandeur que celle admise pour les souches du sous-type C (Abecasis *et al*, 2009; Rousseau *et al*, 2007; Travers *et al*, 2004). Par exemple, Rousseau *et al*. (2007) l'estiment à 1961 [1947 ; 1962] et le taux de substitution à $5{,}1 \times 10^{-3}$ [$3{,}9 \times 10^{-3}$; $5{,}3 \times 10^{-3}$] substitutions par site et par année ; mais à partir de génomes presque complets. Dalai *et al*. (2009), quant à eux, estiment avec BEAST la date de l'ancêtre commun aux souches du sous-type C collectées au Zimbabwe à 1972 [1969-1974] et un taux de substitution moyen à $2{,}33 \times 10^{-3}$ substitutions par site et par année sur des séquences *pol*, suggérant que notre taux de substitution est encore une fois plus important que celui de BEAST ; mais les données sont bien différentes. Sur ce jeu de données, *Root-to-tip* estime un taux de substitution de $8{,}04 \times 10^{-4}$ substitutions par site et par année. Ce taux est très nettement inférieur à ceux mentionnés ci-dessus qui semblent être plus en accord avec l'estimation d'ULS. La date de l'ancêtre commun estimée par *Root-to-Tip*, qui est de 1782, est complètement différente de celle qui est communément admise par la communauté scientifique (début

de la deuxième moitié du xx^{e} siècle) et montre ainsi la limite de cette méthode. Notons cependant que cette date de référence se fonde uniquement sur les estimations moyennes de plusieurs références bibliographiques, revues dans Hemelaar *et al*. (2012), et n'intègre pas l'information des intervalles de confiance associés à ces estimations, qui sont parfois très larges (couvrant la période de 1933 à 1973). Ces intervalles de confiance montrent l'incertitude associée aux estimations ponctuelles qui vient probablement du manque de signal dans les données étudiées.

Cette phylogénie inclut aussi les souches collectées au Sénégal, y compris celles isolées chez les MSM (cf. ci-dessus), et les dates associées à leur ancêtre commun peuvent donc être estimées. La date de l'ancêtre commun aux souches collectées au Sénégal est estimée par ULS à 1973 et celle de l'ancêtre commun aux souches des MSM à 1989 (Figure 41). Ces deux estimations sont tout à fait cohérentes avec celles estimées précédemment en ne considérant que les souches collectées au Sénégal (cf. ci-dessus). Sur ce jeu de données, ULS met moins de 3 minutes à estimer le taux de substitution et la date de l'ancêtre commun, tandis que *Root-to-tip*, dépendant du nombre de feuilles de la phylogénie, met un peu plus de 5 minutes.

4.4 Conclusion

Nous présentons une nouvelle méthode de distances, *Ultrametric Least Squares* (ULS), basée sur le principe des moindres carrés, qui permet d'estimer le taux de substitution sous les hypothèses du modèle *Single Rate Dated Tips* (SRDT) (feuilles hétérochrones et horloge moléculaire stricte). Pour ce faire, elle minimise un critère parabolique par morceaux qui mesure l'ultramétricité d'une distance en $O(n^3 \log n)$, où n est le nombre de feuilles. Un algorithme de type Monte Carlo borne cette complexité, et cela sans perte de précision, à partir d'un certain seuil déterminé en fonction de n et du nombre de feuilles par date d'échantillonnage. Cette méthode est aussi étendue aux modèles *Multiple Rates with Dated Tips* (MRDT) et *Different Rate* (DR) mais seulement avec des horloges moléculaires locales. L'implémentation de cette méthode en langage C fournit aussi l'opportunité d'estimer la date de l'ancêtre commun aux souches du jeu d'entrée, ainsi que d'enraciner une phylogénie en considérant les dates de prélèvement associées à chaque feuille.

Le principe itératif utilisé afin d'adapter ULS au modèle d'horloge moléculaire MRDT et à l'estimation de plusieurs taux de substitution par lignage (horloges moléculaires locales) peut être appliqué à n'importe quelle autre méthode d'estimation de taux de substitution faisant les hypothèses du modèle SRDT. À notre connaissance, seules deux autres méthodes permettent d'estimer le taux de substitution sous le modèle MRDT : sUPGMA, une méthode de distances, et TipDate, une méthode probabiliste (Drummond *et al*, 2001). Les horloges moléculaires locales sont très appréciées

puisqu'elles reflètent mieux la réalité des données surtout quand, par exemple, plusieurs sous-types du VIH sont étudiés en même temps. Toutefois, ces hypothèses de taux variant par branche ou par lignée nécessitent la donnée d'une phylogénie. Notre modèle suppose que l'on connaisse *a priori* les lignées évoluant avec un taux de substitution différent et que chaque lignée évolue avec le même taux. Yoder et Yang (2000) présentent un exemple d'application avec ce modèle. D'autres approches existent, comme celle proposée par Sanderson (1997) qui suppose que chaque branche de la phylogénie évolue avec un taux unique, mais en supposant une auto-corrélation des taux ; il minimise l'écart entre les taux de substitution d'une même lignée. Adapter la méthode ULS à un tel modèle serait un atout supplémentaire pour cette méthode.

La comparaison entre la précision d'estimation d'ULS et celle des méthodes de distances sUPGMA, TREBLE, ainsi que celle des régressions linéaires *Root-to-tip* et *Pairwise-Distance* sur différents jeux de données simulées indique qu'ULS est, en moyenne, la méthode la plus précise. Malgré cela, au cas par cas ULS n'est pas systématiquement la méthode la plus performante et est souvent en concurrence avec la méthode *Root-to-tip*. De plus, les estimations avec des matrices de distances laissent souvent quelque peu à désirer, en particulier sur les jeux de données intra-hôte avec 3 temps d'échantillonnage et des séquences de 300 paires de bases. Ces jeux de données contiennent, en moyenne, de grandes distances et cela suggère que la performance d'ULS est (logiquement) grandement dépendante de la justesse des estimations des distances évolutives qui sont mieux estimées lorsqu'elles sont petites. L'utilisation d'une autre variance, plus appropriée, pourrait balancer en notre faveur. En attendant, l'utilisation d'arbres FastME, rapides à obtenir, semble corriger ce problème.

La différence en précision d'estimation entre ULS et la méthode probabiliste de référence, BEAST, est en moyenne à notre avantage mais reste généralement non significative sur les jeux de données simulées en inter-hôtes. En revanche, la performance d'ULS contre BEAST est autre sur les jeux de données intra-hôtes contenant un taux de mort élevé (995 sur 1 000 à chaque temps d'échantillonnage). Récemment, une nouvelle version de BEAST est disponible et cette version permet d'utiliser un modèle de spéciation, plus adapté que celui du coalescent (seul modèle disponible avec la version 1.6.2 pour des données hétérochrones), considérant un taux de naissance et de mort constant avec des données hétérochrones (Drummond *et al*, 2012; Stadler, 2010). Une comparaison avec cette nouvelle version est nécessaire et permetterait de confirmer ou d'infirmer les résultats présentés dans ce manuscit.

Chapitre 5

Origine géographique et temporelle du sous-type C du VIH-1 au Sénégal

Nous présentons une étude moléculaire visant à connaître l'origine géographique et temporelle de l'épidémie du sous-type C du VIH-1 au Sénégal, avec un intérêt particulier pour les souches circulant chez les hommes ayant des rapports sexuels avec des hommes (MSM). Pour cette étude, nous analysons le gène pol *de toutes les souches virales précédemment publiées et de dix-huit nouvelles souches collectées au Sénégal. Une grande phylogénie contenant plus 3 000 séquences est calculée afin de déterminer les séquences proches de celles du Sénégal. Deux phylogénies (PhyML et MrBayes) sont construites avec l'ensemble des souches du Sénégal et des souches proches afin de déterminer l'origine géographique de l'épidémie du sous-type C au Sénégal. Une analyse bayésienne (BEAST) est menée, mais uniquement avec les souches collectées au Sénégal, pour déterminer l'origine temporelle de cette épidémie. Ces analyses montrent de multiples introductions de ce variant dans la population générale provenant de pays de l'Afrique de l'est et australe, tandis que l'épidémie chez les MSM a connu une introduction unique suivie d'une diffusion efficace originaire d'Afrique australe (probablement de Zambie). L'ancêtre commun aux souches du Sénégal est daté au début des années soixante-dix et celui des séquences des MSM environ dix ans après. Comme cette étude a fait l'objet d'une publication dans une revue internationalle (PLoS One), nous présentons uniquement un résumé détaillé en français et joignons l'article à ce chapitre.*

Sommaire

5.1 Introduction

Les premières recherches d'infection liée au virus de l'immunodéficience humaine (VIH) au Sénégal ont été faites dans des cohortes de prostituées, parce qu'elles sont jugées être un groupe à haut risque d'infection (Meda *et al*, 1999; Barin *et al*, 1985; Van de Perre *et al*, 1985). Les tests sérologiques effectués montraient que ces individus étaient contaminés par le VIH-2 et une forte prévalence de ce variant était observée dans différentes villes, entre 10,0% et 38,1% versus 0,4% et 4,1% pour le VIH-1 (Kanki *et al*, 1992). Dès 1986, les premiers cas d'infection au VIH-1 sont reportés (Kanki *et al*, 1992) et, depuis, la prévalence du VIH-2 a diminué tandis que celle du VIH-1 a augmenté (Marlink, 1996; Hamel *et al*, 2007). L'identification à Dakar de tous les sous-types du groupe M du VIH-1 suggère de multiples introductions du virus dans ce pays (Toure-Kane *et al*, 2000), probablement dues aux activités de commerce ou de voyage avec les autres pays de l'Afrique. Actuellement, les infections au VIH dans l'ouest de l'Afrique, et donc au Sénégal, sont surtout causées par des souches de la forme recombinante circulante CRF02_AG (Buonaguro *et al*, 2007; Sankalé *et al*, 2000; Toure-Kane *et al*, 2000). Au Sénégal, le sous-type A est aussi très présent (Sankalé *et al*, 2000) et le sous-sous-type A3 a été caractérisé pour la première fois dans une cohorte de prostituées résidant à Dakar (Meloni *et al*, 2004a, 2004b).

Les études sur le groupe à risque des hommes ayant des rapports sexuels avec des hommes (MSM, *men having sex with men*), vulnérables aux infections sexuellement transmissibles (Geibel *et al*, 2010), ne se sont faites que bien plus tard à cause de la stigmatisation exercée sur eux dans la plupart des pays africains (Niang *et al*, 2003). En raison de cette répression près de 95% d'entre eux ont des rapports sexuels avec des femmes afin de garder leur double vie secrète (Wade *et al*, 2005). En 2009, une étude sur la distribution de la prévalence des sous-types et des formes recombinantes du VIH-1 a montré une prévalence du sous-type C de 40% chez les MSM, alors qu'elle est à moins de 5% dans la population générale et chez les prostituées (Ndiaye *et al*, 2009). Ce sous-type est également très peu prévalant dans les autres pays de l'Afrique de l'ouest.

Nous présentons la première étude moléculaire visant à connaître l'origine géographique et temporelle de l'ancêtre commun aux souches du sous-type C du VIH-1 circulant dans la population générale sénégalaise, mais aussi de celui circulant chez les MSM. Cette étude a plusieurs objectifs : 1) savoir s'il existe un lien épidémiologique entre les souches des MSM et celles provenant de la population générale ; 2) connaître l'origine géographique de l'épidémie du sous-type C sévissant au Sénégal et chez les MSM de ce pays ; 3) enfin, dater l'origine de l'introduction de cette épidémie chez les MSM ainsi que dans la population générale du Sénégal. Pour cela, nous utilisons des outils bioinformatiques afin d'inférer une phylogénie sur 3 081 séquences. Cette phylogénie met en évidence les

liens épidémiologiques existant entre les souches du sous-type C du Sénégal et celles des autres pays. Par la suite, nous utiliserons uniquement les souches disponibles du sous-type C du Sénégal pour estimer la date de leur ancêtre commun, ainsi que celle de l'ancêtre commun aux souches des MSM.

5.2 Préparation des données

Les séquences sont collectées dans la base de données public du laboratoire national de Los Alamos : *HIV Databases*[5]. Toutes les séquences disponibles du sous-type C du VIH-1, sur la région 2 253-3 263 du génome d'HXB2, et dont la date et le pays de collecte sont connus, sont téléchargées[6]. Cette région code la protéase et une partie de la transcriptase inverse. La vérification d'éventuels recombinants ou de sous-types non-C est faite par l'application web *REGA HIV-1 & 2 Automated Subtyping Tool* (de Oliveira *et al*, 2005). Les séquences non reconnues comme du sous-type C à 100% sont écartées de la suite de nos analyses. La séquence d'HXB2 (sous-type B ; numéro d'accession : K03455) sert d'*outgroup* pour enraciner les arbres de maximum de vraisemblance construits dans cette étude. À cette collection, 18 nouvelles séquences collectées au Sénégal entre 1996 et 2007 sont ajoutées. Elles ont été séquencées par les membres de l'équipe TransVIHMI. Seule une séquence choisie au hasard est conservée parmi celles qui sont identiques ou qui présentent un lien épidémiologique proche (par exemple dans le cas d'une transmission mère-enfant).

Les séquences provenant de la base de données *HIV Databases* sont déjà alignées. Un alignement séquences contre profil du programme MAFFT version 6 (Katoh *et al*, 2002), avec la méthode L-INS-i (Katoh *et al*, 2005), est réalisé afin d'y ajouter les 18 nouvelles séquences. Quelques corrections manuelles sont ensuite apportées à l'aide de MEGA version 5 (Tamura *et al*, 2011) et tous les sites contenant un nombre excessif de gaps ($\geq 50\%$) sont supprimés. Pour éviter tout biais éventuel dû aux mutations de résistance causées par les traitements antirétroviraux, les analyses sont faites en parallèle sur un alignement où 43 codons connus pour être associés à des mutations de résistance majeures sont supprimés (Bennett *et al*, 2009).

Le calcul de l'arbre PhyML (Guindon & Gascuel, 2003) représentant l'histoire évolutive de la totalité des séquences est fait sous le modèle *general time reversible* avec une proportion de sites invariables et une loi gamma de catégorie 4 (GTR+I+Γ4) (Posada & Crandall, 2001). L'option SPR (*subtree pruning and regrafting*) est choisie pour explorer l'espace des arbres. Pour une meilleure estimation, tous les paramètres sont évalués et optimisés par PhyML. Enfin, les supports de branche sont déterminés par la méthode *approximate likelihood ratio test* (aLRT) (Anisimova & Gascuel, 2006), option

[5] http://www.hiv.lanl.gov/content/index
[6] Accédé le 11 avril 2011

SH-like. Puis, un second arbre de vraisemblance est inféré sous le même modèle, mais contenant uniquement les séquences du Sénégal et celles (proches) contenues dans chaque sous-arbre ayant pour racine le nœud ancestral de deuxième génération à chaque séquence provenant du Sénégal (d'après le premier arbre). Sur ce dernier, nous estimons aussi les supports de branche obtenus par la méthode du *bootstrap* (100 itérations). La topologie et les résultats sont vérifiés à l'aide d'un arbre bayésien, calculé par MrBayes version 3.1 (Ronquist & Huelsenbeck, 2003).

Les estimations du taux de substitution et des dates des ancêtres communs sont réalisées avec BEAST v1.6.1 (Drummond & Rambaut, 2007). Seules les 56 séquences du Sénégal sont considérées dans ces analyses. Le modèle de substitution utilisé est choisi en adéquation avec celui des arbres de maximum de vraisemblance (GTR+I+Γ4). Les estimations sont faites sous l'hypothèse de trois horloges moléculaires : stricte, relâchée en log-normal et en exponentiel (Drummond et al., 2006). Avec l'horloge moléculaire relâchée en log-normal, chaque taux de substitution suit une loi log-normale de moyenne *ucld.mean* et d'écart-type *ucld.stdev*, chaque taux de substitution de l'horloge relâchée en exponentiel suit une loi exponentielle de moyenne *uced.mean*, et le taux de substitution associé à l'horloge moléculaire stricte est constant et dépend du paramètre *strict.clock*. L'histoire démographique est calculée sous le modèle *Bayesian Skyride* avec l'option *Time-aware* (Minin *et al*, 2008). Quatre *priors* différentes sont utilisées pour les paramètres *ucld.mean*, *uced.mean* et *strict.clock*. La première, non informative, suit une loi uniforme entre 0 et 1. Les suivantes suivent une loi normale de moyenne $2{,}5 \times 10^{-3}$ (d'après Dalai *et al*. (2009) et Path-O-Gen v1.3[7]) et d'écart-type 10×10^{-4}, $7{,}5 \times 10^{-4}$ et 5×10^{-4} respectivement. La distribution de *ucld.stdev* suit une loi exponentielle de paramètre 0,1 (d'après une communication personnelle avec Alexei DRUMMOND). Le nombre de générations pour les chaînes de Markov avec technique de Monte Carlo (*Markov chain Monte Carlo*, MCMC) est de $2{,}5 \times 10^{8}$ avec un échantillonnage toutes les $2{,}5 \times 10^{5}$ générations. La convergence est vérifiée avec le logiciel Tracer v1.5, tout comme l'extraction des résultats et les estimations des facteurs de Bayes.

5.3 Résultats

L'origine géographique de l'épidémie du VIH-1 sous-type C au Sénégal est initialement explorée à l'aide d'un arbre de maximum de vraisemblance, contenant toutes les séquences *pol* disponibles plus 18 nouvelles collectées au Sénégal (soit un total de 3 081 séquences). Sur les deux phylogénies obtenues (une contenant les sites associés à des mutations de résistance, l'autre sans), la plupart des séquences échantillonnées en Asie et en Amérique sont regroupées dans deux clades. Les séquences restantes sont dispersées ou forment des clades marginaux. Les souches collectées en Afrique de

[7] http://tree.bio.ed.ac.uk/software/pathogen/

l'est sont aussi regroupées, tandis que les souches collectées en Europe sont disséminées dans l'arbre, tout comme les souches du Sénégal provenant de la population générale. Les souches des MSM forment un clade net et distinct, et des souches collectées en Afrique australe se positionnent à sa racine.

Afin de mieux discerner l'origine géographique des différentes souches du Sénégal, un second arbre contenant uniquement les séquences proches à celles du Sénégal est inféré. Ces dernières proviennent essentiellement du continent africain (147 sur 177, soit 83,05%). Cette phylogénie confirme l'idée de multiples introductions du virus dans la population générale sénégalaise ; introductions qui semblent provenir de deux zones géographiques distinctes. Une provenant de l'Afrique australe et l'autre de l'Afrique de l'est. Des souches de l'Afrique australe (dont beaucoup proviennent de Zambie) se placent à proximité du clade contenant les souches des MSM, suggérant que l'ancêtre commun est originaire de l'Afrique australe. Malgré des topologies légèrement différentes, les conclusions épidémiologiques sont aussi confirmées sur l'arbre MrBayes, que ce soit à partir de l'alignement contenant les sites associés aux mutations de résistance ou non.

Les estimations des dates des ancêtres communs sont faites avec le logiciel BEAST sous trois horloges moléculaires différentes (stricte, relâchée en exponentiel et relâchée en log-normal), chacune associée au modèle de croissance démographique *Bayesian Skyride*. Les facteurs de Bayes estimés avec Tracer indiquent que l'horloge moléculaire relâchée en exponentiel (resp. en log-normal) est un peu mieux adaptée aux données que l'horloge moléculaire relâchée en log-normal (resp. stricte). L'horloge moléculaire relâchée en exponentiel sur les deux *priors* les moins informatives montrent des estimations ayant de grands intervalles de confiance (plusieurs siècles pour certaines estimations temporelles). De ce fait, les résultats ne sont pas interprétables et nous utilisons dans la suite le modèle log-normal. Très peu de différences sont à noter entre les résultats des deux alignements (avec ou sans les sites associés aux mutations de résistance). Les estimations du taux de substitution sont assez similaires entre les horloges stricte et relâchée en log-normal et semblent avoisiner les $1{,}75 \times 10^{-3}$ substitutions par site et par année. Les estimations des dates des ancêtres communs sont aussi relativement similaires. Comme valeur consensus, la date de l'ancêtre commun aux souches du Sénégal est estimée au début des années soixante-dix et celle de l'ancêtre commun aux souches des MSM au début des années quatre-vingt, environ dix ans après.

À la section 4.3.2, page 120, les estimations de BEAST du taux de substitution et des dates des ancêtres communs sont comparées à celles d'ULS. Brièvement, les estimations d'ULS sont obtenues à partir d'une phylogénie inférée par PhyML (en utilisant l'alignement complet, c'est-à-dire celui contenant les codons associés à des mutations de résistance), sous le modèle d'évolution GTR+I+Γ4, et

ne contenant que les souches collectées au Sénégal. Les intervalles de confiance sont calculés par *bootstrap*. La date de l'ancêtre commun des souches collectées au Sénégal est estimée à 1972 [1969 ; 1979] et celle des souches isolées chez les MSM à 1988 [1982 ; 1991]. L'estimation d'ULS de la date de l'ancêtre commun des souches isolées dans la population générale du Sénégal est assez similaire à celle estimée par BEAST sous une horloge moléculaire relâchée en log-normal et avec une *prior* informative (1974 [1964 ; 1982] ; cf. Figure 41). Notons qu'avec une *prior* non informative, l'estimation de BEAST est sensiblement plus ancienne (1967 [1950 ; 1983]). En revanche, les différentes estimations de BEAST de la date de l'ancêtre commun des souches isolées chez les MSM sont assez proches entre elles (p. ex. 1983 [1976 ; 1989] avec la *prior* informative), mais, comme attendu, avec un intervalle de confiance plus large pour la *prior* non informative (1979 [1965 ; 1989]). ULS estime plutôt la date de ce même ancêtre commun vers la fin des années quatre-vingt (1988 [1982 ; 1991]). Remarquons, encore une fois, que les intervalles de confiance sont largement recouvrants. Observons aussi que, à chaque fois, l'amplitude des intervalles de confiance des estimations d'ULS est moindre par rapport à ceux de BEAST (p. ex. pour la date de l'ancêtre commun des souches isolées de patients MSM, il est de 9 ans pour ULS, 13 ans pour BEAST avec une *prior* informative et 24 ans avec une *prior* non informative).

5.4 Conclusion

Nous présentons la première étude moléculaire visant à connaître l'origine géographique et temporelle de l'épidémie du sous-type C du VIH-1 au Sénégal. Les résultats obtenus montrent que l'épidémie du sous-type C chez les MSM provient d'un évènement fondateur et que l'ancêtre commun est originaire d'un pays d'Afrique australe, probablement de la Zambie. Cette épidémie est assez récente (début des années quatre-vingt) comparée à celle en Éthiopie (milieu des années soixante) (Tully & Wood, 2010) ou celle au Malawi (fin des années soixante) (Travers *et al*, 2004). Les souches du sous-type C de la population générale proviennent d'introductions multiples et d'origines géographiques différentes (Afrique de l'est et australe). Cela montre les liens établis par cette population avec les autres pays africains (Toure-Kane *et al*, 2000). Leur ancêtre commun est daté au début des années soixante-dix, environ dix ans avant la date de l'ancêtre commun aux souches des MSM.

L'utilisation d'un alignement avec ou sans les sites associés à des mutations de résistance majeures montre un faible impact sur la formation de clusters ou sur les estimations des dates des ancêtres communs et des taux de substitution (Hué *et al*, 2004). L'analyse des 3 081 séquences du sous-type C collectées à travers le monde fournit une représentation de la diversité globale du sous-type C, ainsi que des informations additionnelles sur l'épidémie du sous-type C. Nos analyses confirment le lien épidémiologique entre le Brésil et l'Afrique de l'est, précédemment établie par Bello *et al*.

(2008), et suggèrent un lien épidémiologique entre l'Afrique australe (Dietrich *et al*, 1993) et l'Inde ainsi que de nombreuses interactions entre l'Europe et l'Afrique. À cause du nombre important de migrations et de voyages, la distribution géographique des sous-types est en constante évolution et le mélange entre les variants du VIH-1 est inévitable. Ces changements continueront d'être un challenge pour les stratégies thérapeutiques et la recherche d'un vaccin.

Malgré les multiples introductions du sous-type C du VIH-1 dans la population générale, seulement une expansion majeure de ce variant est observé chez les MSM, soulignant le fait qu'ils sont une population à risque pour les maladies sexuellement transmissibles (Geibel *et al*, 2010). Comme plus de 90% des MSM du Sénégal disent avoir des relations sexuelles avec des femmes (Wade *et al*, 2005), ils peuvent servir de pont envers la population générale et diffuser des variants endémiques à ce groupe. D'ailleurs, des séquences du sous-type C récemment isolées chez des femmes, se placent à l'intérieur du clade formé par les souches des MSM (Coumba Toure-Kane, données non publiées). Les programmes ciblant les MSM doivent aussi prendre en considération les pratiques hétérosexuelles de ces individus, afin d'éviter la propagation d'épidémies à des populations plus larges (Larmarange *et al*, 2010).

OPEN ACCESS Freely available online

PLoS one

The Origin and Evolutionary History of HIV-1 Subtype C in Senegal

Matthieu Jung[1,2], Nafissatou Leye[1,3], Nicole Vidal[1], Denis Fargette[4], Halimatou Diop[3], Coumba Toure Kane[3], Olivier Gascuel[2]*, Martine Peeters[1]*

1 UMI233, TransVIHMI, IRD (Institut de Recherche pour le Développement) and University of Montpellier 1, Montpellier, France, **2** UMR 5506, Méthodes et Algorithmes pour la Bioinformatique, Laboratoire d'Informatique, de Robotique et de Microélectronique de Montpellier, CNRS and University of Montpellier 2, Montpellier, France, **3** Laboratory of Bacteriology and Virology, Le Dantec University Teaching Hospital, Dakar, Senegal, **4** UMR RPB, Institut de Recherche pour le Développement, La Recherche Agronomique pour le Développement and University of Montpellier 2, Montpellier, France

Abstract

Background: The classification of HIV-1 strains in subtypes and Circulating Recombinant Forms (CRFs) has helped in tracking the course of the HIV pandemic. In Senegal, which is located at the tip of West Africa, CRF02_AG predominates in the general population and Female Sex Workers (FSWs). In contrast, 40% of Men having Sex with Men (MSM) in Senegal are infected with subtype C. In this study we analyzed the geographical origins and introduction dates of HIV-1 C in Senegal in order to better understand the evolutionary history of this subtype, which predominates today in the MSM population

Methodology/Principal Findings: We used a combination of phylogenetic analyses and a Bayesian coalescent-based approach, to study the phylogenetic relationships in *pol* of 56 subtype C isolates from Senegal with 3,025 subtype C strains that were sampled worldwide. Our analysis shows a significantly well supported cluster which contains all subtype C strains that circulate among MSM in Senegal. The MSM cluster and other strains from Senegal are widely dispersed among the different subclusters of African HIV-1 C strains, suggesting multiple introductions of subtype C in Senegal from many different southern and east African countries. More detailed analyses show that HIV-1 C strains from MSM are more closely related to those from southern Africa. The estimated date of the MRCA of subtype C in the MSM population in Senegal is estimated to be in the early 80's.

Conclusions/Significance: Our evolutionary reconstructions suggest that multiple subtype C viruses with a common ancestor originating in the early 1970s entered Senegal. There was only one efficient spread in the MSM population, which most likely resulted from a single introduction, underlining the importance of high-risk behavior in spread of viruses.

Citation: Jung M, Leye N, Vidal N, Fargette D, Diop H, et al. (2012) The Origin and Evolutionary History of HIV-1 Subtype C in Senegal. PLoS ONE 7(3): e33579. doi:10.1371/journal.pone.0033579

Editor: Chiyu Zhang, Jiangsu University, China

Received September 26, 2011; **Accepted** February 15, 2012; **Published** March 28, 2012

Funding: MJ was supported by a PhD grant from the Région Languedoc-Roussillon and from the University of Montpellier 2, France. Nafissatou Leye has a PhD grant from S.C.A.C. (Service de Coopération et d'Action Culturelle) of the French Embassy in Senegal. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: martine.peeters@ird.fr (MP); gascuel@lirmm.fr (OG)

Introduction

HIV-1 group M, which predominates in the global HIV/AIDS epidemic, can be further subdivided into subtypes (A–D, F–H, J, K), sub-subtypes (A1 to A4, F1 and F2), circulating recombinant forms (CRF01 to CRF51) and numerous unique recombinant forms (URFs) (www.hiv.lanl.gov). This genetic diversity has an impact on almost all aspects of the management of this infection going from identification and monitoring of infected persons, to treatment efficacy and vaccine design [1–3]. The classification of HIV strains has also helped in tracking the course of the HIV pandemic [4]. Numerous molecular epidemiological studies showed a heterogeneous geographic distribution of the different HIV-1 M subtypes and CRFs. The initial diversification of group M most likely occurred within or near the Democratic Republic of Congo (DRC) [5,6], where the highest diversity of group M strains has been observed and the earliest cases of HIV-1 infection (1959 and 1960) have been documented in Kinshasa, the capital city [7]. Different HIV variants have then spread across the world, and the epidemics in the different continents and countries are the result of different founder effects. Today, subtype C accounts for 50% of all infections [8]. The majority of subtype C infections are found in southern Africa where they represent almost 100% of circulating HIV-1 strains. Subtype C also predominates in India, Ethiopia and southern China, and has entered East Africa, Brazil, and many European countries. With increasing mobility and human migration, HIV-1 variants inevitably intermix in different parts of the world and the distribution of the different HIV-1 variants is a dynamic process.

In Senegal, which is located at the tip of West Africa, both AIDS viruses, HIV-1 and HIV-2, co-circulate. HIV-2 was first described in Senegal, but like in other West African countries, the prevalence of HIV-2 remained low and is decreasing [9,10]. Today HIV-1 predominates and since the description of the first HIV-1 AIDS case in 1986, HIV-1 seroprevalence remains below 1% in the general population but can reach up to 20% in

population groups with high risk behavior like female sex workers (FSWs) or men having sex with men (MSM) [11]. Several studies showed that CRF02_AG predominates in Senegal, representing 50–70% of circulating strains in the general population and FSWs, but in contrast to surrounding west African countries, a wide diversity of other HIV-1 variants co-circulate; subtypes A1, A3, B, D, F, G, H, CRF01, CRF06, CRF09, CRF11, CRF45 and HIV-1 group O have all been documented [10,12–14]. As mentioned above, the distribution of HIV-1 subtypes/CRFs can differ between geographic origins and between population groups. Recently our studies showed that 40% of MSM in Senegal are infected with subtype C, which is in strong contrast with 4% to 10% in the general population and FSWs [10,12–15]. The factors associated with the rapid spread of subtype C and its predominance in the global epidemic are not entirely known, but in certain regions where it has been introduced, subtype C has overtaken other HIV-1 variants [16]. The high prevalence and the rapid spread of subtype C among MSM needs thus particular attention because this could also lead to an increase overtime of subtype C in the general population because more than 90% of MSM recognize having sex with women [17].

Using a combination of phylogenetic analyses and a Bayesian coalescent-based approach, we studied the phylogenetic relationships of subtype C isolates from Senegal with other subtype C strains that were sampled worldwide, in order to define the origin and onset of the subtype C epidemic in MSM in Senegal.

Results

Origin of subtype C sequences in Senegal

Among the HIV-1 subtype C pol sequences that were downloaded, we first eliminated all sequences that were not identified as subtype C (i.e. intersubtype recombinants) by the REGA-subtyping tool and kept only one isolate per patient. The final dataset includes a total of 3,081 sequences spanning a 1,011 bp fragment in pol between positions 2,253 and 3,263 on the HXB2 genome, including 56 (among which 24 MSM and 18 newly sequenced) strains from Senegal (Table 1 and Table S1). Sequences were included from 4 different continents and 61 countries: Africa (22 countries), the Americas (7 countries), Asia (9 countries) and Europe (23 countries) (Table 2). The majority (67.73%) of the sequences are from Africa and more precisely from southern Africa (55.14%) that is South Africa (22.36%) and Zambia (20.55%), and to a lower extent Botswana (4.32%), Mozambique (3.18%), Malawi (2.30%), Swaziland (1.53%), and Zimbabwe (0.91%). Subtype C sequences from Asia are predominantly from India (355 sequences on a total of 380) and those from the Americas mainly from southern Brazil (253 sequences on a total of 299). Subtype C sequences from Europe represent 10.22% of the dataset and are collected from 23 different countries, without a single country or area that predominates in the dataset.

The maximum likelihood (PhyML) tree of the 3,081 subtype C sequences is shown in Figure 1. The strains from Senegal are highlighted in red, those from southern Africa (South Africa, Zambia, Zimbabwe, Malawi, Mozambique, Botswana, and Swaziland) in orange and those from the other African countries, which are predominantly from East Africa, in yellow. Strains from Asia, the Americas, and Europe are highlighted in green, purple and blue respectively. The sequences from Senegal are interspersed with the other African strains, but one significant cluster (98.9% aLRT support), which comprised all sequences obtained from MSM from Senegal, was identified. The phylogenetic tree shows also separate clades for subtype C strains from southern Africa and one from eastern Africa (cluster B, 75.9% aLRT support), each of which contains sequences from Senegal. The tree shows the presence of two other major clusters, one for the majority of South American (cluster A, purple) and one for the Asian strains (cluster C, green), each apparently resulting from different single introductions, but no strain from Senegal was observed in these clusters. The clusters from South America and Asia are each supported by 72.7% and 82.3% aLRT values, respectively. No significant cluster of European subtype C was observed, they are all interspersed with strains from different geographic origins mainly in Africa and in Asia and southern America. In order to exclude the possibility of artifactual phylogenetic clustering due to drug induced convergent evolution, especially for the clades from Senegal, the phylogenetic tree analysis was repeated on an alignment where 43 (i.e. 129 nt, ~12.7% of the full alignment) codon positions known to be associated with major resistance mutations were removed. This analysis shows the same subtype C clusters (Figure S1).

The above analysis showed that subtype C was introduced into Senegal at multiple occasions. Figure 2 shows in more details the subtype C sequences that are most closely related to those observed in Senegal. As described in Materials and Methods, only sequences that branched with one or more sequences from Senegal until the second ancestral node in the phylogenetic tree of the 3,081 sequences, were used for this subtree. In addition to the 56 sequences from Senegal, 121 other subtype C sequences were included (Table S2), representing 5.7% of the total alignment. Figure 2 shows the tree obtained by PhyML with strains colored according to their geographic origin (the same tree with strain names is available in Figure S2). HIV-1 strains from Zambia are represented by a separate color in this tree because strains from this country are frequently present. The majority of the subtype C strains from Senegal and those from the MSM cluster (node C) are falling in clusters (aLRT >85%) which are mainly represented by strains from Zambia and other countries from southern Africa (for example node A, E and F). Nevertheless, some strains from Senegal are related to subtype C from east African countries (majority Ethiopia: node D). Although the exact country at the origin of the most recent common ancestor of the MSM strains remains uncertain, this was most likely in southern Africa. The first ancestral node to the MSM cluster (node B) suggests an origin in Zambia, but this node is only supported with 83.7% aLRT and 11% bootstrap values. The first ancestral node (node A), supported by an aLRT value of 94.7% and a bootstrap value of 49%, contains mainly strains from Zambia but also from other southern African countries. The Bayesian phylogenetic tree analysis performed with MrBayes shows similar results (Figure S3).

Dating the subtype C epidemic in Senegal and MSM population

We used a Bayesian MCMC approach implemented in BEASTv1.6.1 to estimate the dates of the most recent common ancestors (MRCAs) for the subtype C sequences from Senegal in the general population and for the subtype C epidemic in the MSM population. We used the Bayesian skyride population growth model associated to three molecular clock models: strict, relaxed uncorrelated lognormal, and relaxed uncorrelated exponential. Moreover, we used four different priors on the average substitution rate among branches with varying informative levels. Figure 3 shows the resulting estimations of the MRCA dates for the different models and priors used. More details are provided in Table S3, including substitution rate estimations.

Bayes factors (BF) indicate that the relaxed exponential model has a small advantage (BF in the 3 to 5 range) over the relaxed

Table 1. HIV-1 subtype C strains from Senegal included in this study.

Strain identification	Accession Number	Year of isolation	Population group	Reference
90SN-90SE364	AY713416	1990	general population	[53]
98SN-66HPD	AJ583722	1998	general population	[54]
99SN-159HALD	AJ583716	1999	general population	[54]
99SN-142HPD	AJ583715	1999	general population	[54]
98SN-39HALD	AJ287005	1998	general population	[55]
99SN-86HPD	AJ583739	1999	general population	[54]
04SN-MS003	FM210753	2004	MSM	[15]
04SN-MS883	FM210752	2004	MSM	[15]
04SN-MS855	FM210749	2004	MSM	[15]
04SN-MS835	FM210745	2004	MSM	[15]
04SN-MS821	FM210741	2004	MSM	[15]
04SN-MS816	FM210740	2004	MSM	[15]
04SN-MS779	FM210737	2004	MSM	[15]
04SN-MS700	FM210736	2004	MSM	[15]
04SN-MS540	FM210726	2004	MSM	[15]
04SN-MS522	FM210725	2004	MSM	[15]
04SN-MS492	FM210723	2004	MSM	[15]
04SN-MS048	FM210722	2004	MSM	[15]
04SN-MS481	FM210718	2004	MSM	[15]
04SN-MS477	FM210717	2004	MSM	[15]
04SN-MS475	FM210716	2004	MSM	[15]
04SN-MS448	FM210712	2004	MSM	[15]
04SN-MS422	FM210709	2004	MSM	[15]
04SN-MS245	FM210699	2004	MSM	[15]
04SN-MS029	FM210691	2004	MSM	[15]
04SN-MS015	FM210689	2004	MSM	[15]
04SN-MS011	FM210687	2004	MSM	[15]
04SN-MS010	FM210686	2004	MSM	[15]
04SN-MS007	FM210685	2004	MSM	[15]
04SN-MS002	FM210684	2004	MSM	[15]
03SN-980HALD	FN599776	2003	general population	[14]
03SN-965HALD	FN599773	2003	general population	[14]
02SN-510HALD	FN599737	2002	general population	[14]
99SN-67HDP	FN599718	1999	general population	[14]
09SN-SNA3-366	HM002544	2009	not known	unpublished
08SN-SNA3-220	HM002517	2008	not known	unpublished
08SN-SNA3-191	HM002515	2008	not known	unpublished
07SN-SNA3-107	HM002507	2007	not known	unpublished
02SN-260HALD	HE588158	2002	general population	this study
03SN-154HALD	HE588157	2003	general population	this study
03SN-321HALD	HE588156	2003	general population	this study
03SN-L065	HE588149	2003	general population	this study
06SN-463HALD	HE588155	2006	general population	this study
07SN-2658HALD	HE588150	2007	general population	this study
07SN-2909HALD	HE588151	2007	general population	this study
07SN-2911HALD	HE588152	2007	general population	this study
07SN-2936HALD	HE588153	2007	general population	this study
07SN-3076HALD	HE588154	2007	general population	this study
00SN-102HALD	HE588159	2000	general population	this study
97SN-1119	HE588162	1997	general population	this study

Table 1. Cont.

Strain identification	Accession Number	Year of isolation	Population group	Reference
02SN-478HALD	HE588163	2002	general population	this study
97SN-14Fann	HE588165	1997	general population	this study
97SN-25Fann	HE588164	1997	general population	this study
96SN-1083	HE588166	1996	general population	this study
97SN-1186	HE588161	1997	general population	this study
97SN-1189	HE588160	1997	general population	this study

doi:10.1371/journal.pone.0033579.t001

lognormal model, which in turn is slightly better (BF in the 3 to 6 range) than the strict molecular clock. However, the relaxed exponential model becomes non-informative when non- or poorly informative priors on the substitution rate are used (U[0,1] and N[2.5×10^{-3}, 10×10^{-4}], see Materials and Methods), which reveals spurious peaks leading to very large (up to ~400 years) 95% Highest Posterior Density (HPD) intervals and unrealistic estimates. Except in these two cases, the results with all models and priors are quite consistent. As expected, when we used more informative priors we obtained more restricted 95% HPD intervals. Nevertheless, the median date estimates of the MRCAs of subtype C in the general population of Senegal and for the MSM cluster are similar for all models and priors, indicating likely epidemic origins in the early 80's, in the MSM population. The MRCA for the subtype C strains that entered at multiple occasions into the general population (i.e. heterosexual or mother to child transmission), is estimated in the early 70's.

To illustrate in more detail the MRCA of the subtype C strains in the MSM population and their relation to the other HIV-1 C strains from Senegal, the maximum clade credibility (MCC) tree with time scale obtained from BEAST is shown in Figure 4. We see the same MSM cluster as in the phylogeny of Figure 2 (see also Figure S2 and S3), and the early 70's and 80's dates for the MRCAs of general and MSM population respectively.

We verified whether presence of drug resistance mutations could have an impact on MRCA dates and substitution rate estimations. Therefore calculations were repeated on the three different molecular clock models and for the four priors on an alignment where 43 codon positions known to be associated with major resistance mutations were removed. This analysis showed no significant difference, compared to the results obtained with the complete alignment (Table S3 for details on estimations and Figure S4 for the MCC tree with time scale).

Finally, our reconstruction of the demographic history of HIV-1 C in Senegal identified an initial, slow growth phase until the end of the 70's followed by a period of quick exponential-like growth at the end of the 90's where the epidemic growth became slower (Figure 5).

Discussion

In this study we analyzed the geographical origins and introduction dates of HIV-1 subtype C in Senegal in order to better understand the evolutionary history of this subtype which predominates today in the MSM population [15]. Our evolutionary reconstructions suggest that multiple subtype C viruses with a common ancestor originating in the early 1970s entered the country, followed by a sharp growth of the effective number of infections over the next decade.

This analysis of more than 3,000 globally collected reference sequences most likely provides an adequate representation of global subtype C diversity, and provides also additional information on the subtype C epidemic in other continents. The phylogenetic tree analysis showed several major clusters of subtype C sequences, mainly related to the continent of origin, like Asia, Southern America or Africa, except for Europe. Interestingly, among the African strains, a separate cluster of strains derived from patients living in east African countries was observed [18], and subtype C strains from Europe do not form a separate cluster and are interspersed among the different continents and major clusters. Our data also confirm the previously reported link of the subtype C epidemic in Brazil with east Africa [19–22].

Our analyses with various methods (PhyML, MrBayes and BEAST) showed a significantly well-supported cluster which contained all subtype C strains that circulate among MSM in Senegal. The MSM cluster and other strains from Senegal are widely dispersed among the different subclusters of African strains, suggesting multiple introductions of subtype C into Senegal from many different southern and also eastern African countries. More detailed analyses showed that the majority of the HIV-1 C strains from Senegal, including those circulating among MSM, are more closely related to strains from southern African countries, mainly Zambia. The cluster of subtype C strains derived from the MSM population includes also strains from HIV-1 infected men from Senegal, who were not identified as MSM. Homosexuality is illegal in Senegal and male-to-male sex is condemned by political and religious authorities and by the general population, therefore most MSM keep their sexual life secret, including from their own family and more than 90% of MSM reported having sex also with women [17]. Thus, these additional strains in the MSM cluster are most likely from individuals with male-to-male sex activities. Subtype C in MSM may have its origin directly from southern Africa but it is also possible that the ancestor of this subtype C cluster circulated already for a certain period in the general population in Senegal before it was introduced into the MSM group.

The wide diversity and multiple introductions of subtype C fit also with the distribution of the HIV-1 variants in the general population in Senegal. Several studies showed that in addition to CRF02_AG, many other HIV-1 subtypes and CRFs are also present in the country, reflecting multiple introductions [10,12–14]. This is most likely related to the important trading activity and travel links of the country with many other African countries [23,24]. Our estimates suggest that the MRCA of the subtype C strains that entered Senegal was in the early 1970's, about 10–15 years before the description of the first HIV-1 AIDS case in the country or the first HIV-1 subtype C strain in 1988 in Senegal [25]. The MRCA date estimate of subtype C in Senegal is

Table 2. Numbers of HIV-1 subtype C strains from different countries that were included in this study.

Continent	Country	Number	%
Africa		*2087*	*67.73*
	Botswana	133	4.32
	Burundi	91	2.95
	Democratic Republic of Congo	19	0.62
	Djibouti	1	0.03
	Equatorial Guinea	1	0.03
	Eritrea	2	0.06
	Ethiopia	99	3.21
	Gabon	1	0.03
	Kenya	4	0.13
	Malawi	71	2.30
	Mali	1	0.03
	Mozambique	98	3.18
	Niger	4	0.13
	Senegal	56	1.82
	Somalia	1	0.03
	South Africa	689	22.36
	Sudan	10	0.32
	Swaziland	47	1.53
	Tanzania	82	2.66
	Uganda	16	0.52
	Zambia	633	20.55
	Zimbabwe	28	0.91
America		*299*	*9.71*
	Argentina	8	0.26
	Brazil	253	8.21
	Cuba	25	0.81
	Honduras	1	0.03
	United States of America	9	0.29
	Uruguay	2	0.06
	Venezuela	1	0.03
Asia		*380*	*12.33*
	China	7	0.23
	India	355	11.52
	Israël	5	0.16
	Myanmar	1	0.03
	Philippines	1	0.03
	Russia	1	0.03
	South Korea	2	0.06
	Taiwan	1	0.03
	Yemen	7	0.23
Europe		*315*	*10.22*
	Austria	3	0.10
	Belgium	35	1.14
	Cyprus	8	0.26
	Czech Republic	11	0.36
	Danmark	21	0.68
	Finland	6	0.19
	France	7	0.23
	Georgia	1	0.03
	Germany	7	0.23
	Greece	3	0.10
	Italy	22	0.71
	Luxemburg	3	0.10
	Norway	16	0.52
	Poland	2	0.06
	Portugal	28	0.91
	Roumania	35	1.14
	Slovakia	1	0.03
	Spain	26	0.84
	Sweden	64	2.08
	Switzerland	2	0.06
	The Netherlands	8	0.26
	Ukraine	3	0.10
	United Kingdom	3	0.10
Total		*3081*	

doi:10.1371/journal.pone.0033579.t002

relatively close to those estimated in other African countries, like 1966 for subtype C in Ethiopia [26], beginning of the 70's for Zimbabwe [27] or in the late 60's for Malawi [28]. As expected, we found that MRCA of subtype C in Senegal is not specific, because multiple introductions occurred, and our MRCA date estimate corresponds most likely to those of subtype C strains outside Senegal. In contrast to southern African countries, subtype C did not become the predominant strain in Senegal and did only spread efficiently in the MSM population, underlining the importance of high risk behavior in spread of viruses [29]. The MRCA of subtype C in the MSM population is estimated in the early 80's and is the result of a single introduction. This estimate coincides with the period where the HIV-1 C epidemic started a quick exponential-like growth phase in Senegal for nearly 15 years according to the Bayesian skyride analysis.

Our study showed also that analysis of alignments with or without codons that are associated with drug resistance did not have a significant impact on phylogenetic clustering or on MRCA date and substitution rate estimations. Among the different molecular clock models used, Bayes factors suggested the use of the relaxed exponential molecular clock above the most frequently used relaxed lognormal molecular clock. However, the very large confidence intervals and convergence problems with the exponential model with poorly informative priors, and the almost similar results with informative priors for both models are probably at the basis for the preferential use of the relaxed lognormal molecular clock model for HIV.

Previous studies suggest that subtype C could spread more efficiently due to the predominance of CCR5 variants or a stronger predisposition for localization in the female genital mucosa than other subtypes, which may facilitate both vertical and heterosexual transmission [30–33]. Increase of subtype C could also have implications on treatment because other subtype C specific mutations have been documented and commercial drug resistance assays cannot correctly test subtype C infections [2,34–

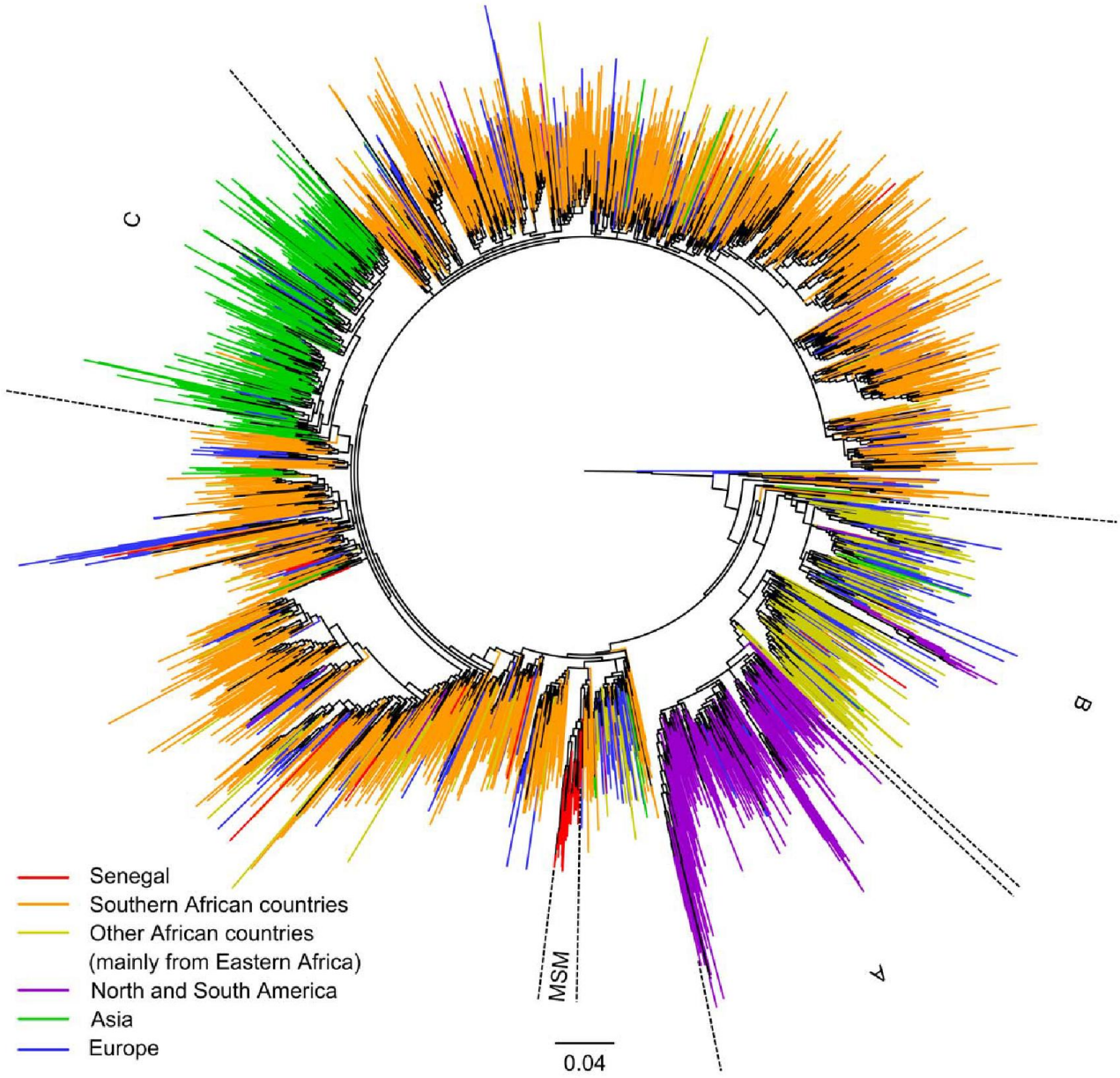


Figure 1. Maximum likelihood phylogenetic tree based on 3,081 HIV-1 subtype C *pol* sequences. Maximum likelihood (PhyML) phylogenetic tree based on 1,011 nucleotide sites of *pol* gene sequence (nucleotides 2,253–3,263 of HXB2 coordinates) from 3,081 HIV-1 subtype C isolates. Sequences were isolated in the countries shown in Table 2. Sequences are colored to their region of origin: Senegal in red, Southern African countries (South-Africa, Botswana, Malawi, Mozambique, Swaziland, Zambia and Zimbabwe) in orange, other African countries (mainly from the East) in yellow, North and South America in purple, Asia in green and Europe in blue. The branch support (aLRT) of clade A, B, C and MSM are of 73%, 76%, 82% and 99% respectively.
doi:10.1371/journal.pone.0033579.g001

36]. A cross-sectional study of women in Kenya indicated that women infected with subtype C had a higher viral load and lower CD4 counts than those infected with subtypes A and D, which could also have an impact on pathogenesis and transmission [37]. Therefore, it is important to continue to monitor HIV-1 subtype/CRF distribution among different population groups in Senegal. However, in order to be able to compare trends over time, such studies should be organized in a standardized way. For example, WHO proposed standardized protocols for surveillance of drug resistance mutations in recently infected individuals [38]. These studies can be combined with subtype/CRF characterization.

Because MSM reported having sex also with women, they could potentially serve as a bridge between high-risk men and low-risk women. This sexual mixing pattern might contribute in the future to the subsequent increase of subtype C in the general population. An increase from 4% in 2000 to almost 10% between 2000 and 2010 among the general population in Senegal has already been observed, and subtype C sequences recently obtained from HIV-1 C infected women in 2011 that cluster within the clade of strains from the MSM population have now been observed (Coumba Toure Kane, unpublished results). Understanding the origins and dispersal patterns of HIV-1 clades at regional and country levels is

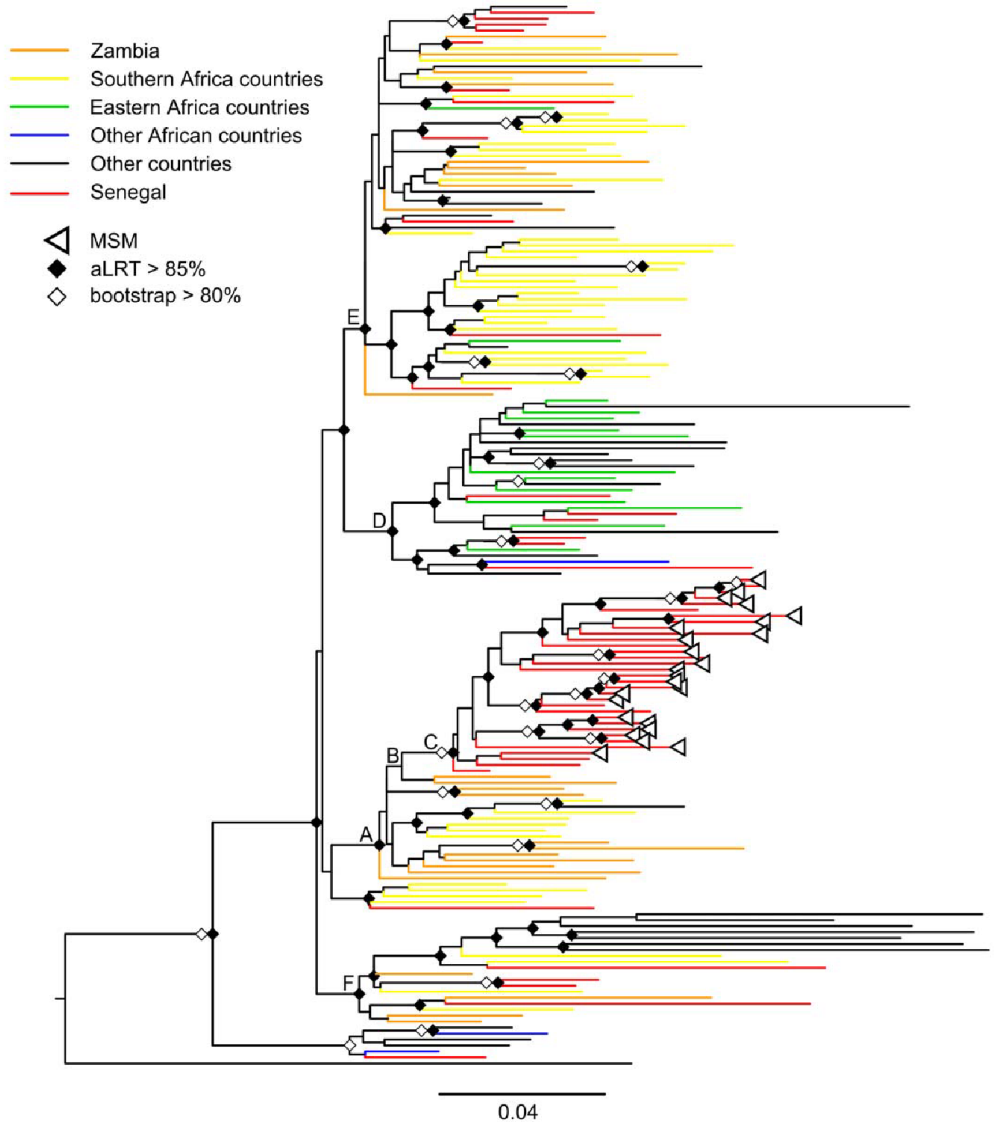


Figure 2. Maximum likelihood phylogenetic tree constructed from 56 HIV-1 C *pol* sequences from Senegal and 121 close relatives. Detailed maximum likelihood (PhyML) phylogenetic tree constructed using 1,011 nucleotide sites of *pol* gene sequence (nucleotides 2,253–3,263 of HXB2 coordinates) from 177 HIV-1 subtype C isolates from Senegal and close relatives (see text). Branch support values (bootstrap and aLRT) are displayed (see figure legend). Colors indicate the geographic origin and sequences were isolated in the following countries: 56 in red from Senegal, 25 in orange from Zambia, 49 in yellow from southern Africa (Botswana 6; Mozambique 5; Swaziland 2; South Africa 35; Zimbabwe 1), 12 in green from East Africa (Burundi 2; Ethiopia 9; Kenya 1; Sudan 2), 3 in blue from other African countries (DRC 1; Equatorial Guinea 1; Gabon 1) and 30 in black from European and Asian countries (Belgium 4; China 1; Germany 2; Denmark 1; Spain 5; France 1; Greece 1; Israel 1; India 1; Italia 1; Luxembourg 1; Norway 2; Portugal 2; Sweden 7).
doi:10.1371/journal.pone.0033579.g002

useful to improve the characterization and control of HIV spread. Continuous monitoring of HIV variants seems necessary to adapt treatment and vaccine strategies to be efficient against local and contemporary circulating HIV variants.

Materials and Methods

Nucleotide sequence dataset

In order to increase the number of sequences and to cover a wide geographic range, we used the *pol* region for our analysis. *Pol* sequences are highly studied because they are the target of antiretroviral drugs. A total of 56 subtype C *pol* gene sequences from Senegal were used in this study. Thirty-eight were obtained from the Los Alamos HIV sequence database (www.hiv.lanl.gov) from previously published reports and eighteen were newly characterized from ongoing molecular epidemiology and/or drug resistance studies mainly in Dakar, the capital city of Senegal (Table 1). We downloaded only sequences that were at least 1,000 nucleotides in length and spanning the genomic region which covers protease and majority of RT in *pol* between positions 2,253–3,263 on the HXB2 genome. Sequences were from blood samples collected between 1990 and 2009. In addition, all available subtype C sequences spanning the same genomic region and for which country of origin and sampling year were known, were also downloaded from the Los Alamos HIV database (www.hiv.lanl.gov). We then submitted all the sequences to the REGA

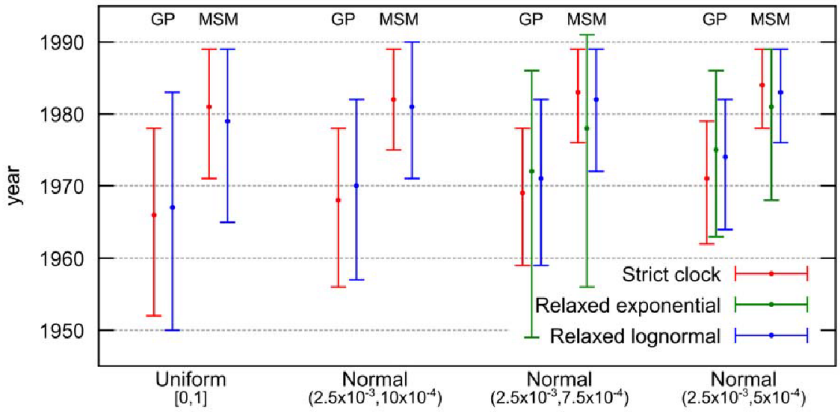


Figure 3. Dating the subtype C epidemic in general and MSM populations in Senegal. Coalescent based estimations (BEAST) and 95% highest posterior density (HPD) intervals of the MRCA dates of 56 HIV-1 subtype C *pol* sequences obtained from the general and the MSM population. Results are displayed for all tested substitution rate priors and molecular clock models, except for relaxed exponential with both less informative priors which provides very large 95% HPD intervals and shows convergence problems (see Table S3 for detailed results, including substitution rate estimations).
doi:10.1371/journal.pone.0033579.g003

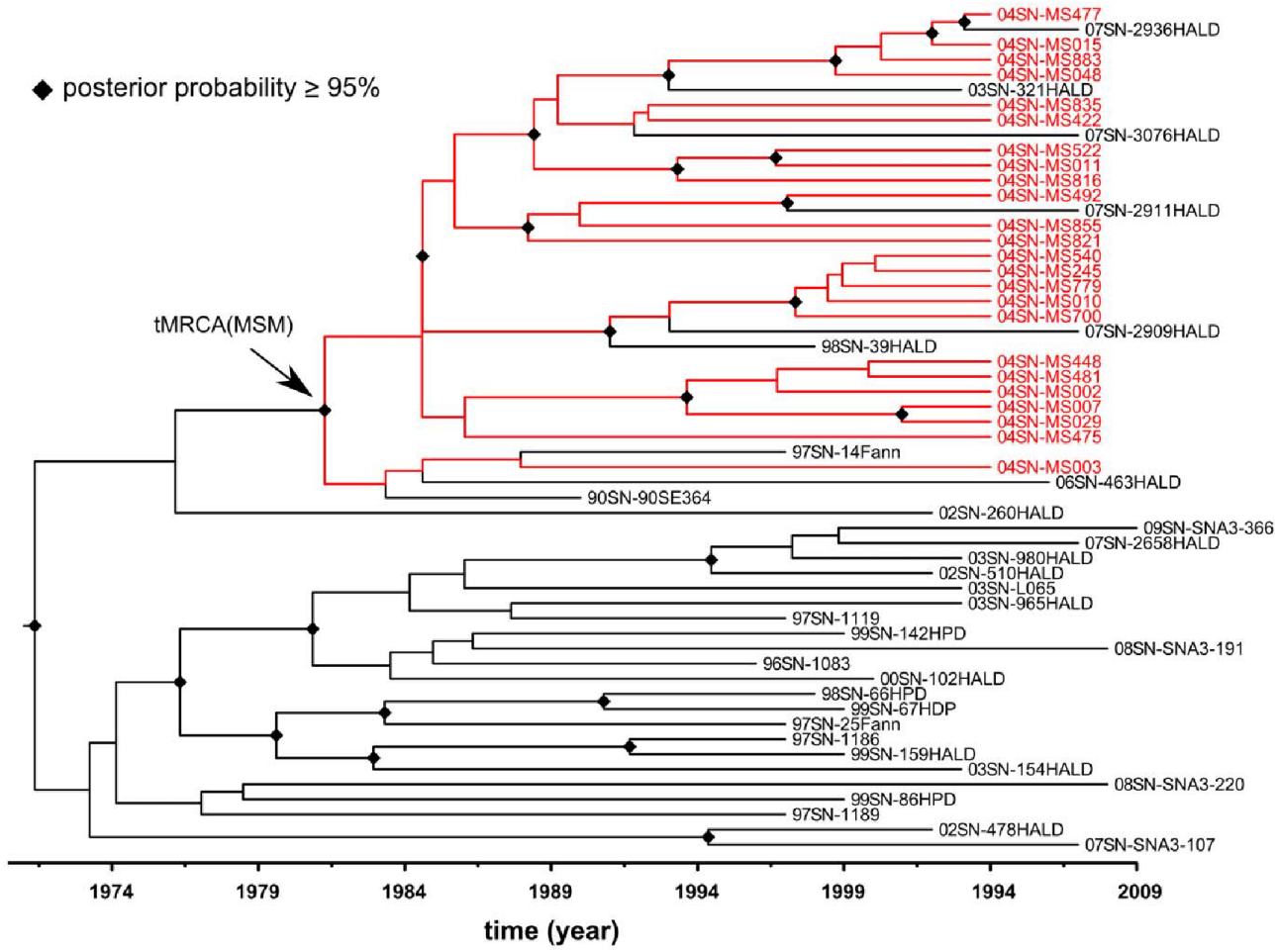


Figure 4. Bayesian tree with timescale of 56 HIV-1 C *pol* sequences from Senegal. Maximum clade credibility tree with time scale obtained with BEAST using 1,011 nucleotide sites of *pol* gene sequences (nucleotides 2,253–3,263 of HXB2 coordinates) from 56 HIV-1 subtype C isolates from Senegal. This tree is obtained using the relaxed uncorrelated lognormal molecular clock model and moderately informative substitution rate prior (Normal: 2.5×10^{-3},7.5×10^{-4}). Clades with posterior probabilities ≥95% are indicated by diamonds. MSM isolates are colored in red.
doi:10.1371/journal.pone.0033579.g004

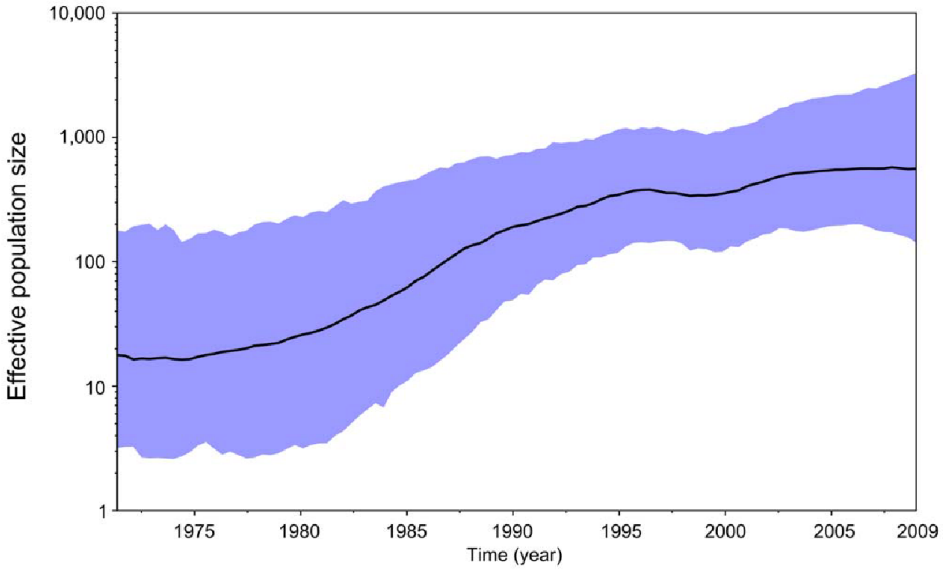


Figure 5. Bayesian skyride plot of HIV-1 C demographic growth in Senegal using 56 *pol* sequences. Estimates of HIV-1 C effective number of infections (N_e) over time from 56 Senegalese *pol* sequences using a Bayesian skyride plot in BEAST with relaxed uncorrelated lognormal molecular clock and moderately informative substitution rate prior (ucld.mean Normal: 2.5×10^{-3}, 7.5×10^{-4}). The X-axis represents the time in year. The Y-axis represents the HIV-1 effective number of infections ($\log_{10}$ scale). The black line marks the median estimate for N_e and the blue shadow region displays the 95% highest posterior density (HPD) interval.
doi:10.1371/journal.pone.0033579.g005

subtyping tool v.2 to confirm subtype assignments and to eliminate eventual intersubtype recombinants [39,40]. We selected one sequence per individual when sequential sequences were available or when sequences were epidemiologically linked by direct donor–recipient transmission.

HIV-1 pol sequencing

The 18 new HIV-1 *pol* sequences were obtained with an in-house technique as previously described [41]. Briefly, RNA was extracted using the QIAamp Viral RNA extraction kit (Qiagen SA, Courtabeauf, France) and processed for reverse transcription polymerase chain reaction (RT-PCR) with the integrase specific primer IN3 5′-TCTATBCCATCTAAAAATAGTACTTTCCT-GATTCC-3′ using the Expand reverse transcriptase (Roche Diagnostics, Meylan, France) according to the manufacturer's instructions. The resulting cDNA served as template in the subsequent nested PCR reaction during which a 1,865 base pairs fragment, corresponding to the protease and the first 440 amino acids of the reverse transcriptase region of the *pol* gene, was amplified with previously described primers and cycling conditions using the Expand Long Template PCR system (Roche Diagnostics, Meylan, France). The amplified HIV-1 nucleic acid fragments were purified using the Geneclean Turbo Kit (Q-Biogen, MPbiomedicals, France) and directly sequenced with primers encompassing the *pol* region using BigDye Terminator version 3.1 (Applied Biosystems, Courtaboeuf, France) according to the manufacturer's instructions. Electrophoresis and data collection were done on an Applied Biosystems 3130XL Genetic Analyzer. The sequenced fragments from both strands were reconstituted using Seqman II from the DNAstar package v5.08 (Lasergene, Madison, WI, USA).

Sequence alignment and phylogenetic tree analysis

The 18 newly obtained sequences were aligned with the alignment of subtype C sequences downloaded from the Los Alamos HIV database, using the L-INS-i method from MAFFT [42,43], and then manually edited with MEGA5 [44]. The HXB2 subtype B prototype strain was used as outgroup. In order to study potential bias due to drug-induced convergent evolution, all our analysis were also repeated on an alignment for which we removed 43 codon positions known to be associated with major resistance mutations according to the WHO-list of 2009 [45]. The following positions were excluded for protease (23, 24, 30, 32, 46, 47, 48, 50, 53, 54, 73, 76, 82, 83, 84, 85, 88, 90) and RT (41, 65, 67, 69, 70, 74, 75, 77, 100, 101, 103, 106, 115, 116, 151, 179, 181, 184, 188, 190, 210, 215, 219, 225, 230), leaving 882 nt in the final alignment. Both complete (1,011 nt) and restricted (882 nt) sequence alignments are available from the authors upon request. Maximum Likelihood phylogenies were inferred using the GTR+I+Γ4 nucleotide substitution model recommended by [46] and implemented in PhyML v3.0 [47]. The SPR option was selected to search the tree space and aLRT SH-like branch supports were used to assess confidence in topology [48]. The phylogenetic tree was drawn with FIGTREE (tree.bio.ed.ac.uk/software/figtree/).

In order to better determine and visualize the relationship of the subtype C sequences from Senegal to those from other geographic areas, another phylogenetic analysis was performed with less sequences. For this subtree, we collected from the large, previous phylogenetic tree, all descendant sequences of nodes that are first or second level ancestor of at least one sequence from Senegal (i.e., all Senegalese sequences plus their sisters and close relatives). A phylogeny was then inferred, using the same method and options as described above, but in addition to aLRT we ran a non-parametric bootstrap with 100 replicates to obtain a second assessment of branch supports. A phylogenetic analysis on this subset of sequences was also inferred using MrBayes v3.1 [49] with the same substitution model as for the maximum likelihood tree, and with chain length and tree sampling frequency of 5×10^7 and 1×10^4 generations, respectively. A burn-in of 2,000 sampled trees (i.e. ~40%) was selected. By the end of the run, the average

standard deviation of split frequencies was below 0.01 and the potential scale reduction factor of every parameter was in the range [0.999, 1.001], except the parameter pinvar which is at 1.002, proving the convergence of the Markov chains (see MrBayes manual).

Dating the introduction of subtype C in Senegal and MSM population

Estimates of the substitution rate and dates of the most recent common ancestor (MRCA) of subtype C in Senegal and in the sub-epidemic in MSM were obtained using BEAST v1.6.1 [50]. The 56 *pol* gene subtype C sequences from Senegal were analyzed under a GTR+I+Γ4 substitution process (as for phylogenetic analyzes). We used three different molecular clock models (strict clock, relaxed uncorrelated exponential and relaxed uncorrelated lognormal) [51] as implemented in BEAST with a Bayesian skyride tree prior as a coalescent demographic model with time-aware smoothing [52]. For the parameters of each molecular clock model (ucld.mean, uced.mean and clock.rate for the relaxed lognormal, relaxed exponential and strict molecular clock respectively) we tested a total of four different priors, one non-informative prior based on a uniform distribution (between 0.0 and 1.0) and three priors with varying information levels based on normal distribution with a mean of 2.5×10^{-3} (based on estimations from a previous study [27] in the same genomic region and as estimated by Path-O-Gen: tree.bio.ed.ac.uk/software/pathogen/) and standard deviations of 10×10^{-4}, 7.5×10^{-4}, and 5.0×10^{-4}, respectively. For the ucld.stdev parameter (representing the variability of the rates among branches for the relaxed lognormal molecular clock) we used a prior based on an exponential distribution with mean of 0.1 (personal communication with A. Drummond). MCMC simulations were run for 2.5×10^{8} chain steps with sub-sampling every 2.5×10^{5} steps. Convergence of the chains was inspected using Tracer v.1.5. For each tested prior and for each parameter, effective sample size (ESS) values were always above 300. The Bayes Factor was calculated to compare molecular clock models, using marginal likelihood as implemented in Tracer v.1.5. The Maximum Clade Credibility with time scale (MCC) tree was obtained by TreeAnnotator v1.6.1 with a burn-in of the first hundred trees.

Supporting Information

Figure S1 Maximum likelihood phylogenetic tree based on 3,081 HIV-1 subtype C *pol* sequences, without codons associated to drug resistance in PR and RT. Maximum likelihood phylogenetic tree (PhyML, with the same options as for the tree in Figure 1) based on 882 nucleotide sites of *pol* gene sequence from 3,081 HIV-1 subtype C isolates; nucleotide sites with coordinates 2,253–3,263 of HXB2 are included, but codon positions known to be associated with major resistance mutations according to the WHO-list of 2009 were removed (see Materials and Methods). Sequences were isolated in the countries shown in Table 2. Sequences are colored according to their region of origin: Senegal in red, Southern African countries (South-Africa, Botswana, Malawi, Mozambique, Swaziland, Zambia and Zimbabwe) in orange, other African countries (mainly from the East) in yellow, North and South America in purple, Asia in green and Europe in blue. The branch support (aLRT) of clades A, B, C and MSM are respectively of 94%, 92%, 83% and 96%.
(JPG)

Figure S2 Maximum likelihood phylogenetic tree constructed of 56 HIV-1 C *pol* sequences from Senegal and 121 close relatives. Detailed maximum likelihood (PhyML) phylogenetic tree constructed using 1,011 nucleotide sites of *pol* gene sequence (nucleotides 2,253–3,263 of HXB2 coordinates) from 177 HIV-1 subtype C isolates from Senegal and close relatives (see Materials and Methods) as shown in Figure 2 but names of the strains are added. Branch support values (bootstrap and aLRT) are displayed (see figure legend). Colors indicate the geographic origin and sequences were isolated in the following countries: 56 in red from Senegal, 25 in orange from Zambia, 49 in yellow from southern Africa (Botswana 6; Mozambique 5; Swaziland 2; South Africa 35; Zimbabwe 1), 12 in green from East Africa (Burundi 2; Ethiopia 9; Kenya 1; Sudan 2), 3 in blue from other African countries (DRC 1; Equatorial Guinea 1; Gabon 1) and 30 in black from European and Asian countries (Belgium 4; China 1; Germany 2; Denmark 1; Spain 5; France 1; Greece 1; Israel 1; India 1; Italia 1; Luxembourg 1; Norway 2; Portugal 2; Sweden 7).
(TIFF)

Figure S3 Bayesian phylogenetic tree of 56 HIV-1 C *pol* sequences from Senegal and 121 close relatives. Detailed Bayesian phylogenetic tree (MrBayes, same model and similar options as for the tree in Figure 2, see Materials and Methods) constructed using 1,011 nucleotide sites of *pol* gene sequence (nucleotides 2,253–3,263 of HXB2 coordinates) from 177 HIV-1 subtype C isolates from Senegal and close relatives. Clades with posterior probabilities ≥95% are shown. Colors indicate the geographic origin of the sequences, which were isolated in the following countries: 56 in red from Senegal, 25 in orange from Zambia, 49 in yellow from southern Africa (Botswana 6; Mozambique 5; Swaziland 2; South Africa 35; Zimbabwe 1), 12 in green from East Africa (Burundi 2; Ethiopia 9; Kenya 1; Sudan 2), 3 in blue from other African countries (DRC 1; Equatorial Guinea 1; Gabon 1) and 30 in black from European and Asian countries (Belgium 4; China 1; Germany 2; Denmark 1; Spain 5; France 1; Greece 1; Israel 1; India 1; Italia 1; Luxembourg 1; Norway 2; Portugal 2; Sweden 7).
(TIFF)

Figure S4 Bayesian tree with timescale of 56 HIV-1 C *pol* sequences from Senegal, without sites associated to major, known resistance in PR and RT. Maximum clade credibility tree with time scale obtained with BEAST using 1,011 nucleotide sites of *pol* gene sequences (nucleotides 2,253–3,263 of HXB2 coordinates) from 56 HIV-1 subtype C isolates from Senegal. This tree is obtained using the relaxed uncorrelated lognormal molecular clock model and moderately informative substitution rate prior (Normal: 2.5×10^{-3}, 7.5×10^{-4}). Clades with posterior probabilities ≥95% are indicated by diamonds. MSM isolates are colored in red.
(TIFF)

Table S1 Genbank accession numbers per country of subtype C HIV-1 strains included in the study.
(DOC)

Table S2 Details of the strains included in the restricted phylogenetic tree analysis from Figures 2, S2 and S3.
(PDF)

Table S3 Dating the subtype C epidemic in general and MSM populations in Senegal. Coalescent based estimations (BEAST) and 95% highest posterior density (HPD) intervals of the MRCA dates and substitution rates of 56 HIV-1 subtype C *pol* sequences obtained from the general and the MSM population. Results are displayed for all tested substitution rate priors and molecular clock models.
(PDF)

Author Contributions

Conceived and designed the experiments: MP OG. Performed the experiments: MJ NL NV. Analyzed the data: MJ NL NV DF HD CTK OG MP. Contributed reagents/materials/analysis tools: HD CTK. Wrote the paper: MJ DF OG MP.

References

1. Thomson MM, Pérez-Alvarez L, Nájera R (2002) Molecular epidemiology of HIV-1 genetic forms and its significance for vaccine development and therapy. Lancet Infect Dis 2: 461–71. Review.
2. Peeters M, Aghokeng AF, Delaporte E (2010) Genetic diversity among human immunodeficiency virus-1 non-B subtypes in viral load and drug resistance assays. Clin Microbiol Infect 16: 1525–31. Review.
3. Gamble LJ, Matthews QL (2010) Current progress in the development of a prophylactic vaccine for HIV-1. Drug Des Devel Ther 5: 9–26. Review.
4. Tebit DM, Arts EJ (2011) Tracking a century of global expansion and evolution of HIV to drive understanding and to combat disease. Lancet Infect Dis 11: 45–56. Review.
5. Vidal N, Peeters M, Mulanga-Kabeya C, Nzilambi N, Robertson D, et al. (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. J Virol 74: 10498–507.
6. Rambaut A, Robertson DL, Pybus OG, Peeters M, Holmes EC (2001) Human immunodeficiency virus. Phylogeny and the origin of HIV-1. Nature 410: 1047–8.
7. Worobey M, Gemmel M, Teuwen DE, Haselkorn T, Kunstman K, et al. (2008) Direct evidence of extensive diversity of HIV-1 in Kinshasa by 1960. Nature 455: 661–4.
8. Hemelaar J, Gouws E, Ghys PD, Osmanov S, WHO-UNAIDS Network for HIV Isolation and Characterisation (2011) Global trends in molecular epidemiology of HIV-1 during 2000–2007. AIDS 25: 679–89.
9. Barin F, M'Boup S, Denis F, Kanki P, Allan JS, et al. (1985) Serological evidence for virus related to simian T-lymphotropic retrovirus III in residents of west Africa. Lancet 2: 1387–9.
10. Hamel DJ, Sankalé JL, Eisen G, Meloni ST, Mullins C, et al. (2007) Twenty years of prospective molecular epidemiology in Senegal: changes in HIV diversity. AIDS Res Hum Retroviruses 23: 1189–96.
11. UNAIDS website. Available: www.unaids.org/en/regionscountries/countries/senegal/. Accessed 2011 Aug 23.
12. Toure-Kane C, Montavon C, Faye MA, Gueye PM, Sow PS, et al. (2000) Identification of all HIV type 1 group M subtypes in Senegal, a country with low and stable seroprevalence. AIDS Res Hum Retroviruses 16: 603–9.
13. Ayouba A, Lien TT, Nouhin J, Vergne L, Aghokeng AF, et al. (2009) Low prevalence of HIV type 1 drug resistance mutations in untreated, recently infected patients from Burkina Faso, Côte d'Ivoire, Senegal, Thailand, and Vietnam: the ANRS 12134 study. AIDS Res Hum Retroviruses 25: 1193–6.
14. Diop-Ndiaye H, Toure-Kane C, Leye N, Ngom-Gueye NF, Montavon C, et al. (2010) Antiretroviral drug resistance mutations in antiretroviral-naive patients from Senegal. AIDS Res Hum Retroviruses 26: 1133–8.
15. Ndiaye HD, Toure-Kane C, Vidal N, Niama FR, Niang-Diallo PA, et al. (2009) Surprisingly high prevalence of subtype C and specific HIV-1 subtype/CRF distribution in men having sex with men in Senegal. J Acquir Immune Defic Syndr 52: 249–52.
16. Soares EA, Martinez AM, Souza TM, Santos AF, Da Hora V, et al. (2005) HIV-1 subtype C dissemination in southern Brazil. AIDS 19: Suppl 4S81–86.
17. Wade AS, Kane CT, Diallo PAN, Diop AK, Gueye K, et al. (2005) HIV infection and sexually transmitted infections among men who have sex with men in Senegal. AIDS 19: 2133–2140.
18. Thomson MM, Fernández-García A (2011) Phylogenetic structure in African HIV-1 subtype C revealed by selective sequential pruning. Virology 415: 30–8.
19. Fontella R, Soares MA, Schrago CG (2008) On the origin of HIV-1 subtype C in South America. AIDS 22: 2001–11.
20. Bello G, Passaes CP, Guimarães ML, Lorete RS, Matos Almeida SE, et al. (2008) Origin and evolutionary history of HIV-1 subtype C in Brazil. AIDS 22: 1993–2000.
21. de Oliveira T, Pillay D, Gifford RJ, UK Collaborative Group on HIV Drug Resistance (2010) The HIV-1 subtype C epidemic in South America is linked to the United Kingdom. PLoS One 5(2): e9311.
22. Véras NM, Gray RR, Brígido LF, Rodrigues R, Salemi M (2011) High-resolution phylogenetics and phylogeography of human immunodeficiency virus type 1 subtype C epidemic in South America. J Gen Virol 92: 1698–709.
23. Kane F, Alary M, Ndoye I, Coll AM, M'boup S, et al. (1993) Temporary expatriation is related to HIV-1 infection in rural Senegal. AIDS 9: 1261–5.
24. Kanki PJ, Peeters M, Gueye-Ndiaye A (1997) Virology of HIV-1 and HIV-2: implications for Africa. AIDS 11 Suppl B: S33–4.
25. Kanki PJ, Hamel DJ, Sankalé JL, Hsieh C, Thior I, et al. (1999) Human immunodeficiency virus type 1 subtypes differ in disease progression. J Infect Dis 179: 68–73.
26. Tully DC, Wood C (2010) Chronology and evolution of the HIV-1 subtype C epidemic in Ethiopia. AIDS 24: 1577–82.
27. Dalai SC, de Oliveira T, Harkins GW, Kassaye SG, Lint J, et al. (2009) Evolution and molecular epidemiology of subtype C HIV-1 in Zimbabwe. AIDS 23: 2523–32.
28. Travers SA, Clewley JP, Glynn JR, Fine PE, Crampin AC, et al. (2004) Timing and reconstruction of the most recent common ancestor of the subtype C clade of human immunodeficiency virus type 1. J Virol 78: 10501–6.
29. McDaid LM, Hart GJ (2010) Sexual risk behaviour for transmission of HIV in men who have sex with men: recent findings and potential interventions. Curr Opin HIV AIDS 5: 311–5. Review.
30. Abraha A, Nankya IL, Gibson R, Demers K, Tebit DM, et al. (2009) CCR5-and CXCR4-tropic subtype C human immunodeficiency virus type 1 isolates have a lower level of pathogenic fitness than other dominant group M subtypes: implications for the epidemic, J Virol 83: 5592–5605.
31. Ball SC, Abraha A, Collins KR, Marozsan AJ, Baird H, et al. (2003) Comparing the ex vivo fitness of CCR5-tropic human immunodeficiency virus type 1 isolates of subtypes B and C. J Virol 77: 1021–38.
32. Renjifo B, Gilbert P, Chaplin B, Msamanga G, Mwakagile D, et al. (2004) Preferential in-utero transmission of HIV-1 subtype C as compared to HIV-1 subtype A or D. AIDS 18: 1629–1636.
33. John-Stewart GC, Nduati RW, Rousseau CM, Mbori-Ngacha DA, Richardson BA, et al. (2005) Subtype C is associated with increased vaginal shedding of HIV-1. J Infect Dis 192: 492–496.
34. Martinez-Cajas JL, Pai NP, Klein MB, Wainberg MA (2009) Differences in resistance mutations among HIV-1 non-subtype B infections: A systematic review of evidence (1996–2008). J Int AIDS Soc 12: 11.
35. Vergne L, Snoeck J, Aghokeng A, Maes B, Valea D, et al. (2006) Genotypic drug resistance interpretation algorithms display high levels of discordance when applied to non-B strains from HIV-1 naive and treated patients. FEMS Immunol Med Microbiol 46: 53–62.
36. Snoeck J, Kantor R, Shafer RW, Van Laethem K, Deforche K, et al. (2006) Discordances between interpretation algorithms for genotypic resistance to protease and reverse transcriptase inhibitors of human immunodeficiency virus are subtype dependent. Antimicrob Agents Chemother 50: 694–701.
37. Neilson JR, John GC, Carr JK, Lewis P, Kreiss JK, et al. (1999) Subtypes of human immunodeficiency virus type 1 and disease stage among women in Nairobi, Kenya. J Virol 73: 4393–4403.
38. Bennett DE, Bertagnolio S, Sutherland D, Gilks CF (2008) The World Health Organization's global strategy for prevention and assessment of HIV drug resistance. Antivir Ther 13 Suppl 2: 1–13.
39. de Oliveira T, Deforche K, Cassol S, Salminen M, Paraskevis D, et al. (2005) An automated genotyping system for analysis of HIV-1 and other microbial sequences. Bioinformatics 21: 3797–800.
40. Alcantara LC, Cassol S, Libin P, Deforche K, Pybus OG, et al. (2009) Standardized framework for accurate, high-throughput genotyping of recombinant and non-recombinant viral sequences. Nucleic Acids Res 37: 634–42.
41. Vergne L, Diagbouga S, Kouanfack C, Aghokeng A, Butel C, et al. (2006) HIV-1 drug-resistance mutations among newly diagnosed patients before scaling-up programmes in Burkina Faso and Cameroon. Antivir Ther 11: 575–9.
42. Katoh K, Misawa K, Kuma K, Miyata T (2002) MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic Acids Res 30: 3059–3066.
43. Katoh K, Kuma K, Toh H, Miyata T (2005) MAFFT version 5: improvement in accuracy of multiple sequence alignment. Nucleic Acids Res 33: 511–8.
44. Tamura K, Peterson D, Peterson N, Stecher G, Nei M, et al. (2011) MEGA5: Molecular Evolutionary Genetics Analysis using Maximum Likelihood, Evolutionary Distance, and Maximum Parsimony Methods. Mol Biol Evol 28: 2731–9.
45. Bennett DE, Camacho RJ, Otelea D, Kuritzkes DR, Fleury H, et al. (2009) Drug resistance mutations for surveillance of transmitted HIV-1 drug-resistance: 2009 update. PLoS One. e4724 p.
46. Posada D, Crandall KA (2001) Selecting models of nucleotide substitution: an application to human immunodeficiency virus 1 (HIV-1). Mol Biol Evol 18: 897–906.
47. Guindon S, Dufayard JF, Lefort V, Anisimova M, Hordijk W, et al. (2010) New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. Syst Biol 59: 307–21.
48. Anisimova M, Gascuel O (2006) Approximate likelihood-ratio test for branches: A fast, accurate, and powerful alternative. Syst Biol 55: 539–52.
49. Ronquist F, Huelsenbeck JP (2003) MrBayes 3: Bayesian phylogenetic inference under mixed models. Bioinformatics 19: 1572–1574.
50. Drummond AJ, Rambaut A (2007) BEAST: Bayesian evolutionary analysis by sampling trees. BMC Evol Biol 7: 214.
51. Drummond AJ, Ho SYW, Phillips MJ, Rambaut A (2006) Relaxed phylogenetics and dating with confidence. PLoS Biol 4: e88.

52. Minin VN, Bloomquist EW, Suchard MA (2008) Smooth skyride through a rough skyline: Bayesian coalescent-based inference of population dynamics. Mol Biol Evol 25: 1459–71.
53. Brown BK, Darden JM, Tovanabutra S, Oblander T, Frost J, et al. (2005) Biologic and genetic characterization of a panel of 60 human immunodeficiency virus type 1 isolates, representing clades A, B, C, D, CRF01_AE, and CRF02_AG, for the development and assessment of candidate vaccines. J Virol 79: 6089–101.
54. Vergne L, Kane CT, Laurent C, Diakhaté N, Gueye NF, et al. (2003) Low rate of genotypic HIV-1 drug-resistant strains in the Senegalese government initiative of access to antiretroviral therapy. AIDS 17 Suppl 3: S31–8.
55. Vergne L, Peeters M, Mpoudi-Ngole E, Bourgeois A, Liegeois F, et al. (2000) Genetic diversity of protease and reverse transcriptase sequences in non-subtype-B human immunodeficiency virus type 1 strains: evidence of many minor drug resistance mutations in treatment-naïve patients. J Clin Microbiol 38: 3919–25.

Chapitre 6

Histoire épidémiologique du sous-type C du VIH-1 dans la pandémie mondiale

Le sous-type C du virus de l'immunodéficience humaine de type 1 (VIH-1) est responsable de près de 50% des infections mondiales au VIH-1, mais il est surtout prévalent en Afrique australe, en Afrique de l'est, en Inde et au sud du Brésil. Certaines études d'épidémiologie moléculaire montrent que ce variant génétique s'est propagé au Brésil et en Inde à partir du Burundi et de l'Afrique du Sud respectivement. Ces études se basent systématiquement sur un échantillon réduit des souches disponibles et les migrations du sous-type C au sein même de l'Afrique restent mal connues. Nous proposons une étude visant à connaître l'origine géographique de l'épidémie du sous-type C, ainsi que ses migrations dans le monde entier, incluant l'Afrique, en utilisant toutes les souches disponibles de ce variant plus 528 souches séquencées par l'équipe TransVIHMI. La phylogénie obtenue, comprenant plus de 3 600 souches, est difficilement interprétable « à la main ». Plusieurs indices basés sur les transitions entre pays (reconstruites par parcimonie) sont proposés afin de donner une vision synthétique des flux migratoires du sous-type C à l'échelle mondiale. Le logiciel PhyloType est ensuite utilisé pour mettre en valeur des liens entre les évènements fondateurs probables. La plupart des flux migratoires du sous-type C décrits dans la littérature sont observés, par exemple le lien entre le Brésil et le Burundi, et d'autres sont différents, par exemple, la Zambie est suggérée être à l'origine de l'épidémie en Inde. En Afrique, ce variant se propage indépendamment de la Zambie, épicentre de l'épidémie, vers l'Afrique australe et vers l'Afrique de l'est.

Sommaire

6.1 Introduction

Les erreurs lors de la rétrotranscription, les phénomènes de recombinaison, la pression de sélection immunitaire et médicamenteuse, un fort taux de réplication virale ont donné lieu à de nombreuses variantes génétiques du virus de l'immunodéficience humaine (VIH) que l'on nomme sous-types ou formes recombinantes circulantes (*circulating recombinant forms*, CRF) (Rambaut *et al*, 2004). Ceci est le résultat de l'adaptation du virus à son environnement (Brun-Vézinet *et al*, 1999).

De ce fait, les souches du groupe pandémique du VIH-1 (groupe M) présentent une diversité génétique importante. Elles sont répertoriées dans 9 sous-types (A à D, F à H, J et K), 6 sous-sous-types (A1 à A4, F1 et F2) et 51 CRF (CRF01_AE à CRF51_01B) ; sans compter les nombreuses formes recombinantes uniques découvertes (*unique recombinant forms*, URF). Le variant génétique du groupe M le plus répandu est, sans conteste, le sous-type C (VIH-1C). Ce sous-type est responsable de 48,23% des infections mondiales liées au VIH-1[8], mais sa distribution géographique est hétérogène et en constante évolution. Cette distribution géographique hétérogène est la résultante de nombreux facteurs, tant biologiques que sociologiques (Perrin *et al*, 2003). En effet, le VIH-1C a une virulence moindre par rapport à celle des autres sous-types (Abraha *et al*, 2009) et donc une phase asymptomatique plus longue, laissant plus d'opportunité de transmission (Ariën *et al*, 2007). Le sous-type C a aussi une prédisposition plus élevée à se localiser dans les muqueuses génitales des femmes (Walter *et al*, 2009) favorisant ainsi la transmission par contact hétérosexuel.

L'épidémie du VIH en Afrique australe, en Éthiopie et en Inde est presque exclusivement due au sous-type C. Pour ces pays, les études épidémiologiques estiment respectivement que 98,31%,

[8] D'après Hemelaar *et al*. (2011), sur la période 2004-2007.

97,44% et 97,77% des infections au VIH-1 sont dues au sous-type C[8]. Elles ont aussi montré que deux épidémies différentes du sous-type C (C et C') co-circulent en Éthiopie (Abebe *et al*, 2000), probablement d'origine différente. Dans le reste de l'Afrique, le sous-type C est aussi observé à l'est, où il est responsable de 22,97% des infections (Éthiopie exclue)[8], avec une forte prévalence au Burundi, où il est responsable de plus de 80% des infections (Vidal *et al*, 2007; Koch *et al*, 2001), et au centre où il est responsable de 5,75% des infections[8], notamment à Lubumbashi et à Mbuji-Mayi, deux villes situées au sud de la République Démocratique du Congo (RDC), où il est responsable de 51,3% et de 16,3% des infections respectivement (Vidal *et al*, 2005, 2000). À l'ouest et au nord de l'Afrique le sous-type C est assez rare (<1%[8]). Toutefois, il est retrouvé avec une prévalence de 40% chez les MSM sénégalais (Ndiaye *et al*, 2009) (cf. Chapitre 5). Sur le continent américain, le sous-type C est surtout prévalant au sud du Brésil (Soares *et al*, 2005) et il est aussi observé dans quelques autres pays voisins, comme l'Argentine ou l'Uruguay (Carrion *et al*, 2004). En Amérique du nord et centrale des cas sont observés mais restent rares (Sides *et al*, 2005; Cuevas *et al*, 2002). En Asie (sauf Inde), la prévalence du sous-type C est faible (2,93%[8,9]), mais elle est élevée en Océanie (particulièrement aux îles Fidji (Ryan *et al*, 2009) et en Papouasie-Nouvelle-Guinée (Ryan *et al*, 2007)) où le VIH-1C est responsable de 66,34% des infections[8]. En Europe, le sous-type C est très peu prévalant et est souvent observé chez des patients qui ont des liens avec l'Afrique (Paraschiv *et al*, 2011; Giuliani *et al*, 2009; Vercauteren *et al*, 2008; Tatt *et al*, 2004; Couturier *et al*, 2000; Alaeus *et al*, 1997).

L'émergence ou l'introduction d'un nouveau variant dans une population donnée, où un autre variant génétique est déjà prédominant, peut provoquer des phénomènes de recombinaison lors des co- ou surinfections. C'est le cas de l'introduction du sous-type C chez les utilisateurs de drogues intraveineuses (*intravenous drug users*, IDU) en Asie de l'est, où le sous-type B était prédominant, qui engendra l'épidémie des CRF08_BC et CRF07_BC (Takebe *et al*, 2010). Au sud du Brésil, l'introduction du sous-type C a produit l'épidémie du CRF31_BC (Passaes *et al*, 2009).

L'origine et la diffusion des virus restent d'un intérêt majeur pour les épidémiologistes, car la diversité génétique peut avoir des conséquences sur l'efficacité d'outils diagnostiques (sérologiques et/ou moléculaires), le développement de résistances aux antirétroviraux, la pathogénicité virale ou la possibilité de développement d'un vaccin. Depuis l'avènement de la phylogénie moléculaire, de nombreuses méthodes permettent aujourd'hui de répondre aux questions relatives à la dynamique des épidémies, sur la base des séquences nucléotidiques. La plupart de ces méthodes utilisent ou infèrent un arbre phylogénétique et déduisent, à partir de celui-ci, les régions géographiques correspondantes aux nœuds ancestraux de l'arbre connaissant celles associées aux souches contempo-

[9] Ce chiffre prend en considération quelques pays de l'Europe de l'est.

raines (représentées par les feuilles dans l'arbre). Par exemple, Véras *et al*. (2011a) utilisent le principe de parcimonie (décrit à la fin du Chapitre 1) et Faria *et al*. (2011) utilisent une méthode bayésienne implémentée dans la suite de logiciels BEAST. Cette branche de la phylogénie moléculaire porte le nom de phylogéographie dont Avise (2000) donne la définition suivante : « *field of study concerned with the principles and processes governing the geographical distributions of geographical lineages, especially those within and among closely related species* ».

Des études phylogénétiques de ce type décrivent déjà les migrations de l'épidémie du VIH-1C (de Oliveira *et al*, 2010; Bello *et al*, 2008; Fontella *et al*, 2008; Qiu *et al*, 2005; Gehring *et al*, 1997; Dietrich *et al*, 1995). L'épidémie du sous-type C en Afrique du Sud s'est propagée en Inde suite à un évènement fondateur d'origine inconnue (Shen *et al*, 2011; Dietrich *et al*, 1995, 1993), puis s'est introduite en Chine où quelques souches virales se sont recombinées avec du sous-type B (plus précisément avec un sous-cluster particulier du B, appelé B' ou Thai B) pour former les recombinants CRF08_BC et CRF07_BC qui circulent chez les IDU (Qiu *et al*, 2005). L'épidémie du sous-type C qui sévit en Éthiopie s'est répandue en Israël en 1991 à la suite de l'opération *Salomon* qui permit à plus de 14 000 juifs d'Éthiopie de rejoindre l'Israël (Gehring *et al*, 1997). Notons que ce variant est resté endémique à cette communauté sur le territoire israélien. En Amérique du sud, l'épidémie du sous-type C a pour centre de dispersion le sud du Brésil (Véras *et al*, 2011a; de Oliveira *et al*, 2010; Jones *et al*, 2009; Bello *et al*, 2008; Fontella *et al*, 2008), puis s'est répandue en Argentine et en Uruguay par le biais d'immigrants et de touristes (Carrion *et al*, 2004). Dans la littérature, l'introduction du sous-type C sur le territoire d'Amérique Latine a connu plusieurs origines géographiques différentes. Cependant l'hypothèse la plus citée reste celle de l'Afrique de l'est (le Burundi est souvent mentionné). Récemment, de Oliveira *et al*. (2010) suggèrent que l'épidémie s'est propagée du Burundi vers l'Angleterre puis de l'Angleterre vers le Brésil, mais Véras *et al*. (2011a) apportent une controverse à cette théorie. En revanche, toutes ces études s'accordent sur le fait que l'origine de l'épidémie du VIH-1C en Amérique du sud est monophylétique (ou avec un nombre faible d'introductions marginales).

Malgré le nombre grandissant d'études de ce genre, à notre connaissance, aucune ne montre ou ne discute de l'origine exacte de l'épidémie du VIH-1C, qui, au vu des informations relevées dans la littérature, semble être en Afrique. De plus, aucune donnée générale sur le mouvement de l'épidémie du sous-type C en Afrique n'est disponible. Seule observation à noter, les souches collectées en Afrique de l'est se regroupent dans un cluster (Thomson & Fernández-García, 2011).

La plupart de ces études moléculaires utilisent seulement une partie des souches disponibles. En effet, ces études utilisent généralement des méthodes probabilistes, lourdes en temps de calcul, qui

permettent rarement de dépasser quelques centaines de séquences. Cependant, la sélection de souches peut introduire un biais dans les résultats. L'exemple du rôle de l'Angleterre dans l'origine de l'épidémie du VIH-1C en Amérique du sud est caractéristique : de Oliveira *et al*. (2010) se sont restreint à certaines séquences et en ont déduit que l'Angleterre a eu un rôle dans la diffusion de l'épidémie au Brésil. L'étude de Véras *et al*. (2011a) utilise les mêmes souches d'Angleterre mais avec une couverture plus large de souches du VIH-1C collectées en Amérique du sud et trouve des conclusions différentes.

Nous présentons ici la première étude de phylogénie moléculaire visant à retracer l'histoire épidémiologique mondiale et l'origine du VIH-1C. Pour ce faire, nous inférons une phylogénie contenant toutes les souches du VIH-1C disponibles, sans aucune sélection aléatoire ou arbitraire. Elles proviennent de la base de données sur le VIH du laboratoire de Los Alamos, soit 3 081 séquences utilisées dans cette étude (cf. Chapitre 5), auxquelles nous ajoutons 528 séquences collectées en Afrique, continent suspecté d'être à l'origine de l'épidémie du VIH en général et donc aussi du sous-type C. À l'aide de cette phylogénie et de la connaissance de l'origine géographique des souches nos objectifs sont de : 1) déterminer l'origine géographique de l'épidémie du VIH-1C ; 2) connaître les voies ayant mené à la diffusion de ce variant à travers le monde. Toutefois, une telle phylogénie, associée aux origines géographiques des séquences, est très difficilement interprétable de façon manuelle et nécessite l'utilisation d'outils ou logiciels permettant les analyses phylogénétiques d'un grand nombre de souches dans un temps relativement court. Aussi, nous proposons deux manières différentes mais complémentaires pour analyser ces données, utilisant toutes deux le principe de parcimonie (détaillé à la fin du Chapitre 1). La première utilise des indices basés sur les transitions entre pays (reconstruites par parcimonie) pour donner une information synthétique retraçant les grandes tendances (pays donneurs ou receveurs, symétrie des échanges, etc.). Ces indices sont proches de ceux proposés par Slatkin et Maddison (1989) ou encore Salemi *et al*. (2005) aussi définis à partir du nombre de transitions entre annotations. Associés à des sorties graphiques appropriées, ces indices permettent très rapidement de se faire une idée globale sur l'épidémie. La seconde approche est basée sur l'outil PhyloType (Chevenet, Jung, de Oliveira et Gascuel, en cours de soumission) qui permet d'identifier des évènements fondateurs probables, comme par exemple l'introduction du VIH-1C au Brésil ou chez les MSM du Sénégal. La section suivante présente les méthodes et logiciels utilisés pour la préparation des données, la définition des indices utilisés permettant de synthétiser l'information de la phylogénie, une présentation sommaire du logiciel PhyloType et enfin son paramétrage. Les deux dernières sections présentent respectivement les résultats obtenus et leurs discutions.

6.2 Préparation des données

6.2.1 Conception de l'alignement

Pour cette étude, nous réutilisons l'alignement du Chapitre 5 contenant tous les sites, même ceux associés à des mutations de résistance (Hué *et al*, 2004). Pour mémoire, il contient 3 081 séquences du sous-type C du VIH-1, collectées à travers le monde et à différentes dates, d'une longueur de 1 011 sites et correspondant à la région génomique 2 253-3 263 (*pol*) d'HXB2 (codant l'intégralité de la protéase et le début de la transcriptase inverse).

À cet alignement, 528 nouvelles séquences, collectées et séquencées par l'équipe TransVIHMI, sont ajoutées. Parmi ces 528 séquences, 199 sont collectées au Burundi en 2008 et 20 en 2010, 1 en République Démocratique du Congo (RDC) en 2007 et 66 en 2008, 1 en République Centrafricaine en 2006, 2 en République du Congo en 2007, 1 en Éthiopie en 1999 et 238 au Swaziland en 2008. Ces séquences sont toutes confirmées comme appartenant au sous-type C soit par l'application web RE-GA HIV-1 & 2 *Automated Subtyping Tool* (de Oliveira *et al*, 2005), soit à l'aide d'analyses de similarité et *bootscan* réalisés avec le logiciel SimPlot (Lole *et al*, 1999). La séquence HXB2 (sous-type B ; numéro d'accession : K03455) sert d'*outgroup* pour enraciner l'arbre de maximum de vraisemblance construit dans cette étude.

Un alignement séquences contre profil est effectué afin d'ajouter les 528 nouvelles séquences à l'alignement initial de 3 081 séquences. La méthode d'alignement et le logiciel utilisés sont les mêmes qu'au Chapitre 5, à savoir MAFFT version 6 (Katoh *et al*, 2002) avec la méthode L-INS-i (Katoh *et al*, 2005). Quelques corrections manuelles sont apportées avec MEGA version 5 (Tamura *et al*, 2011) et les sites contenant un nombre excessif de gaps ($\geq 50\%$) sont supprimés.

Au final, nous obtenons un alignement de 1 011 sites contenant 3 609 séquences collectées dans 63 pays différents entre 1986 et 2010. Le Tableau 4 liste le nombre de souches pour chaque pays présents dans cette étude.

6.2.2 Inférence phylogénétique

Les mêmes paramètres et les mêmes options sont utilisés pour calculer l'arbre de maximum de vraisemblance, que pour celui contenant les 3 081 souches du Chapitre 5. Pour rappel, il est inféré sous le modèle *general time reversible* (GTR) avec des sites invariants et une loi gamma discrète avec 4 catégories de taux (GTR+I+Γ4), comme conseillé par Posada et Crandall (2001), avec le logiciel PhyML v3.0 (Guindon *et al*, 2010). L'option *subtree pruning and regrafting* (SPR) est choisie pour explorer l'espace des arbres. Tous les paramètres sont évalués et optimisés par PhyML. Les supports

de branche sont déterminés par la méthode *approximate likelihood ratio test* (aLRT) (Anisimova & Gascuel, 2006), option SH-like.

Tableau 4. Liste des pays utilisés dans cette étude, ainsi que le nombre de séquences associées en nombre et en pourcentage.

Afrique				**2 615**	**72,46%**
Afrique du Sud	689	19,09%	Mozambique	98	2,72%
Botswana	133	3,69%	Niger	4	0,11%
Burundi	310	8,59%	Ouganda	16	0,44%
Congo	2	0,06%	Rép. Centrafricaine	1	0,03%
Djibouti	1	0,03%	Rép. Démo. du Congo	86	2,38%
Érythrée	2	0,06%	Sénégal	56	1,55%
Éthiopie	100	2,77%	Somalie	1	0,03%
Gabon	1	0,03%	Soudan	10	0,28%
Guinée Équatoriale	1	0,03%	Swaziland	285	7,90%
Kenya	4	0,11%	Tanzanie	82	2,27%
Malawi	71	1,97%	Zambie	633	17,54%
Mali	1	0,03%	Zimbabwe	28	0,76%
Amérique				**299**	**8,28%**
Argentine	8	0,22%	Honduras	1	0,03%
Brésil	253	7,01%	Uruguay	2	0,06%
Cuba	25	0,69%	Venezuela	1	0,03%
États-Unis	9	0,80%			
Asie				**380**	**10,53%**
Birmanie	1	0,03%	Israël	5	0,14%
Chine	7	0,19%	Philippines	1	0,03%
Corée du Sud	2	0,06%	Taiwan	1	0,03%
Inde	355	9,84%	Yémen	7	0,19%
Europe				**315**	**8,73%**
Allemagne	7	0,19%	Luxembourg	3	0,08%
Autriche	3	0,08%	Norvège	16	0,44%
Belgique	35	0,97%	Pays-Bas	8	0,22%
Chypre	8	0,22%	Pologne	2	0,06%
Danemark	21	0,58%	Portugal	28	0,78%
Espagne	26	0,72%	Rép. Tchèque	11	0,30%
Finlande	6	0,17%	Roumanie	35	0,97%
France	7	0,19%	Russie	1	0,03%
Géorgie	1	0,03%	Slovaquie	1	0,03%
Grande-Bretagne	3	0,08%	Suède	64	1,77%
Grèce	3	0,08%	Suisse	2	0,06%
Italie	22	0,61%	Ukraine	3	0,08%

Compte tenu du nombre important de séquences, la méthode du *bootstrap* n'est pas utilisée. De même, aucune approche bayésienne n'est possible.

6.2.3 Reconstruction des états ancestraux

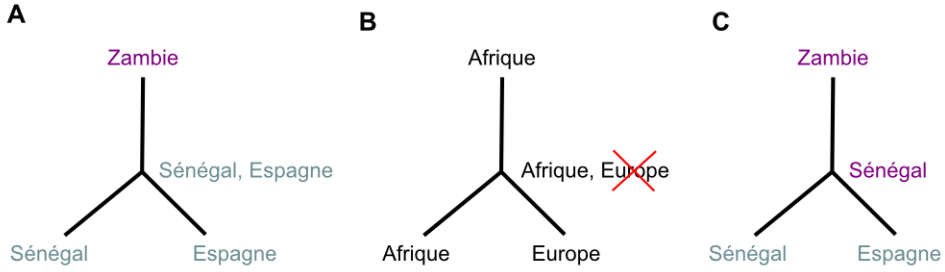
Afin de comprendre le mouvement de l'épidémie du VIH-1C dans son intégralité, les états géographiques ancestraux de chaque nœud de la phylogénie sont calculés à partir de l'information sur les

pays de collecte des souches contemporaines. La méthode choisie pour inférer les états ancestraux est la parcimonie (Hartigan, 1973; Fitch, 1971) parce qu'elle nécessite peu de temps de calcul comparé aux méthodes probabilistes où l'utilisation d'une telle quantité de données rend le temps de calcul prohibitif. Par ailleurs, il n'existe pas (comme pour les séquences) de modèle simple et consensuel pour les mouvements géographiques et le passage d'un pays à un autre.

Deux souches virales sont initialement écartées de cette analyse. L'*outgroup* (83FR-HXB2) et l'isolat 02ZA-1752 (numéro d'accession EF602195). Cet isolat a été collecté en 2002 au Cap en Afrique du Sud par Jacobs *et al*. (2008) et il est ancestral à toutes les souches de la phylogénie hormis 83FR-HXB2. Cependant, dans la phylogénie de l'étude sur le VIH-1C au Sénégal (cf. Chapitre 5) et dans les phylogénies de Jacobs *et al*. (2008), cet isolat se situe dans un clade contenant d'autres souches d'Afrique du Sud. Donc, cette souche semble être une *rogue taxon* (Trautwein *et al*, 2011), c'est-à-dire une souche « avec un placement phylogénétique incertain et variable qui a généralement un effet négatif sur la reconstruction topologique et les valeurs supports ». Pour cette raison, elle n'est pas considérée dans l'interprétation de la phylogénie.

La première phase du calcul des états ancestraux est effectuée avec l'algorithme UPPASS, décrit à la section 1.6. Les calculs par parcimonie nécessitent l'utilisation d'une seconde phase afin que chaque nœud interne (exception faite du nœud racine) exploite l'information de toute la phylogénie. Les algorithmes DOWNPASS (Maddison & Maddison, 2003), ACCTRAN (Fitch, 1971) et DELTRAN (Swofford & Maddison, 1987) sont choisis pour cette seconde phase. Ils sont aussi décrits à la section 1.6. Comme beaucoup de nœuds internes sont ambigus (plus d'un état est assigné) à la fin de la seconde phase, deux règles sont ajoutées pour résoudre ces ambiguïtés au cours de cette seconde phase. La première utilise la parcimonie pour résoudre l'ambiguïté d'un nœud, mais en utilisant non plus les pays, mais les continents auxquels ils appartiennent. Ainsi, les états attribués à ce nœud sont uniquement les pays appartenant au(x) continent(s) inféré(s) par parcimonie (Figure 42). La seconde règle choisit aléatoirement un état parmi ceux restants (s'il en reste plusieurs) et l'assigne comme état final. De ce fait, chaque nœud de la phylogénie contient exactement une seule annotation. Lorsque le choix aléatoire est utilisé pour résoudre les ambiguïtés, la procédure complète est répétée 1 000 fois et les nombres finals de transitions sont obtenus par des moyennes sur les 1 000 cas. Notons que les ambiguïtés au niveau du nœud racine peuvent uniquement être résolues avec le choix aléatoire (du moins si comme ici on décide de ne pas prendre en compte l'*outgroup*, très éloigné phylogénétiquement de l'*ingroup*). Les algorithmes de parcimonie utilisant ces deux règles seront précédés par le terme « Rand » afin de les distinguer des algorithmes originaux.

Figure 42. Illustration de la première règle pour la résolution de nœuds ambigus.
Les annotations correspondantes à la phase ascendante de la parcimonie sont surlignées en bleu, tandis que les annotations de la phase descendante en mauve. La figure A montre un nœud où, par exemple, la parcimonie ACCTRAN ne peut résoudre l'ambiguïté puisque l'annotation Zambie n'est pas associée au nœud interne. La figure B montre ce même nœud mais en regardant les continents associés aux pays. Maintenant la parcimonie ACCTRAN peut résoudre l'ambiguïté et calcule que l'annotation correspondante est le continent africain. La figure C montre les annotations associées au nœud après sa résolution à l'aide de la règle des continents.



6.2.4 Mesure des taux de migrations entre pays

Une fois la reconstruction des états ancestraux achevée, chaque nœud interne de la phylogénie contient un unique état correspondant à la localisation géographique (dans notre cas, le pays) la plus parcimonieuse de l'ancêtre commun représenté par ce nœud. Certaines branches de la phylogénie, dont les localisations aux deux extrémités diffèrent, symbolisent alors les migrations, aussi appelées transitions, du VIH-1C au cours de son histoire évolutive. Nous proposons ici des indices, basés sur ces transitions, qui visent à donner une vision synthétique des migrations du VIH-1C. Une méthode de ré-échantillonnage aléatoire (ou *shuffling*) permet de dégager la significativité de ces mesures, et de s'affranchir (au moins pour une part) des effets liés aux tailles variables d'échantillon par pays (Wallace *et al*, 2007). Autrement, il serait nécessaire d'utiliser des tailles similaires d'échantillon par pays, et par conséquent réduire considérablement le nombre de séquences étudiées, afin d'éviter tout biais potentiel sur les estimations de ces mesures (Véras *et al*, 2011a). Également, en procédant ainsi, on considère que le nombre de souches disponibles dans chaque pays représente peu ou prou la prévalence du sous-type C dans le pays. Avec des échantillons ramenés à la même taille cette information disparaît.

Notations

Soit $\mathcal{E}$ l'ensemble de toutes les annotations (pays) et soient a et b deux annotations de cet ensemble. Le nombre de transitions de l'annotation a vers l'annotation b se note $N_{a\to b}$, c'est-à-dire qu'il représente le nombre de branches dont le nœud adjacent le plus proche de la racine est annoté a et le plus éloigné b. Considérons, de plus, le nombre de transitons de l'annotation b vers l'annotation a, $N_{b\to a}$, le nombre de fixations (branches dont les deux nœuds aux extrémités ont la même annotation) de l'annotation a, $N_{a\to a}$, et le nombre de fixations de l'annotation b, $N_{b\to b}$. De manière générale, $N_{a\to X} = \sum_{u\in\mathcal{E},u\neq a} N_{a\to u}$ désigne le nombre de toutes les transitions sortantes de a et $N_{X\to a} = \sum_{u\in\mathcal{E},u\neq a} N_{u\to a}$ le nombre de toutes les transitions entrantes dans a. Si $N_{X\to a} > 0$, le pays

a est dit receveur (ou IN) et si $N_{a\to X} > 0$, le pays a est dit donneur (ou OUT). Paraskevis *et al*. (2009) utilisent une terminologie similaire. Les indices proposés dans la suite vont nous servir à quantifier les tendances principales (plutôt donneur ? vers quel pays ? plutôt receveur ? de quel pays ?). Notons aussi par $|a| = N_{a\to X} + N_{a\to a}$, le nombre de transitions ayant l'annotation a en entrée et supposons que la phylogénie contient n arêtes.

Indice de dispersion

Nous proposons un premier indice D_a qui indique le degré de dispersion d'une annotation $a \in \mathcal{E}$ au sein de la phylogénie, c'est-à-dire si les feuilles annotées a sont plus ou moins regroupées (dans le cas extrême, elles forment un clade) ou dispersées (dans le cas extrême, aucune feuille annotée a n'est à côté d'une autre feuille a). Cet indice est construit à partir de l'indice $D_{a\to X}$, qui normalise le nombre de transitions sortantes de a (OUT), et de l'indice $D_{X\to a}$, qui normalise le nombre de transitions entrantes dans a (IN). Posons

$$D_{a\to X} = \frac{N_{a\to X}}{N_{a\to X} + N_{X\to a} + N_{a\to a} - 1},$$

$$D_{X\to a} = \max\left\{0, \frac{N_{X\to a} - 1}{N_{a\to X} + N_{X\to a} + N_{a\to a} - 1}\right\}$$

et

$$D_a = D_{a\to X} + D_{X\to a}.$$

La fonction maximum de l'indice $D_{X\to a}$ permet d'éviter les valeurs négatives lorsque l'annotation a n'est pas receveuse (quand la racine est annotée a et que cette annotation est uniquement donneuse), c'est-à-dire lorsque $N_{X\to a} = 0$. Si l'annotation a est totalement dispersée au sein de la phylogénie et qu'aucun état ancestral a n'a pu être inféré, alors $N_{a\to a} = N_{a\to X} = 0$ et donc $D_{X\to a} = D_a = 1$. Au contraire, lorsque l'annotation a est totalement régionalisée dans l'arbre au sein d'un clade unique, la mesure $N_{a\to X} = 0$ et $N_{X\to a} = 1$ (sauf si a est l'unique annotation de la phylogénie), donc $D_{a\to X} = D_{X\to a} = D_a = 0$ (Figure 43). Le dénominateur est nul lorsque l'annotation a est uniquement représentée par une feuille. Dans ce cas (peu intéressant), cette mesure n'est pas utilisée.

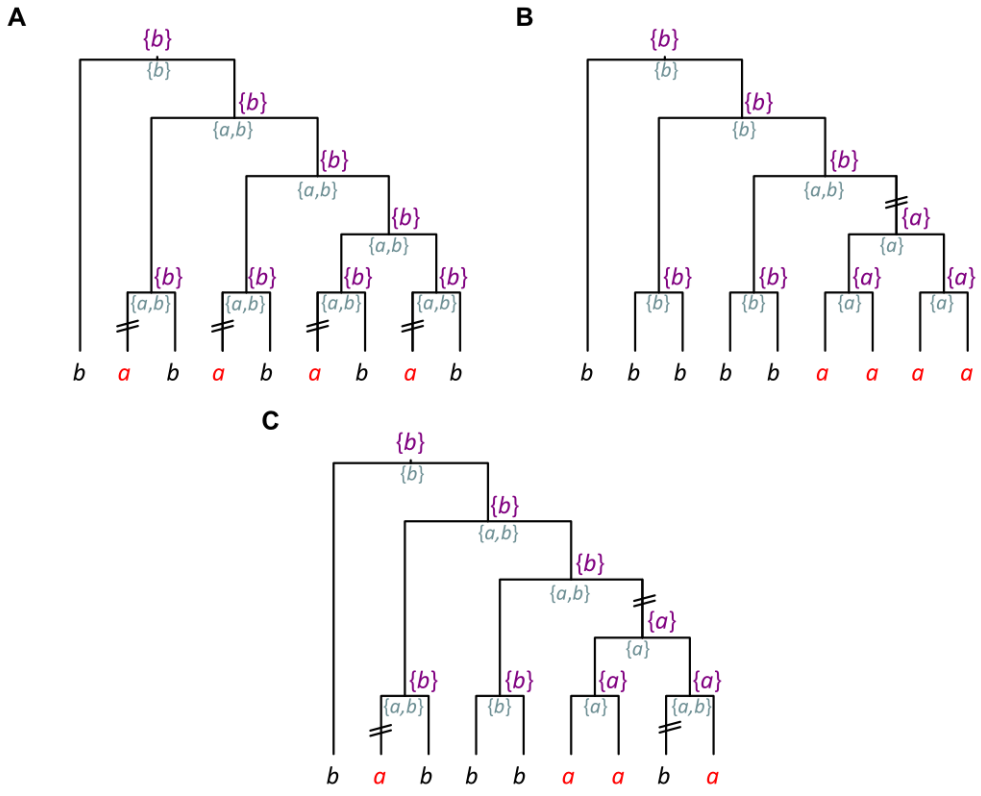
Indice de flux

Nous proposons un second indice $F_{a\to b}$ qui renseigne sur la fréquence des transitions de l'annotation a vers l'annotation b. Cet indice est largement utilisé par d'autres (Salemi *et al*, 2008; Wallace *et al*, 2007) et certains logiciels proposent de le calculer (Maddison & Maddison, 2003). Toutefois, nous lui ajoutons une normalisation afin d'augmenter la lisibilité des graphiques. L'indice de flux $F_{a\to b}$, avec $a, b \in \mathcal{E}$, $a \neq b$, est défini par

$$F_{a\to b} = \frac{N_{a\to b}}{\sum_{i\in\mathcal{E}, i\neq b} N_{i\to b}}.$$

Lorsqu'un pays b n'est receveur que d'un seul pays, disons a, la mesure $F_{a\to b}$ vaut 1 et $F_{i\to b}$ vaut 0 quel que soit le pays i. Pour tout b, la somme des mesures $F_{i\to b}$ vaut 1, $i \in \mathcal{E}$, $i \neq b$. Cet indice représente donc la proportion du nombre de transitions de a vers b parmi toutes les transitions vers b. Notons que si a est uniquement un pays receveur alors $F_{a\to i} = 0$, pour tout $i \in \mathcal{E}$, $i \neq a$. Il est donc inutile de reporter ces mesures. En exemple, la phylogénie A de la Figure 43 montre que $F_{a\to b} = 0$ et la mesure $F_{b\to a} = 1$.

Figure 43. Exemples d'application avec l'indice de dispersion.
Ces deux figures illustrent le comportement de l'indice de dispersion. Les annotations ancestrales calculées lors de la phase ascendante de la parcimonie sont indiquées en bleu, celles obtenues lors de la phase descendante (par ACCTRAN) en mauve. Les doubles barres obliques indiquent les branches où une transition se produit. La figure A montre un cas où l'annotation a est totalement dispersée dans la phylogénie, $D_a = D_{X\to a} = 1$ et $D_{a\to X} = 0$, la figure B un cas où les feuilles associées à l'annotation a forment un clade, $D_a = D_{a\to X} = D_{X\to a} = 0$, et la figure C un cas intermédiaire $D_{a\to X} = D_{X\to a} = 1/6$ et $D_a = 1/3$.



Indice de symétrie

Le dernier indice introduit indique la quantité de transitions échangées entre deux pays a et b de $\mathcal{E}$. Il permet de vérifier si l'échange est symétrique (autant de transitions de a vers b que de b vers a), unidirectionnel (que des transitions de a vers b ou que des transitions de b vers a) ou bidirectionnel (des transitions en quantité variable de a vers b et de b vers a). Afin d'intégrer les effectifs

correspondant à chaque annotation a et b, il s'inspire dans sa définition d'un processus de Markov stationnaire réversible dans le temps, c'est-à-dire que pour tout a et b de $\mathcal{E}$, $a \neq b$,

$$p_a p_{a\to b} = p_b p_{b\to a}$$

où p_a (respectivement p_b) est la probabilité d'observer un nœud interne annoté a (resp. b) et $p_{a\to b}$ (resp. $p_{b\to a}$) la probabilité d'observer une transition de a vers b (resp. b vers a). Or $p_i = |i|/n$, où n représente le nombre d'arêtes de la phylogénie, et $p_{i\to y} = N_{i\to y}/|i|$. Après simplification, la relation devient tout simplement

$$N_{a\to b} = N_{b\to a}.$$

L'indice de symétrie est donc défini pour tout a et b par (Véras *et al*, 2011a)

$$S_{a\leftrightarrow b} = N_{a\to b} - N_{b\to a}.$$

Remarquons que $S_{a\leftrightarrow b} = -S_{b\leftrightarrow a}$. Si l'échange est parfaitement symétrique alors $N_{a\to b} = N_{b\to a}$ et la mesure vaut 0. Si la mesure est positive (resp. négative), alors il y a plus de transitions de a vers b (resp. b vers a) que l'inverse. Si l'échange est unidirectionnel alors $|S_{a\leftrightarrow b}/(N_{a\to b} + N_{b\to a})| = 1$. Il est évident que cet indice est uniquement appliqué s'il existe des échanges entre a et b.

Test de ré-échantillonnage aléatoire

Afin de vérifier la significativité statistique des résultats observés, la procédure de ré-échantillonnage aléatoire (*shuffling*) est utilisée 1 000 fois pour comparer les valeurs observées à celles de l'hypothèse nulle ou panmixie (cf. Chapitre 1). Pour chaque paire de pays et chaque indice, la valeur observée est comparée à la distribution des valeurs aléatoires obtenues. Un indice observé est jugé statistiquement significatif avec une p-valeur de 5%, s'il est plus grand (ou plus petit, suivant que les valeurs remarquables de l'indice sont élevées ou au contraire faibles) que le quantile à 95% (ou 5%) de cette distribution. Ce test permet de comparer les valeurs de l'indice de flux F à celles de l'hypothèse nulle ; on s'attend à des valeurs plus grandes que celles obtenues aléatoirement, et on se positionne donc par rapport au quantile à 95%. Au contraire, pour juger de la significativité de la dispersion D, dont on s'attend à ce qu'elle soit plus faible en raison de réalité des frontières entre pays que dans l'hypothèse nulle, on se positionne donc au quantile à 5%. Enfin, les valeurs observées par l'indice S peuvent être grandes (proches de 1) ou petites (proches de -1) et on doit faire un test « *two-sided* » en se positionnant par rapport aux quantiles à 2,5% et 97,5%.

6.2.5 Recherche d'évènements fondateurs à l'aide de PhyloType

Les indices décrits ci-dessus permettent de décrire les grands flux géographiques. On décrit ici la méthode PhyloType[10] (Chevenet, Jung, de Oliveira et Gascuel, en cours de soumission), qui va nous permettre de rechercher les grands événements fondateurs expliquant l'essentiel de la pandémie.

6.2.5.1 Présentation de PhyloType

En épidémiologie, un évènement fondateur correspond à l'introduction d'un nouvel élément pathogène dans une population donnée et à sa diffusion au sein de celle-ci. Ce genre d'évènement peut être observé à l'aide d'une phylogénie. En effet, les séquences de l'agent pathogène collectées après sa diffusion sont toutes issues d'une même séquence ancestrale, celle à l'origine de l'évènement fondateur. Ainsi, les feuilles d'une phylogénie associées à ces séquences forment un clade. L'identification de clades dans une phylogénie reste, en pratique, assez aisée. Mais cela se complique lorsque le nombre de séquences dans la phylogénie est important, si les séquences étudiées proviennent de plusieurs évènements fondateurs différents, imbriqués ou non les uns dans les autres, ou si une quantité non négligeable de séquences parasites y sont mêlées, par exemple, celles provenant de plusieurs chaines de transmission marginales entre plusieurs populations, ou bien en raison d'erreurs de reconstruction.

La méthode PhyloType (Chevenet *et al*) aide à la localisation d'évènements fondateurs (clades parfaits ou imbriqués) sur une phylogénie, celle-ci doit être racinée afin de connaître l'orientation du temps, et on doit avoir la connaissance des pays de collecte associés à chaque séquence échantillonnée. Les groupes de séquences mis en valeur par ce logiciel sont appelées des *phylotypes*. Un *phylotype* est un sous-ensemble de souches dont chaque souche x et leur ancêtre commun ρ partagent la même annotation A, et telle que A est conservée le long du chemin de ρ à x (Figure 44). Par la suite nous utiliserons le terme « membre » pour désigner les souches appartenant à un phylotype. Pour faire cela, PhyloType doit connaître les annotations ancestrales associées à chaque nœud interne de la phylogénie. Il les calcule par parcimonie (ACCTRAN ou DELTRAN).

Des critères de sélection sont utilisés pour restreindre le nombre de *phylotypes* et leur garantir de fortes propriétés spécifiques. Par exemple, en limittant le nombre de séquences à l'intérieur du clade définissant le *phylotype* ayant une annotation différente de celle du *phylotype*. Le choix des critères à utiliser et leur seuil de validité sont choisis par l'utilisateur. Ils sont définis récursivement et leur complexité en temps de calcul est en $O(n)$, où n est le nombre de feuilles dans la phylogénie. Notons par P un phylotype potentiel d'annotation A, par L et R les nœuds racines de ses sous-arbres gauche

[10] Mise en œuvre dans un serveur web (http://lamarck.lirmm.fr/phylotype)

et droit respectivement et par F son nœud père. Neuf critères, plus trois qui sont des rapports entre deux autres critères, sont proposés par PhyloType :

- *size* (Sz) : ce critère correspond au nombre de membres du *phylotype* (et non au nombre de feuilles contenu dans le clade de même racine que le *phylotype*). Il est défini par

> *Size*(P, A)
> Si Annotation(P) est différent de A, alors 0
> Sinon si P est une feuille, alors 1
> Sinon *Size*(L, A) + *Size*(R, A) ;

- *different* (Df) : ce critère compte le nombre de sous-arbres et/ou de feuilles inclus dans le *phylotype* et ayant une annotation différente de ce dernier (dans le cas d'un sous-arbre, seule l'annotation à la racine est considérée). Il est défini par

Different(P, A)
Si Annotation(P) est différent de A, alors 1
Sinon si P est une feuille, alors 0
Sinon *Different*(L, A) + *Different*(R, A) ;

- *total* (Tt) : ce critère compte le nombre total de feuilles incluses dans le clade de même racine que celle du *phylotype*. Il est défini par

Total(P)
Si P est une feuille, alors 1
Sinon *Total*(L) + *Total*(R) ;

- *persistence* (Ps) : ce critère mesure le degré de conservation de l'annotation d'un *phylotype*, de la racine de celui-ci jusqu'à ses descendants. Il débute à la racine du *phylotype* et est égal au nombre minimum de générations où l'annotation est conservée dans chaque lignée. Il est défini par

Persistence(P, A)
Si (P est une feuille) ou (Annotation(L) est différent de A) ou (Annotation(R) est différent de A), alors 0
Sinon 1 + Min{*Persistence*(L, A), *Persistence*(R, A)} ;

- *local separation* (Sl) : ce critère correspond à la longueur de la branche parente à la racine du *phylotype*, c'est-à-dire la longueur de la branche qui sépare le *phylotype* du reste de la phylogénie ;
- *global separation* (Sg) : ce critère est utile lorsque la longueur de la branche parente de la racine du *phylotype* est courte, mais que les longueurs des autres branches qui séparent la racine du *phylotype* et la racine de la phylogénie sont grandes, indiquant ainsi une grande séparation du *phylotype* par rapport au reste de la phylogénie. Il est défini par

  Global_separation(P)
  Si P n'est pas la racine, alors
  Local_separation(P) + *Total*(P) × *Global_separation*(F)/*Total*(F)
  Sinon 0 ;

- *diversity* (Dv) : ce critère mesure la diversité génétique des membres du *phylotype*. Il est défini par

  Diversity(P)
  Sum(P, Annotation(P))/*Size*(P, Annotation(P)),

  avec (où l et r sont respectivement les longueurs des branches parentes à L et R)

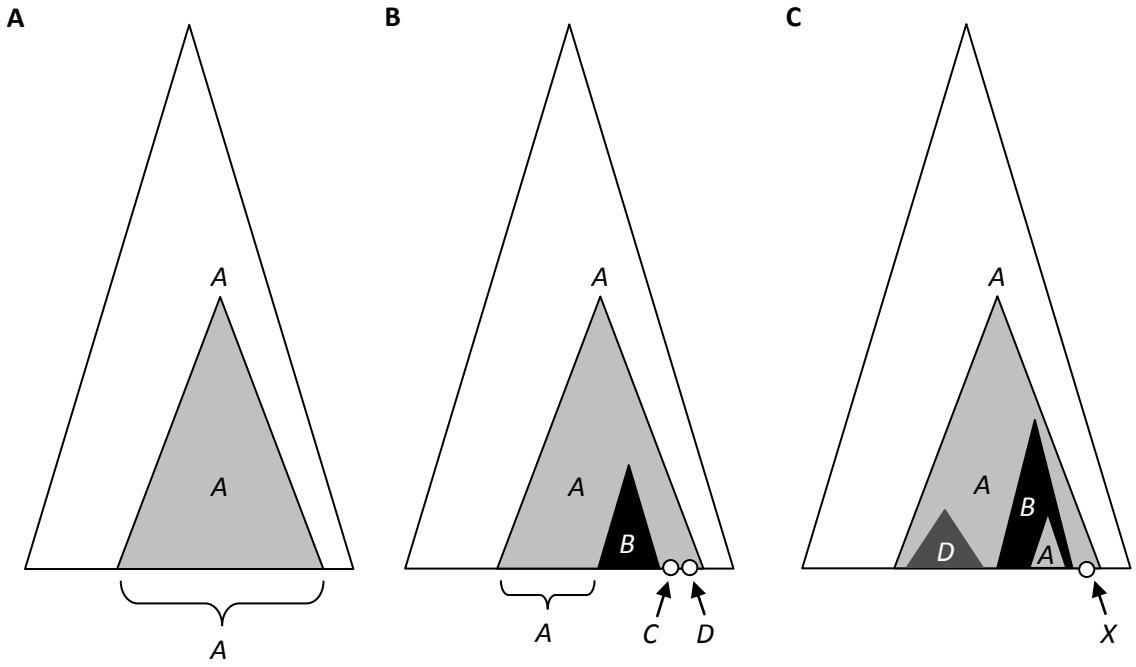
  Sum(P, A)
  Si Annotation(P) est A et si P n'est pas une feuille, alors
  Sum(L, A) + l × *Size*(L, A) + *Sum*(R, A) + r × *Size*(R, A)
  Sinon 0 ;

- *support* (Sp) : ce critère renvoie le support de branche (lorsqu'il est présent) associé à la racine du *phylotype* ;
- *global support* (SpG) : ce critère renvoie la plus forte valeur entre *support* et une pondération entre les supports des branches se trouvant sur le chemin reliant la racine du *phylotype* à celle de la phylogénie. Il est défini par

  Global_support(P)
  Si P n'est pas la racine, alors
  Max{*Support*(P), Total(P) × *Global_support*(F)/*Total*(F)}
  Sinon 0.

Les trois autres critères restant sont les rapports *size*/*different* (Sz/Df), *local separation*/*diversity* (Sl/Dv) et *global separation*/*diversity* (Sg/Dv). Par exemple, le critère *size*/*different* permet à un *phylotype* donné de contenir un nombre de sous-arbres et/ou de feuilles ayant une annotation différente et qui varie en fonction de la taille du *phylotype* donnée.

Figure 44. Exemples de *phylotypes*.
La figure A montre la définition la plus simple d'un *phylotype* (un clade) où la racine du *phylotype* et toutes ses feuilles ont la même annotation (A). La figure B montre un *phylotype* annoté A qui contient un *phylotype* annoté B et deux feuilles annotées C et D. La figure C montre un *phylotype* annoté A contenant deux *phylotypes* annotés D et B, ainsi qu'une feuille annotée X, et le *phylotype* annoté B contient un autre *phylotype* annoté A.
Extrait de Chevenet et al.



PhyloType possède aussi une procédure statistique qui permet de savoir si les résultats obtenus sont statistiquement significatifs ou non, grâce à une p-valeur associée à chaque critère (sélectionné par l'utilisateur) de chaque *phylotype*. Cette p-valeur est obtenue par ré-échantillonnage aléatoire (*shuffling*) (cf. Chapitre 1). Par exemple, pour un *phylotype* donné, si une valeur de 2 est obtenue pour le critère *size* avec une p-valeur de 3/1 000, cela signifie que 3 *shufflings* sur les 1 000 ont au moins un *phylotype* de même annotation avec une valeur supérieure ou égale à 2 pour le critère s*ize*. En pratique, si ce nombre est supérieur à 5% du nombre de *shufflings*, alors le *phylotype* résultant est généralement considéré comme non significatif.

En plus de cela, PhyloType propose une interface permettant d'enraciner la phylogénie de l'utilisateur (au cas où elle ne l'est pas) de trois manières différentes à l'aide d'un logiciel que j'ai développé pour cette occasion. La première manière positionne la racine sur le point qui minimise la variance de la distance séparant chaque feuille de la racine, de sorte à rendre la phylogénie la plus

ultramétrique possible (toutes les feuilles sont plus ou moins à égale distance de la racine). Si l'utilisateur dispose de dates de collecte associées à chaque feuille et que les souches étudiées proviennent d'une population à évolution mesurable (MEP) (Drummond *et al*, 2003b), une deuxième méthode est proposée. Il s'agit de la régression linéaire *Root-to-tip* pour laquelle la minimisation de la somme des résidus permet d'obtenir l'emplacement optimal de la racine (cf. Chapitre 2). La dernière méthode proposée localise la racine en fonction de séquences sélectionnées (supposées *outgroup*) par l'utilisateur.

6.2.5.2 Association de certains pays afin de favoriser l'apparition de *phylotypes*

PhyloType est une méthode qui met en exergue des *phylotypes* correspondant à diverses annotations et les inclusions de *phylotypes* au cœur de la phylogénie enracinée suggèrent les migrations successives du virus au cours du temps ou « chaines de transmission ». Les annotations dont les feuilles sont dispersées dans la phylogénie ont peu de chance d'être interprétées par PhyloType, puisque les *phylotypes* sont dérivés de clades dans lesquels une annotation donnée est très représentée. En revanche, PhyloType permet de grouper deux ou plusieurs annotations ensembles dans le but de favoriser l'apparition de *phylotypes* correspondant à l'union de ces annotations. Ces groupements doivent avoir un sens épidémiologique, et typiquement correspondre à des pays voisins et/ou ayant des échanges forts et identifiés.

Dans ce but, nous proposons un indice qui renseigne sur la régionalisation d'une annotation, que l'on pourra comparer avec celle de l'union de deux annotations ou plus. En accord avec PhyloType, le principe de parcimonie est utilisé (même procédé qu'à la section 6.2.3) mais en ne considérant que deux annotations dans la phylogénie (l'annotation ou les annotations étudiées a et la réunion des autres annotations $X = \neg a$). En réutilisant les mêmes notations qu'à la section 6.2.4, l'indice de régionalisation R_a, pour une annotation a, est définie par

$$R_a = \frac{N_{a\to X} + N_{X\to a} - 1}{|a| - 1},$$

où $|a|$ représente ici le nombre de feuilles annotées a. Cette mesure vaut 1 lorsque l'annotation a est totalement dispersée ($N_{a\to X} = 0$ et $N_{X\to a} = |a|$), et vaut 0 lorsqu'elle est totalement régionalisée ($N_{a\to X} = 0$ et $N_{X\to a} = 1$). Cet indice ne peut pas être utilisé avec des annotations représentées par une seule souche (le dénominateur est nul). Pour vérifier si l'union de deux annotations a et b est plus régionalisée que les deux annotations prises séparément, il suffit de comparer $R_{a\cup b}$ à $\min\{R_a, R_b\}$. Si la comparaison est favorable (c'est-à-dire $R_{a\cup b} < \min\{R_a, R_b\}$) alors le *phylotype* de l'union a plus de chance d'apparaître que ceux correspondant aux annotations a et b séparées.

En plus d'aboutir à une meilleure régionalisation, une association de pays est uniquement proposée s'ils partagent une frontière géographique (les migrations ou le commerce, et donc le transport de germe viral, en est facilité) et s'ils sont trop peu représentés pour qu'on puisse espérer obtenir des *phylotypes* pertinents en considérant chacun de ces pays pris séparément (par exemple, s'il y a moins de 20 souches par annotation et si le critère *size* est supérieur ou égal à 20).

6.2.5.3 Paramétrage de PhyloType

Trois analyses PhyloType successives sont faites pour chaque option de parcimonie disponible (ACCTRAN et DELTRAN) et en faisant varier la taille (*size*) minimale des *phylotypes* : 20, 10 et 5 ; ce qui correspond donc à des analyses plus ou moins détaillées, avec des niveaux d'exigence variables. Les trois autres critères choisis sont fixes et sont *persistence* ≥ 1, *size*/*different* ≥ 1 et *support* ≥ 70% (valeur aLRT minimum pour la branche aboutissant à la racine du *phylotype*). Mille s*hufflings* sont calculés pour chaque analyse et les *phylotypes* dont la p-valeur est supérieure à 10/1 000 (1%) pour le critère s*ize* ne sont pas considérés dans les résultats.

6.3 Résultats

6.3.1 Séquences *pol* du VIH-1C incluses dans l'étude

Les 3 081 séquences *pol*, couvrant plus de 1 000 paires de bases, de l'étude sur l'origine géographique et temporelle du VIH-1C au Sénégal (cf. Chapitre 5) sont incluses dans cette étude. À celles-ci, 528 nouvelles séquences sont ajoutées dont 219 sont collectées au Burundi, 67 en RDC, 1 en République Centrafricaine, 2 en République du Congo, 1 en Éthiopie et 238 au Swaziland.

L'ensemble des séquences provient de 63 pays différents listés dans le Tableau 4. Le continent africain reste le plus représenté ; il contient à lui seul 72% du nombre de séquences et 75% des souches collectées en Afrique proviennent de l'Afrique australe. L'Afrique du Sud et la Zambie sont toujours les deux pays les plus représentés (37% du nombre total de séquences). Aucune souche collectée en Amérique ou en Asie n'est ajoutée à cette étude. Ces continents sont donc toujours majoritairement représentés par le Brésil (253 séquences sur 299) et l'Inde (355 séquences sur 380) respectivement. Le continent européen représente seulement 9% du nombre total de séquences et aucun pays de collecte ne se démarque fortement en nombre de souches disponibles.

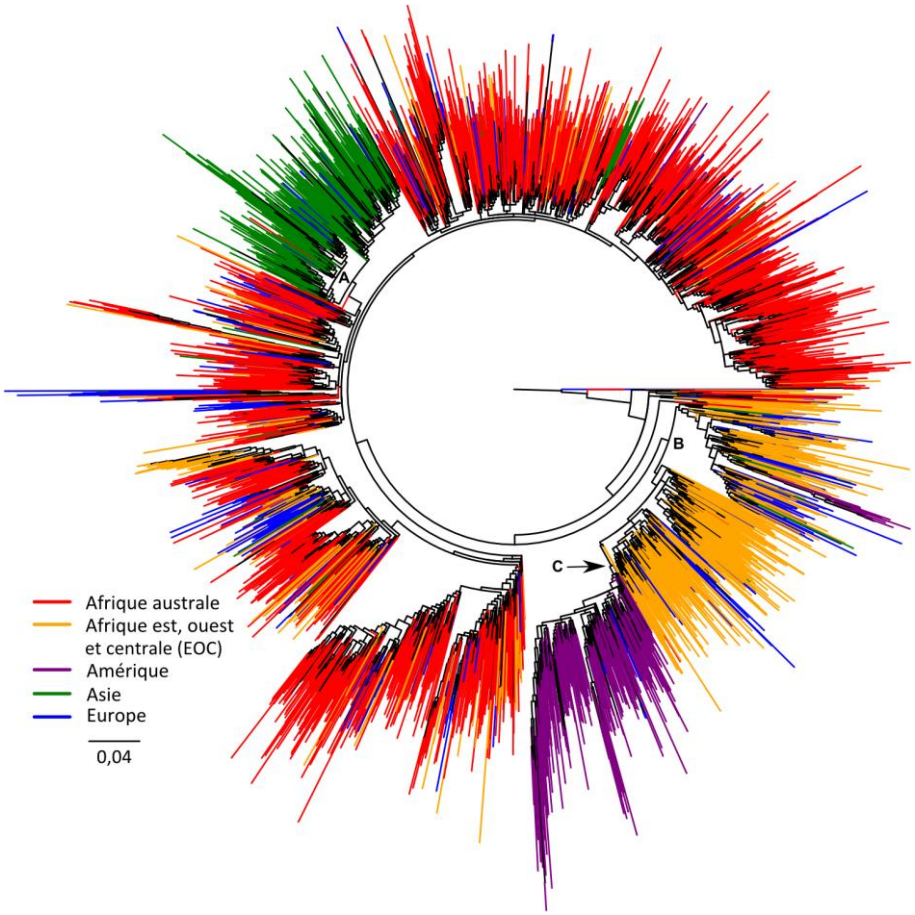
6.3.2 Phylogénie des séquences *pol* du VIH-1C

La phylogénie du maximum de vraisemblance (PhyML) des 3 609 souches du VIH-1C collectées à travers le monde est présentée à la Figure 45. Les souches sont coloriées en fonction de leur pays de collecte : l'Afrique australe (Afrique du Sud, Botswana, Malawi, Mozambique, Swaziland, Zambie et

Zimbabwe) en rouge, les pays africains de l'est, de l'ouest et centrale (EOC) en orange, l'Amérique en mauve, l'Asie en vert et l'Europe en bleu. À l'instar de la phylogénie du Chapitre 5, trois clusters importants contenant la majorité des souches de l'Amérique (essentiellement des souches du Brésil), de l'Asie (essentiellement des souches de l'Inde) et de l'Afrique EOC (majoritairement des souches du Burundi) sont observés. Ils sont respectivement supportés en valeur aLRT à 99,2% (nœud C), 74,0% (nœud A) et 86,1% (nœud B). Très peu de souches de l'Afrique australe sont visibles dans le cluster de l'Afrique EOC. Les souches collectées en Europe sont dispersées dans toute la phylogénie, sans formation de clusters importants. Aucune souche d'Afrique n'est visible à l'intérieur des clusters formés par les souches de l'Amérique et de l'Asie (hormis une souche collectée en Zambie qui se situe à l'intérieur du cluster asiatique). Ces observations suggèrent de multiples introductions du sous-type C en Europe, provenant principalement de pays africains, et une introduction majeure de ce variant, suivie d'une diffusion efficace, aussi bien en Afrique EOC (majoritairement représentée par le Burundi) que sur les continents américain (majoritairement représenté par le Brésil) et asiatique (majoritairement représenté par l'Inde).

Tout comme les souches d'Inde et du Brésil, certaines souches de pays africains forment des clusters d'intérêt au sein de la phylogénie. La Figure 46 montre la phylogénie où seulement les souches appartenant aux pays africains d'intérêt sont coloriées : l'Afrique du Sud en jaune, le Burundi en bleu, l'Éthiopie en vert, la Tanzanie en orange et la Zambie en rouge. Les souches collectées en Afrique du Sud sont disséminées dans la phylogénie mais un cluster remarquable, supporté en valeur aLRT à 88,4% (nœud A), apparaît. Cela suggère une introduction majeure du sous-type C dans ce pays, accompagnée de multiples introductions mineures. Aucune souche du Burundi n'est mélangée à l'Afrique australe et la majorité de celles-ci forme un cluster, supporté en valeur aLRT à 76,2% (nœud C), dans lequel un cluster de souches collectées en Tanzanie apparaît, supporté en valeur aLRT à 87,9% (nœud D). Ceci suggère une chaîne de transmission depuis l'origine épidémique, sans doute la Zambie (cf. ci-dessous) vers le Burundi, et ensuite la Tanzanie. Les souches de l'Éthiopie sont aussi bien retrouvées parmi des souches de l'Afrique EOC que de l'Afrique australe, mais elles y sont régionalisées (nœud B et E supportés respectivement en valeur aLRT à 88,2% et 90,6%), indiquant que l'épidémie du sous-type C en Éthiopie proviendrait de deux origines géographiques différentes. Malgré le nombre important de souches collectées en Zambie, aucun cluster important n'apparaît dans la phylogénie. Cependant, elles sont uniquement mélangées avec d'autres souches de l'Afrique australe (une souche est également vue en Asie). Ceci indique une origine possible de l'épidémie en Zambie, qui sera confirmée par les analyses suivantes basées sur nos indices et PhyloType.

Figure 45. Phylogénie basée sur le gène *pol* des 3 609 souches du VIH-1C.
Cette figure montre la phylogénie obtenue par maximum de vraisemblance (PhyML) des 3 609 souches *pol* d'une longueur de 1 011 paires de bases. Les souches appartenant au continent africain sont séparées en deux groupes. Un groupe contient toutes les souches de l'Afrique australe (en rouge), région géographique où l'épidémie du VIH-1C est très intense, et l'autre contient toutes celles de l'Afrique de l'est, de l'ouest et centrale (EOC, en orange) qui forment un clade. Le continent américain (en mauve) montre une introduction majeure du sous-type C sur ce continent, tout comme sur le continent asiatique (en vert). D'autres introductions sont visibles mais sont marginales. Les souches du continent européen (en bleu) sont mélangées dans la phylogénie, suggérant des introductions multiples de ce virus en Europe.



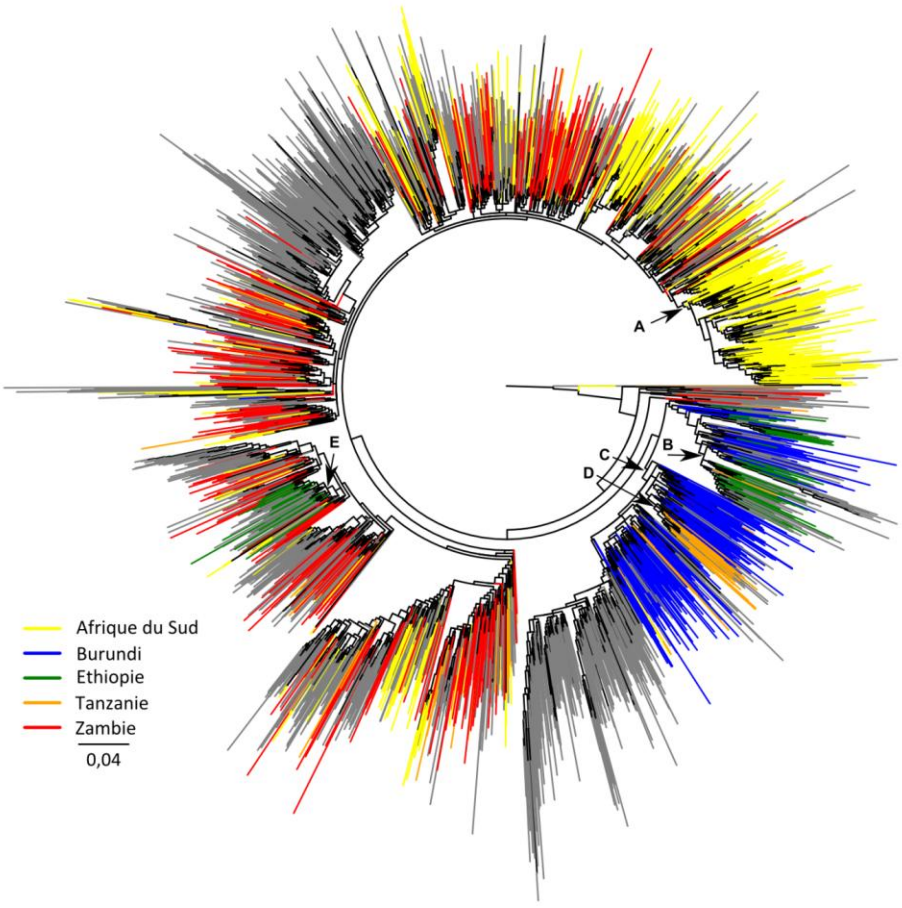
La Figure 47 montre plus en détail les souches situées à proximité de la racine de la phylogénie des 3 609 souches *pol* du VIH-1C. Les annotations ancestrales déduites des parcimonies ACCTRAN (en bleu), DELTRAN (en rouge) et DOWNPASS (en vert) sont reportées sur chaque nœud interne de la phylogénie ; les estimations communes sont en noir. Les souches 83FR-HXB2 (*outgroup*) et 02ZA-1752 (*rogue taxa*) sont écartées de l'étude (cf. section 6.2.3). L'annotation correspondante à l'ancêtre commun de toutes les souches du VIH-1C est donc RDC/Tanzanie/Zambie et cela quelle que soit la parcimonie utilisée. Les annotations RDC et Tanzanie sont dues au groupe de trois souches (deux collectées en RDC et une en Tanzanie) qui sont ancestrales aux autres souches du VIH-1C. Hormis ce groupe de trois séquences et pour les trois parcimonies, l'annotation résultante à l'ancêtre commun des autres souches du VIH-1C est la Zambie. Les différentes parcimonies renvoient des estimations assez similaires sur les autres nœuds proches de la racine, seulement six nœuds sont discordants. Enfin, un cluster remarquable de treize souches collectées en RDC, et supporté à 63,7% en aLRT, apparaît (nœud A). L'origine géographique de l'épidémie du VIH-1C est donc indéterminée, elle

est soit en RDC, soit en Zambie ou soit en Tanzanie. Notons que ces trois pays africains sont limitrophes.

6.3.3 Étude des flux migratoires du VIH-1C

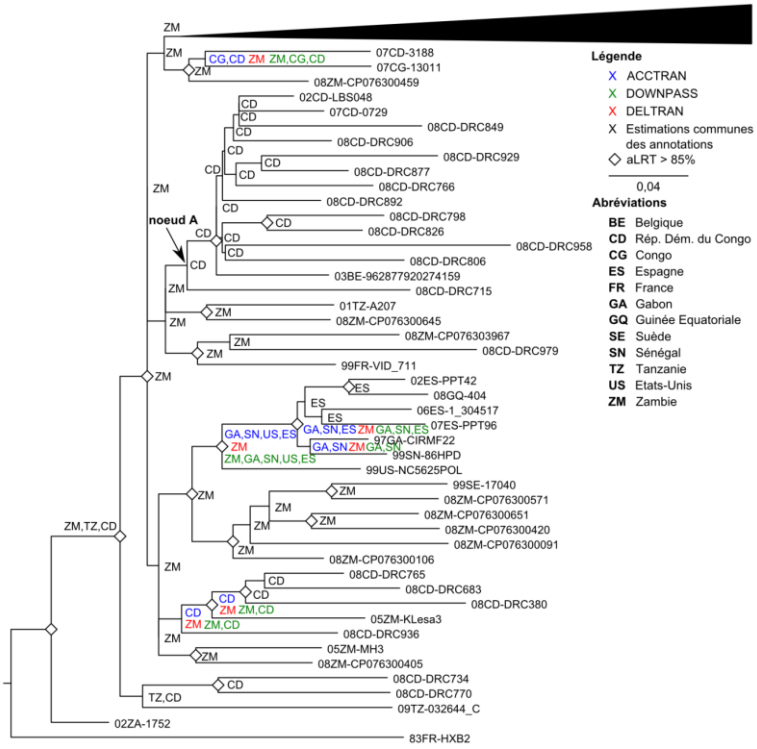
Afin de synthétiser les flux migratoires du VIH-1C de la phylogénie, nous avons développé trois indices basés sur les transitions entre pays. Ces transitions sont obtenues de trois manières différentes, en utilisant les algorithmes de parcimonie RandACCTRAN, RandDELTRAN et RandDOWNPASS (cf. section 6.2.3). Une méthode de ré-échantillonnage aléatoire, le *shuffling*, est utilisée pour mesurer la significativité statistique de ces mesures. Seuls les résultats correspondant à la parcimonie RandDOWNPASS sont présentés dans ce chapitre ; cette méthode étant la moins arbitraire des trois. Les résultats des autres parcimonies sont disponibles dans l'Annexe A.

Figure 46. Phylogénie basée sur le gène *pol* des 3 609 souches du VIH-1C (zoom sur les pays africains d'intérêt).
Cette figure montre la phylogénie des 3 609 souches où seulement les souches appartenant aux pays d'intérêt sont coloriées. Les souches appartenant à l'Afrique du Sud sont en jaune, celles du Burundi en bleu, celles de l'Éthiopie en vert, celles de Tanzanie en orange et celles de la Zambie en rouge. Les souches des autres pays sont grisées. La majorité des souches du Burundi et de l'Afrique du Sud forment deux clusters importants. Les souches de Tanzanie forment un cluster au sein des souches du Burundi et les souches de l'Éthiopie forment deux clusters. Le premier est à proximité des souches du Burundi et l'autre parmi les souches de l'Afrique australe. Enfin, les souches de Zambie sont ubiquitaires au sein des souches de l'Afrique australe.



Nous présentons d'abord les graphiques des mesures obtenues par les trois indices, puis nous discutons des résultats.

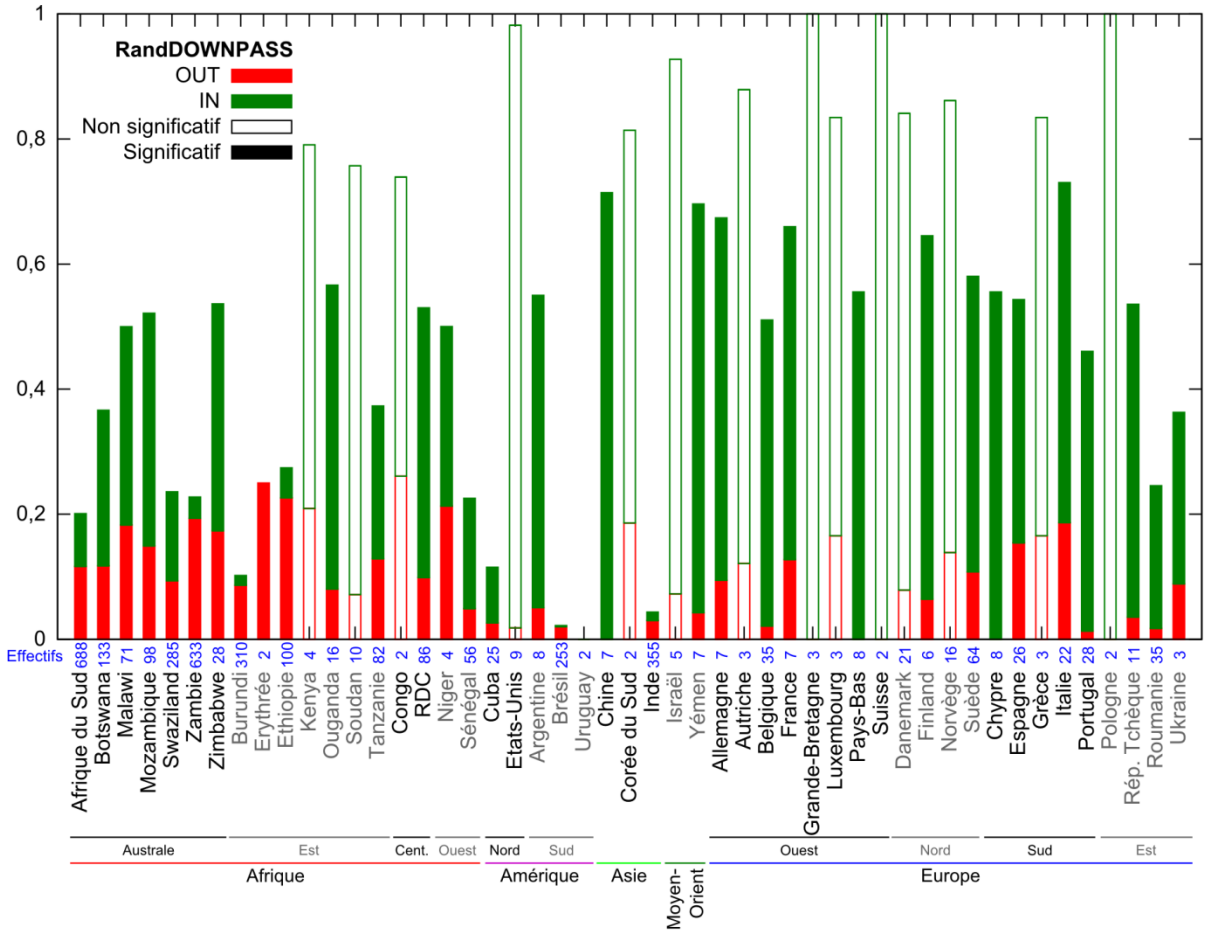
Figure 47. Souches à proximité de la racine.
Phylogénie du maximum de vraisemblance (PhyML) des 3 609 souches du VIH-1C où uniquement les souches à proximité de la racine sont représentées. Les annotations ancestrales inférées par les parcimonies ACCTRAN (en bleu), DELTRAN (en rouge) et DOWNPASS (en vert) sont indiquées sur chaque nœud. Les estimations communes sont en noir. Les nœuds avec un losange blanc indiquent des valeurs supports (aLRT) plus grandes que 85%. Chaque nom de souche est précédé de l'année de collecte (par exemple, 02 pour 2002 ou 99 pour 1999) puis d'une abréviation représentant le pays de collecte (SN pour Sénégal, etc.). La liste complète des abréviations utilisées est présentée dans la figure.



La Figure 48 montre les estimations de l'indice de dispersion D pour chaque pays étudié représenté par plus d'une souche et pour la parcimonie RandDOWNPASS. Ces pays sont organisés par région géographique, puis par ordre alphabétique. Les deux composantes IN et OUT de cet indice sont respectivement représentées en vert et rouge. Les mesures non significatives ont un intérieur vide (obtenus en se positionnant par rapport aux quantiles à 5%) et les mesures significatives sont représentées par une barre pleine. Le nombre de souches associées à chaque pays est donné en bleu en abscisse. Plus l'indice de dispersion est grand (proche de 1), plus les souches sont éparpillées dans la phylogénie, plus il est petit (proche de 0), plus elles sont régionalisées et forment un clade lorsque cet indice vaut 0. Par exemple, l'indice de dispersion de l'Uruguay vaut zéro, ce qui signifie que les souches forment une cerise (cette phylogénie ne contient que deux souches collectées en Uruguay), tandis que les souches collectées en Pologne sont isolées les une des autres (l'indice vaut 1). Par définition de l'indice de dispersion, un pays donneur ne peut être totalement dispersé dans la phylogénie. Les Figures 54 et 55 de l'Annexe A correspondent aux mesures de l'indice de dispersion pour les méthodes de parcimonie RandACCTRAN et RandDELTRAN respectivement. Outre le fait de rensei-

gner sur la régionalisation des souches dans la phylogénie (hauteur des barres), comme pour l'Inde et le Brésil, cet indice permet aussi d'identifier les pays fortement donneurs (donc probablement à l'origine de la diffusion de l'épidémie) et les pays fortement receveurs (donc au bout d'une chaîne de transmission). On voit ici nettement que les pays africains sont plutôt donneurs, donc à l'origine de la diffusion de l'épidémie du sous-type C, alors que, par exemple, les pays européens, sont globalement receveurs. Cependant, cet indice n'indique pas vers quels pays est transmise l'épidémie ou de quels pays elle provient ; c'est le rôle de l'indice de flux.

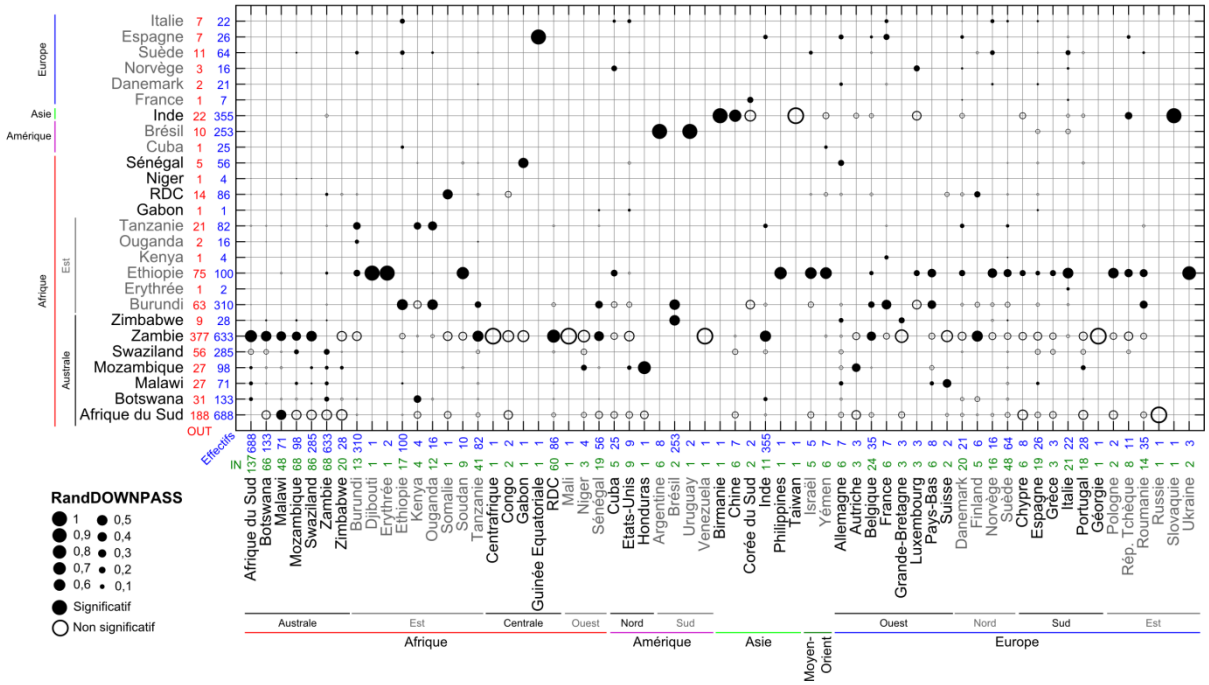
Figure 48. Estimations de l'indice de dispersion avec la parcimonie RandDOWNPASS.
Le graphique indique, pour chaque pays de collecte ayant au moins deux souches, les valeurs correspondantes à l'indice de dispersion IN (en vert) et OUT (en rouge) ; leur somme correspondant à l'indice de dispersion totale. Ces résultats sont obtenus avec la parcimonie RandDOWNPASS. Les mesures significatives de l'indice de dispersion sont représentées par des barres pleines tandis que les mesures non significatives par des barres vides. Les pays sont regroupés par zone géographique, puis par ordre alphabétique. Une mesure totale de 1 signifie que les souches sont totalement dispersées dans la phylogénie, sans possibilité de formation d'annotations ancestrales (par exemple la Grande-Bretagne). Une mesure de zéro signifie que toutes les souches d'un pays forment un clade monophylétique (seul exemple, l'Uruguay). Pour chaque pays, le nombre de souches présentes dans la phylogénie est rappelé en bleu en abscisse.



La Figure 49 indique les estimations de l'indice de flux F obtenues pour chaque couple de pays avec la parcimonie RandDOWNPASS. Chaque point reflète la proportion des transitions reçues par le pays en abscisse, du pays en ordonnée. Par exemple, le Brésil a deux introductions une provenant du Burundi et l'autre du Zimbabwe, tandis que la Roumanie ne donne qu'une seule fois vers la Grèce. Les

cercles vides représentent les mesures non significatives (obtenus en se positionnant par rapport aux quantiles à 95%) et les cercles pleins les mesures significatives. Par définition, la somme des mesures d'une colonne vaut 1. Le nombre de souches (en bleu), le nombre de transitions OUT (en rouge) et le nombre de transitions IN (en vert) sont indiqués pour chaque pays en dessous ou à côté des axes correspondants. Seuls les pays dont le nombre de transitions OUT est supérieur ou égal à 1 sont représentés en ordonnée. Les pays sont classés de la même façon qu'à la Figure 48. Les résultats des méthodes de parcimonie RandACCTRAN et RandDELTRAN sont disponibles aux Figures 56 et 57 de l'Annexe A. Cet indice est utile pour identifier une partie des chaînes de transmission, par exemple, de quel(s) pays est venue l'épidémie du VIH-1C en Éthiopie ? Et à quelle proportion ? On voit pour l'Éthiopie une source principale significative issue du Burundi, une deuxième source moindre non significative issue de la Zambie, et des sources accessoires venant de nombreux pays (dont l'Europe). Cet indice permet aussi d'identifier l'épicentre de l'épidémie du sous-type C : c'est le pays qui est le plus souvent donneur, donc probablement la Zambie.

Figure 49. Estimations de l'indice de flux avec la parcimonie RandDOWNPASS.
Chaque point sur le graphique reflète la proportion de transitions IN pour le pays en abscisse issues du pays en ordonnée. La somme des points d'une colonne vaut 1. Les pays sont ordonnés par zone géographique, puis par ordre alphabétique. Le nombre de souches de chaque pays est indiqué sur les deux axes en face des pays concernés. Sur l'axe des abscisses, la mesure IN indique le nombre de transitions entrantes dans ce pays. Sur l'axe des ordonnées, la mesure OUT indique le nombre de transitions sortantes de ce pays. Les mesures significatives sont représentées par un cercle plein et les mesures non significatives par un cercle vide. Les pays avec un nombre de transitions OUT inférieur à 1 ne sont pas représentés en ordonnée. Le graphique présenté correspond à la parcimonie RandDOWNPASS.



La Figure 50 montre les mesures obtenues par l'indice de symétrie S avec la parcimonie RandDOWNPASS. Cet indice renseigne sur la symétrie des échanges entre pays donneurs. Par commodité, la mesure reportée sur le graphique entre un couple d'annotations a et b correspond à $S_{a\leftrightarrow b}/(N_{a\to b}+$

$N_{b\to a}$). Lorsque le point est rouge (respectivement bleu) il y a plus de mouvement du pays en ordonnée vers le pays en abscisse (resp. du pays en abscisse vers le pays en ordonnée) que l'inverse. Les cercles vides représentent des échanges non statistiquement supportés (obtenus en se positionnant par rapport aux quantiles à 2,5% et 97,5%) et les cercles pleins de mesures statistiquement significatives. Les croix indiquent qu'il n'y a aucune transition entre le couple de pays en question. Plus la taille des points est proche de 1, plus les échanges sont asymétriques, tandis que plus la taille des points est proche de zéro plus les échanges sont symétriques. Par exemple, l'échange entre la Zambie et le Danemark est parfaitement asymétrique (le flux migratoire va uniquement de la Zambie vers le Danemark), tandis que l'échange entre le Niger et l'Afrique du Sud est parfaitement symétrique (autant de flux migratoires du Niger vers l'Afrique du Sud que de l'Afrique du Sud vers le Niger). Les pays sont organisés de la même manière qu'aux Figures 47 et 48. Le nombre de souches correspondant à chaque pays est rappelé en bleu en abscisse et en ordonnée. Le graphique est évidemment symétrique (cf. section 6.2.4). Les Figures 58 et 59 de l'Annexe A correspondent aux estimations de l'indice de symétrie des méthodes de parcimonie RandACCTRAN et RandDELTRAN respectivement. Cet indice est utile pour identifier le pays ou les pays à l'épicentre de l'épidémie puisque le flux doit logiquement être plus important en sortie qu'en entrée, et, dans notre cas, deux pôles impotants sont clairement identifiés : la Zambie dont les flux sont tous majoritairement sortants, même si certains ne sont pas significatifs, et l'Afrique du Sud. Remarquons tout de même que la Zambie est bien l'épicentre de l'épidémie (en accord avec les observations précédentes) puisque le flux épidémique est plus intense de la Zambie vers l'Afrique du Sud.

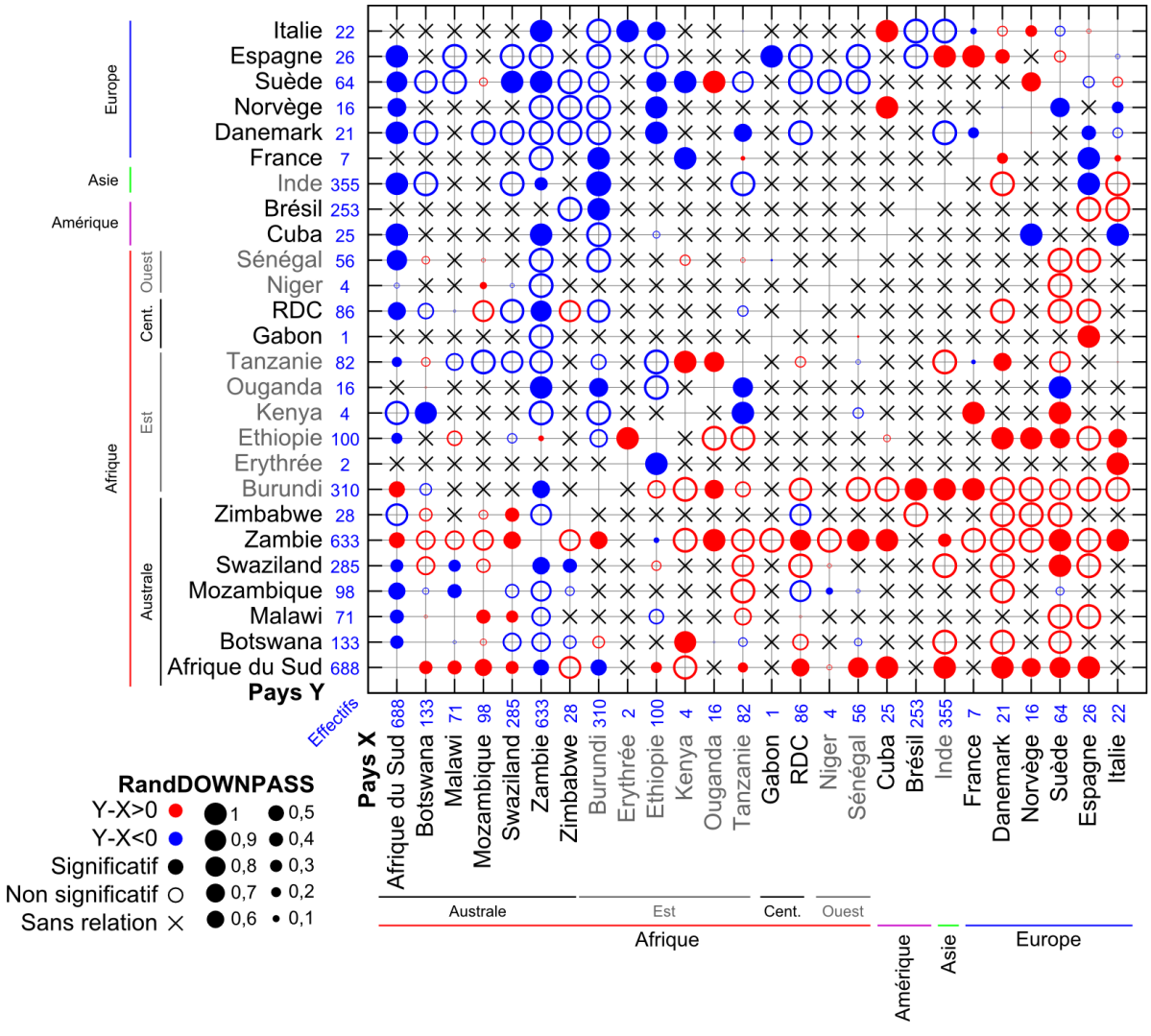
Origine de l'épidémie du sous-type C du VIH-1

L'origine géographique de l'épidémie du VIH-1C est déterminée par l'annotation associée au nœud racine de la phylogénie et les différentes méthodes de parcimonie s'accordent sur la même incertitude en hésitant entre la Zambie, la Tanzanie et la RDC (cf. section 6.3.2). Toutefois, il reste possible d'identifier l'épicentre de cette épidémie pouvant, à l'instar de l'épidémie du VIH-1 (cf. Chapitre 5), être différent du pays d'origine. L'épicentre de l'épidémie est délicat à observer puisque l'annotation correspondante se situe sur les nœuds internes de la phylogénie, à proximité de la racine et suffisamment éloigné des feuilles de celle-ci.

Les mesures de l'indice de dispersion obtenues par la méthode de parcimonie RandDELTRAN montrent que les pays donneurs sont majoritairement des pays africains localisés dans la région australe et est (12 sur 19 donneurs), suggérant que l'épicentre de l'épidémie se situe en Afrique (Figure 56 de l'Annexe A). Le nombre de transitions OUT et le nombre de pays différents vers lesquels la Zambie donne suggèrent que ce pays est l'épicentre de l'épidémie du sous-type C du VIH-1 (518 tran-

sitions OUT sur un total de 976, soit 53% du nombre total de transitions, vers 46 pays sur 62, soit 74% ; Figure 58). Notons toutefois que seule une portion de ces mesures est significative (12 flux sur 46, dont 5 vers d'autres pays de l'Afrique australe), sans doute en raison du nombre très important de souches de Zambie (633). Ce résultat est confirmé par l'indice de symétrie qui indique que la Zambie est le seul pays où le flux migratoire est le plus important en sortie qu'en entrée, quel que soit le deuxième pays auquel on se réfère (Figure 59). L'Afrique du Sud peut être considérée comme un deuxième pôle épidémique, puisque, pour ce pays aussi, le flux est plus important en sortie qu'en entrée. Mais le flux entre la Zambie et l'Afrique du Sud reste plus important du premier vers le second pays, confortant l'hypothèse d'un épicentre en Zambie.

Figure 50. Estimations de l'indice de symétrie avec la parcimonie RandDOWNPASS.
Ce graphique renseigne sur la symétrie des échanges entre les pays donneurs pour la parcimonie RandDOWNPASS. Pour en faciliter la lecture, la mesure $S_{a\leftrightarrow b}/(N_{a\to b}+N_{b\to a})$ est reportée sur le graphique pour tout a et b. Si la mesure est représentée par un point rouge (resp. bleu) cela signifie qu'il y a plus de mouvement du pays en ordonnée (resp. abscisse) vers le pays en abscisse (resp. ordonnée), que l'inverse. Lorsque la mesure vaut 1 et que le point est rouge (resp. bleu), seuls des mouvements du pays Y (resp. X) vers le pays X (resp. Y) sont observés. Lorsqu'elle vaut 0, l'échange est symétrique. Les cercles vides montrent les mesures non significatives et les cerlces pleins les mesures significatives. Les croix indiquent deux pays sans relations. Les pays sont ordonnés par zone géographique, puis par ordre alphabétique, et leur nombre de souches est rappelé sur les axes. Seuls les pays dont le nombre de transitions OUT est supérieur à 1 sont représentés.



L'indice de dispersion de la méthode de parcimonie RandDOWNPASS montre aussi que, pour les pays africains, il y a un flux plus important en sortie qu'en entrée (100% des pays africains sont donneurs), suggérant donc également que l'épidémie est originaire du continent africain (Figure 48). L'indice de flux montre, tout comme celui de la méthode RandDELTRAN, que la Zambie est le pays donnant le plus souvent (377 transitions OUT sur un total de 965, soit 39% du nombre total de transitions, vers de nouveau 46 pays sur 62, soit 74%, Figure 49), mais cette fois-ci 11 flux (sur les 46) sont significatifs, au lieu de 12, et ils ne correspondent pas forcément au même pays receveur. Par exemple, avec RandDELTRAN (resp. RandDOWNPASS), le flux de la Zambie vers le Mozambique (resp. Belgique) est significatif alors qu'il ne l'est pas avec RandDOWNPASS (resp. RandDELTRAN). L'indice de symétrie (Figure 50), quant à lui, montre qu'à nouveau la Zambie est le seul pays dont les flux sont plus importants en sortie qu'en entrée (sauf avec l'Éthiopie ou les flux sont quasiment symétriques), mais la majorité des points (19 sur 24) est non significative. Une observation similaire à celle apportée précédemment à l'Afrique du Sud peut être faite.

Une conclusion identique (Zambie épicentre de l'épidémie du sous-type C) est obtenue avec la méthode de parcimonie RandACCTRAN, dont les résultats sont assez semblables à ceux de la méthode de parcimonie RandDOWNPASS. Cela provient du fait que les nombres de nœuds internes ambigus résultant de l'application des parcimonies RandDOWNPASS (239 nœuds ambigus) et RandACCTRAN (119 nœuds ambigus) sont tous deux élevés, alors que ce nombre est bien plus faible avec la parcimonie RandDELTRAN (11 nœuds ambigus). Les deux premières méthodes sont donc proches, car elles utilisent largement les choix aléatoires, alors que ceux-ci sont bien plus rares avec RandDELTRAN. Notons toutefois qu'avec RandACCTRAN, il y a 335 transitions OUT pour la Zambie sur 969, soit 35% du nombre total de transitions OUT (Figure 57), contre 39% avec RandDOWNPASS (et 53% avec RandDELTRAN). C'est un effet attendu, compte tenu des choix (plus ou moins arbitraires) effectués par ACCTRAN et DELTRAN de « pousser » les transitions vers les feuilles ou la racine, alors que DOWNPASS ne tranche pas. Malgré ces différences quantitatives, les trois méthodes s'accordent sur l'essentiel, à savoir que la Zambie est probablement l'épicentre de l'épidémie du sous-type C.

Flux migratoires du sous-type C du VIH-1

L'observation des flux migratoires se fait à l'aide de l'indice de flux qui indique précisément que tel pays donne à tel autre pays. Globalement, les mesures de cet indice montrent de fortes interactions entre les pays de l'Afrique australe ou pratiquement chaque pays donne et reçoit de tous les autres pays de l'Afrique australe. Une observation identique peut être faite avec les pays de l'Afrique de l'est, mais le signal est moins soutenu. Ces deux observations indiquent que les échanges semblent se faire principalement entre pays géographiquement proches. Les pays européens reçoivent

majoritairement des pays africains, mais des échanges marginaux entre pays européens sont aussi à noter. Précisément, les flux significatifs entre pays représentés par au moins 20 souches et pour lesquels le pays en entrée a au moins 10 transitions IN, sont présentés dans le Tableau 5. Ces flux indiquent donc des mouvements épidémiques importants (plusieurs introductions) entre deux pays, qui peuvent être à l'origine d'évènements fondateurs ou de cas isolés (les conséquences de ces introductions ne peuvent pas être connues avec cet indice). Les résultats montrent bien les pays ayant des liens épidémiques forts, notamment les pays de l'Afrique australe (11 flux sur 19 sont identifiés entre deux pays de l'Afrique australe).

Tableau 5. Flux significatifs déduits des mesures de l'indice de flux.
Les flux présentés dans ce tableau proviennent de pays représentés par au moins 20 souches, avec un minimum de 10 transitions IN. Le nombre de transitions IN est indiqué pour chaque algorithme de parcimonie, et surligné en gris lorsque le flux est significatif. Seuls les flux significatifs pour au moins un algorithme sont indiqués. Les mesures entre parenthèses ne respectent pas la condition du minium de 10 transitions IN, mais sont données à titre indicatif.

Pays		Nombre de transitions IN		
De	Vers	RandDELTRAN	RandACCTRAN	RandDOWNPASS
Zambie	Afrique du Sud	114	80	87
	Botswana	51	27	32
	Malawi	28	17	20
	Mozambique	41	21	25
	Swaziland	54	39	43
	Tanzanie	28	20	22
	RDC	52	43	45
	Inde	10	(4)	(6)
	Sénégal	12	(7)	(8)
Afrique du Sud	Malawi	22	20	22
Botswana	Zambie	(1)	10	(1)
	Afrique du Sud	10	(9)	(9)
Malawi	Afrique du Sud	(6)	11	(6)
Swaziland	Afrique du Sud	11	19	16
	Zambie	(6)	13	10
Éthiopie	Suède	15	14	14
	Italie	12	10	10
Burundi	Éthiopie	11	(8)	(9)
	Tanzanie	11	(6)	(7)

On voit à nouveau dans le Tableau 5 une origine ou épicentre probable en Zambie, avec cependant des flux significatifs de retour vers la Zambie de souches venant du Botswana et du Malawi. La présence de flux IN et OUT significatifs est également visible entre le Malawi et l'Afrique du Sud qui est la plaque tournante du transport (notamment aérien) en Afrique australe et donc logiquement aussi pour la diffusion de l'épidémie. On retrouve des transmissions connues (par exemple du Burundi vers l'Éthiopie) ou historiquement explicable (par exemple l'Éthiopie vers l'Italie).

Le Tableau 6 donne les mesures entre les pays a et b donneurs étant chacun représentés par plus de 20 séquences et pour lesquels le nombre de transitions $N_{a\to b} + N_{b\to a}$ est supérieur ou égal à 10, afin d'observer la tendance du mouvement de l'épidémie entre ces pays (plutôt de a vers b ou plutôt

de b vers a ?). Ces mesures montrent presque tout le temps des mouvements unidirectionnels, par exemple, le flux migratoire est plus intense en sortie de la Zambie qu'en entrée (>0,4), confirmant un mouvement épidémique de la Zambie vers l'Afrique australe (Afrique du Sud, Botswana et Swaziland), la RDC et l'Inde. Les mouvements épidémiques avec l'Afrique du Sud sont nettement moins unidirectionnels, en particulier avec RandACCTRAN (<0,4), suggérant des échanges épidémiques réguliers, dans les deux sens, avec les pays de l'Afrique australe (Botswana, Malawi et Swaziland), comme déjà discuté ci-dessus. Il y a cependant une exception avec le Mozambique pour lequel le flux est largement unidirectionnel (>0,5). Enfin, les mouvements de l'Éthiopie vers les pays européens (Italie et Suède) sont aussi unidirectionnels (>0,5), tout comme ceux du Burundi vers les pays de l'Afrique de l'est (Éthiopie et Tanzanie ; >0,6), suggérant une chaîne de transmission du Burundi vers l'Éthiopie, puis de l'Éthiopie vers la Suède et Italie (cf. ci-dessus).

Tableau 6. Mesures significatives et remarquables de l'indice de symétrie.
Ce tableau présente les mesures de l'indice de symétrie entre les pays a et b donneurs à fort effectif (>20 séquences) et pour lesquels le nombre de transitions $N_{a\to b} + N_{b\to a} \geq 10$. La mesure correspondante à l'indice de symétrie est présentée pour chaque algorithme de parcimonie, et surlignée en gris lorsqu'elle est significative. Seules les mesures significatives pour au moins un algorithme sont indiquées. Les mesures entre parenthèses ne respectent pas la condition $N_{a\to b} + N_{b\to a} \geq 10$ et sont données à titre indicatif.

Pays		Mesure de l'indice de symétrie		
De	Vers	RandDELTRAN	RandACCTRAN	RandDOWNPASS
Zambie	Afrique du Sud	0,69	0,43	0,51
	Botswana	0,96	0,44	0,96
	RDC	0,99	0,80	0,87
	Swaziland	0,80	0,50	0,61
	Inde	0,67	(0,19)	(0,67)
Afrique du Sud	Botswana	0,11	0,37	0,35
	Malawi	0,57	0,27	0,40
	Swaziland	0,48	0,26	0,33
	Mozambique	0,72	0,50	0,59
Éthiopie	Italie	-	0,59	0,69
	Suède	0,76	0,86	0,79
Burundi	Éthiopie	0,69	0,47	0,55
	Tanzanie	0,69	0,25	0,41

Les mesures de flux présentées ici sont globales. Elles peuvent résulter d'un évènement fondateur unique et bien visible (par exemple du Brésil vers l'Uruguay), ou bien de transmissions multiples, sans qu'on discerne clairement les effets fondateurs, s'il y en a (par exemple entre la Zambie et l'Afrique du Sud). L'utilisation du logiciel PhyloType va précisément relever les chaînes de transmission complètes issues d'évènements fondateurs.

6.3.4 Recherche des chaînes de transmission majeures du VIH-1C avec PhyloType

Les chaînes de transmission du VIH-1C sont déterminées à l'aide de la méthode PhyloType qui met en exergue des *phylotypes*, reflets d'évènements fondateurs probables, mais surtout les liens qui les unissent. Ces liens sont difficilement observables avec les indices présentés ci-avant, en particulier si les chaînes de transmission traversent plus de deux pays ou correspondent à des flux faibles mais essentiels (fondateurs). Cette analyse vient donc en complément de celles présentées précédemment. Mais avant cela, nous proposons une analyse visant à déterminer quels regroupements peuvent être faits afin d'intégrer le maximum d'information (de feuilles) dans l'analyse PhyloType, puisque nous nous limiterons à des *phylotypes* d'une certaine taille, excluant d'office certaines annotations peu représentées.

6.3.4.1 Associations d'annotations pour l'analyse avec PhyloType

Les chaînes de transmission de l'épidémie du VIH-1C sont déduites de la phylogénie des 3 609 séquences à l'aide du logiciel PhyloType. Auparavant, les souches des pays peu régionalisées (probabilité faible de formation de *phylotypes*) et en faible effectif (représentés par moins de 20 souches) sont étudiées afin de proposer des associations d'annotations permettant l'émergence de *phylotypes* portant sur ces combinaisons d'annotation. Sans ces regroupements, il y a peu de chance qu'un *phylotype* avec ces annotations apparaisse et, donc, l'information fournie par ces souches est perdue. La Figure 51 présente l'indice de régionalisation R (cf. section 6.2.5.2) appliquée pour chaque pays africain (en bleu) et pour chaque paire de pays africains (en vert et rouge). Des tableaux similaires sont donnés en annexe pour les autres continents. Lorsque le point est rouge la régionalisation est strictement inférieure au minimum des régionalisations des deux pays. Une croix noire symbolise deux pays ayant une frontière géographique commune, règle importante dans le choix des associations. On trouve ainsi 20 points remarquables (rouges), pouvant relever d'une association. Sur ceux-ci, 15 sont observés sur des couples de pays appartenant à la même région géographique (Afrique australe, est, ouest, etc.) et 13 correspondent à un couple de pays partageant une frontière géographique. Ce résultat est particulièrement frappant et montre que la géographie des pays africains et la phylogénie sont fortement corrélées, confirmant ainsi que les échanges se font avant tout entre pays proches et que nos indices sont à même de détecter ce signal, bien que simples et basés sur des calculs rapides de parcimonie. Aucune différence entre les résultats obtenus par ACCTRAN, DELTRAN ou DOWNPASS n'est à noter puisque, pour cet indice, seulement deux annotations sont considérées (a versus $\neg a$, cf. section 6.2.5.2). Cela a pour effet que le nombre total de transitions de chaque méthode de parcimonie est égal. Plusieurs associations sont suggérées par ce graphique, comme la République Démocratique du Congo (RDC) avec la République du Congo ou encore la Tanzanie avec

l'Ouganda et/ou le Kenya. Remarquons qu'avec ce dernier cas, si nous souhaitons associer ces trois pays ensembles, il est nécessaire de calculer l'indice de régionalisation correspondant à l'union des trois annotations afin de le comparer aux indices de régionalisation de la Tanzanie, de l'Ouganda et du Kenya de manière à garantir un gain global en régionalisation.

À cet effet, le Tableau 7 récapitule la liste de toutes les associations successives suggérées par l'indice de régionalisation R, mais uniquement pour les pays à faible effectif (<20) et pour ceux qui partagent une frontière géographique. Les lignes grisées indiquent les associations retenues pour l'analyse PhyloType où le critère *size* est supérieur ou égal à 20, tandis que les autres lignes indiquent des associations temporaires ayant permis de les obtenir. Ce tableau est construit itérativement de la manière suivante :

1. Chaque annotation ayant un faible effectif (<20) est associée avec celles partageant une frontière géographique ;
2. L'indice de régionalisation est calculé pour les deux annotations et pour leur union ;
3. Si l'indice de l'union indique une meilleure régionalisation alors l'union de ces deux annotations est considérée par la suite, sinon l'association n'est pas retenue ;
4. Une fois la procédure finie, la première étape est répétée avec les nouvelles associations et cela jusqu'à convergence.

Les associations retenues par cette procédure sont donc :

- le Congo et la RDC ;
- la Birmanie, la Chine et l'Inde ;
- l'Espagne et la France ;
- l'Argentine et le Brésil ;
- le Kenya, l'Ouganda et la Tanzanie ;
- la Norvège et la Suède ;
- Djibouti, l'Érythrée, l'Éthiopie et le Soudan.

Toutefois, certaines de ces associations perdent leur intérêt lorsque le critère *size* est supérieur ou égal à 10, puisque certaines annotations sont alors suffisamment représentées. Ainsi, les associations retenues lorsque le critère *size* est supérieur ou égal à 10 sont :

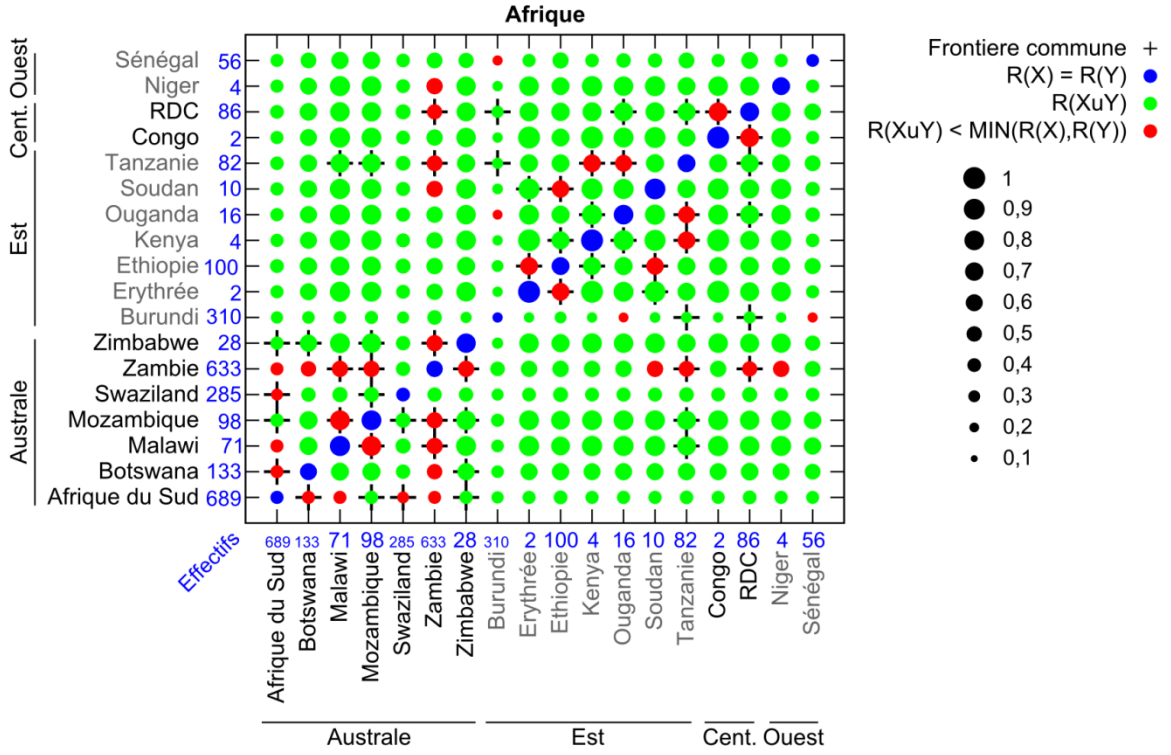
- le Congo et la RDC ;
- la Birmanie, la Chine et l'Inde ;
- l'Espagne et la France ;

- l'Argentine et le Brésil ;
- le Kenya et la Tanzanie ;
- l'Érythrée et l'Éthiopie.

Et lorsqu'il est supérieur ou égal à 5 :

- le Congo et la RDC ;
- la Birmanie et l'Inde ;
- le Kenya et la Tanzanie ;
- l'Érythrée et l'Éthiopie.

Figure 51. Estimations de l'indice de régionalisation entre les souches de pays africains.
Ce graphique renseigne sur la possibilité de grouper deux pays ensembles lors de l'analyse avec PhyloType. Les couples de pays ayant une croix sont ceux qui partagent une frontière géographique commune. Les points en bleu indiquent l'indice de régionalisation. Les points verts et rouges indiquent l'indice de régionalisation de l'union des deux pays situés sur l'axe des ordonnées et des abscisses. Lorsque ce point est en rouge la régionalisation de l'union est meilleure que la régionalisation des deux pays pris séparément. Par exemple, la mesure pour le couple RDC/Congo indique que l'union des deux pays est plus régionalisée (point rouge) que celle des pays pris séparément. De plus, ces deux pays partagent une frontière géographique commune (une croix), il est donc conseillé de les grouper ensembles lors de l'analyse avec PhyloType afin de maximiser les chances d'apparition de *phylotypes*. Les pays sont regroupés par zone géographique et seuls les pays ayant au moins deux souches sont représentés. Seuls des groupements entre pays d'un même continent sont envisagés. Il n'y a pas de différence entre les méthodes DELTRAN, ACCTRAN et DOWNPASS. Voir l'Annexe A pour les graphiques des autres continents.



6.3.4.2 Analyse des chaînes de transmission du VIH-1C avec PhyloType

Les chaînes de transmission du VIH-1C sont étudiées et analysées avec l'outil PhyloType (Chevenet *et al*) qui met en exergue des *phylotypes*, reflets d'évènements fondateurs, ainsi que les liens

épidémiologiques qui les unissent. Trois analyses sont faites (*size* ≥ 20, 10 et 5) avec les parcimonies ACCTRAN et DELTRAN respectivement. La parcimonie DOWNPASS n'est pas disponible dans PhyloType car générant trop d'ambigüités sur les annotations ancestrales. Quelles que soient les analyses, les autres critères utilisés sont *persistence*, *size*/*different* et *support* respectivement supérieurs ou égaux à 1, 1 et 70%. Enfin, pour mesurer la significativité statistique des résultats, 1 000 *shufflings* sont calculés et les *phylotypes* avec une p-valeur strictement supérieure à 10/1 000 (= 1%) pour le critère *size* ne sont pas conservés. À nouveau, les résultats des analyses PhyloType sont d'abord présentés, puis nous les discuterons.

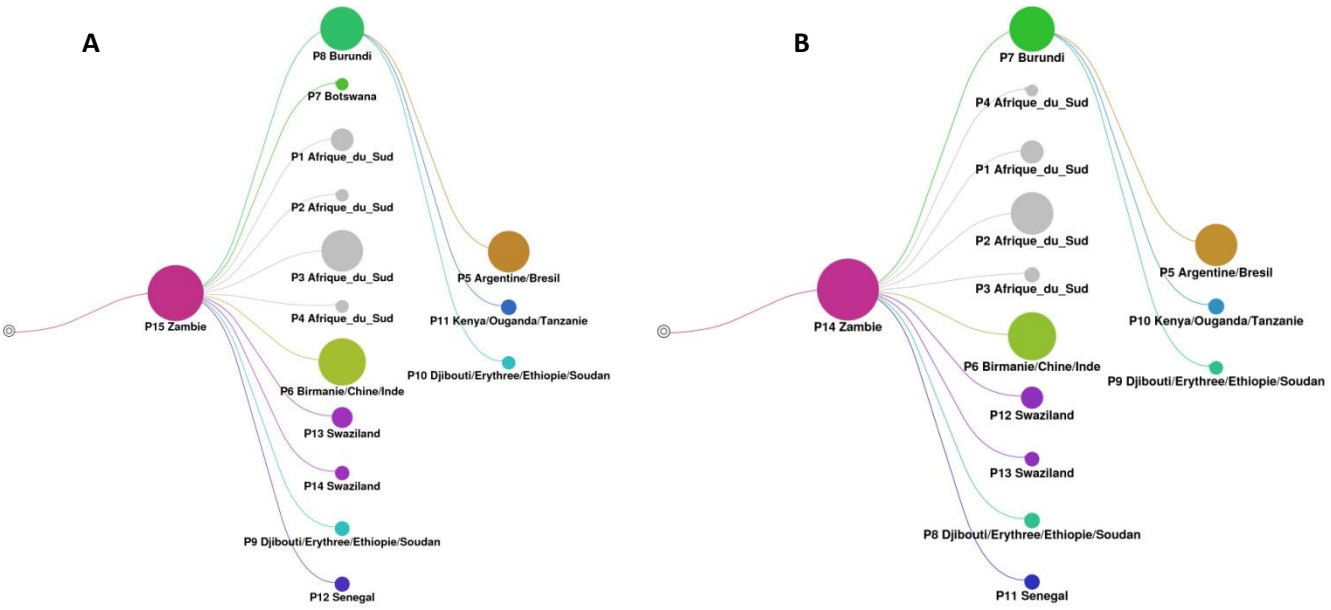
Tableau 7. Liste des associations suggérées par l'indice de régionalisation.
Ce tableau indique toutes les associations successives suggérées par l'indice de régionalisation. Les effectifs et l'indice de régionalisation correspondant à chaque annotation sont rappelés et l'estimation de leur union indiquée. Les lignes grisées montrent les associations retenues pour *size* ≥ 20, tandis que les lignes non grisées indiquent les associations temporaires. Les mesures sont identiques entre les parcimonies DELTRAN, ACCTRAN et DOWNPASS.

Association		**Effectifs**		**Régionalisation**		
$\boldsymbol{A}$	$\boldsymbol{B}$	$\boldsymbol{A}$	$\boldsymbol{B}$	$\boldsymbol{A}$	$\boldsymbol{B}$	$\boldsymbol{A} \cup \boldsymbol{B}$
Birmanie	Inde	1	355	-	0,090	0,087
Congo	RDC	2	86	1,000	0,741	0,736
Érythrée	Éthiopie	2	100	1,000	0,667	0,663
Kenya	Tanzanie	4	82	1,000	0,667	0,659
France	Espagne	7	26	0,833	0,800	0,781
Argentine	Brésil	8	253	0,714	0,044	0,023
Norvège	Suède	16	64	1,000	0,825	0,823
Chine	Birmanie/Inde	7	356	0,833	0,087	0,080
Soudan	Érythrée/Éthiopie	10	102	0,889	0,663	0,622
Ouganda	Kenya/Tanzanie	16	86	0,800	0,659	0,594
Djibouti	Érythrée/Éthiopie/Soudan	1	112	-	0,622	0,616

Les Figures 51 et 52 montrent un diagramme où les différents *phylotypes* significatifs observés, représentés par des cercles de surface proportionnelle à leur taille (*size*), sont disposés en fonction de leur apparition le long de la phylogénie. Les inclusions sont représentées par des arêtes reliant deux phylotypes. Ces figures correspondent aux résultats obtenus lorsque le critère *size* est respectivement supérieur ou égal à 20 et à 5. La Figure 52A montre les résultats de la parcimonie ACCTRAN, tandis que la Figure 52B et la Figure 53 donnent les résultats de la parcimonie DELTRAN. Les résultats de la parcimonie ACCTRAN, lorsque le critère *size* est supérieur ou égal à 10 et à 5, sont disponibles dans l'Annexe A et correspondent respectivement aux Figures 61 et 63. Les résultats de la parcimonie DELTRAN lorsque le critère *size* est supérieur ou égal à 10 sont aussi disponibles dans l'Annexe A (Figure 63). Tous les *phylotypes* représentés sont significatifs ($p \leq 1\%$ pour le critère *size*). Un numéro d'identification est attribué à chaque *phylotype*. Il est ainsi possible de connaître explicitement

les valeurs associées à chaque critère proposé par PhyloType (cf. section 6.2.5.1) en se reportant dans le tableau correspondant.

Figure 52. Cartes des liens entre les *phylotypes* des analyses avec *size* ≥ 20 pour ACCTRAN et DELTRAN.
Cartes des analyses PhyloType (ACCTRAN en figure A et DELTRAN en figure B) avec *size* ≥ 20, *persistence* ≥ 1, *size/different* ≥ 1 et *support* ≥ 70%. Tous les *phylotypes* sont statistiquement supportés (p-valeur inférieure ou égale à 1% pour le critère *size*). La taille des cercles est proportionnelle à la valeur du critère *size* pour le *phylotype* en question. Chaque couleur correspond à une annotation donnée, spécifiée au bas de chaque cercle. Un numéro d'identification est attribué à chaque *phylotype* et est indiqué avant l'annotation correspondante au *phylotype*.



Le Tableau 8 (resp. Tableau 9 de l'Annexe A) liste les *phylotypes* significatifs obtenus ($p \leq 1\%$ pour le critère *size*), les valeurs correspondantes à chaque critère et les p-valeurs (indiquées en rose) associées uniquement aux critères choisis (en gras) pour la parcimonie DELTRAN (resp. ACCTRAN) et lorsque le critère *size* est supérieur ou égal à 20. Les résultats de la parcimonie ACCTRAN (resp. DELTRAN) où le critère *size* est supérieur ou égal à 10 et à 5 correspondent respectivement aux Tableaux 10 et 12 (resp. Tableaux 11 et 13) de l'Annexe A. Le numéro d'identification de chaque phylotype est donné dans la colonne P. Lorsque le critère *different* (resp. *size/different*) est à 0 (resp. infini) alors le *phylotype* en question définit un clade. Par exemple, le *phylotype* n°11 du Tableau 8, représentant les souches des MSM de l'étude sur le Sénégal (cf. Chapitre 5), est un clade.

Origine de l'épidémie du sous-type C du VIH-1

Toutes les analyses PhyloType s'accordent à identifier un *phylotype* annoté Zambie (racine supportée à 88,8% en valeur aLRT (numéro 14 dans le Tableau 8 et numéro 15 dans le Tableau 5) à l'origine de l'épidémie du VIH-1C. Remarquons toutefois que le nombre total de souches incluses dans ce *phylotype* (critère *total*) n'englobe pas toutes les séquences de la phylogénie (3 605 souches sur 3 608 séquences étudiées), signifiant que la racine de ce *phylotype* ne correspond pas à la racine de la phylogénie comme déjà discuté plus haut. Les analyses PhyloType ne révèlent donc pas avec certitude l'origine de l'épidémie (indéterminée entre Zambie, RDC et Tanzanie, d'après les analyses

précédentes) mais plutôt l'épicentre de celle-ci, c'est-à-dire la région géographique à l'origine de la diffusion massive du VIH-1C. Ce résultat est en accord avec (et conforte) ceux observés précédemment.

Chaînes de transmission du sous-type C du VIH-1

Les chaînes de transmission majeures du VIH-1C sont facilement observables à l'aide des graphiques générés automatiquement par PhyloType qui synthétisent les liens entre les *phylotypes* significatifs observés. Les grands flux géographiques sont donnés par les analyses où le critère *size* est supérieur ou égal à 20. Ils sont aussi observés sur les autres analyses mais de nombreux *phylotypes* secondaires (de moindre importance en termes de nombre de membres) compliquent leur interprétation. Mais tous les *phylotypes* observés lorsque le critère *size* est supérieur ou égal à 20, le sont aussi lorsqu'il est supérieur ou égal à 10 et ceux observés lorsqu'il est supérieur ou égal à 10 sont, quant à eux, aussi observés lorsqu'il est supérieur ou égal à 5.

D'après l'analyse de la parcimonie DELTRAN présentant les grands flux migratoires (*size* ≥ 20), l'épidémie du VIH-1C se diffuse indépendamment de la Zambie vers l'Afrique australe (Swaziland [*phylotypes* n°13 et n°14] et Afrique du Sud [*phylotypes* n°1, n°2, n°3 et n°4]), vers l'Afrique de l'est (Burundi [*phylotype* n°7] et Djibouti/Érythrée/Éthiopie/Soudan [*phylotype* n°8]), vers le Sénégal (*phylotype* n°11 ; contenant les souches des MSM de l'étude sur le Sénégal) et vers le continent asiatique (Birmanie/Chine/Inde [*phylotype* n°6]) (Figure 52B). Du Burundi (*phylotype* contenant la presque totalité des souches collectées au Burundi, soit 300 sur 310, Tableau 8) l'épidémie se diffuse en Afrique de l'est (Kenya/Ouganda/Tanzanie [*phylotype* n°10]) et à nouveau vers les pays de la Corne de l'Afrique et le Soudan (Djibouti/Érythrée/Éthiopie/Soudan [*phylotype* n°9]). L'analyse avec ACCTRAN donne des résultats très similaires, mais ajoute un *phylotype* annoté Botswana (n°7), inclus dans le *phylotype* principal annoté Zambie (n°15) (Figure 52A). On retrouve notamment dans ces deux approches la double origine de l'épidémie en Éthiopie et dans les pays proches, issue directement de Zambie et du Burundi, et correspondant probablement aux variants C et C' référencés par Abebe *et al*. (2000).

Les analyses de la parcimonie DELTRAN où le critère *size* est moins restrictif (p. ex. Figure 53) montrent la formation de deux *phylotypes* annotés Roumanie (n°37 et n°36) (resp. Cuba [n°25 et n°26]) d'origine géographique différente (Burundi et Zambie). Hormis la Roumanie, les *phylotypes* représentant des pays européens (particulièrement la Belgique [*phylotype* n°14], le Portugal [*phylotype* n°35] et la Suède [*phylotype* n°39]) ont tous pour origine le *phylotype* principal Zambie. À nouveau les analyses avec la parcimonie ACCTRAN confirment ces observations, sauf en ce qui concerne un *phylotype* annoté Roumanie (n°40) qui n'a plus pour origine le *phylotype* principal Zambie (n°53)

mais un *phylotype* annoté Botswana (n°21) ayant pour origine le *phylotype* principal Zambie (Figure 64 de l'Annexe A).

Figure 53. Carte des liens entre *phylotypes* lorsque *size* ≥ 5 avec DELTRAN.
Carte de l'analyse PhyloType lorsque *size* ≥ 5, *persistence* ≥ 1, *size/different* ≥ 1 et *support* ≥ 70% avec la parcimonie DELTRAN. Tous les *phylotypes* présentés sont statistiquement supportés (p-valeur inférieure ou égale à 1% pour le critère *size*). La taille des cercles est proportionnelle à la valeur du critère *size* pour le *phylotype* en question. Chaque couleur correspond à une annotation donnée, spécifiée au bas de chaque cercle. Un numéro d'identification est attribué à chaque *phylotype* et est indiqué avant l'annotation correspondante au *phylotype*.

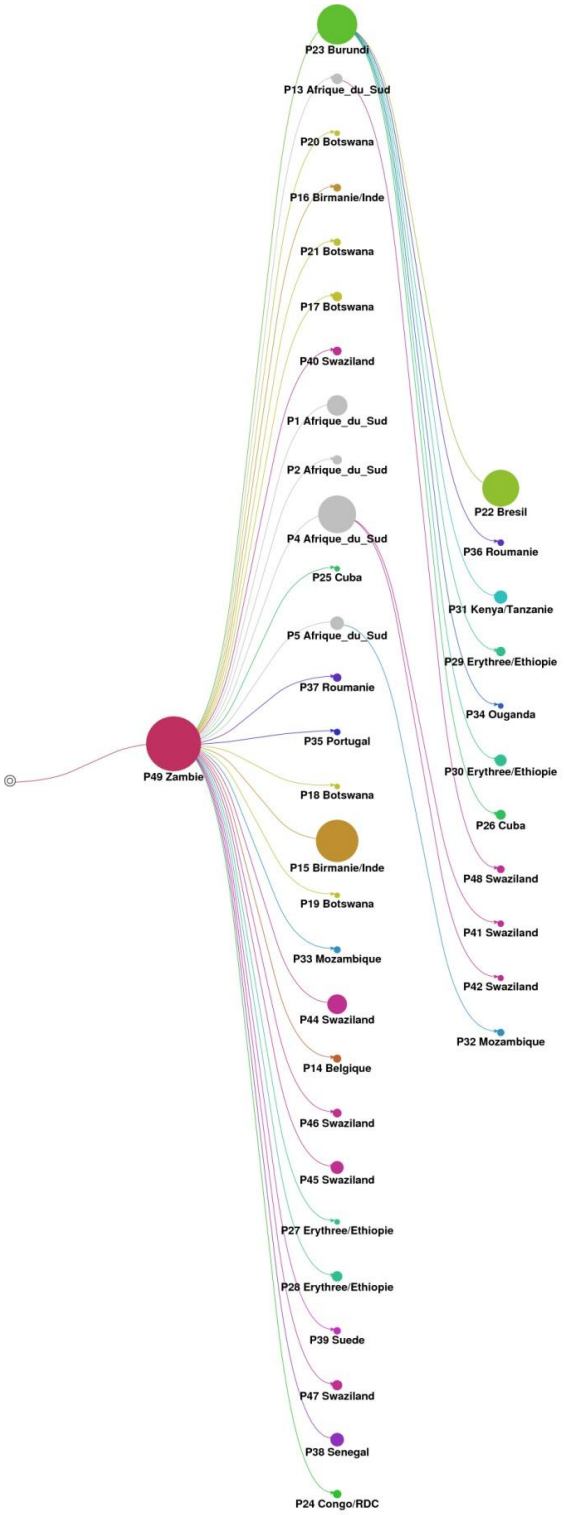


Tableau 8. Valeurs associées à chaque critère pour tous les *phylotypes* significatifs observés lorsque *size* ≥ 20 avec DELTRAN.

Ce tableau contient la liste de tous les *phylotypes* obtenus lorsque *size* ≥ 20, *persistence* ≥ 1, *size/different* ≥ 1 et *support* ≥ 70% avec la parcimonie DELTRAN, ainsi que toutes les valeurs associées à chacun des critères que l'on peut choisir. Les valeurs en rose représentent les p-valeur obtenues par le *shuffling*. Seuls les *phylotypes* dont la p-valeur est strictement inférieure à 11/1 000 pour le critère *size* sont représentés. Les abréviations des pays sont les suivantes : AR, Argentine ; BI, Burundi ; BR, Brésil ; CN, Chine ; DJ, Djibouti ; ER, Érythrée ; ET, Éthiopie ; IN, Inde ; KE, Kenya ; MM, Birmanie ; SD, Soudan ; SN, Sénégal ; SZ, Swaziland ; TZ, Tanzanie ; UG, Ouganda ; ZA, Afrique du Sud ; ZM, Zambie. Les abréviations des titres sont les suivantes : P, identifiant du *phylotype* ; A, *annotation* ; Tt, *total* ; Sz, *size* ; Ps, *persistence* ; Df, *different* ; Sl, *local separation* ; Sg, *global separation* ; Dv, *diversity* ; Sp, *support* ; Spg, *global support*.

P	A	Tt	Sz	Ps	Df	Sz/Df	Sl	Sg	Dv	Sl/Dv	Sg/Dv	Sp	Spg
1	ZA	86	**76**	**2**	9	**8,444**	0,005	0,007	0,083	0,064	0,079	**0,855**	0,855
			0/1000	0/1000		0/1000						0/1000	
2	ZA	311	**265**	**3**	33	**8,030**	0,002	0,014	0,089	0,027	0,154	**0,831**	0,843
			0/1000	0/1000		0/1000						0/1000	
3	ZA	80	**33**	**2**	29	**1,138**	0,007	0,010	0,075	0,092	0,137	**0,876**	0,876
			0/1000	0/1000		0/1000						0/1000	
4	ZA	51	**20**	**2**	16	**1,250**	0,005	0,008	0,071	0,069	0,108	**0,874**	0,874
			0/1000	0/1000		0/1000						0/1000	
5	AR/BR	269	**260**	**2**	5	**52,000**	0,018	0,029	0,106	0,172	0,269	**0,992**	0,992
			0/1000	0/1000		0/1000						0/1000	
6	MM/CN/IN	356	**339**	**2**	14	**24,214**	0,004	0,022	0,081	0,044	0,265	**0,740**	0,882
			0/1000	0/1000		0/1000						0/1000	
7	BI	829	**300**	**3**	75	**4,000**	0,006	0,010	0,093	0,065	0,106	**0,861**	0,861
			0/1000	0/1000		0/1000						0/1000	
8	DJ/ER/ET/SD	71	**34**	**3**	34	**1,000**	0,005	0,014	0,051	0,095	0,264	**0,906**	0,906
			0/1000	0/1000		0/1000						0/1000	
9	DJ/ER/ET/SD	47	**26**	**2**	17	**1,529**	0,003	0,023	0,059	0,042	0,394	**0,773**	0,773
			0/1000	0/1000		0/1000						0/1000	
10	KE/UG/TZ	43	**36**	**3**	7	**5,143**	0,009	0,014	0,083	0,104	0,171	**0,926**	0,926
			0/1000	0/1000		0/1000						0/1000	
11	SN	33	**33**	**1**	0	∞	0,018	0,033	0,075	0,240	0,438	**0,980**	0,980
			0/1000	0/1000		0/1000						0/1000	
12	SZ	87	**70**	**3**	13	**5,385**	0,003	0,045	0,077	0,035	0,594	**0,749**	0,766
			0/1000	0/1000		0/1000						0/1000	
13	SZ	33	**30**	**2**	3	**10,000**	0,002	0,021	0,059	0,041	0,352	**0,781**	0,781
			0/1000	0/1000		0/1000						0/1000	
14	ZM	3605	**564**	**4**	490	**1,151**	0,014	0	0,114	0,120	0,000	**0,880**	0,880
			0/1000	0/1000		0/1000						0/1000	

6.4 Conclusion

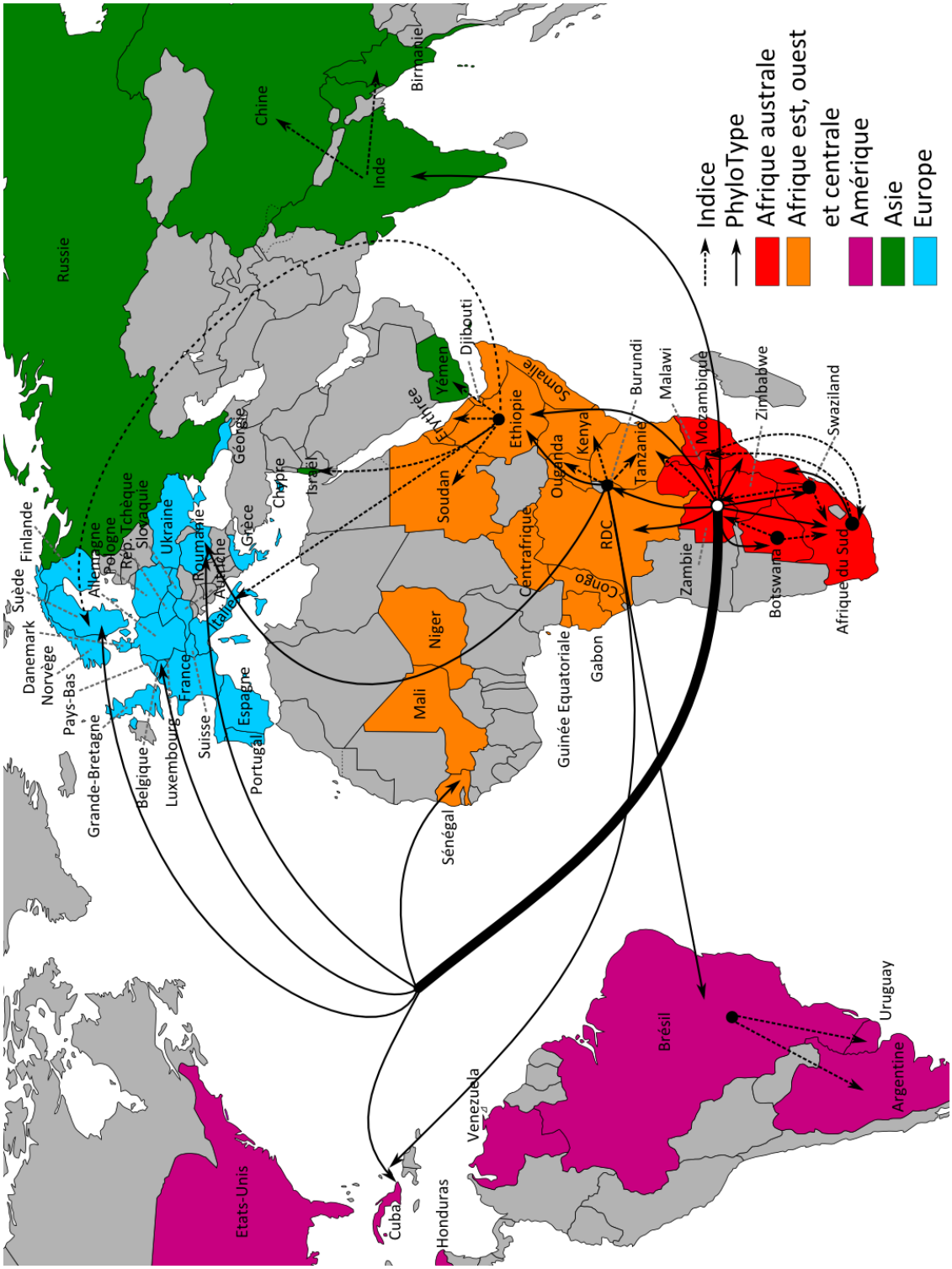
Nous présentons la première étude moléculaire visant à retracer l'histoire épidémiologique du sous-type C du VIH-1 à l'échelle mondiale en s'aidant du maximum de souches disponibles. Pour cela un arbre de maximum de vraisemblance est calculé et comprend 3 609 souches *pol* (dont 528 sont nouvelles) collectées à travers le monde (63 pays différents) et à différentes époques (entre 1986 et 2010). Cette phylogénie est exploitée de deux manières différentes mais complémentaire, toutes deux basées sur le principe de parcimonie. La première utilise des indices basés sur les transitions entre pays (reconstruites par parcimonie) pour synthétiser le mouvement de l'épidémie décrit par la phylogénie. La deuxième utilise le logiciel PhyloType qui met en exergue des *phylotypes*, groupes de séquences reflétant des évènements fondateurs probables, afin d'observer les chaînes de transmission majeures de l'épidémie du sous-type C du VIH-1.

En supposant que les séquences utilisées dans cette étude fournissent une représentation adéquate de la diversité globale du sous-type C du VIH-1 et que la phylogénie obtenue soit la plus juste possible, notre étude suggère que l'épicentre de l'épidémie du sous-type C du VIH-1 se situe en Zambie (Figure 54). Cette épidémie s'est ensuite diffusée indépendamment dans les pays de l'Afrique australe, au Burundi, en Éthiopie, au Sénégal et en Inde. Du Burundi, l'épidémie du sous-type C s'est diffusée dans les pays à l'est (Ouganda, Kenya et Tanzanie), à nouveau en Éthiopie et au Brésil (Véras *et al*, 2011a; de Oliveira *et al*, 2010; Jones *et al*, 2009; Bello *et al*, 2008; Fontella *et al*, 2008). De l'Éthiopie l'épidémie s'est répandue dans les autres pays de la Corne de l'Afrique (Érythrée, Djibouti), au Soudan et dans les pays du Moyen-Orient (Israël, Yémen) (Gehring *et al*, 1997). Du Brésil, l'épidémie s'est propagée dans le sud de l'Amérique Latine (Argentine et Uruguay) (Carrion *et al*, 2004) et de l'Inde dans les pays de l'Asie de l'est (Birmanie, Chine, Corée du Sud et Taiwan) (Lau *et al*, 2007). Enfin, notons les nombreuses introductions de ce variant en Europe provenant de pays africains et qui évoquent les liens sociologiques entre les pays européens et africains créés durant l'époque coloniale (Perrin *et al*, 2003).

Notre étude n'a pu déterminer avec exactitude l'origine géographique du sous-type C du VIH-1. Toutefois, les différentes méthodes de parcimonie utilisées (DELTRAN, ACCTRAN et DOWNPASS) s'accordent sur l'incertitude à donner au nœud racine (RDC, Zambie ou Tanzanie) (Figure 47). Ces trois pays se situent sur le continent africain et sont frontaliers. Au vu de l'étonnante diversité génétique présente en RDC (Vidal *et al*, 2000) et sachant que deux souches isolées à partir de matériels anciens (1958 et 1960) en RDC, présentaient déjà une diversité génétique étonnante (Worobey *et al*, 2008; Zhu *et al*, 1998), la RDC est généralement vue comme l'épicentre de l'épidémie du VIH-1 et il serait donc probable, dans cette hypothèse, que l'origine géographique du sous-type C se situe également en RDC. Ceci d'autant plus qu'un nombre important de séquences (21 sur 45, soit 47%) collectées en RDC se trouvent à proximité de la racine (Figure 47). Malgré cela, l'hypothèse d'une origine zambienne est aussi très probable, puisque, premièrement, ce pays est identifié comme l'épicentre de l'épidémie du sous-type C et, deuxièmement, la prévlence du sous-type C en RDC est surtout observée au sud du pays (Vidal *et al*, 2005), à proximité de la frontière avec la Zambie. D'ailleurs, sur les 21 souches de RDC situées à proximité de la racine, 20 sont collectées à Mbuji-Mayi ou Lumbumbashi, deux villes au sud de la RDC. Au vu de la situation géographique particulière entre la RDC et la Zambie (l'appendice au sud-est de la RDC traverse pratiquement la Zambie), facilitant certainement les migrations de populations entre ces deux pays (par exemple, pour traverser la Zambie d'est en ouest), l'argumentation en faveur de l'un ou l'autre pays devient difficile et il serait plus vraisemblable de supposer que l'origine de l'épidémie du sous-type C se situe au niveau de la région frontalière entre ces deux pays ; zone riche en industrie minière (p. ex. cuivre ou cobalt) et où

les mouvements de populations sont donc nombreuses. En revanche, l'hypothèse d'une origine tanzanienne reste assez peu probable, mais elle ne peut être complètement rejetée.

Figure 54. Planisphère résumant la diffusion de l'épidémie du sous-type C du VIH-1.
Les flèches représentent les mouvements de l'épidémie du sous-type C du VIH-1 dans le monde entier. Les flèches en pointillé indiquent les flux identifiés avec les indices, tandis que les flèches en trait continu indiquent ceux identifiés avec PhyloType et avec ou sous les indices. Les cercles indiquent que l'épidémie se diffuse de manière indépendante dans plusieurs pays différents. Le cercle blanc indique l'épicentre de l'épidémie. Seuls les pays représentés dans cette étude sont mentionnés et coloriés. Les pays de l'Afrique australe sont en rouge, ceux de l'Afrique de l'est, ouest et centrale en orange, ceux du continent asiatique en vert, ceux du continent américain en mauve et ceux du continent européen en bleu.



À la section 4.3.2, page 120, nous présentons une étude, basée sur cette phylogénie, visant à déterminer l'origine temporelle de l'épidémie du sous-type C du VIH-1. Deux méthodes de distances sont utilisées : la méthode ULS, présentée au Chapitre 4, et la régression linéaire *Root-to-tip* (cf. Chapitre 2). Pour mémoire, la méthode ULS date l'ancêtre commun aux souches du sous-type C à 1964, estimation qui semble cohérente avec celles retrouvées dans la littérature (Hemelaar, 2012), et *Root-to-tip* à 1782, estimation complètement différente de celles communément admises. Ici, et pour des raisons de temps de calculs, les intervalles de confiance ne sont pas calculés, mais rappelons que ceux associés aux estimations publiées dans la littérature sont larges (allant de 1933 à 1973), indiquant une grande incertitude, vraisemblablement liée à un faible signal global (cf. discussion à la section 4.3.2).

Nos analyses ont révélé deux origines géographiques différentes aux souches collectées en Éthiopie, probablement l'explication de l'observation de deux variants C et C' décrit précédemment (Abebe *et al*, 2000). La souche 86ET-ETH2220, présente dans cette étude, se situe dans le groupe qui a pour origine géographique le Burundi. Dans d'autres études, elle se place en-dehors du sous-cluster C' (Kassu *et al*, 2007; Abebe *et al*, 2000), ce qui suggère que les souches appartenant au sous-cluster C' ont pour origine épidémique la Zambie, tandis que le C (pour ce qui concerne l'Éthiopie) viendrait du Burundi. Toutefois, nous ne pouvons pas confirmer ces inférences sur la base d'informations publiées. Il faudrait aller plus loin dans les recoupements et disposer de plus de données. Notons toutefois que Kassu *et al*. (2007) n'observent pas la formation du sous-cluster C' sur la protéase et sur la transcriptase inverse, mais uniquement sur les gènes *gag* et *env*.

Plusieurs études indépendantes ont suggéré un lien épidémiologique entre l'Inde et l'Afrique du Sud (Shen *et al*, 2011; Dietrich *et al*, 1995, 1993). Toutefois, Dietrich *et al*. (1993) utilisent très peu de souches provenant de l'Afrique (peu disponibles à l'époque), tandis que le jeu de séquences de Shen *et al*. (2011) comprend en totalité 312 séquences réparties sur 27 pays différents (en-dehors de celles collectées en Inde). Ce chiffre représente moins de la moitié des souches collectées en Zambie ou en Afrique du Sud utilisées dans cette étude. Nos analyses qui intègrent une plus grande quantité de séquences collectées en Zambie et en Afrique du Sud (respectivement 633 et 689 souches), suggèrent un lien direct entre la Zambie et l'Inde et non avec l'Afrique du Sud. Notons que ces études utilisent le gène *env* et qu'une seule de leurs souches pertinentes (03ZA-PS057MB2) est dans notre analyse. Elle se place à l'intérieur d'un groupe contenant d'autres souches d'Afrique du Sud.

L'absence significative de séquences collectées au Royaume-Uni (due à la faible quantité de séquences disponibles publiquement) sur la région génomique considérée, ne permet pas de confirmer les liens épidémiologiques établies par de Oliveira *et al*. (2010), entre le Brésil, l'Afrique de l'est et le

Royaume-Uni. En effet, cette dernière étude suggère que l'épidémie du sous-type C s'est d'abord diffusée au Royaume-Uni avant d'être introduite au Brésil, suite à un évènement fondateur. Théorie en contradiction avec une étude récente (Véras *et al*, 2011a) et d'autres études anciennes (Bello *et al*, 2008; Fontella *et al*, 2008) qui suggèrent une introduction du sous-type C au Brésil directement par le Burundi ou un pays à proximité. Cette dernière version est corroborée par nos analyses. Notons que Bello *et al*. (2008) et Fontella *et al*. (2008) n'utilisent pas de souches d'Angleterre dans leurs études. Les quelques souches d'Angleterre considérées dans notre étude permettent uniquement de montrer le lien épidémiologique avec l'Afrique australe, laissant sous-entendre qu'elles proviennent probablement d'individus originaires ou ayant voyagé en Afrique (Dougan *et al*, 2005; Hughes *et al*, 2009). Rendre public toutes les séquences disponibles permettrait de considérer systématiquement un ensemble de séquences plus vaste et ainsi de mettre en évidence des liens épidémiologiques plus complexes, précis et exhaustifs.

Cette dernière remarque rappelle le problème du temps de calcul nécessaire pour inférer des phylogénies par des méthodes probabilistes (considérées comme les plus précises) et pose un double challenge aux chercheurs qui doivent développer des outils permettant l'inférence de phylogénies de plus en plus grandes, en des temps raisonnables, mais aussi de les visualiser et les interpréter simplement et rapidement.

Conclusion

Les travaux effectués durant cette thèse se regroupent en trois projets distincts mais complémentaires utilisant chacun, en plus de l'information fournie par les séquences nucléotidiques, celle des dates de prélèvement et/ou celles des pays de collecte. Le premier projet est le développement d'une méthode de distances, *Ultrametic Least Squares* (ULS), qui permet de calculer la vitesse évolutive d'un ensemble de séquences hétérochrones. Le deuxième projet a pour objectif de déterminer l'origine géographique et temporelle de l'épidémie du sous-type C du VIH-1 (VIH-1C) au Sénégal, particulièrement chez les hommes ayant des rapports sexuels avec des hommes (MSM), à l'aide d'outils phylogénétiques classiques. Le dernier projet essaie de retracer les principaux flux migratoires du VIH-1C à la surface du globe, ainsi que son origine géographique, avec de nouveaux outils informatiques pouvant synthétiser l'information contenue dans de grandes phylogénies (plus de 3 600 feuilles dans notre étude).

La méthode de distances ULS permet d'estimer la vitesse évolutive sous les hypothèses du modèle *Single Rate Dated Tips* (SRDT ; séquences hétérochrones et horloge moléculaire stricte). Pour cela, elle utilise un algorithme, basé sur le principe des moindres carrés pondérés, souvent utilisé par les méthodes de distances d'inférence phylogénétique (Felsenstein, 1997; Bulmer, 1991; Fitch & Margoliash, 1967), qui minimise un critère quadratique mesurant l'ultramétricité d'une distance en $O(n^3 \log n)$, où n est le nombre de taxa. L'utilisation du choix aléatoire de triplets permet de borner cette complexité, et cela sans perte de précision. ULS est ensuite étendue aux modèles *Multiple Rates Dated Tips* (un taux de substitution entre chaque intervalle de temps obtenu sur deux temps de collecte successifs) et *Different Rate* dans le cas d'horloges moléculaires locales (un taux de substitution global avec la possibilité de taux de substitution différents par lignage) par un principe itératif qui peut être adapté à toutes méthodes estimant le taux de substitution sous le modèle SRDT. Les tests de performance montrent qu'ULS est plus précise que les méthodes de distances sUPGMA (Rodrigo *et al*, 2007; Drummond & Rodrigo, 2000), TREBLE (Yang *et al*, 2007) et les régressions linéaires *Pairwise-Distance* et *Root-to-tip* (Drummond *et al*, 2003a). Notons que la précision d'estimation de la

méthode *Root-to-tip* est souvent similaire à la notre et, dans ce cas, il est difficile de les départager. La précision d'estimation est aussi comparée à la méthode probabiliste de référence, BEAST (Drummond & Rambaut, 2007). Contre cette dernière, ULS semble équivalente en précision d'estimation sur des jeux de données simulant le comportement inter-hôte du VIH, alors qu'elle est nettement supérieure à celle de BEAST sur des jeux de données simulant le comportement intra-hôte. Rappelons que le modèle démographique utilisé par BEAST est peu adapté à notre modèle de simulation, mais qu'aucun modèle approprié n'est disponible avec la version 1.6.2 de BEAST.

La méthode ULS utilise les hypothèses du modèle SRDT afin d'estimer le taux de substitution. Elle a ensuite été étendue au modèle MRDT et au modèle par lignage (horloges moléculaires locales). Toutefois, il est évident que le modèle le plus adapté à la réalité est celui où chaque branche de la phylogénie à son propre taux de substitution. Développer un tel modèle est très difficile (voire impossible), en particulier lorsqu'il n'y a aucune corrélation entre les différents taux de substitution, mais serait une prochaine étape à réaliser. La dernière version de BEAST, sortie au moment de la rédaction de ce manuscrit, permet d'utiliser un modèle démographique plus adapté à nos simulations. Le temps de calcul prohibitif nécessaire à BEAST ne nous a pas permis de refaire les estimations à temps mais l'utilisation de cette dernière version est primordiale avant toute publication et est le prochain point à réaliser (Drummond *et al*, 2012; Stadler, 2010).

L'étude moléculaire visant à déterminer l'origine géographique et temporelle de l'épidémie du VIH-1C au Sénégal utilise uniquement des outils phylogénétiques préalablement publiés (Jung *et al*, 2012). Ce genre d'étude foisonne et est pratiquement disponible pour la plupart des variants génétiques du VIH (Véras *et al*, 2011b; Passaes *et al*, 2009; Salemi *et al*, 2008; Tee *et al*, 2008). Dans cette étude nous avons récupéré toutes les séquences *pol* du sous-type C disponibles dans la base de données sur le VIH du laboratoire national de Los Alamos, et pour lesquelles la date de prélèvement et le pays de collecte sont connus. Un arbre de maximum de vraisemblance (PhyML) est calculé avec toutes celles dont REGA confirme l'appartenance au sous-type C et celles sans liens épidémiologiques proches (par exemple, mère-enfant). Outre les observations de clusters géographiques importants (Brésil, Inde, Afrique de l'est) (Shen *et al*, 2011; Thomson & Fernández-García, 2011; Véras *et al*, 2011a), cet arbre a permis d'identifier les séquences épidémiologiquement proches de celles du Sénégal. Un autre arbre de maximum de vraisemblance (PhyML) est uniquement calculé avec ces dernières, ainsi qu'un arbre bayésien (MrBayes), afin de déterminer l'origine géographique de l'épidémie du VIH-1C sévissant au Sénégal et particulièrement chez les MSM. Ces deux dernières phylogénies montrent de multiples introductions de ce variant dans la population générale mais de deux origines géographiques différentes (une de l'Afrique de l'est et l'autre de l'Afrique australe), confirmant les liens établis entre le Sénégal et les autres pays de l'Afrique (Toure-Kane *et al*, 2000).

Les souches isolées chez les MSM sont regroupées dans un cluster. Ce résultat suggère une introduction unique de ce variant, provenant d'Afrique australe (probablement de Zambie), suivie d'une diffusion rapide chez les MSM (évènement fondateur). L'origine temporelle du VIH-1C est étudiée avec le logiciel BEAST. Différents modèles d'horloges moléculaires (stricte, relâchée en exponentiel et en log-normal) et différentes *priors* sont utilisés pour estimer la date de l'ancêtre commun aux souches de la population générale et celle de l'ancête commun des souches des MSM. Les différentes analyses révèlent que l'ancêtre commun aux souches du VIH-1C collectées au Sénégal est daté au début des années soixante-dix (les méthodes de distances ULS et *Root-to-tip* estiment aussi une date similaire, cf. section 4.3.2) et celui des souches isolées chez les MSM environ dix ans après. Ces estimations sont cohérentes avec celles de l'ancêtre commun du sous-type C daté au début de la second moitié du xx^{e} siècle (Hemelaar, 2012).

Cette étude a été initiée suite à l'observation d'une forte prévalence du sous-type C chez les MSM (40%), alors que ce variant génétique est presque absent dans la population générale du Sénégal et dans les pays voisins où c'est la forme recombinante circulante CRF02_AG qui prédomine (Hemelaar *et al*, 2011; Ndiaye *et al*, 2009). Une prochaine étude, similaire à celle-ci, peut donc être menée sur le CRF02_AG au Sénégal, sachant que c'est déjà le cas pour d'autres pays, comme, par exemple, avec le Cameroun (Faria *et al*, 2011; Véras *et al*, 2011b).

Devant le nombre croissant de séquences disponibles dans les bases de données biologiques, les études moléculaires basées sur de grandes phylogénies deviendront de plus en plus courantes. Nous présentons, lors d'une étude visant à déterminer les flux migratoires mondiaux de l'épidémie du VIH-1C, ainsi que son origine géographique, des outils, basés sur le principe de parcimonie, permettant de synthétiser l'information contenue dans de grandes phylogénies en des temps raisonnables. L'indice de symétrie, identifiant le sens d'échange le plus intense entre deux pays, et l'indice de flux, identifiant la proportion de transitions entrantes entre deux pays, développés à cet effet sont déjà utilisés par d'autres (Véras *et al*, 2011a; Maddison & Maddison, 2003), tandis que l'indice de dispersion, qui permet de mesurer la régionalisation d'un pays, est, à notre connaissance, nouveau ; ces indices sont tous basés sur celui de Slatkin et Maddison (1989) qui compte le nombre total de transitions dans une phylogénie. Un second logiciel, PhyloType (Chevenet *et al*), pour lequel j'ai apporté ma contribution en développant un logiciel d'enracinement, est utilisé afin de déterminer les chaînes de transmission principales du VIH-1C, reflets d'évènements fondateurs probables. Ces outils ont permis d'identifier la Zambie comme étant l'épicentre mondial de l'épidémie du VIH-1C, pays frontalier et situé au sud de la République Démocratique du Congo, lui-même épicentre de la pandémie mondiale (Vidal *et al*, 2000). De ce pays, l'épidémie s'est ensuite diffusée en Afrique australe, au Bu-

rundi, en Éthiopie et en Inde. Du Burundi elle s'est diffusée dans l'est de l'Afrique, au Brésil et à nouveau en Éthiopie. De ce dernier elle s'est diffusée dans les pays de la Corne de l'Afrique, au Soudan et au Moyen-Orient (Israël, Yémen). Enfin, de l'Inde et du Brésil, l'épidémie s'est diffusée en Asie et en Amérique du sud respectivement. La plupart de ces flux migratoires corroborent ceux identifiés indépendamment dans d'autres études (Véras *et al*, 2011a; Lau *et al*, 2007; Gehring *et al*, 1997; Carrion *et al*, 2004), sauf pour l'Inde qui est supposé être liée avec l'Afrique du Sud dans d'autres études (Shen *et al*, 2011; Dietrich *et al*, 1995, 1993). Toutefois, ces dernières utilisent très peu de souches collectées en Afrique, et donc en Zambie. L'origine géographique exacte de l'épidémie du VIH-1C reste indéterminée, mais nos résultats suggèrent qu'elle se situe dans la zone frontalière entre la Zambie et la RDC. De nombreuses études similaires à celle-ci sont régulièrement réalisées par d'autres équipes, mais souvent à l'échelle d'un pays ou d'une région et rarement d'un continent ou du monde entier, par exemple en Europe pour le sous-type B (Paraskevis *et al*, 2009), puisque les méthodes qu'elles utilisent sont généralement lourdes en temps de calcul et de ce fait, des chaînes de transmission globales ne peuvent être étudiées.

Un des résultats majeurs de cette étude est que la Zambie soit l'épicentre de l'épidémie du sous-type C du VIH-1. Ce résultat est peut-être biaisé par le nombre important de souches collectées en Zambie, mais qui est considéré dans nos analyses car représentatif de la prévalence de ce variant en ce pays. Afin d'éviter tout artéfact dû à la taille des échantillons, il faudrait refaire les analyses mais en ne considérant, cette fois-ci, qu'une partie (choisie aléatoirement) des souches de Zambie. Si les conclusions sont similaires, ce résultat sera fiable. La parcimonie utilisée dans cette étude utilise le choix aléatoire afin de résoudre les ambiguïtés de certains nœuds internes, facilitant ainsi le calcul des indices. Nakano et al. (2004) préfèrent, à tort ou à raison, ne pas considérer les transitions avec un nœud ambigu. Cette approche, bien que n'utilisant qu'une partie de l'information, permettrait peut-être de supprimer le bruit généré par le choix aléatoire. Dans un autre registre, il serait aussi possible d'adapter l'indice de flux non par rapport aux transitions entrantes mais par rapport à l'impact qu'elles ont sur un pays. Par exemple, dans le cas du Brésil, deux transitions entrantes sont à noter, donc représentées graphiquement avec la même proportion. Mais seulement une de ces transitions (celle venant du Burundi) est à l'origine d'un évènement fondateur, l'autre est uniquement observée avec une souche. Il est donc impossible de dire laquelle de ces deux transitions est à l'origine de l'évènement fondateur, sans la lecture de la phylogénie ou sans PhyloType. Pour cela, il faudrait en plus considérer, dans la définition de cet indice, le nombre de souches collectées au Brésil issues de cette transition afin de faire varier sa valeur en fonction de la conséquence épidémiologique.

Bibliographie

Abebe A, Pollakis G, Fontanet AL, Fisseha B, Tegbaru B, Kliphuis A, Tesfaye G, Negassa H, Cornelissen M, Goudsmit J & Rinke de Wit TF (2000) Identification of a genetic subcluster of HIV type 1 subtype C (C') widespread in Ethiopia. *AIDS Res. Hum. Retroviruses* **16:** 1909-1914

Abecasis AB, Vandamme A-M & Lemey P (2009) Quantifying differences in the tempo of human immunodeficiency virus type 1 subtype evolution. *J. Virol* **83:** 12917-12924

Abraha A, Nankya IL, Gibson R, Demers K, Tebit DM, Johnston E, Katzenstein D, Siddiqui A, Herrera C, Fischetti L, Shattock RJ & Arts EJ (2009) CCR5- and CXCR4-tropic subtype C human immunodeficiency virus type 1 isolates have a lower level of pathogenic fitness than other dominant group M subtypes: implications for the epidemic. *J. Virol.* **83:** 5592-5605

Adelson-Velskii G & Landis EM (1962) An algorithm for the organization of information. *Doklady Akademii Nauk SSSR* **146:** 263-266

Agnarsson I & Miller JA (2008) Is ACCTRAN better than DELTRAN? *Cladistics* **24:** 1032-1038

Alaeus A, Leitner T, Lidman K & Albert J (1997) Most HIV-1 genetic subtypes have entered Sweden. *AIDS* **11:** 199-202

Alaeus A, Lidman K, Björkman A, Giesecke J & Albert J (1999) Similar rate of disease progression among individuals infected with HIV-1 genetic subtypes A-D. *AIDS* **13:** 901-907

An W & Telesnitsky A (2002) HIV-1 genetic recombination: experimental approaches and observations. *AIDS Rev* **4:** 195-212

Ancelle R, Bletry O, Baglin AC, Brun-Vezinet F, Rey MA & Godeau P (1987) Long incubation period for HIV-2 infection. *Lancet* **1:** 688-689

Anderson CNK, Ramakrishnan U, Chan YL & Hadly EA (2005) Serial SimCoal: a population genetics model for data from multiple populations and points in time. *Bioinformatics* **21:** 1733-1734

Anisimova M & Gascuel O (2006) Approximate likelihood-ratio test for branches: A fast, accurate, and powerful alternative. *Syst. Biol* **55:** 539-552

Ariën KK, Vanham G & Arts EJ (2007) Is HIV-1 evolving to a less virulent form in humans? *Nat. Rev. Microbiol* **5:** 141-151

Auvert B, Taljaard D, Lagarde E, Sobngwi-Tambekou J, Sitta R & Puren A (2005) Randomized, controlled intervention trial of male circumcision for reduction of HIV infection risk: the ANRS 1265 Trial. *PLoS Med.* **2:** e298

Avise JC (2000) Phylogeography: The History and Formation of Species Harvard University Press

Ayouba A, Mauclère P, Martin PM, Cunin P, Mfoupouendoun J, Njinku B, Souquières S & Simon F (2001) HIV-1 group O infection in Cameroon, 1986 to 1998. *Emerging Infect. Dis.* **7:** 466-467

Ayouba A, Souquières S, Njinku B, Martin PM, Müller-Trutwin MC, Roques P, Barré-Sinoussi F, Mauclère P, Simon F & Nerrienet E (2000) HIV-1 group N among HIV-1-seropositive individuals in Cameroon. *AIDS* **14:** 2623-2625

Azuonwu O, Erhabor O & Obire O (2012) HIV Among Military Personnel in the Niger Delta of Nigeria. *J Community Health* **37:** 25-31

Bailey RC, Moses S, Parker CB, Agot K, Maclean I, Krieger JN, Williams CFM, Campbell RT & Ndinya-Achola JO (2007) Male circumcision for HIV prevention in young men in Kisumu, Kenya: a randomised controlled trial. *Lancet* **369:** 643-656

Barin F, M'Boup S, Denis F, Kanki P, Allan JS, Lee TH & Essex M (1985) Serological evidence for virus related to simian T-lymphotropic retrovirus III in residents of west Africa. *Lancet* **2:** 1387-1389

Barré-Sinoussi F (2010) HIV: a discovery opening the road to novel scientific knowledge and global health improvement. *Virology* **397:** 255-259

Barré-Sinoussi F, Chermann JC, Rey F, Nugeyre MT, Chamaret S, Gruest J, Dauguet C, Axler-Blin C, Vézinet-Brun F, Rouzioux C, Rozenbaum W & Montagnier L (1983) Isolation of a T-lymphotropic retrovirus from a patient at risk for acquired immune deficiency syndrome (AIDS). *Science* **220:** 868-871

Barthélemy JP & Guénoche A (1988) les arbres et les representations de proximités Masson. Paris

Belda FJ, Barlow KL, Murphy G, Parry JV & Clewley JP (1998) A dual subtype B/E HIV type 1 infection with a novel V3 loop crown motif among infections acquired in Thailand and imported into England. *AIDS Res. Hum. Retroviruses* **14:** 911-916

Bello G, Passaes CP, Guimarães ML, Lorete RS, Matos Almeida SE, Medeiros RM, Alencastro PR & Morgado MG (2008) Origin and evolutionary history of HIV-1 subtype C in Brazil. *AIDS* **22:** 1993-2000

Bennett DE, Camacho RJ, Otelea D, Kuritzkes DR, Fleury H, Kiuchi M, Heneine W, Kantor R, Jordan MR, Schapiro JM, Vandamme A-M, Sandstrom P, Boucher CAB, van de Vijver D, Rhee S-Y, Liu TF, Pillay D & Shafer RW (2009) Drug resistance mutations for surveillance of transmitted HIV-1 drug-resistance: 2009 update. *PLoS ONE* **4:** e4724

Berman PW, Huang W, Riddle L, Gray AM, Wrin T, Vennari J, Johnson A, Klaussen M, Prashad H, Köhne C, deWit C & Gregory TJ (1999) Development of bivalent (B/E) vaccines able to neutralize CCR5-dependent viruses from the United States and Thailand. *Virology* **265:** 1-9

Bobkov A, Cheingsong-Popov R, Selimova L, Ladnaya N, Kazennova E, Kravchenko A, Fedotov E, Saukhat S, Zverev S, Pokrovsky V & Weber J (1997) An HIV type 1 epidemic among injecting

drug users in the former Soviet Union caused by a homogeneous subtype A strain. *AIDS Res. Hum. Retroviruses* **13:** 1195-1201

Bodelle P, Vallari A, Coffey R, McArthur CP, Beyeme M, Devare SG, Schochetman G & Brennan CA (2004) Identification and genomic sequence of an HIV type 1 group N isolate from Cameroon. *AIDS Res. Hum. Retroviruses* **20:** 902-908

Bolognesi DP & Matthews TJ (1998) HIV vaccines. Viral envelope fails to deliver? *Nature* **391:** 638-639

Bouchaud O, Fontanet A & Niyongabo T (2011) Caractéristiques épidémiologiques et cliniques de l'infection VIH en région tropiclae. In *VIH édition 2011* p 647-664. France: Pierre-Marie Girard, Christine Katlama, Gilles Pialoux

Brennan CA, Yamaguchi J, Vallari AS, Hickman RK & Devare SG (1997) Genetic variation in human immunodeficiency virus type 2: identification of a unique variant from human plasma. *AIDS Res. Hum. Retroviruses* **13:** 401-404

Bromham L & Penny D (2003) The modern molecular clock. *Nat. Rev. Genet.* **4:** 216-224

Bruno WJ, Socci ND & Halpern AL (2000) Weighted neighbor joining: a likelihood-based approach to distance-based phylogeny reconstruction. *Mol. Biol. Evol.* **17:** 189-197

Brun-Vézinet F, Damond F & Simon F (1999) [Variability of human immunodeficiency virus type 1]. *Bull Soc Pathol Exot* **92:** 261-263

Bulmer M (1991) Use of the Method of Generalized Least Squares in Reconstructing Phylogenies from Sequence Data. *Molecular Biology and Evolution* **8:** 868

Buneman P (1971) The recovery of trees from measures of dissimilarity. In *Mathematics in the Archeological and Historical Sciences* Hodson F et al.

Buonaguro L, Tornesello ML & Buonaguro FM (2007) Human immunodeficiency virus type 1 subtype distribution in the worldwide epidemic: pathogenetic and therapeutic implications. *J. Virol* **81:** 10209-10219

Bwayo J, Plummer F, Omari M, Mutere A, Moses S, Ndinya-Achola J, Velentgas P & Kreiss J (1994) Human immunodeficiency virus infection in long-distance truck drivers in east Africa. *Arch. Intern. Med.* **154:** 1391-1396

Caraux G, Gascuel O, Andrieu G & Levy D (1995) Méthodes informatiques pour la reconstruction phylogenetique. *Technique et Science Informatiques* **14:** 113-139

Carrion G, Eyzaguirre L, Montano SM, Laguna-Torres V, Serra M, Aguayo N, Avila MM, Ruchansky D, Pando MA, Vinoles J, Perez J, Barboza A, Chauca G, Romero A, Galeano A, Blair PJ, Weissenbacher M, Birx DL, Sanchez JL, Olson JG, *et al* (2004) Documentation of subtype C HIV Type 1 strains in Argentina, Paraguay, and Uruguay. *AIDS Res. Hum. Retroviruses* **20:** 1022-1025

Charneau P, Borman AM, Quillent C, Guétard D, Chamaret S, Cohen J, Rémy G, Montagnier L & Clavel F (1994) Isolation and envelope sequence of a highly divergent HIV-1 isolate: definition of a new HIV-1 group. *Virology* **205:** 247-253

Chen JH-K, Wong K-H, Chan KC-W, To SW-C, Chen Z & Yam W-C (2011) Phylodynamics of HIV-1 subtype B among the men-having-sex-with-men (MSM) population in Hong Kong. *PLoS ONE* **6:** e25286

Chen Z, Luckay A, Sodora DL, Telfer P, Reed P, Gettie A, Kanu JM, Sadek RF, Yee J, Ho DD, Zhang L & Marx PA (1997) Human immunodeficiency virus type 2 (HIV-2) seroprevalence and characterization of a distinct HIV-2 genetic subtype from the natural range of simian immunodeficiency virus-infected sooty mangabeys. *J. Virol.* **71:** 3953-3960

Chen Z, Telfier P, Gettie A, Reed P, Zhang L, Ho DD & Marx PA (1996) Genetic characterization of new West African simian immunodeficiency virus SIVsm: geographic clustering of household-derived SIV strains with human immunodeficiency virus type 2 subtypes and genetically diverse viruses from a single feral sooty mangabey troop. *J. Virol.* **70:** 3617-3627

Chevenet F, Jung M, de Oliveira T & Gascuel O PhyloType: combining annotations and phylogenies ; application to viruses.

De Cock KM, Adjorlolo G, Ekpini E, Sibailly T, Kouadio J, Maran M, Brattegaard K, Vetter KM, Doorly R & Gayle HD (1993) Epidemiology and transmission of HIV-2. Why there is no HIV-2 pandemic. *JAMA* **270:** 2083-2086

Coffin J, Haase A, Levy JA, Montagnier L, Oroszlan S, Teich N, Temin H, Toyoshima K, Varmus H & Vogt P (1986) Human immunodeficiency viruses. *Science* **232:** 697

Couturier E, Damond F, Roques P, Fleury H, Barin F, Brunet JB, Brun-Vézinet F & Simon F (2000) HIV-1 diversity in France, 1996-1998. The AC 11 laboratory network. *AIDS* **14:** 289-296

Cuevas MT, Ruibal I, Villahermosa ML, Díaz H, Delgado E, Parga EV, Pérez-Alvarez L, de Armas MB, Cuevas L, Medrano L, Noa E, Osmanov S, Nájera R & Thomson MM (2002) High HIV-1 genetic diversity in Cuba. *AIDS* **16:** 1643-1653

Curlin ME, Gottlieb GS, Hawes SE, Sow PS, Ndoye I, Critchlow CW, Kiviat NB & Mullins JI (2004) No evidence for recombination between HIV type 1 and HIV type 2 within the envelope region in dually seropositive individuals from Senegal. *AIDS Res. Hum. Retroviruses* **20:** 958-963

Dalai SC, de Oliveira T, Harkins GW, Kassaye SG, Lint J, Manasa J, Johnston E & Katzenstein D (2009) Evolution and molecular epidemiology of subtype C HIV-1 in Zimbabwe. *AIDS* **23:** 2523-2532

Damond F, Worobey M, Campa P, Farfara I, Colin G, Matheron S, Brun-Vézinet F, Robertson DL & Simon F (2004) Identification of a highly divergent HIV type 2 and proposal for a change in HIV type 2 classification. *AIDS Res. Hum. Retroviruses* **20:** 666-672

Daniel MD, Letvin NL, King NW, Kannagi M, Sehgal PK, Hunt RD, Kanki PJ, Essex M & Desrosiers RC (1985) Isolation of T-cell tropic HTLV-III-like retrovirus from macaques. *Science* **228:** 1201-1204

Delaporte E, Janssens W, Peeters M, Buvé A, Dibanga G, Perret JL, Ditsambou V, Mba JR, Courbot MC, Georges A, Bourgeois A, Samb B, Henzel D, Heyndrickx L, Fransen K, van der Groen G & Larouzé B (1996) Epidemiological and molecular characteristics of HIV infection in Gabon, 1986-1994. *AIDS* **10:** 903-910

Delaugerre C, De Oliveira F, Lascoux-Combe C, Plantier J-C & Simon F (2011) HIV-1 group N: travelling beyond Cameroon. *Lancet* **378:** 1894

Descamps D, Apetrei C, Collin G, Damond F, Simon F & Brun-Vézinet F (1998) Naturally occurring decreased susceptibility of HIV-1 subtype G to protease inhibitors. *AIDS* **12:** 1109-1111

Desper R & Gascuel O (2002) Fast and accurate phylogeny reconstruction algorithms based on the minimum-evolution principle. *J. Comput. Biol* **9:** 687-705

Desper R & Gascuel O (2004) Theoretical foundation of the balanced minimum evolution method of phylogenetic inference and its relationship to weighted least-squares tree fitting. *Mol. Biol. Evol.* **21:** 587-598

Dessimoz C, Gil M, Schneider A & Gonnet GH (2006) Fast estimation of the difference between two PAM/JTT evolutionary distances in triplets of homologous sequences. *BMC Bioinformatics* **7:** 529

Dietrich U, Grez M, von Briesen H, Panhans B, Geissendörfer M, Kühnel H, Maniar J, Mahambre G, Becker WB & Becker ML (1993) HIV-1 strains from India are highly divergent from prototypic African and US/European strains, but are linked to a South African isolate. *AIDS* **7:** 23-27

Dietrich U, Maniar JK & Rübsamen-Waigmann H (1995) The epidemiology of HIV in India. *Trends Microbiol* **3:** 17-21

Dimmic MW, Rest JS, Mindell DP & Goldstein RA (2002) rtREV: an amino acid substitution matrix for inference of retrovirus and reverse transcriptase phylogeny. *J. Mol. Evol.* **55:** 65-73

Djoko CF, Rimoin AW, Vidal N, Tamoufe U, Wolfe ND, Butel C, LeBreton M, Tshala FM, Kayembe PK, Muyembe JJ, Edidi-Basepeo S, Pike BL, Fair JN, Mbacham WF, Saylors KE, Mpoudi-Ngole E, Delaporte E, Grillo M & Peeters M (2011) High HIV type 1 group M pol diversity and low rate of antiretroviral resistance mutations among the uniformed services in Kinshasa, Democratic Republic of the Congo. *AIDS Res. Hum. Retroviruses* **27:** 323-329

Domingo E (1998) Quasispecies and the implications for virus persistence and escape. *Clin Diagn Virol* **10:** 97-101

Douady CJ, Delsuc F, Boucher Y, Doolittle WF & Douzery EJP (2003) Comparison of Bayesian and maximum likelihood bootstrap measures of phylogenetic reliability. *Mol. Biol. Evol.* **20:** 248-254

Dougan S, Gilbart VL, Sinka K & Evans BG (2005) HIV infections acquired through heterosexual intercourse in the United Kingdom: findings from national surveillance. *BMJ* **330:** 1303-1304

Drummond A, Forsberg R & Rodrigo AG (2001) The inference of stepwise changes in substitution rates using serial sequence samples. *Mol. Biol. Evol* **18:** 1365-1371

Drummond A, Pybus OG & Rambaut A (2003a) Inference of viral evolutionary rates from molecular sequences. *Adv. Parasitol.* **54:** 331-358

Drummond A & Rodrigo AG (2000) Reconstructing genealogies of serial samples under the assumption of a molecular clock using serial-sample UPGMA. *Mol. Biol. Evol.* **17:** 1807-1815

Drummond AJ, Ho SYW, Phillips MJ & Rambaut A (2006) Relaxed phylogenetics and dating with confidence. *PLoS Biol* **4:** e88

Drummond AJ, Nicholls GK, Rodrigo AG & Solomon W (2002) Estimating mutation parameters, population history and genealogy simultaneously from temporally spaced sequence data. *Genetics* **161:** 1307-1320

Drummond AJ, Pybus OG, Rambaut A, Forsberg R & Rodrigo AG (2003b) Measurably evolving populations. *Trends in Ecology & Evolution* **18:** 481-488

Drummond AJ & Rambaut A (2007) BEAST: Bayesian evolutionary analysis by sampling trees. *BMC Evol. Biol* **7:** 214

Drummond AJ, Suchard MA, Xie D & Rambaut A (2012) Bayesian phylogenetics with BEAUti and the BEAST 1.7. *Molecular Biology and Evolution*

Dunham EJ & Holmes EC (2007) Inferring the timescale of dengue virus evolution under realistic models of DNA substitution. *J. Mol. Evol.* **64:** 656-661

Eshleman SH, Becker-Pergola G, Deseyve M, Guay LA, Mracna M, Fleming T, Cunningham S, Musoke P, Mmiro F & Jackson JB (2001) Impact of human immunodeficiency virus type 1 (hiv-1) subtype on women receiving single-dose nevirapine prophylaxis to prevent hiv-1 vertical transmission (hiv network for prevention trials 012 study). *J. Infect. Dis.* **184:** 914-917

Etienne L & Peeters M (2010) Origine du VIH, une réussite émergentielle. *Virologie* **14:** 171-174

Faria NR, Suchard MA, Abecasis A, Sousa JD, Ndembi N, Bonfim I, Camacho RJ, Vandamme A-M & Lemey P (2011) Phylodynamics of the HIV-1 CRF02_AG clade in Cameroon. *Infection, Genetics and Evolution: Journal of Molecular Epidemiology and Evolutionary Genetics in Infectious Diseases*

Farris JS (1970) Methods for Computing Wagner Trees. *Systematic Zoology* **19:** 83-92

Felsenstein J (1978) Cases in which Parsimony or Compatibility Methods Will be Positively Misleading. *Systematic Zoology* **27:** 401-410

Felsenstein J (1981) Evolutionary trees from DNA sequences: a maximum likelihood approach. *J. Mol. Evol.* **17:** 368-376

Felsenstein J (1984) Distance Methods for Inferring Phylogenies: A Justification. *Evolution* **38:** 16-24

Felsenstein J (1985) Confidence Limits on Phylogenies: An Approach Using the Bootstrap. *Evolution* **39:** 783-791

Felsenstein J (1989) PHYLIP - Phylogeny Inference Package (Version 3.2). *Cladistics* **5:** 164-166

Felsenstein J (1993) PHYLIP (phylogeny inference package) version 3.5c.

Felsenstein J (1997) An alternating least squares approach to inferring phylogenies from pairwise distances. *Syst. Biol* **46:** 101-111

Felsenstein J (2003) Inferring Phylogenies 2^{e} éd. Sinauer Associates

Felsenstein J (2007) Trees of genes in populations. In *Reconstructing Evolution* p 3-29. New York: Olivier Gascuel et Mike Steel

Fitch W & Margoliash E (1967) Construction of Phylogenetic Trees. *Science* **155:** 279-284

Fitch WM (1971) Toward Defining the Course of Evolution: Minimum Change for a Specific Tree Topology. *Systematic Zoology* **20:** 406-416

Foley B (2000) An overview of the molecular phylogeny of lentiviruses. In *HIV Sequence Compendium 2000* p 35-43. Los Alamos National Laboratory, Los Alamos, NM: Kuiken C, McCutchan F, Foley B, Mellors JW, Hahn B, Mullins J, Marx P, Wolinsky S

Fonjungo PN, Mpoudi EN, Torimiro JN, Alemnji GA, Eno LT, Nkengasong JN, Gao F, Rayfield M, Folks TM, Pieniazek D & Lal RB (2000) Presence of diverse human immunodeficiency virus type 1 viral variants in Cameroon. *AIDS Res. Hum. Retroviruses* **16:** 1319-1324

Fontella R, Soares MA & Schrago CG (2008) On the origin of HIV-1 subtype C in South America. *AIDS* **22:** 2001-2011

Gallo RC (2006) A reflection on HIV/AIDS research after 25 years. *Retrovirology* **3:** 72

Galvin SR & Cohen MS (2004) The role of sexually transmitted diseases in HIV transmission. *Nat. Rev. Microbiol.* **2:** 33-42

Gao F, Bailes E, Robertson DL, Chen Y, Rodenburg CM, Michael SF, Cummins LB, Arthur LO, Peeters M, Shaw GM, Sharp PM & Hahn BH (1999) Origin of HIV-1 in the chimpanzee Pan troglodytes troglodytes. *Nature* **397:** 436-441

Gao F, Robertson DL, Carruthers CD, Morrison SG, Jian B, Chen Y, Barré-Sinoussi F, Girard M, Srinivasan A, Abimiku AG, Shaw GM, Sharp PM & Hahn BH (1998) A comprehensive panel of near-full-length clones and reference sequences for non-subtype B isolates of human immunodeficiency virus type 1. *J. Virol.* **72:** 5680-5698

Gao F, Vidal N, Li Y, Trask SA, Chen Y, Kostrikis LG, Ho DD, Kim J, Oh MD, Choe K, Salminen M, Robertson DL, Shaw GM, Hahn BH & Peeters M (2001) Evidence of two distinct subsubtypes within the HIV-1 subtype A radiation. *AIDS Res. Hum. Retroviruses* **17:** 675-688

Gao F, Yue L, Robertson DL, Hill SC, Hui H, Biggar RJ, Neequaye AE, Whelan TM, Ho DD & Shaw GM (1994) Genetic diversity of human immunodeficiency virus type 2: evidence for distinct sequence subtypes with differences in virus biology. *J. Virol.* **68:** 7433-7447

Gaschen B, Taylor J, Yusim K, Foley B, Gao F, Lang D, Novitsky V, Haynes B, Hahn BH, Bhattacharya T & Korber B (2002) Diversity considerations in HIV-1 vaccine selection. *Science* **296:** 2354-2360

Gascuel O (1997) BIONJ: an improved version of the NJ algorithm based on a simple model of sequence data. *Mol. Biol. Evol.* **14:** 685-695

Gascuel O (2000) Data Model and Classification by Trees: The Minimum Variance Reduction (MVR) Method. *Journal of Classification* **17:** 67-99

Gascuel O & Steel M (2006) Neighbor-joining revealed. *Mol. Biol. Evol.* **23:** 1997-2000

Gehring S, Maayan S, Ruppach H, Balfe P, Juraszczyk J, Yust I, Vardinon N, Rimlawi A, Polak S, Bentwich Z, Rübsamen-Waigmann H & Dietrich U (1997) Molecular epidemiology of HIV in Israel. *J. Acquir. Immune Defic. Syndr. Hum. Retrovirol* **15:** 296-303

Geibel S, Tun W, Tapsoba P & Kellerman S (2010) HIV vulnerability of men who have sex with men in developing countries: Horizons studies, 2001-2008. *Public Health Rep* **125:** 316-324

Giuliani M, Montieri S, Palamara G, Latini A, Alteri C, Perno CF, Santoro MM, Rezza G & Ciccozzi M (2009) Non-B HIV type 1 subtypes among men who have sex with men in Rome, Italy. *AIDS Res. Hum. Retroviruses* **25:** 157-164

Gojobori T, Moriyama EN & Kimura M (1990) Molecular clock of viral evolution, and the neutral theory. *Proc. Natl. Acad. Sci. U.S.A* **87:** 10015-10018

Gojobori T, Yamaguchi Y, Ikeo K & Mizokami M (1994) Evolution of pathogenic viruses with special reference to the rates of synonymous and nonsynonymous substitutions. *Jpn. J. Genet.* **69:** 481-488

Gojobori T & Yokoyama S (1985) Rates of evolution of the retroviral oncogene of Moloney murine sarcoma virus and of its cellular homologues. *Proc. Natl. Acad. Sci. U.S.A.* **82:** 4198-4201

Goode M & Rodrigo AG (2004) Using PEBBLE for the evolutionary analysis of serially sampled molecular sequences. *Curr Protoc Bioinformatics* **Chapter 6:** Unit 6.8

Gottlieb GS, Sow PS, Hawes SE, Ndoye I, Coll-Seck AM, Curlin ME, Critchlow CW, Kiviat NB & Mullins JI (2003) Molecular epidemiology of dual HIV-1/HIV-2 seropositive adults from Senegal, West Africa. *AIDS Res. Hum. Retroviruses* **19:** 575-584

Gray RH, Kigozi G, Serwadda D, Makumbi F, Watya S, Nalugoda F, Kiwanuka N, Moulton LH, Chaudhary MA, Chen MZ, Sewankambo NK, Wabwire-Mangen F, Bacon MC, Williams CFM, Opendi P, Reynolds SJ, Laeyendecker O, Quinn TC & Wawer MJ (2007) Male circumcision for HIV prevention in men in Rakai, Uganda: a randomised trial. *Lancet* **369:** 657-666

Grmek MD (Mirko D (1990) Histoire du sida Nouv. éd. Payot

Gu X, Fu YX & Li WH (1995) Maximum likelihood estimation of the heterogeneity of substitution rate among nucleotide sites. *Mol. Biol. Evol.* **12:** 546-557

Gueudin M, Plantier J-C, Damond F, Roques P, Mauclère P & Simon F (2003) Plasma viral RNA assay in HIV-1 group O infection by real-time PCR. *J. Virol. Methods* **113:** 43-49

Guindon S, Dufayard J-F, Lefort V, Anisimova M, Hordijk W & Gascuel O (2010) New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Syst. Biol* **59:** 307-321

Guindon S & Gascuel O (2003) A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst. Biol* **52:** 696-704

Gürtler LG, Hauser PH, Eberle J, von Brunn A, Knapp S, Zekeng L, Tsague JM & Kaptue L (1994) A new subtype of human immunodeficiency virus type 1 (MVP-5180) from Cameroon. *J. Virol.* **68:** 1581-1585

Gutell RR, Larsen N & Woese CR (1994) Lessons from an evolving rRNA: 16S and 23S rRNA structures from a comparative perspective. *Microbiol. Rev.* **58:** 10-26

Hahn BH, Shaw GM, Taylor ME, Redfield RR, Markham PD, Salahuddin SZ, Wong-Staal F, Gallo RC, Parks ES & Parks WP (1986) Genetic variation in HTLV-III/LAV over time in patients with AIDS or at risk for AIDS. *Science* **232:** 1548-1553

Hamel DJ, Sankalé J-L, Eisen G, Meloni ST, Mullins C, Gueye-Ndiaye A, Mboup S & Kanki PJ (2007) Twenty years of prospective molecular epidemiology in Senegal: changes in HIV diversity. *AIDS Res. Hum. Retroviruses* **23:** 1189-1196

Hampl H, Sawitzky D, Stöffler-Meilicke M, Groh A, Schmitt M, Eberle J & Gürtler L (1995) First case of HIV-1 subtype 0 infection in Germany. *Infection* **23:** 369-370

Hanna GJ & D'Aquila RT (2001) Clinical use of genotypic and phenotypic drug resistance testing to monitor antiretroviral chemotherapy. *Clin. Infect. Dis.* **32:** 774-782

van Harmelen J, Wood R, Lambrick M, Rybicki EP, Williamson AL & Williamson C (1997) An association between HIV-1 subtypes and mode of transmission in Cape Town, South Africa. *AIDS* **11:** 81-87

Hartigan JA (1973) Minimum Mutation Fits to a Given Tree. *Biometrics* **29:** 53-65

Hasegawa M, Kishino H & Yano T (1985) Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. *J. Mol. Evol* **22:** 160-174

Hastings WK (1970) Monte Carlo sampling methods using Markov chains and their applications. *Biometrika* **57:** 97 -109

Heled J & Drummond AJ (2010) Bayesian inference of species trees from multilocus data. *Mol. Biol. Evol.* **27:** 570-580

Hemelaar J (2012) The origin and diversity of the HIV-1 pandemic. *Trends Mol Med* **18:** 182-192

Hemelaar J, Gouws E, Ghys PD & Osmanov S (2011) Global trends in molecular epidemiology of HIV-1 during 2000-2007. *AIDS* **25:** 679-689

Henrickson RV, Maul DH, Osborn KG, Sever JL, Madden DL, Ellingsworth LR, Anderson JH, Lowenstine LJ & Gardner MB (1983) Epidemic of acquired immunodeficiency in rhesus monkeys. *Lancet* **1:** 388-390

Van Heuverswyn F, Li Y, Bailes E, Neel C, Lafay B, Keele BF, Shaw KS, Takehisa J, Kraus MH, Loul S, Butel C, Liegeois F, Yangda B, Sharp PM, Mpoudi-Ngole E, Delaporte E, Hahn BH & Peeters M (2007) Genetic diversity and phylogeographic clustering of SIVcpzPtt in wild chimpanzees in Cameroon. *Virology* **368:** 155-171

Hey J (1992) Using Phylogenetic Trees to Study Speciation and Extinction. *Evolution* **46:** 627-640

Heyndrickx L, Alary M, Janssens W, Davo N & van der Groen G (1996) HIV-1 group O and group M dual infection in Bénin. *Lancet* **347:** 902-903

Hirsch MS, Brun-Vézinet F, D'Aquila RT, Hammer SM, Johnson VA, Kuritzkes DR, Loveday C, Mellors JW, Clotet B, Conway B, Demeter LM, Vella S, Jacobsen DM & Richman DD (2000) Antiretroviral drug resistance testing in adult HIV-1 infection: recommendations of an International AIDS Society-USA Panel. *JAMA* **283:** 2417-2426

Holmes EC (2004) The phylogeography of human viruses. *Mol. Ecol.* **13:** 745-756

Holmes EC (2008) Evolutionary history and phylogeography of human viruses. *Annu. Rev. Microbiol.* **62:** 307-328

Hué S, Clewley JP, Cane PA & Pillay D (2004) HIV-1 pol gene variation is sufficient for reconstruction of transmissions in the era of antiretroviral therapy. *AIDS* **18:** 719-728

Huet T, Cheynier R, Meyerhans A, Roelants G & Wain-Hobson S (1990) Genetic organization of a chimpanzee lentivirus related to HIV-1. *Nature* **345:** 356-359

Hughes GJ, Fearnhill E, Dunn D, Lycett SJ, Rambaut A & Leigh Brown AJ (2009) Molecular phylodynamics of the heterosexual HIV epidemic in the United Kingdom. *PLoS Pathog* **5:** e1000590

Ibe S, Yokomaku Y, Shiino T, Tanaka R, Hattori J, Fujisaki S, Iwatani Y, Mamiya N, Utsumi M, Kato S, Hamaguchi M & Sugiura W (2010) HIV-2 CRF01_AB: first circulating recombinant form of HIV-2. *J. Acquir. Immune Defic. Syndr.* **54:** 241-247

Identification of HIV-1 group O infection--Los Angeles county, California, 1996 (1996) *MMWR Morb. Mortal. Wkly. Rep.* **45:** 561-565

Jacobs GB, Laten A, van Rensburg EJ, Bodem J, Weissbrich B, Rethwilm A, Preiser W & Engelbrecht S (2008) Phylogenetic diversity and low level antiretroviral resistance mutations in HIV type 1 treatment-naive patients from Cape Town, South Africa. *AIDS Res. Hum. Retroviruses* **24:** 1009-1012

Jin L & Nei M (1990) Limitations of the evolutionary parsimony method of phylogenetic analysis. *Mol. Biol. Evol.* **7:** 82-102

Jonassen TO, Stene-Johansen K, Berg ES, Hungnes O, Lindboe CF, Frøland SS & Grinde B (1997) Sequence analysis of HIV-1 group O from Norwegian patients infected in the 1960s. *Virology* **231:** 43-47

Jones DT, Taylor WR & Thornton JM (1992) The rapid generation of mutation data matrices from protein sequences. *Comput. Appl. Biosci.* **8:** 275-282

Jones LR, Dilernia DA, Manrique JM, Moretti F, Salomón H & Gomez-Carrillo M (2009) In-depth analysis of the origins of HIV type 1 subtype C in South America. *AIDS Res. Hum. Retroviruses* **25:** 951-959

Jukes T & Cantor C (1969) Evolution of protein molecules. In *Mammalian Protein Metabolism* p 21-132. New York: HN Munro

Jung M, Leye N, Vidal N, Fargette D, Diop H, Toure Kane C, Gascuel O & Peeters M (2012) The Origin and Evolutionary History of HIV-1 Subtype C in Senegal. *PLoS ONE* **7:** e33579

Kaleebu P, French N, Mahe C, Yirrell D, Watera C, Lyagoba F, Nakiyingi J, Rutebemberwa A, Morgan D, Weber J, Gilks C & Whitworth J (2002) Effect of human immunodeficiency virus (HIV) type 1 envelope subtypes A and D on disease progression in a large cohort of HIV-1-positive persons in Uganda. *J. Infect. Dis.* **185:** 1244-1250

Kane F, Alary M, Ndoye I, Coll AM, M'boup S, Guèye A, Kanki PJ & Joly JR (1993) Temporary expatriation is related to HIV-1 infection in rural Senegal. *AIDS* **7:** 1261-1265

Kanki P, M'Boup S, Marlink R, Travers K, Hsieh CC, Gueye A, Boye C, Sankalé JL, Donnelly C & Leisenring W (1992) Prevalence and risk determinants of human immunodeficiency virus

type 2 (HIV-2) and human immunodeficiency virus type 1 (HIV-1) in west African female prostitutes. *Am. J. Epidemiol* **136:** 895-907

Kanki PJ, Hamel DJ, Sankalé JL, Hsieh C c, Thior I, Barin F, Woodcock SA, Guèye-Ndiaye A, Zhang E, Montano M, Siby T, Marlink R, NDoye I, Essex ME & MBoup S (1999) Human immunodeficiency virus type 1 subtypes differ in disease progression. *J. Infect. Dis* **179:** 68-73

Kanki PJ, Travers KU, MBoup S, Hsieh CC, Marlink RG, Gueye-NDiaye A, Siby T, Thior I, Hernandez-Avila M & Sankalé JL (1994) Slower heterosexual spread of HIV-2 than HIV-1. *Lancet* **343:** 943-946

Kassu A, Fujino M, Matsuda M, Nishizawa M, Ota F & Sugiura W (2007) Molecular epidemiology of HIV type 1 in treatment-naive patients in north Ethiopia. *AIDS Res. Hum. Retroviruses* **23:** 564-568

Katoh K, Kuma K, Toh H & Miyata T (2005) MAFFT version 5: improvement in accuracy of multiple sequence alignment. *Nucleic Acids Res* **33:** 511-518

Katoh K, Misawa K, Kuma K & Miyata T (2002) MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res* **30:** 3059-3066

Keele BF, Van Heuverswyn F, Li Y, Bailes E, Takehisa J, Santiago ML, Bibollet-Ruche F, Chen Y, Wain LV, Liegeois F, Loul S, Ngole EM, Bienvenue Y, Delaporte E, Brookfield JFY, Sharp PM, Shaw GM, Peeters M & Hahn BH (2006) Chimpanzee reservoirs of pandemic and nonpandemic HIV-1. *Science* **313:** 523-526

Kidd KK & Sgaramella-Zonta LA (1971) Phylogenetic analysis: concepts and methods. *Am. J. Hum. Genet.* **23:** 235-252

Kils-Hütten L, Cheynier R, Wain-Hobson S & Meyerhans A (2001) Phylogenetic reconstruction of intrapatient evolution of human immunodeficiency virus type 1: predominance of drift and purifying selection. *J. Gen. Virol.* **82:** 1621-1627

Kimura M (1980) A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *J. Mol. Evol.* **16:** 111-120

Kingman J (1982) The coalescent. *Stochastic Processes and their Applications* **13:** 235-248

Kishino H & Hasegawa M (1989) Evaluation of the maximum likelihood estimate of the evolutionary tree topologies from DNA sequence data, and the branching order in hominoidea. *J. Mol. Evol* **29:** 170-179

Kluge AG & Farris JS (1969) Quantitative Phyletics and the Evolution of Anurans. *Systematic Zoology* **18:** 1-32

Koch N, Ndihokubwayo JB, Yahi N, Tourres C, Fantini J & Tamalet C (2001) Genetic analysis of hiv type 1 strains in bujumbura (burundi): predominance of subtype c variant. *AIDS Res. Hum. Retroviruses* **17:** 269-273

Korber B, Muldoon M, Theiler J, Gao F, Gupta R, Lapedes A, Hahn BH, Wolinsky S & Bhattacharya T (2000) Timing the ancestor of the HIV-1 pandemic strains. *Science* **288:** 1789-1796

Kumar S (2005) Molecular clocks: four decades of evolution. *Nat. Rev. Genet.* **6:** 654-662

Lal RB, Chakrabarti S & Yang C (2005) Impact of genetic diversity of HIV-1 on diagnosis, antiretroviral therapy & vaccine development. *Indian J. Med. Res.* **121:** 287-314

Lambert DM, Ritchie PA, Millar CD, Holland B, Drummond AJ & Baroni C (2002) Rates of evolution in ancient DNA from Adélie penguins. *Science* **295:** 2270-2273

Lanave C, Preparata G, Saccone C & Serio G (1984) A new method for calculating evolutionary substitution rates. *J. Mol. Evol.* **20:** 86-93

Langley CH & Fitch WM (1974) An examination of the constancy of the rate of molecular evolution. *J. Mol. Evol.* **3:** 161-177

Larmarange J, Wade AS, Diop AK, Diop O, Gueye K, Marra A & du Loû AD (2010) Men who have sex with men (MSM) and factors associated with not using a condom at last sexual intercourse with a man and with a woman in Senegal. *PLoS ONE* **5:** e13189

Lasky M, Perret JL, Peeters M, Bibollet-Ruche F, Liegeois F, Patrel D, Molinier S, Gras C & Delaporte E (1997) Presence of multiple non-B subtypes and divergent subtype B strains of HIV-1 in individuals infected after overseas deployment. *AIDS* **11:** 43-51

Lau KA, Wang B & Saksena NK (2007) Emerging trends of HIV epidemiology in Asia. *AIDS Rev* **9:** 218-229

Laurent C, Bourgeois A, Mpoudi M, Butel C, Peeters M, Mpoudi-Ngolé E & Delaporte E (2004) Commercial logging and HIV epidemic, rural Equatorial Africa. *Emerging Infect. Dis.* **10:** 1953-1956

Leendertz SAJ, Locatelli S, Boesch C, Kücherer C, Formenty P, Liegeois F, Ayouba A, Peeters M & Leendertz FH (2011) No evidence for transmission of SIVwrc from western red colobus monkeys (Piliocolobus badius badius) to wild West African chimpanzees (Pan troglodytes verus) despite high exposure through hunting. *BMC Microbiol.* **11:** 24

Leitner T & Albert J (1999) The molecular clock of HIV-1 unveiled through analysis of a known transmission history. *Proc. Natl. Acad. Sci. U.S.A.* **96:** 10752-10757

Leitner T & Fitch WM (1999) The Phylogenetics of Known Transmission Histories. In *The Evolution of HIV* p 315-345. Baltimore: Keith A. Crandall

Lemey P, Rambaut A, Drummond AJ & Suchard MA (2009a) Bayesian phylogeography finds its roots. *PLoS Comput. Biol* **5:** e1000520

Lemey P, Rambaut A, Welch JJ & Suchard MA (2010) Phylogeography takes a relaxed random walk in continuous space and time. *Mol. Biol. Evol* **27:** 1877-1885

Lemey P, Salemi M & Vandamme A-M (2009b) The Phylogenetic Handbook: A Practical Approach to Phylogenetic Analysis and Hypothesis Testing 2nd Revised edition. Cambridge University Press

De Leys R, Vanderborght B, Vanden Haesevelde M, Heyndrickx L, van Geel A, Wauters C, Bernaerts R, Saman E, Nijs P & Willems B (1990) Isolation and partial characterization of an unusual human immunodeficiency retrovirus from two persons of west-central African origin. *J. Virol.* **64:** 1207-1216

Li WH & Tanimura M (1987) The molecular clock runs more slowly in man than in apes and monkeys. *Nature* **326:** 93-96

Li WH, Tanimura M & Sharp PM (1988) Rates and dates of divergence between AIDS virus nucleotide sequences. *Mol. Biol. Evol.* **5:** 313-330

Loemba H, Brenner B, Parniak MA, Ma'ayan S, Spira B, Moisi D, Oliveira M, Detorio M & Wainberg MA (2002) Genetic divergence of human immunodeficiency virus type 1 Ethiopian clade C reverse transcriptase (RT) and rapid development of resistance against nonnucleoside inhibitors of RT. *Antimicrob. Agents Chemother.* **46:** 2087-2094

Lole KS, Bollinger RC, Paranjape RS, Gadkari D, Kulkarni SS, Novak NG, Ingersoll R, Sheppard HW & Ray SC (1999) Full-length human immunodeficiency virus type 1 genomes from subtype C-infected seroconverters in India, with evidence of intersubtype recombination. *J. Virol.* **73:** 152-160

Loussert-Ajaka I, Chaix ML, Korber B, Letourneur F, Gomas E, Allen E, Ly TD, Brun-Vézinet F, Simon F & Saragosti S (1995) Variability of human immunodeficiency virus type 1 group O strains isolated from Cameroonian patients living in France. *J. Virol.* **69:** 5640-5649

Loussert-Ajaka I, Ly TD, Chaix ML, Ingrand D, Saragosti S, Couroucé AM, Brun-Vézinet F & Simon F (1994) HIV-1/HIV-2 seronegativity in HIV-1 subtype O infected patients. *Lancet* **343:** 1393-1394

Maddison D & Maddison W (2003) MacClade 4.

Maljkovic Berry I, Athreya G, Kothari M, Daniels M, Bruno WJ, Korber B, Kuiken C, Ribeiro RM & Leitner T (2009) The evolutionary rate dynamically tracks changes in HIV-1 epidemics: application of a simple method for optimizing the evolutionary rate in phylogenetic trees with longitudinal data. *Epidemics* **1:** 230-239

Maljkovic Berry I, Ribeiro R, Kothari M, Athreya G, Daniels M, Lee HY, Bruno W & Leitner T (2007) Unequal evolutionary rates in the human immunodeficiency virus type 1 (HIV-1) pandemic: the evolutionary rate of HIV-1 slows down when the epidemic rate increases. *J. Virol.* **81:** 10625-10635

Marechal V, Jauvin V, Selekon B, Leal J, Pelembi P, Fikouma V, Gabrie P, Heredeibona LS, Goumba C, Serdouma E, Ayouba A & Fleury H (2006) Increasing HIV type 1 polymorphic diversity but no resistance to antiretroviral drugs in untreated patients from Central African Republic: a 2005 study. *AIDS Res. Hum. Retroviruses* **22:** 1036-1044

Marlink R (1996) Lessons from the second AIDS virus, HIV-2. *AIDS* **10:** 689-699

Marlink R, Kanki P, Thior I, Travers K, Eisen G, Siby T, Traore I, Hsieh CC, Dia MC & Gueye EH (1994) Reduced rate of disease development after HIV-2 infection as compared to HIV-1. *Science* **265:** 1587-1590

McGrath KM, Hoffman NG, Resch W, Nelson JA & Swanstrom R (2001) Using HIV-1 sequence variability to explore virus biology. *Virus Res* **76:** 137-160

Meda N, Ndoye I, M'Boup S, Wade A, Ndiaye S, Niang C, Sarr F, Diop I & Caraël M (1999) Low and stable HIV infection rates in Senegal: natural course of the epidemic or evidence for success of prevention? *AIDS* **13:** 1397-1405

Meloni ST, Kim B, Sankalé J-L, Hamel DJ, Tovanabutra S, Mboup S, McCutchan FE & Kanki PJ (2004a) Distinct human immunodeficiency virus type 1 subtype A virus circulating in West Africa: sub-subtype A3. *J. Virol* **78:** 12438-12445

Meloni ST, Sankalé J-L, Hamel DJ, Eisen G, Guéye-Ndiaye A, Mboup S & Kanki PJ (2004b) Molecular epidemiology of human immunodeficiency virus type 1 sub-subtype A3 in Senegal from 1988 to 2001. *J. Virol* **78:** 12455-12461

Metropolis N, Rosenbluth AW, Rosenbluth MN, Teller AH & Teller E (1953) Equation of State Calculations by Fast Computing Machines. *The Journal of Chemical Physics* **21:** 1087

Minaka N (1993) Algebraic properties of the most parsimonious reconstructions of the hypothetical ancestors on a given tree. *Forma* **8:** 277-296

Minin VN, Bloomquist EW & Suchard MA (2008) Smooth skyride through a rough skyline: Bayesian coalescent-based inference of population dynamics. *Mol. Biol. Evol* **25:** 1459-1471

Mooers A, Harmon L, Blum M, Wong D & Heard S (2007) Some models of phylogenetic tree shape. In *Reconstructing Evolution* p 149-170. New York: O Gascuel and M Steel

Morton BR & Clegg MT (1995) Neighboring base composition is strongly correlated with base substitution bias in a region of the chloroplast genome. *J. Mol. Evol.* **41:** 597-603

Nakano T, Lu L, Liu P & Pybus OG (2004) Viral gene sequences reveal the variable history of hepatitis C virus infection among countries. *J. Infect. Dis.* **190:** 1098-1108

Ndiaye HD, Toure-Kane C, Vidal N, Niama FR, Niang-Diallo PA, Dièye T, Gaye-Diallo A, Wade AS, Peeters M & Mboup S (2009) Surprisingly high prevalence of subtype C and specific HIV-1 subtype/CRF distribution in men having sex with men in Senegal. *J. Acquir. Immune Defic. Syndr* **52:** 249-252

Neilson JR, John GC, Carr JK, Lewis P, Kreiss JK, Jackson S, Nduati RW, Mbori-Ngacha D, Panteleeff DD, Bodrug S, Giachetti C, Bott MA, Richardson BA, Bwayo J, Ndinya-Achola J & Overbaugh J (1999) Subtypes of human immunodeficiency virus type 1 and disease stage among women in Nairobi, Kenya. *J. Virol.* **73:** 4393-4403

Ng OT, Eyzaguirre LM, Carr JK, Chew KK, Lin L, Chua A, Leo YS, Redd AD, Quinn TC & Laeyendecker O (2011) Identification of New CRF51_01B in Singapore Using Full Genome Analysis of Three HIV Type 1 Isolates. *AIDS Res. Hum. Retroviruses* **28:** 527-530

Niama FR, Toure-Kane C, Vidal N, Obengui P, Bikandou B, Ndoundou Nkodia MY, Montavon C, Diop-Ndiaye H, Mombouli JV, Mokondzimobe E, Diallo AG, Delaporte E, Parra H-J, Peeters M & Mboup S (2006) HIV-1 subtypes and recombinants in the Republic of Congo. *Infect. Genet. Evol* **6:** 337-343

Niang CI, Tapsoba P, Weiss E, Diagne M, Niang Y, Moreau AM, Gomis D, Wade AS, Seck K & Castle C (2003) « It »s Raining Stones': Stigma, Violence and HIV Vulnerability among Men Who Have Sex with Men in Dakar, Senegal. *Culture, Health & Sexuality* **5:** 499-512

Nkengasong J, Sylla-Koko F, Peeters M, Ellenberger D, Sassan-Morokro M, Ekpini RA, Msellati P, Greenberg AE, Combe P & Rayfield M (1998) HIV-1 group O virus infection in Abidjan, Côte d'Ivoire. *AIDS* **12:** 1565-1566

Nowak MA (1992) What is a quasispecies? *Trends Ecol. Evol. (Amst.)* **7:** 118-121

O'Brien JD, She Z-S & Suchard MA (2008) Dating the time of viral subtype divergence. *BMC Evol. Biol.* **8:** 172

de Oliveira T, Deforche K, Cassol S, Salminen M, Paraskevis D, Seebregts C, Snoeck J, van Rensburg EJ, Wensing AMJ, van de Vijver DA, Boucher CA, Camacho R & Vandamme A-M (2005) An automated genotyping system for analysis of HIV-1 and other microbial sequences. *Bioinformatics* **21:** 3797-3800

de Oliveira T, Pillay D & Gifford RJ (2010) The HIV-1 subtype C epidemic in South America is linked to the United Kingdom. *PLoS ONE* **5:** e9311

Olsen GJ, Matsuda H, Hagstrom R & Overbeek R (1994) fastDNAmL: a tool for construction of phylogenetic trees of DNA sequences using maximum likelihood. *Comput. Appl. Biosci.* **10:** 41-48

Omland KE (1999) The Assumptions and Challenges of Ancestral State Reconstructions. *Systematic Biology* **48:** 604-611

ONUSIDA (2009) Le point sur l'épidemie de sida, décembre 2009 Genève

ONUSIDA (2010) Rapport ONUSIDA sur l'épidémie mondiale de SIDA 2010 Genève

Ou CY, Ciesielski CA, Myers G, Bandea CI, Luo CC, Korber BT, Mullins JI, Schochetman G, Berkelman RL & Economou AN (1992) Molecular epidemiology of HIV transmission in a dental practice. *Science* **256:** 1165-1171

Palella FJ Jr, Delaney KM, Moorman AC, Loveless MO, Fuhrer J, Satten GA, Aschman DJ & Holmberg SD (1998) Declining morbidity and mortality among patients with advanced human immunodeficiency virus infection. HIV Outpatient Study Investigators. *N. Engl. J. Med.* **338:** 853-860

Paraschiv S, Foley B & Otelea D (2011) Diversity of HIV-1 subtype C strains isolated in Romania. *Infect. Genet. Evol* **11:** 270-275

Paraskevis D, Pybus O, Magiorkinis G, Hatzakis A, Wensing AM, van de Vijver DA, Albert J, Angarano G, Asjö B, Balotta C, Boeri E, Camacho R, Chaix M-L, Coughlan S, Costagliola D, De Luca A, de Mendoza C, Derdelinckx I, Grossman Z, Hamouda O, *et al* (2009) Tracing the HIV-1 subtype B mobility in Europe: a phylogeographic approach. *Retrovirology* **6:** 49

Passaes CPB, Bello G, Lorete RS, Matos Almeida SE, Junqueira DM, Veloso VG, Morgado MG & Guimarães ML (2009) Genetic characterization of HIV-1 BC recombinants and evolutionary history of the CRF31_BC in Southern Brazil. *Infect. Genet. Evol* **9:** 474-482

Pearson WR, Robins G & Zhang T (1999) Generalized neighbor-joining: more reliable phylogenetic tree reconstruction. *Mol. Biol. Evol.* **16:** 806-816

Peeters M, Aghokeng AF & Delaporte E (2010) Genetic diversity among human immunodeficiency virus-1 non-B subtypes in viral load and drug resistance assays. *Clin. Microbiol. Infect.* **16:** 1525-1531

Peeters M, Courgnaud V, Abela B, Auzel P, Pourrut X, Bibollet-Ruche F, Loul S, Liegeois F, Butel C, Koulagna D, Mpoudi-Ngole E, Shaw GM, Hahn BH & Delaporte E (2002) Risk to human health

from a plethora of simian immunodeficiency viruses in primate bushmeat. *Emerging Infect. Dis.* **8:** 451-457

Peeters M, Gaye A, Mboup S, Badombena W, Bassabi K, Prince-David M, Develoux M, Liegeois F, van der Groen G, Saman E & Delaporte E (1996) Presence of HIV-1 group O infection in West Africa. *AIDS* **10:** 343-344

Peeters M, Gueye A, Mboup S, Bibollet-Ruche F, Ekaza E, Mulanga C, Ouedrago R, Gandji R, Mpele P, Dibanga G, Koumare B, Saidou M, Esu-Williams E, Lombart JP, Badombena W, Luo N, Vanden Haesevelde M & Delaporte E (1997) Geographical distribution of HIV-1 group O viruses in Africa. *AIDS* **11:** 493-498

Peeters M, Honoré C, Huet T, Bedjabaga L, Ossari S, Bussi P, Cooper RW & Delaporte E (1989) Isolation and partial characterization of an HIV-related virus occurring naturally in chimpanzees in Gabon. *AIDS* **3:** 625-630

Peeters M, Liegeois F, Torimiro N, Bourgeois A, Mpoudi E, Vergne L, Saman E, Delaporte E & Saragosti S (1999) Characterization of a highly replicative intergroup M/O human immunodeficiency virus type 1 recombinant isolated from a Cameroonian patient. *J. Virol* **73:** 7368-7375

Peeters M, Lobe V, Nkengasong J, Willems B, Delforge ML, Van Renterghem L, Revets H, Sprecher S & van der Groen G (1995) HIV-1 group O infection in Belgium. *Acta Clin Belg* **50:** 171-173

Peeters M, Toure-Kane C & Nkengasong JN (2003) Genetic diversity of HIV in Africa: impact on diagnosis, treatment, vaccine development and trials. *AIDS* **17:** 2547-2560

Perelson AS, Neumann AU, Markowitz M, Leonard JM & Ho DD (1996) HIV-1 dynamics in vivo: virion clearance rate, infected cell life-span, and viral generation time. *Science* **271:** 1582-1586

Van de Perre P, Clumeck N, Carael M, Nzabihimana E, Robert-Guroff M, De Mol P, Freyens P, Butzler JP, Gallo RC & Kanyamupira JB (1985) Female prostitutes: a risk group for infection with human T-cell lymphotropic virus type III. *Lancet* **2:** 524-527

Perrière G & Brochier-Armanet C (2010) Concepts et Méthodes En Phylogénie Moléculaire 1er éd. Springer Verlag France

Perrin L, Kaiser L & Yerly S (2003) Travel and the spread of HIV-1 genetic variants. *Lancet Infect Dis* **3:** 22-27

de Pinna MCC (1991) Concepts and Tests of Homology in the Cladistic Paradigm. *Cladistics* **7:** 367-394

Plantier J-C, Leoz M, Dickerson JE, De Oliveira F, Cordonnier F, Lemée V, Damond F, Robertson DL & Simon F (2009) A new human immunodeficiency virus derived from gorillas. *Nat. Med* **15:** 871-872

Pneumocystis pneumonia--Los Angeles (1981) *MMWR Morb. Mortal. Wkly. Rep* **30:** 250-252

Posada D & Crandall KA (2001) Selecting models of nucleotide substitution: an application to human immunodeficiency virus 1 (HIV-1). *Mol. Biol. Evol* **18:** 897-906

Poulsen AG, Aaby P, Gottschau A, Kvinesdal BB, Dias F, Mølbak K & Lauritzen E (1993) HIV-2 infection in Bissau, West Africa, 1987-1989: incidence, prevalences, and routes of transmission. *J. Acquir. Immune Defic. Syndr.* **6:** 941-948

Preston BD & Dougherty JP (1996) Mechanisms of retroviral mutation. *Trends Microbiol.* **4:** 16-21

Qiu Z, Xing H, Wei M, Duan Y, Zhao Q, Xu J & Shao Y (2005) Characterization of five nearly full-length genomes of early HIV type 1 strains in Ruili city: implications for the genesis of CRF07_BC and CRF08_BC circulating in China. *AIDS Res. Hum. Retroviruses* **21:** 1051-1056

Quiñones-Mateu M, Ball S & Arts E (2000) Role of Human Immunodeficiency Virus Type 1 Group O in the AIDS Pandemic. *AIDS Reviews* **2:** 190-202

Quiñones-Mateu ME, Albright JL, Mas A, Soriano V & Arts EJ (1998) Analysis of pol gene heterogeneity, viral quasispecies, and drug resistance in individuals infected with group O strains of human immunodeficiency virus type 1. *J. Virol.* **72:** 9002-9015

Rambaut A (2000) Estimating the rate of molecular evolution: incorporating non-contemporaneous sequences into maximum likelihood phylogenies. *Bioinformatics* **16:** 395-399

Rambaut A & Grassly NC (1997) Seq-Gen: an application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees. *Comput. Appl. Biosci* **13:** 235-238

Rambaut A, Posada D, Crandall KA & Holmes EC (2004) The causes and consequences of HIV evolution. *Nat. Rev. Genet* **5:** 52-61

Raup DM, Gould SJ, Schopf TJM & Simberloff DS (1973) Stochastic Models of Phylogeny and the Evolution of Diversity. *The Journal of Geology* **81:** 525-542

Reha-Krantz LJ (2010) DNA polymerase proofreading: Multiple roles maintain genome stability. *Biochim. Biophys. Acta* **1804:** 1049-1063

Rerks-Ngarm S, Pitisuttithum P, Nitayaphan S, Kaewkungwal J, Chiu J, Paris R, Premsri N, Namwat C, de Souza M, Adams E, Benenson M, Gurunathan S, Tartaglia J, McNeil JG, Francis DP, Stablein D, Birx DL, Chunsuttiwat S, Khamboonruang C, Thongcharoen P, *et al* (2009) Vaccination with ALVAC and AIDSVAX to prevent HIV-1 infection in Thailand. *N. Engl. J. Med.* **361:** 2209-2220

Roberts JD, Bebenek K & Kunkel TA (1988) The accuracy of reverse transcriptase from HIV-1. *Science* **242:** 1171-1173

Robertson DL, Anderson JP, Bradac JA, Carr JK, Foley B, Funkhouser RK, Gao F, Hahn BH, Kalish ML, Kuiken C, Learn GH, Leitner T, McCutchan F, Osmanov S, Peeters M, Pieniazek D, Salminen M, Sharp PM, Wolinsky S & Korber B (2000) HIV-1 nomenclature proposal. *Science* **288:** 55-56

Rodrigo A, Ewing G & Drummond A (2007) The evolutionary analysis of measurably evolving populations using serially sampled gene sequences. In *Reconstructing Evolution* p 30-61. New York: Olivier Gascuel et Mike Steel

Rodrigo AG, Shpaer EG, Delwart EL, Iversen AK, Gallo MV, Brojatsch J, Hirsch MS, Walker BD & Mullins JI (1999) Coalescent estimates of HIV-1 generation time in vivo. *Proc. Natl. Acad. Sci. U.S.A.* **96:** 2187-2191

Ronquist F & Huelsenbeck JP (2003) MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics* **19:** 1572-1574

Roques P, Robertson DL, Souquière S, Apetrei C, Nerrienet E, Barré-Sinoussi F, Müller-Trutwin M & Simon F (2004) Phylogenetic characteristics of three new HIV-1 N strains and implications for the origin of group N. *AIDS* **18:** 1371-1381

Rouet F & Rouzioux C (2007) HIV-1 viral load testing cost in developing countries: what's new? *Expert Rev. Mol. Diagn.* **7:** 703-707

Rousseau CM, Learn GH, Bhattacharya T, Nickle DC, Heckerman D, Chetty S, Brander C, Goulder PJR, Walker BD, Kiepiela P, Korber BT & Mullins JI (2007) Extensive intrasubtype recombination in South African human immunodeficiency virus type 1 subtype C infections. *J. Virol.* **81:** 4492-4500

Ryan CB, Kama M, Darcy A, Aleksic E, Mirza T, Chaudhary A, Oelrichs RB, Rogers GD & Crowe SM (2009) HIV type 1 in Fiji is caused by subtypes C and B. *AIDS Res. Hum. Retroviruses* **25:** 1355-1358

Ryan CE, Gare J, Crowe SM, Wilson K, Reeder JC & Oelrichs RB (2007) The heterosexual HIV type 1 epidemic in Papua New Guinea is dominated by subtype C. *AIDS Res. Hum. Retroviruses* **23:** 941-944

Rzhetsky A & Nei M (1995) Tests of applicability of several substitution models for DNA sequence data. *Mol. Biol. Evol.* **12:** 131-151

Saitou N & Nei M (1987) The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.* **4:** 406-425

Salemi M, Lamers SL, Yu S, de Oliveira T, Fitch WM & McGrath MS (2005) Phylodynamic analysis of human immunodeficiency virus type 1 in distinct brain compartments provides a model for the neuropathogenesis of AIDS. *J. Virol.* **79:** 11343-11352

Salemi M, de Oliveira T, Ciccozzi M, Rezza G & Goodenow MM (2008) High-resolution molecular epidemiology and evolutionary history of HIV-1 subtypes in Albania. *PLoS ONE* **3:** e1390

Sanderson MJ (1997) A Nonparametric Approach to Estimating Divergence Times in the Absence of Rate Constancy. *Mol Biol Evol* **14:** 1218

Sanderson MJ (2002) Estimating absolute rates of molecular evolution and divergence times: a penalized likelihood approach. *Mol. Biol. Evol.* **19:** 101-109

Sanderson MJ (2003) r8s: inferring absolute rates of molecular evolution and divergence times in the absence of a molecular clock. *Bioinformatics* **19:** 301-302

Sanjuán R & Wróbel B (2005) Weighted least-squares likelihood ratio test for branch testing in phylogenies reconstructed from distance measures. *Syst. Biol* **54:** 218-229

Sankalé JL, Hamel D, Woolsey A, Traoré T, Dia TC, Guèye-Ndiaye A, Essex M, Mboup T & Kanki P (2000) Molecular evolution of human immunodeficiency virus type 1 subtype A in Senegal: 1988-1997. *J. Hum. Virol* **3:** 157-164

Santiago ML, Bibollet-Ruche F, Bailes E, Kamenya S, Muller MN, Lukasik M, Pusey AE, Collins DA, Wrangham RW, Goodall J, Shaw GM, Sharp PM & Hahn BH (2003) Amplification of a complete simian immunodeficiency virus genome from fecal RNA of a wild chimpanzee. *J. Virol.* **77:** 2233-2242

Santiago ML, Range F, Keele BF, Li Y, Bailes E, Bibollet-Ruche F, Fruteau C, Noë R, Peeters M, Brookfield JFY, Shaw GM, Sharp PM & Hahn BH (2005) Simian immunodeficiency virus infection in free-ranging sooty mangabeys (Cercocebus atys atys) from the Taï Forest, Côte d'Ivoire: implications for the origin of epidemic human immunodeficiency virus type 2. *J. Virol.* **79:** 12515-12527

Santiago ML, Rodenburg CM, Kamenya S, Bibollet-Ruche F, Gao F, Bailes E, Meleth S, Soong S-J, Kilby JM, Moldoveanu Z, Fahey B, Muller MN, Ayouba A, Nerrienet E, McClure HM, Heeney JL, Pusey AE, Collins DA, Boesch C, Wrangham RW, *et al* (2002) SIVcpz in wild chimpanzees. *Science* **295:** 465

Schluter D, Price T, Mooers AØ & Ludwig D (1997) Likelihood of Ancestor States in Adaptive Radiation. *Evolution* **51:** 1699-1711

Schultz TR & Churchill GA (1999) The Role of Subjectivity in Reconstructing Ancestral Character States: A Bayesian Approach to Unknown Rates, States, and Transformation Asymmetries. *Systematic Biology* **48:** 651-664

Shafer RW, Stevenson D & Chan B (1999) Human Immunodeficiency Virus Reverse Transcriptase and Protease Sequence Database. *Nucleic Acids Res.* **27:** 348-352

Shen C, Craigo J, Ding M, Chen Y & Gupta P (2011) Origin and Dynamics of HIV-1 Subtype C Infection in India. *PLoS ONE* **6:** e25956

Sides TL, Akinsete O, Henry K, Wotton JT, Carr PW & Bartkus J (2005) HIV-1 subtype diversity in Minnesota. *J. Infect. Dis.* **192:** 37-45

de Silva TI, Cotten M & Rowland-Jones SL (2008) HIV-2: the forgotten AIDS virus. *Trends Microbiol.* **16:** 588-595

da Silva ZJ, Oliveira I, Andersen A, Dias F, Rodrigues A, Holmgren B, Andersson S & Aaby P (2008) Changes in prevalence and incidence of HIV-1, HIV-2 and dual infections in urban areas of Bissau, Guinea-Bissau: is HIV-2 disappearing? *AIDS* **22:** 1195-1202

Simon F, Mauclère P, Roques P, Loussert-Ajaka I, Müller-Trutwin MC, Saragosti S, Georges-Courbot MC, Barré-Sinoussi F & Brun-Vézinet F (1998) Identification of a new human immunodeficiency virus type 1 distinct from group M and group O. *Nat. Med* **4:** 1032-1037

Slatkin M & Maddison WP (1989) A cladistic measure of gene flow inferred from the phylogenies of alleles. *Genetics* **123:** 603-613

Smith SM, Christian D, de Lame V, Shah U, Austin L, Gautam R, Gautam A, Apetrei C & Marx PA (2008) Isolation of a new HIV-2 group in the US. *Retrovirology* **5:** 103

Soares EAJM, Martínez AMB, Souza TM, Santos AFA, Da Hora V, Silveira J, Bastos FI, Tanuri A & Soares MA (2005) HIV-1 subtype C dissemination in southern Brazil. *AIDS* **19 Suppl 4:** S81-86

Sokal R & Michener C (1958) A statistical method for evaluating systematic relationships. *University of Kansas Scientific Bulletin* **28:** 1409-1438

Songok EM, Libondo DK, Rotich MC, Oogo SA & Tukei PM (1996) Surveillance for HIV-1 subtypes O and M in Kenya. *Lancet* **347:** 1700

Soriano V, Gutiérrez M, García-Lerma G, Aguilera O, Mas A, Bravo R, Pérez-Labad ML, Baquero M & González-Lahoz J (1996) First case of HIV-1 group O infection in Spain. *Vox Sang.* **71:** 66

Stadler T (2010) Sampling-through-time in birth-death trees. *J. Theor. Biol.* **267:** 396-404

Stamatakis A (2006) RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics* **22:** 2688-2690

Steel M & McKenzie A (2001) Properties of phylogenetic trees generated by Yule-type speciation models. *Math Biosci* **170:** 91-112

Steel M & Mooers A (2009) Expected length of pendant and interior edges of a Yule tree. *Applied Mathematics Letters* **23:** 6

Studier JA & Keppler KJ (1988) A note on the neighbor-joining algorithm of Saitou and Nei. *Mol. Biol. Evol.* **5:** 729-731

Sullivan PS, Do AN, Ellenberger D, Pau CP, Paul S, Robbins K, Kalish M, Storck C, Schable CA, Wise H, Tetteh C, Jones JL, McFarland J, Yang C, Lal RB & Ward JW (2000) Human immunodeficiency virus (HIV) subtype surveillance of African-born persons at risk for group O and group N HIV infections in the United States. *J. Infect. Dis.* **181:** 463-469

Suzuki Y, Yamaguchi-Kabata Y & Gojobori T (2000) Nucleotide Substitution Rates of HIV-1. *AIDS Rev* **2:** 39-47

Switzer WM, Parekh B, Shanmugam V, Bhullar V, Phillips S, Ely JJ & Heneine W (2005) The epidemiology of simian immunodeficiency virus infection in a large number of wild- and captive-born chimpanzees: evidence for a recent introduction following chimpanzee divergence. *AIDS Res. Hum. Retroviruses* **21:** 335-342

Swofford D & Maddison W (1987) Reconstructing Ancestral character states under Wagner Parsimony. *Mathematical Biosciences* **87:** 199-229

Swofford D, Olsen G, Waddell P & Hillis D (1996) Phylogenetic Inference. In *Molecular Systematics* p 407-509. David Hillis, Graig Moitz, Barbara Mable

Takebe Y, Liao H, Hase S, Uenishi R, Li Y, Li X-J, Han X, Shang H, Kamarulzaman A, Yamamoto N, Pybus OG & Tee KK (2010) Reconstructing the epidemic history of HIV-1 circulating recombinant forms CRF07_BC and CRF08_BC in East Asia: the relevance of genetic diversity and phylodynamics for vaccine strategies. *Vaccine* **28 Suppl 2:** B39-44

Takehisa J, Zekeng L, Ido E, Yamaguchi-Kabata Y, Mboudjeka I, Harada Y, Miura T, Kaptu L & Hayami M (1999) Human immunodeficiency virus type 1 intergroup (M/O) recombination in cameroon. *J. Virol* **73:** 6810-6820

Tamura K & Nei M (1993) Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. *Mol. Biol. Evol.* **10:** 512-526

Tamura K, Peterson D, Peterson N, Stecher G, Nei M & Kumar S (2011) MEGA5: Molecular Evolutionary Genetics Analysis using Maximum Likelihood, Evolutionary Distance, and Maximum Parsimony Methods. *Mol Biol Evol* **28:** 2731-2739

Tatt ID, Barlow KL, Clewley JP, Gill ON & Parry JV (2004) Surveillance of HIV-1 subtypes among heterosexuals in England and Wales, 1997-2000. *J. Acquir. Immune Defic. Syndr* **36:** 1092-1099

Taylor BS, Sobieszczyk ME, McCutchan FE & Hammer SM (2008) The challenge of HIV-1 subtype diversity. *N. Engl. J. Med* **358:** 1590-1602

Tee KK, Pybus OG, Li X-J, Han X, Shang H, Kamarulzaman A & Takebe Y (2008) Temporal and spatial dynamics of human immunodeficiency virus type 1 circulating recombinant forms 08_BC and 07_BC in Asia. *J. Virol.* **82:** 9206-9215

Thompson JD, Linard B, Lecompte O & Poch O (2011) A comprehensive benchmark study of multiple sequence alignment methods: current challenges and future perspectives. *PLoS ONE* **6:** e18093

Thomson MM & Fernández-García A (2011) Phylogenetic structure in African HIV-1 subtype C revealed by selective sequential pruning. *Virology* **415:** 30-38

Toure-Kane C, Montavon C, Faye MA, Gueye PM, Sow PS, Ndoye I, Gaye-Diallo A, Delaporte E, Peeters M & Mboup S (2000) Identification of all HIV type 1 group M subtypes in Senegal, a country with low and stable seroprevalence. *AIDS Res. Hum. Retroviruses* **16:** 603-609

Trautwein MD, Wiegmann BM & Yeates DK (2011) Overcoming the effects of rogue taxa: Evolutionary relationships of the bee flies. *PLoS Curr* **3:** RRN1233

Travers SAA, Clewley JP, Glynn JR, Fine PEM, Crampin AC, Sibande F, Mulawa D, McInerney JO & McCormack GP (2004) Timing and reconstruction of the most recent common ancestor of the subtype C clade of human immunodeficiency virus type 1. *J. Virol* **78:** 10501-10506

Triques K, Bourgeois A, Saragosti S, Vidal N, Mpoudi-Ngole E, Nzilambi N, Apetrei C, Ekwalanga M, Delaporte E & Peeters M (1999) High diversity of HIV-1 subtype F strains in Central Africa. *Virology* **259:** 99-109

Tully DC & Wood C (2010) Chronology and evolution of the HIV-1 subtype C epidemic in Ethiopia. *AIDS* **24:** 1577-1582

Update on acquired immune deficiency syndrome (AIDS)--United States (1982) *MMWR Morb. Mortal. Wkly. Rep.* **31:** 507-508, 513-514

Vallari A, Bodelle P, Ngansop C, Makamche F, Ndembi N, Mbanya D, Kaptué L, Gürtler LG, McArthur CP, Devare SG & Brennan CA (2010) Four new HIV-1 group N isolates from Cameroon: Prevalence continues to be low. *AIDS Res. Hum. Retroviruses* **26:** 109-115

Vallari A, Holzmayer V, Harris B, Yamaguchi J, Ngansop C, Makamche F, Mbanya D, Kaptué L, Ndembi N, Gürtler L, Devare S & Brennan CA (2011) Confirmation of putative HIV-1 group P in Cameroon. *J. Virol* **85:** 1403-1407

Vanden Haesevelde M, Decourt JL, De Leys RJ, Vanderborght B, van der Groen G, van Heuverswijn H & Saman E (1994) Genomic cloning and complete sequence analysis of a highly divergent African human immunodeficiency virus isolate. *J. Virol.* **68:** 1586-1596

Véras NMC, Gray RR, Brígido LF de M, Rodrigues R & Salemi M (2011a) High-resolution phylogenetics and phylogeography of human immunodeficiency virus type 1 subtype C epidemic in South America. *J. Gen. Virol.* **92:** 1698-1709

Véras NMC, Santoro MM, Gray RR, Tatem AJ, Presti AL, Olearo F, Cappelli G, Colizzi V, Takou D, Torimiro J, Russo G, Callegaro A, Salpini R, D'Arrigo R, Perno C-F, Goodenow MM, Ciccozzi M & Salemi M (2011b) Molecular Epidemiology of HIV Type 1 CRF02_AG in Cameroon and African Patients Living in Italy. *AIDS Research and Human Retroviruses* **27:** 1173-1182

Vercauteren J, Derdelinckx I, Sasse A, Bogaert M, Ceunen H, De Roo A, De Wit S, Deforche K, Echahidi F, Fransen K, Goffard J-C, Goubau P, Goudeseune E, Yombi J-C, Lacor P, Liesnard C, Moutschen M, Pierard D, Rens R, Schrooten Y, *et al* (2008) Prevalence and epidemiology of HIV type 1 drug resistance among newly diagnosed therapy-naive patients in Belgium from 2003 to 2006. *AIDS Res. Hum. Retroviruses* **24:** 355-362

Vergne L, Bourgeois A, Mpoudi-Ngole E, Mougnutou R, Mbuagbaw J, Liegeois F, Laurent C, Butel C, Zekeng L, Delaporte E & Peeters M (2003) Biological and genetic characteristics of HIV infections in Cameroon reveals dual group M and O infections and a correlation between SI-inducing phenotype of the predominant CRF02_AG variant and disease stage. *Virology* **310:** 254-266

Vessière A, Rousset D, Kfutwah A, Leoz M, Depatureaux A, Simon F & Plantier J-C (2010) Diagnosis and monitoring of HIV-1 group O-infected patients in Cameroun. *J. Acquir. Immune Defic. Syndr.* **53:** 107-110

Vidal N, Mulanga C, Bazepeo SE, Lepira F, Delaporte E & Peeters M (2006) Identification and molecular characterization of subsubtype A4 in central Africa. *AIDS Res. Hum. Retroviruses* **22:** 182-187

Vidal N, Mulanga C, Bazepeo SE, Mwamba JK, Tshimpaka J-W, Kashi M, Mama N, Laurent C, Lepira F, Delaporte E & Peeters M (2005) Distribution of HIV-1 variants in the Democratic Republic of Congo suggests increase of subtype C in Kinshasa between 1997 and 2002. *J. Acquir. Immune Defic. Syndr* **40:** 456-462

Vidal N, Niyongabo T, Nduwimana J, Butel C, Ndayiragije A, Wakana J, Nduwimana M, Delaporte E & Peeters M (2007) HIV type 1 diversity and antiretroviral drug resistance mutations in Burundi. *AIDS Res. Hum. Retroviruses* **23:** 175-180

Vidal N, Peeters M, Mulanga-Kabeya C, Nzilambi N, Robertson D, Ilunga W, Sema H, Tshimanga K, Bongo B & Delaporte E (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. *J. Virol* **74:** 10498-10507

Vinh LS & von Haeseler A (2005) Shortest triplet clustering: reconstructing large phylogenies using representative sets. *BMC Bioinformatics* **6:** 92

Waddell P & Penny D (1996) Evolutionary trees of apes and humans from dna sequences. In *Handbook of Symbolic Evolution* Oxford: Lock AJ and Peters CR

Wade AS, Kane CT, Diallo PAN, Diop AK, Gueye K, Mboup S, Ndoye I & Lagarde E (2005) HIV infection and sexually transmitted infections among men who have sex with men in Senegal. *AIDS* **19:** 2133-2140

Wainberg MA & Jeang K-T (2008) 25 years of HIV-1 research - progress and perspectives. *BMC Med* **6:** 31

Wallace RG, Hodac H, Lathrop RH & Fitch WM (2007) A statistical phylogeography of influenza A H5N1. *Proc. Natl. Acad. Sci. U.S.A* **104:** 4473-4478

Walter BL, Armitage AE, Graham SC, de Oliveira T, Skinhøj P, Jones EY, Stuart DI, McMichael AJ, Chesebro B & Iversen AK (2009) Functional characteristics of HIV-1 subtype C compatible with increased heterosexual transmissibility. *AIDS* **23:** 1047-1057

Weiss RA (2008) Special anniversary review: twenty-five years of human immunodeficiency virus research: successes and challenges. *Clin. Exp. Immunol* **152:** 201-210

Wertheim JO & Worobey M (2009) Dating the age of the SIV lineages that gave rise to HIV-1 and HIV-2. *PLoS Comput. Biol.* **5:** e1000377

Whelan S & Goldman N (2001) A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. *Mol. Biol. Evol.* **18:** 691-699

Worobey M, Gemmel M, Teuwen DE, Haselkorn T, Kunstman K, Bunce M, Muyembe J-J, Kabongo J-MM, Kalengayi RM, Van Marck E, Gilbert MTP & Wolinsky SM (2008) Direct evidence of extensive diversity of HIV-1 in Kinshasa by 1960. *Nature* **455:** 661-664

Worobey M, Santiago ML, Keele BF, Ndjango J-BN, Joy JB, Labama BL, Dhed'A BD, Rambaut A, Sharp PM, Shaw GM & Hahn BH (2004) Origin of AIDS: contaminated polio vaccine theory refuted. *Nature* **428:** 820

Wright E, Mugaba S, Grant P, Parkes-Ratanshi R, Van der Paal L, Grosskurth H & Kaleebu P (2011) Coreceptor and cytokine concentrations may not explain differences in disease progression observed in HIV-1 clade A and D infected Ugandans. *PLoS ONE* **6:** e19902

Xia X & Yang Q (2011) A distance-based least-square method for dating speciation events. *Mol. Phylogenet. Evol.* **59:** 342-353

Yamaguchi J, Bodelle P, Vallari AS, Coffey R, McArthur CP, Schochetman G, Devare SG & Brennan CA (2004) HIV infections in northwestern Cameroon: identification of HIV type 1 group O and dual HIV type 1 group M and group O infections. *AIDS Res. Hum. Retroviruses* **20:** 944-957

Yamaguchi J, Coffey R, Vallari A, Ngansop C, Mbanya D, Ndembi N, Kaptué L, Gürtler LG, Bodelle P, Schochetman G, Devare SG & Brennan CA (2006a) Identification of HIV type 1 group N infections in a husband and wife in Cameroon: viral genome sequences provide evidence for horizontal transmission. *AIDS Res. Hum. Retroviruses* **22:** 83-92

Yamaguchi J, McArthur CP, Vallari A, Coffey R, Bodelle P, Beyeme M, Schochetman G, Devare SG & Brennan CA (2006b) HIV-1 Group N: evidence of ongoing transmission in Cameroon. *AIDS Res. Hum. Retroviruses* **22:** 453-457

Yamaguchi J, Vallari A, Ndembi N, Coffey R, Ngansop C, Mbanya D, Kaptué L, Gürtler LG, Devare SG & Brennan CA (2008) HIV type 2 intergroup recombinant identified in Cameroon. *AIDS Res. Hum. Retroviruses* **24:** 86-91

Yang Z (1993) Maximum-likelihood estimation of phylogeny from DNA sequences when substitution rates differ over sites. *Mol. Biol. Evol.* **10:** 1396-1401

Yang Z (1994) Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. *J. Mol. Evol.* **39:** 306-314

Yang Z, O'Brien JD, Zheng X, Zhu H-Q & She Z-S (2007) Tree and rate estimation by local evaluation of heterochronous nucleotide data. *Bioinformatics* **23:** 169-176

Ye J-R, Yu S-Q, Lu H-Y, Wang W-S, Xin R-L & Zeng Y (2011) Genetic Diversity of HIV Type 1 Isolated from Newly Diagnosed Subjects (2006-2007) in Beijing, China. *AIDS Research and Human Retroviruses* **28:** 119-123

Yoder AD & Yang Z (2000) Estimation of primate speciation dates using local molecular clocks. *Mol. Biol. Evol.* **17:** 1081-1090

Yule U (1925) A Mathematical Theory of Evolution, Based on the Conclusions of Dr. J. C. Willis, F.R.S. *Philosophical Transactions of the Royal Society of London. Series B, Containing Papers of a Biological Character* **213:** 21-87

Zhang J, Tang LY, Li T, Ma Y & Sapp CM (2000) Most retroviral recombinations occur during minus-strand DNA synthesis. *J. Virol.* **74:** 2313-2322

Zhu T, Korber BT, Nahmias AJ, Hooper E, Sharp PM & Ho DD (1998) An African HIV-1 sequence from 1959 and implications for the origin of the epidemic. *Nature* **391:** 594-597

Zhuang J, Jetzt AE, Sun G, Yu H, Klarmann G, Ron Y, Preston BD & Dougherty JP (2002) Human immunodeficiency virus type 1 recombination: rate, fidelity, and putative hot spots. *J. Virol.* **76:** 11273-11282

Zuckerkandl E & Pauling L (1962) Molecular disease, evolution, and genetic heterogeneity. In *Horizons in Biochemistry* p 189-225. New York: Kasha M and Pullman B

Zuckerkandl E & Pauling L (1965) Evolutionary divergence and convergence in proteins. In *Evolving Genes and Proteins* p 97-166. New York: Bryson V and Vogel HJ

Liste des figures

Liste des tableaux

Annexe A

Matériels supplémentaires à l'étude du Chapitre 6

Table des figures et des tableaux de cette annexe

Figure 55. Estimations de l'indice de dispersion avec la parcimonie RandACCTRAN.
Le graphique indique, pour chaque pays de collecte ayant au moins deux souches, les valeurs correspondantes à l'indice de dispersion IN (en vert) et OUT (en rouge) ; leur somme correspondant à l'indice de dispersion totale. Ces résultats sont obtenus avec la parcimonie RandACCTRAN. Les mesures significatives de l'indice de dispersion sont représentées par des barres pleines tandis que les mesures non significatives par des barres vides. Les pays sont regroupés par zone géographique, puis par ordre alphabétique. Une mesure totale de 1 signifie que les souches sont totalement dispersées dans la phylogénie, sans possibilité de formation d'annotations ancestrales (par exemple la Grande-Bretagne). Une mesure de zéro signifie que toutes les souches d'un pays forment un clade monophylétique (seul exemple, l'Uruguay). Pour chaque pays, le nombre de souches présentes dans la phylogénie est rappelé en bleu en abscisse.

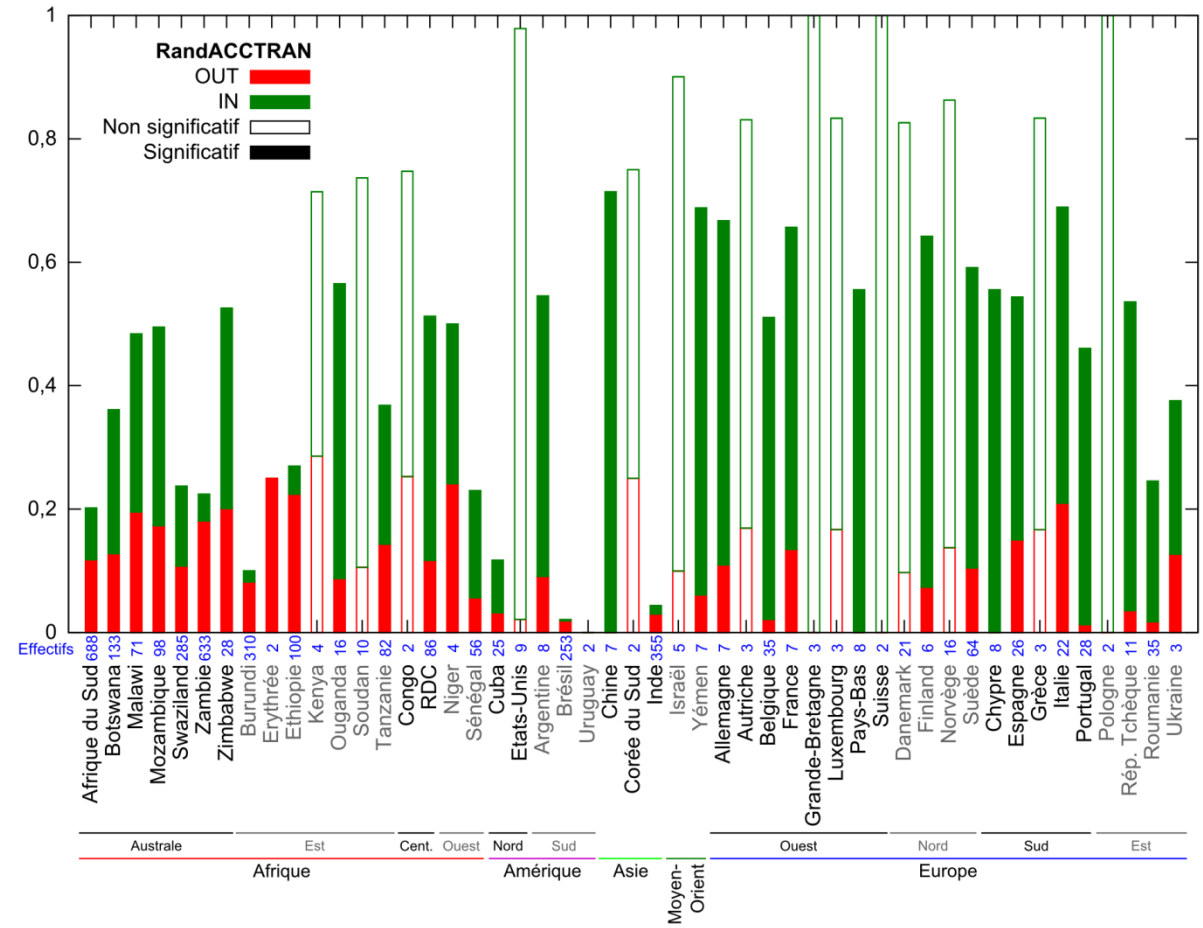


Figure 56. Estimations de l'indice de dispersion avec la parcimonie RandDELTRAN.
Le graphique indique, pour chaque pays de collecte ayant au moins deux souches, les valeurs correspondantes à l'indice de dispersion IN (en vert) et OUT (en rouge) ; leur somme correspondant à l'indice de dispersion totale. Ces résultats sont obtenus avec la parcimonie RandDELTRAN. Les mesures significatives de l'indice de dispersion sont représentées par des barres pleines tandis que les mesures non significatives par des barres vides. Les pays sont regroupés par zone géographique, puis par ordre alphabétique. Une mesure totale de 1 signifie que les souches sont totalement dispersées dans la phylogénie, sans possibilité de formation d'annotations ancestrales (par exemple la Grande-Bretagne). Une mesure de zéro signifie que toutes les souches d'un pays forment un clade monophylétique (seul exemple, l'Uruguay). Pour chaque pays, le nombre de souches présentes dans la phylogénie est rappelé en bleu en abscisse.

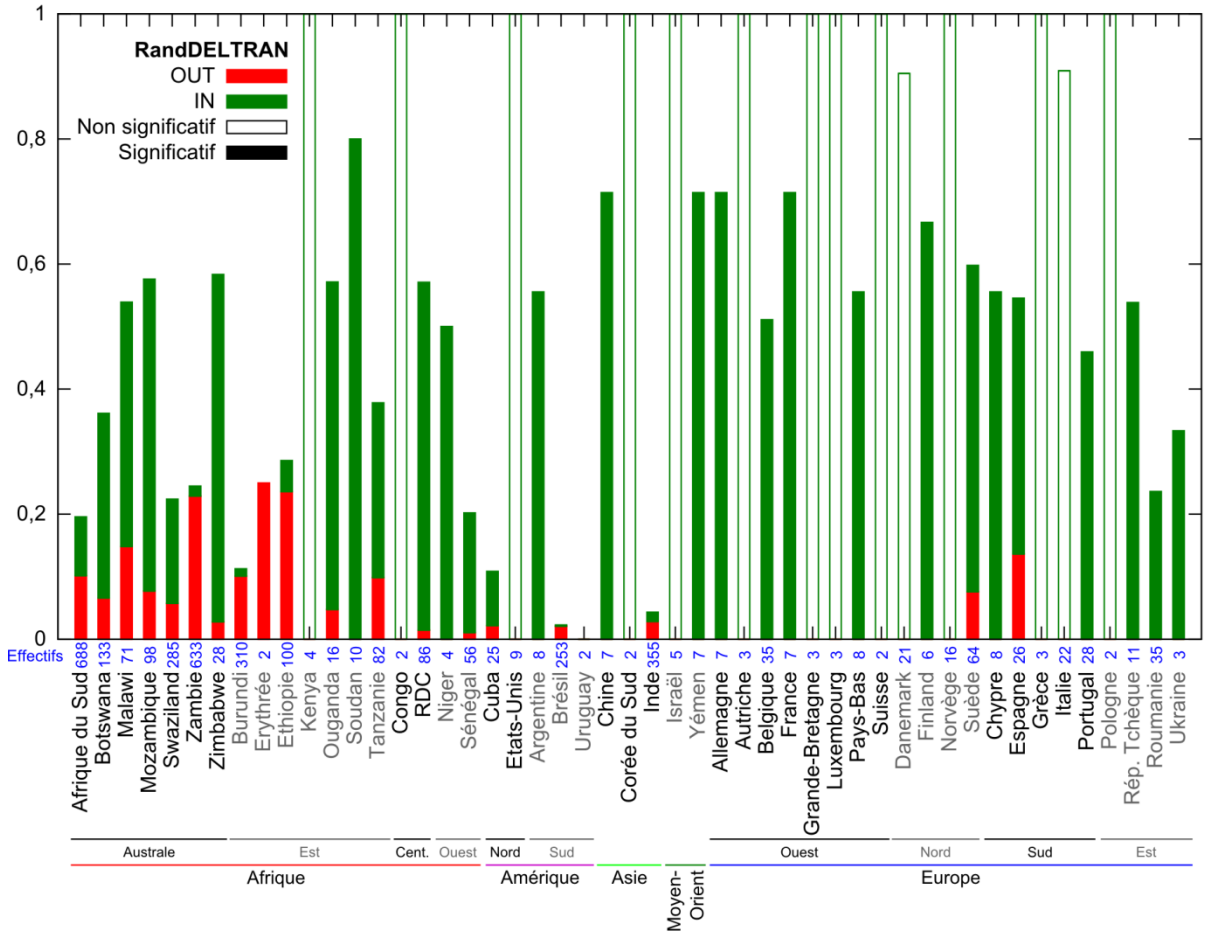


Figure 57. Estimations de l'inde de flux avec la parcimonie RandACCTRAN.
Chaque point sur le graphique reflète la proportion de transitions IN pour le pays en abscisse issues du pays en ordonnée. La somme des points d'une colonne vaut 1. Les pays sont ordonnés par zone géographique, puis par ordre alphabétique. Le nombre de souches de chaque pays est indiqué sur les deux axes en face des pays concernés. Sur l'axe des abscisses, la mesure IN indique le nombre de transitions entrantes dans ce pays. Sur l'axe des ordonnées, la mesure OUT indique le nombre de transitions sortantes de ce pays. Les mesures significatives sont représentées par un cercle plein et les mesures non significatives par un cercle vide. Les pays avec un nombre de transitions OUT inférieur à 1 ne sont pas représentés en ordonnée. Le graphique présenté correspond à la parcimonie RandACCTRAN.

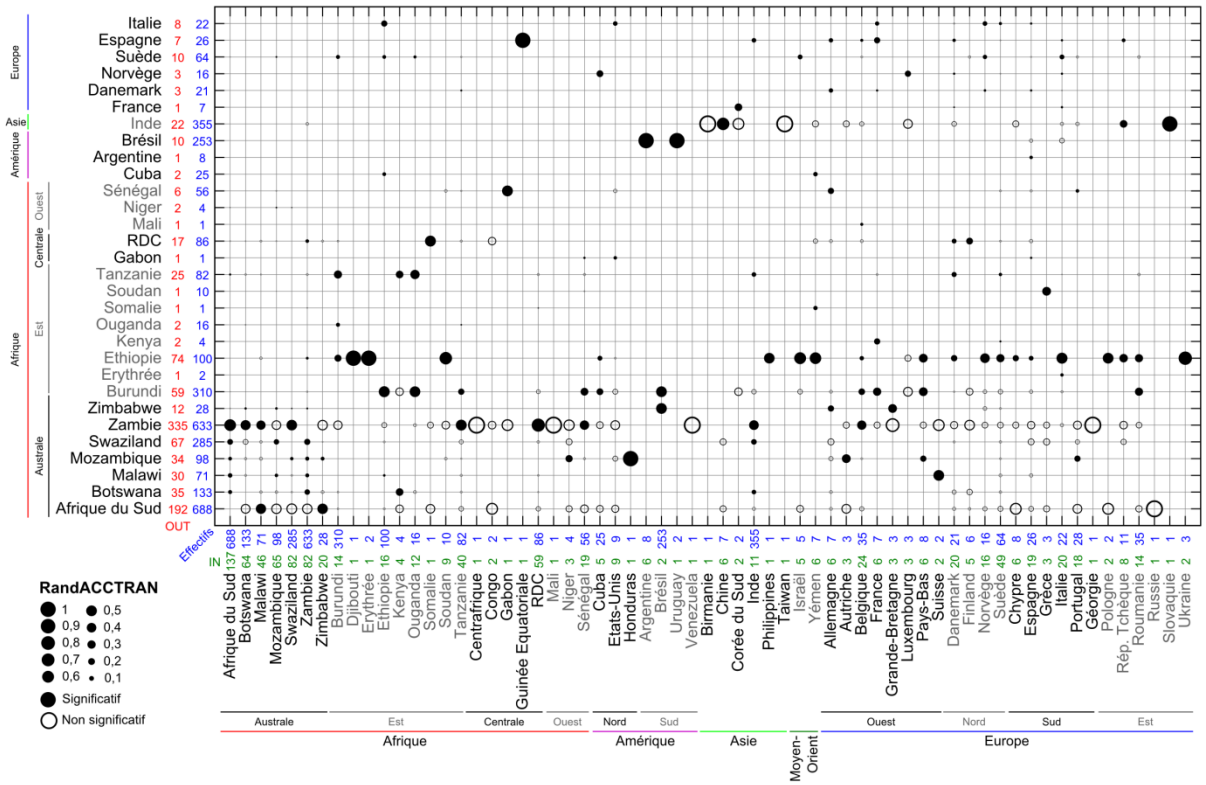


Figure 58. Estimations de l'inde de flux avec la parcimonie RandDELTRAN.
Chaque point sur le graphique reflète la proportion de transitions IN pour le pays en abscisse issues du pays en ordonnée. La somme des points d'une colonne vaut 1. Les pays sont ordonnés par zone géographique, puis par ordre alphabétique. Le nombre de souches de chaque pays est indiqué sur les deux axes en face des pays concernés. Sur l'axe des abscisses, la mesure IN indique le nombre de transitions entrantes dans ce pays. Sur l'axe des ordonnées, la mesure OUT indique le nombre de transitions sortantes de ce pays. Les mesures significatives sont représentées par un cercle plein et les mesures non significatives par un cercle vide. Les pays avec un nombre de transitions OUT inférieur à 1 ne sont pas représentés en ordonnée. Le graphique présenté correspond à la parcimonie RandDELTRAN.

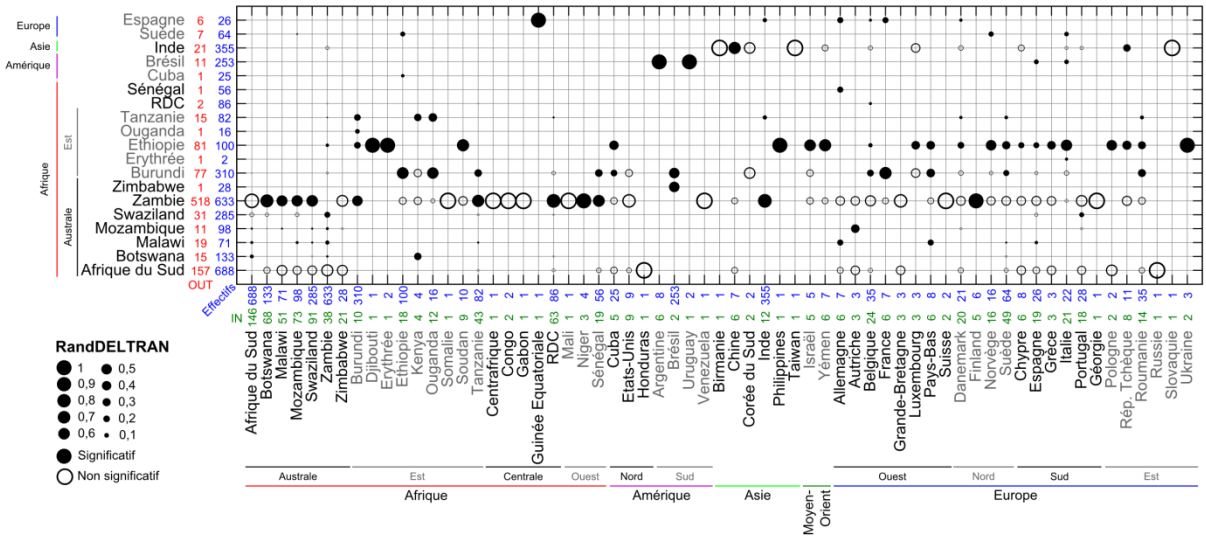


Figure 59. Estimations de l'indice de symétrie avec la parcimonie RandACCTRAN.
Ce graphique renseigne sur la symétrie des échanges entre les pays donneurs pour la parcimonie RandACCTRAN. Pour en faciliter la lecture, la mesure $S_{a\leftrightarrow b}/(N_{a\to b}+N_{b\to a})$ est reportée sur le graphique pour tout a et b. Si la mesure est représentée par un point rouge (resp. bleu) cela signifie qu'il y a plus de mouvement du pays en ordonnée (resp. abscisse) vers le pays en abscisse (resp. ordonnée), que l'inverse. Lorsque la mesure vaut 1 et que le point est rouge (resp. bleu), seuls des mouvements du pays Y (resp. X) vers le pays X (resp. Y) sont observés. Lorsqu'elle vaut 0, l'échange est symétrique. Les cercles vides montrent les mesures non significatives et les cerlces pleins les mesures significatives. Les croix indiquent deux pays sans relations. Les pays sont ordonnés par zone géographique, puis par ordre alphabétique, et leur nombre de souches est rappelé sur les axes. Seuls les pays dont le nombre de transitions OUT est supérieur à 1 sont représentés.

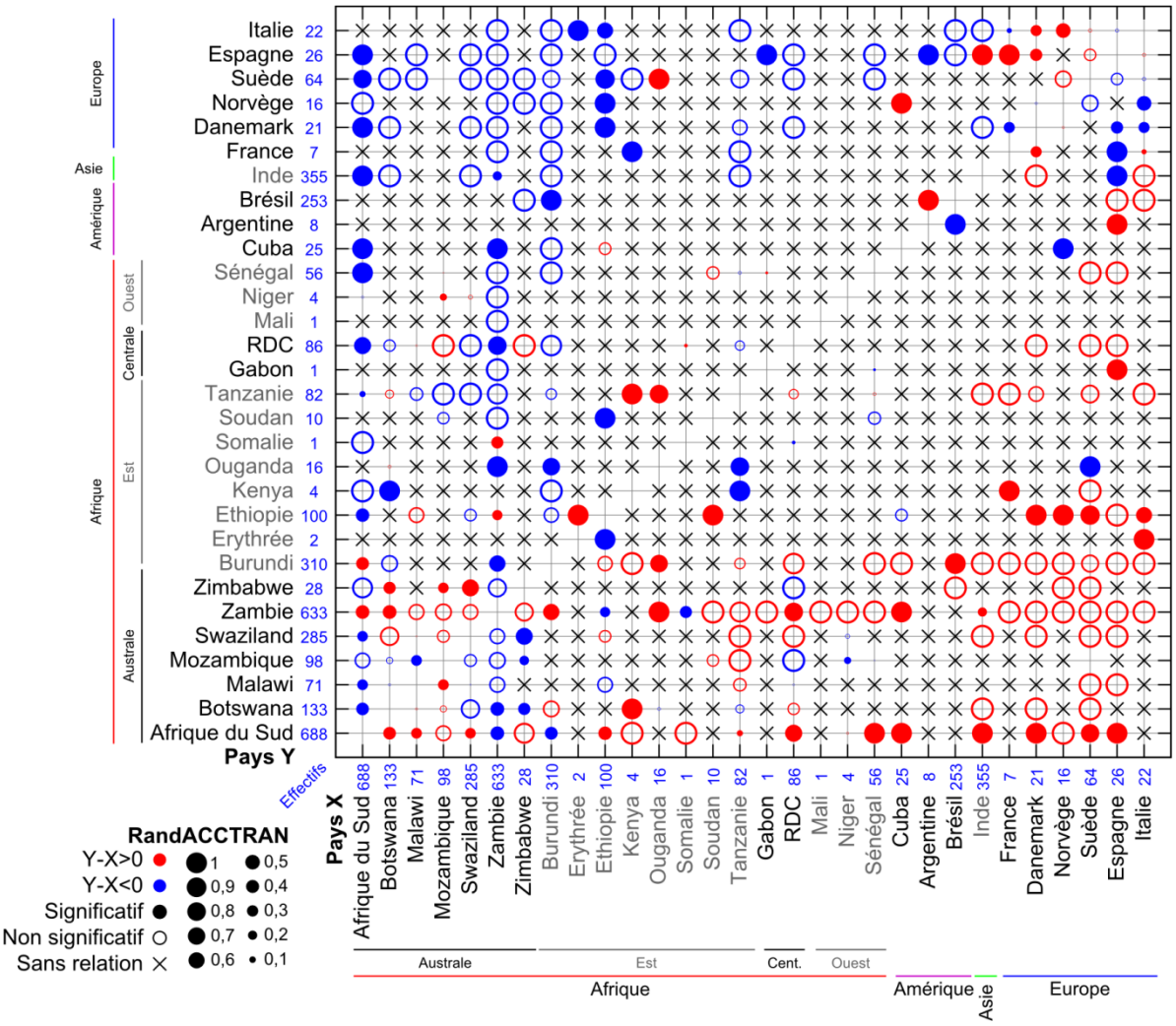


Figure 60. Estimations de l'indice de symétrie avec la parcimonie RandDELTRAN.
Ce graphique renseigne sur la symétrie des échanges entre les pays donneurs pour la parcimonie RandDELTRAN. Pour en faciliter la lecture, la mesure $S_{a\leftrightarrow b}/(N_{a\to b}+N_{b\to a})$ est reportée sur le graphique pour tout a et b. Si la mesure est représentée par un point rouge (resp. bleu) cela signifie qu'il y a plus de mouvement du pays en ordonnée (resp. abscisse) vers le pays en abscisse (resp. ordonnée), que l'inverse. Lorsque la mesure vaut 1 et que le point est rouge (resp. bleu), seuls des mouvements du pays Y (resp. X) vers le pays X (resp. Y) sont observés. Lorsqu'elle vaut 0, l'échange est symétrique. Les cercles vides montrent les mesures non significatives et les cerlces pleins les mesures significatives. Les croix indiquent deux pays sans relations. Les pays sont ordonnés par zone géographique, puis par ordre alphabétique, et leur nombre de souches est rappelé sur les axes. Seuls les pays dont le nombre de transitions OUT est supérieur à 1 sont représentés.

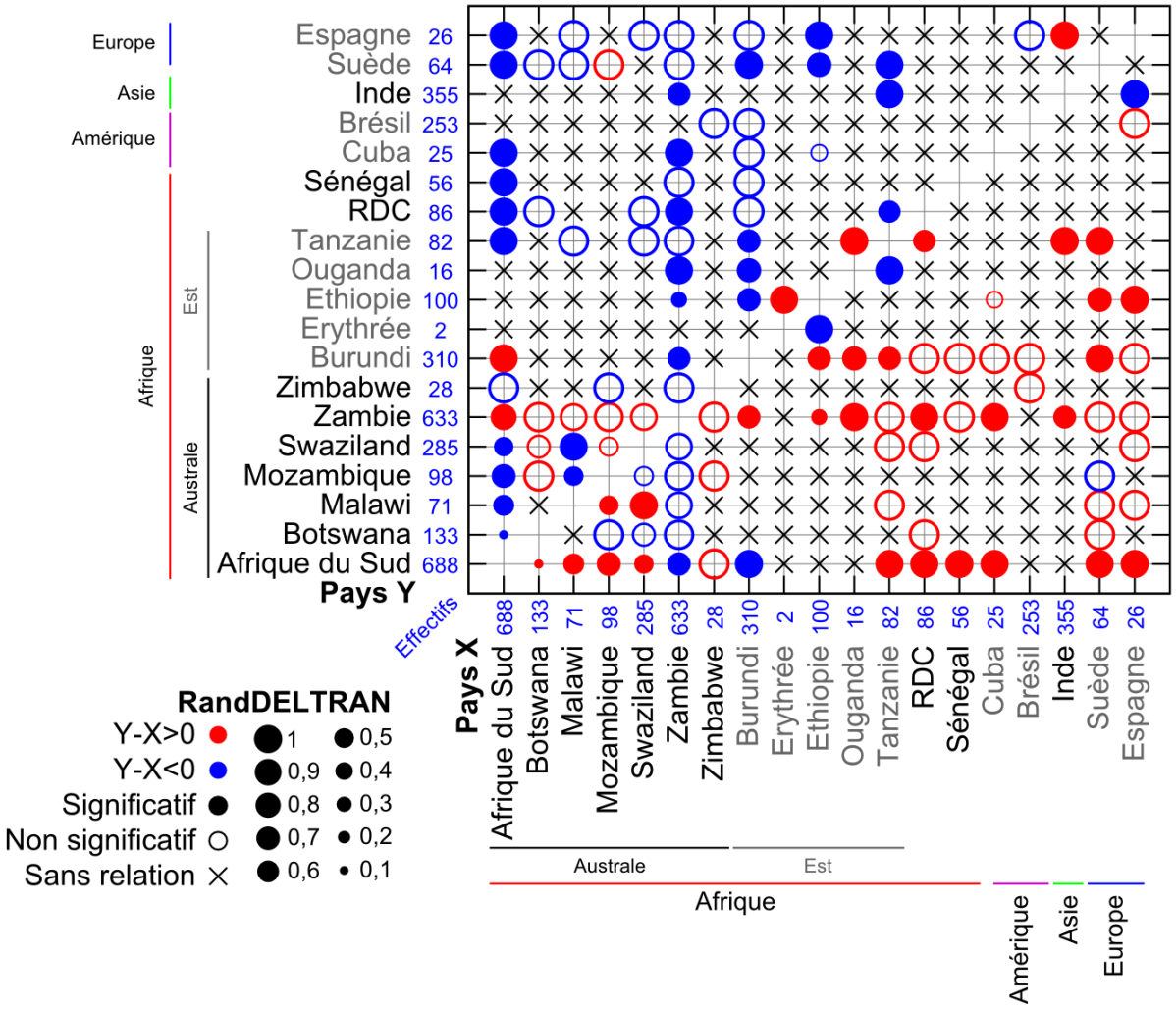


Figure 61. Estimations de l'indice de régionalisation entre les souches de pays européens, asiatiques ou américains.
Ce graphique renseigne sur la possibilité de grouper deux pays ensembles lors de l'analyse avec PhyloType. Les couples de pays ayant une croix sont ceux qui partagent une frontière géographique commune. Les points en bleu indiquent l'indice de régionalisation. Les points verts et rouges indiquent l'indice de régionalisation de l'union des deux pays situés sur l'axe des ordonnées et des abscisses. Lorsque ce point est en rouge la régionalisation de l'union est meilleure que la régionalisation des deux pays pris séparément. Par exemple, la mesure pour le couple Inde/Chine indique que l'union des deux pays est plus régionalisée (point rouge) que celle des pays pris séparément. De plus, ces deux pays partagent une frontière géographique commune (une croix), il est donc conseillé de les grouper ensembles lors de l'analyse avec PhyloType afin de maximiser les chances d'apparition de *phylotypes*. Les pays sont regroupés par zone géographique et seuls les pays ayant au moins deux souches sont représentés. Seuls des groupements entre pays d'un même continent sont envisagés. Il n'y a pas de différence entre les méthodes DELTRAN, ACCTRAN et DOWNPASS.

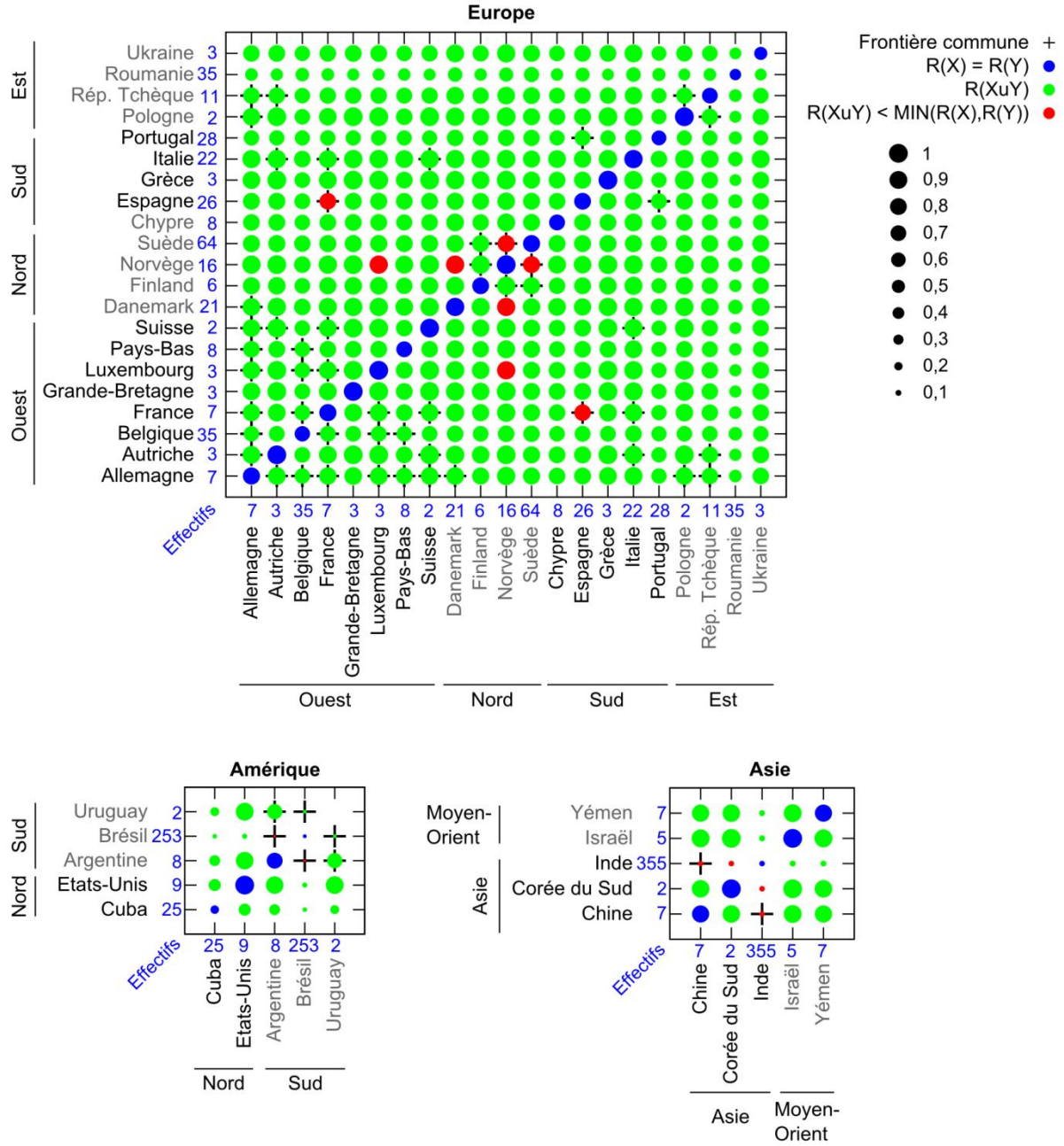


Figure 62. Carte des liens entre *phylotypes* lorsque *size* ≥ 10 avec ACCTRAN.
Carte de l'analyse PhyloType lorsque *size* ≥ 10, *persistence* ≥ 1, *size/different* ≥ 1 et *support* ≥ 70% avec la parcimonie ACCTRAN. Tous les *phylotypes* présentés sont statistiquement supportés (p-valeur inférieure ou égale à 1% pour le critère *size*). La taille des cercles est proportionnelle à la valeur du critère *size* pour le *phylotype* en question. Chaque couleur correspond à une annotation donnée, spécifiée au bas de chaque cercle. Un numéro d'identification est attribué à chaque *phylotype* et est indiqué avant l'annotation correspondante au *phylotype*.

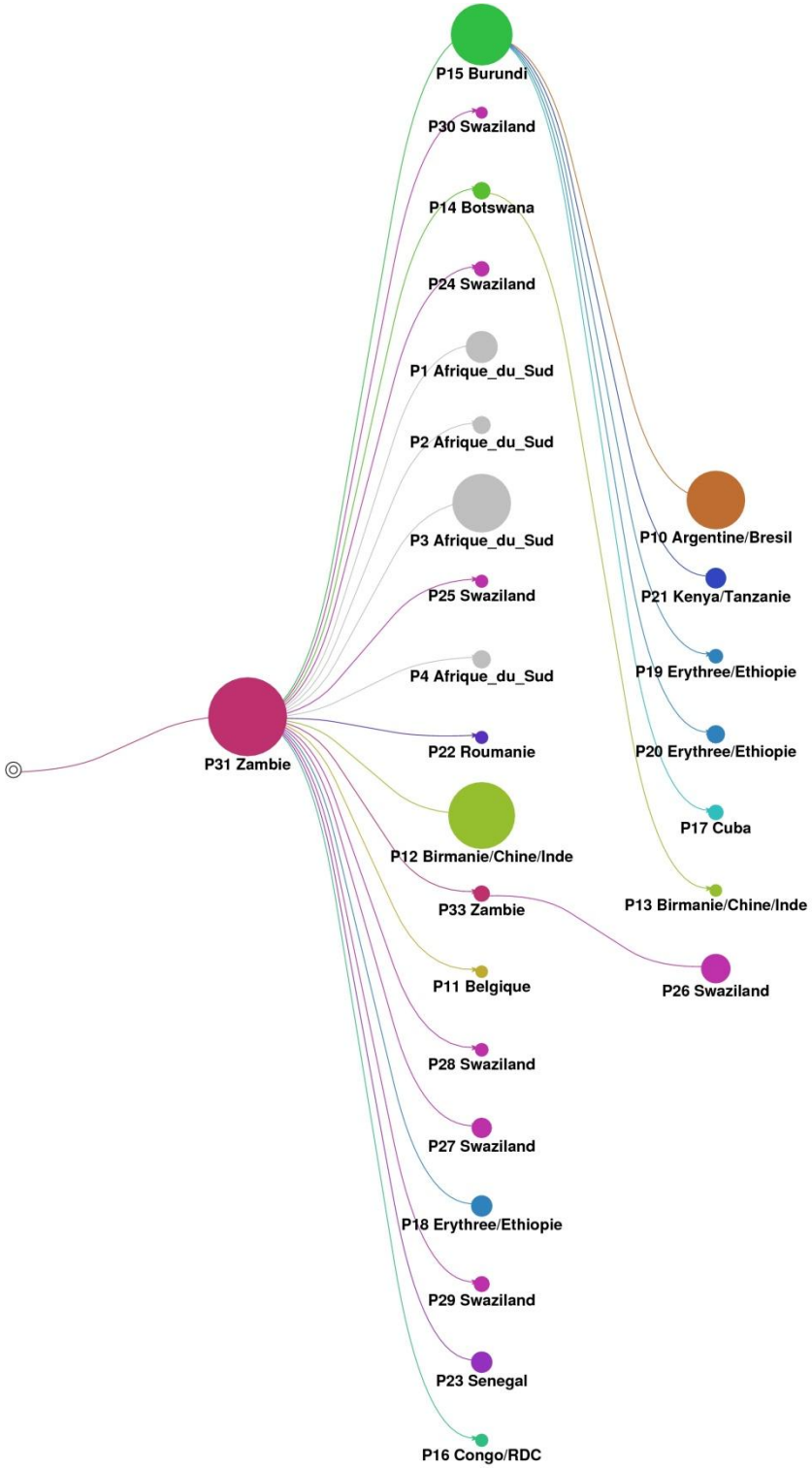


Figure 63. Carte des liens entre *phylotypes* lorsque *size* ≥ 10 avec DELTRAN.
Carte de l'analyse PhyloType lorsque *size* ≥ 10, *persistence* ≥ 1, *size/different* ≥ 1 et *support* ≥ 70% avec la parcimonie DELTRAN. Tous les *phylotypes* présentés sont statistiquement supportés (p-valeur inférieure ou égale à 1% pour le critère *size*). La taille des cercles est proportionnelle à la valeur du critère *size* pour le *phylotype* en question. Chaque couleur correspond à une annotation donnée, spécifiée au bas de chaque cercle. Un numéro d'identification est attribué à chaque *phylotype* et est indiqué avant l'annotation correspondante au *phylotype*.

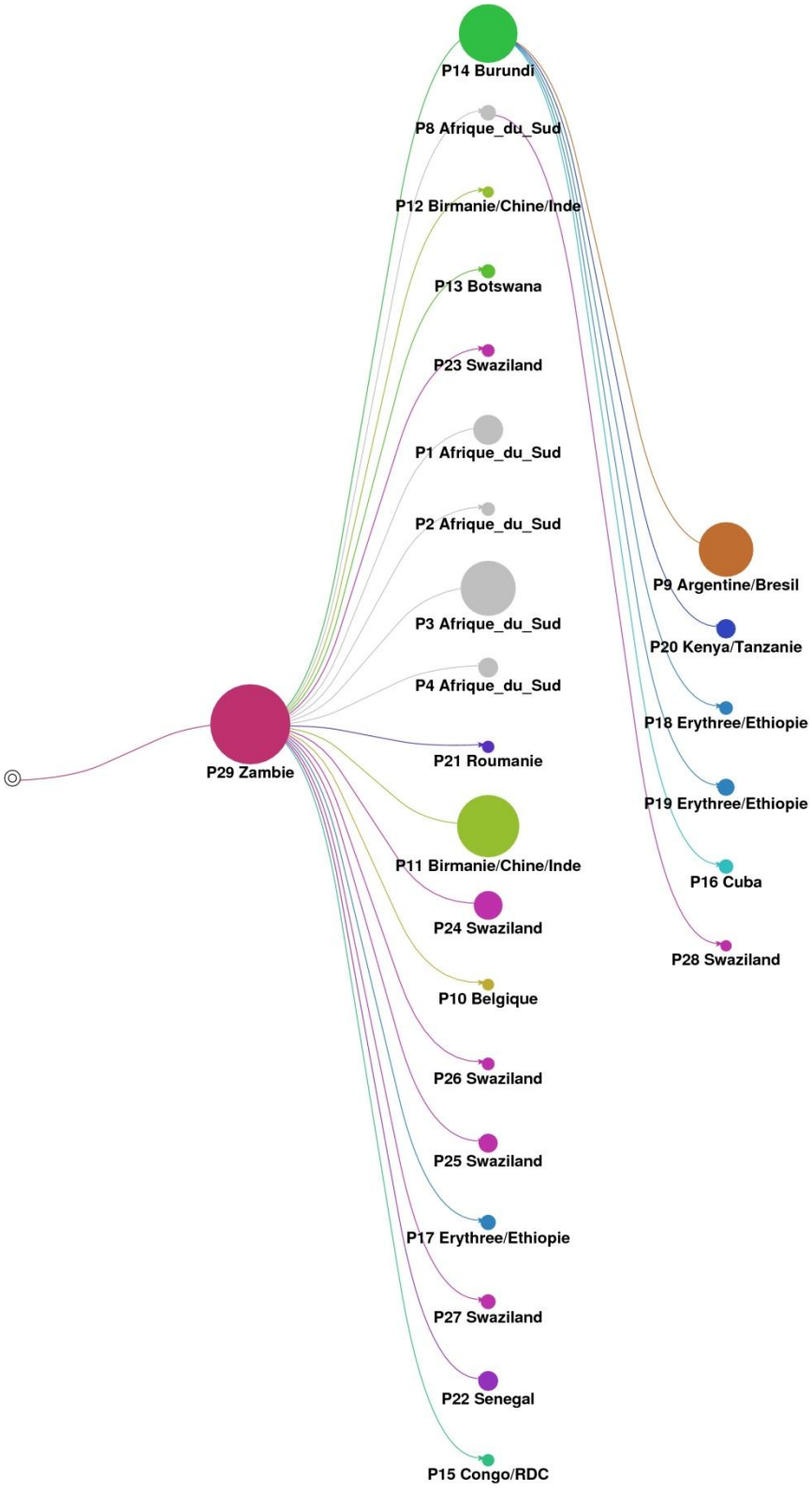


Figure 64. Carte des liens entre *phylotypes* lorsque *size* ≥ 5 avec ACCTRAN.
Carte de l'analyse PhyloType lorsque *size* ≥ 5, *persistence* ≥ 1, *size/different* ≥ 1 et *support* ≥ 70% avec la parcimonie ACCTRAN. Tous les *phylotypes* présentés sont statistiquement supportés (p-valeur inférieure ou égale à 1% pour le critère *size*). La taille des cercles est proportionnelle à la valeur du critère *size* pour le *phylotype* en question. Chaque couleur correspond à une annotation donnée, spécifiée au bas de chaque cercle. Un numéro d'identification est attribué à chaque *phylotype* et est indiqué avant l'annotation correspondante au *phylotype*.

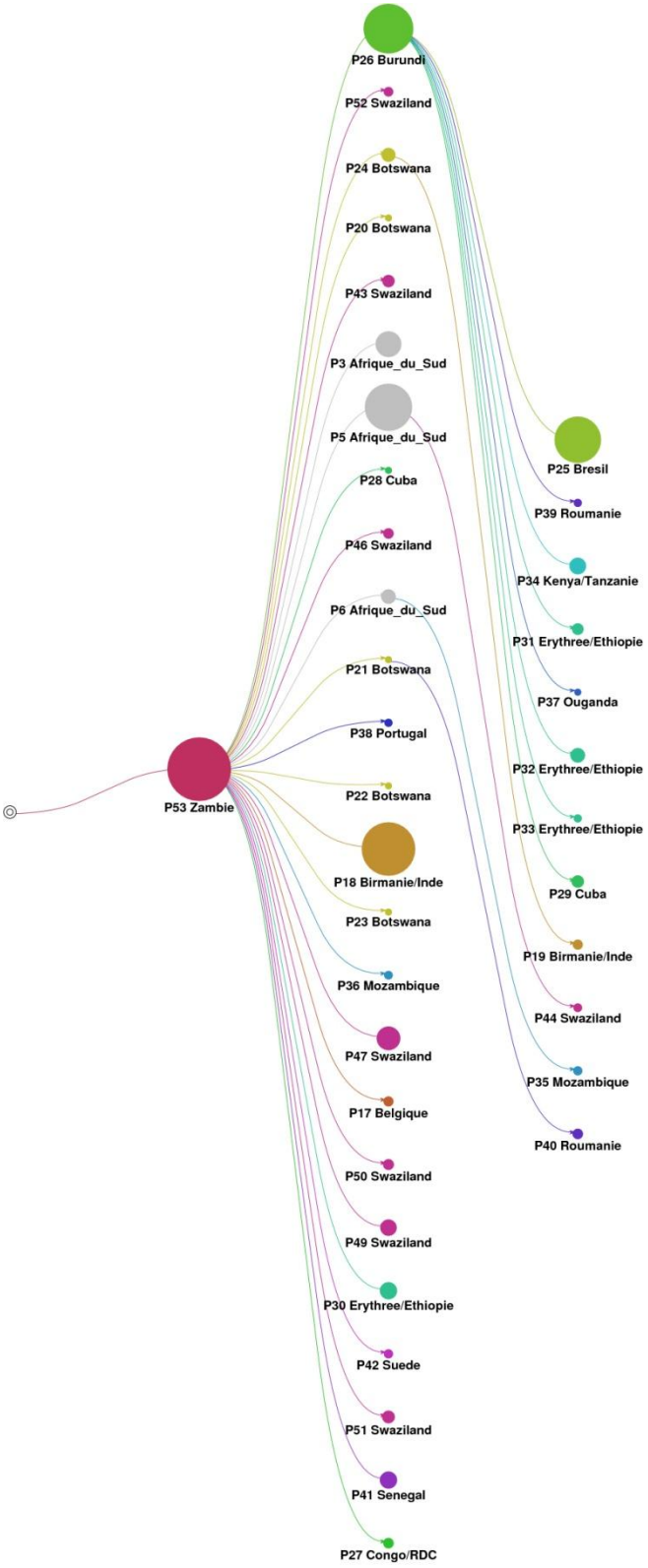


Tableau 9. Valeurs associées à chaque critère pour tous les *phylotypes* significatifs observés lorsque *size* ≥ 20 avec ACCTRAN.

Ce tableau contient la liste de tous les *phylotypes* obtenus lorsque *size* ≥ 20, *persistence* ≥ 1, *size/different* ≥ 1 et *support* ≥ 70% avec la parcimonie ACCTRAN, ainsi que toutes les valeurs associées à chacun des critères que l'on peut choisir. Les valeurs en rose représentent les p-valeur obtenues par le *shuffling*. Seuls les *phylotypes* dont la p-valeur est strictement inférieure à 11/1 000 pour le critère *size* sont représentés. Les abréviations des pays sont les suivantes : AR, Argentine ; BI, Burundi ; BR, Brésil ; BW, Botswana ; CN, Chine ; DJ, Djibouti ; ER, Érythrée ; ET, Éthiopie ; IN, Inde ; KE, Kenya ; MM, Birmanie ; SD, Soudan ; SN, Sénégal ; SZ, Swaziland ; TZ, Tanzanie ; UG, Ouganda ; ZA, Afrique du Sud ; ZM, Zambie. Les abréviations des titres sont les suivantes : P, identifiant du *phylotype* ; A, *annotation* ; Tt, *total* ; Sz, *size* ; Ps, *persistence* ; Df, *different* ; Sl, *local separation* ; Sg, *global separation* ; Dv, *diversity* ; Sp, *support* ; Spg, *global support*.

P	A	Tt	Sz	Ps	Df	Sz/Df	Sl	Sg	Dv	Sl/Dv	Sg/Dv	Sp	Spg
1	ZA	86	**76**	**2**	7	**10,857**	0,005	0,007	0,083	0,064	0,079	**0,855**	0,855
			0/1000	18/1000		0/1000						8/1000	
2	ZA	35	**23**	**2**	4	**5,750**	0,002	0,003	0,082	0,030	0,043	**0,835**	0,835
			10/1000	18/1000		0/1000						11/1000	
3	ZA	311	**261**	**3**	25	**10,440**	0,002	0,014	0,089	0,027	0,155	**0,831**	0,843
			0/1000	1/1000		0/1000						12/1000	
4	ZA	80	**24**	**2**	16	**1,500**	0,007	0,010	0,073	0,095	0,141	**0,876**	0,876
			10/1000	18/1000		0/1000						8/1000	
5	AR/BR	269	**260**	**2**	5	**52,000**	0,018	0,029	0,106	0,172	0,269	**0,992**	0,992
			0/1000	0/1000		0/1000						0/1000	
6	MM/CN/IN	356	**339**	**2**	14	**24,214**	0,004	0,022	0,081	0,044	0,265	**0,740**	0,882
			0/1000	0/1000		0/1000						0/1000	
7	BW	52	**22**	**2**	9	**2,444**	0,002	0,007	0,074	0,033	0,097	**0,756**	0,756
			0/1000	0/1000		0/1000						0/1000	
8	BI	829	**289**	**3**	54	**5,352**	0,006	0,010	0,093	0,065	0,106	**0,861**	0,861
			0/1000	0/1000		0/1000						0/1000	
9	DJ/ER/ET/SD	71	**32**	**2**	25	**1,280**	0,005	0,014	0,049	0,100	0,279	**0,906**	0,906
			0/1000	0/1000		0/1000						0/1000	
10	DJ/ER/ET/SD	47	**26**	**2**	15	**1,733**	0,003	0,023	0,059	0,042	0,394	**0,773**	0,773
			0/1000	0/1000		0/1000						0/1000	
11	KE/UG/TZ	43	**35**	**3**	5	**7,000**	0,009	0,014	0,082	0,104	0,172	**0,926**	0,926
			0/1000	0/1000		0/1000						0/1000	
12	SN	33	**33**	**1**	0	∞	0,018	0,033	0,075	0,240	0,438	**0,980**	0,980
			0/1000	0/1000		0/1000						0/1000	
13	SZ	87	**64**	**3**	11	**5,818**	0,003	0,045	0,074	0,036	0,612	**0,749**	0,766
			0/1000	0/1000		0/1000						0/1000	
14	SZ	33	**30**	**2**	3	**10,000**	0,002	0,021	0,059	0,041	0,352	**0,781**	0,781
			0/1000	0/1000		0/1000						0/1000	
15	ZM	3605	**475**	**3**	281	**1,690**	0,014	0	0,115	0,120	0,000	**0,880**	0,880
			0/1000	2/1000		0/1000						1/1000	

Tableau 10. Valeurs associées à chaque critère pour tous les *phylotypes* significatifs observés lorsque *size* ≥ 10 avec ACCTRAN.

Ce tableau contient la liste de tous les *phylotypes* obtenus lorsque *size* ≥ 10, *persistence* ≥ 1, *size/different* ≥ 1 et *support* ≥ 70% avec la parcimonie ACCTRAN, ainsi que toutes les valeurs associées à chacun des critères que l'on peut choisir. Les valeurs en rose représentent les p-valeur obtenues par le *shuffling*. Seuls les *phylotypes* dont la p-valeur est strictement inférieure à 11/1 000 pour le critère *size* sont représentés. Les abréviations des pays sont les suivantes : AR, Argentine ; BE, Belgique ; BI, Burundi ; BR, Brésil ; BW, Botswana ; CD, République Démocratique du Congo ; CG, Congo ; CN, Chine ; CU, Cuba ; DJ, Djibouti ; ER, Érythrée ; ET, Éthiopie ; IN, Inde ; KE, Kenya ; MM, Birmanie ; RO, Roumanie ; SN, Sénégal ; SZ, Swaziland ; TZ, Tanzanie ; ZA, Afrique du Sud ; ZM, Zambie. Les abréviations des titres sont les suivantes : P, identifiant du *phylotype* ; A, *annotation* ; Tt, *total* ; Sz, *size* ; Ps, *persistence* ; Df, *different* ; Sl, *local separation* ; Sg, *global separation* ; Dv, *diversity* ; Sp, *support* ; Spg, *global support*.

P	A	Tt	Sz	Ps	Df	Sz/Df	Sl	Sg	Dv	Sl/Dv	Sg/Dv	Sp	Spg
1	ZA	86	**76**	**2**	7	**10,857**	0,005	0,007	0,083	0,064	0,079	**0,855**	0,855
			0/1000	441/1000		0/1000						193/1000	
2	ZA	35	**23**	**2**	4	**5,750**	0,002	0,003	0,082	0,030	0,043	**0,835**	0,835
			8/1000	441/1000		0/1000						249/1000	
3	ZA	311	**261**	**3**	25	**10,440**	0,002	0,014	0,089	0,027	0,155	**0,831**	0,843
			0/1000	122/1000		0/1000						255/1000	
4	ZA	80	**24**	**2**	16	**1,500**	0,007	0,010	0,073	0,095	0,141	**0,876**	0,876
			8/1000	441/1000		25/1000						166/1000	
10	AR/BR	269	**260**	**2**	5	**52,000**	0,018	0,029	0,106	0,172	0,269	**0,992**	0,992
			0/1000	0/1000		0/1000						0/1000	
11	BE	11	**11**	**1**	0	**∞**	0,008	0,022	0,047	0,165	0,458	**0,858**	0,858
			0/1000	0/1000		0/1000						0/1000	
12	MM/CN/IN	356	**339**	**2**	14	**24,214**	0,004	0,022	0,081	0,044	0,265	**0,740**	0,882
			0/1000	4/1000		0/1000						4/1000	
13	MM/CN/IN	14	**11**	**2**	3	**3,667**	0,005	0,014	0,063	0,081	0,225	**0,843**	0,843
			2/1000	4/1000		0/1000						0/1000	
14	BW	52	**22**	**2**	9	**2,444**	0,002	0,007	0,074	0,033	0,097	**0,756**	0,756
			0/1000	0/1000		0/1000						0/1000	
15	BI	829	**289**	**3**	54	**5,352**	0,006	0,010	0,093	0,065	0,106	**0,861**	0,861
			0/1000	0/1000		0/1000						1/1000	
16	CG/CD	13	**12**	**1**	1	**12,000**	0,015	0,031	0,081	0,184	0,388	**0,885**	0,885
			0/1000	0/1000		0/1000						0/1000	
17	CU	18	**17**	**2**	1	**17,000**	0,009	0,024	0,097	0,097	0,247	**0,848**	0,848
			0/1000	0/1000		0/1000						0/1000	
18	ER/ET	71	**33**	**2**	30	**1,100**	0,005	0,014	0,052	0,094	0,263	**0,906**	0,906
			0/1000	0/1000		0/1000						0/1000	
19	ER/ET	24	**15**	**2**	8	**1,875**	0,008	0,013	0,080	0,102	0,159	**0,894**	0,894
			0/1000	0/1000		0/1000						0/1000	
20	ER/ET	47	**24**	**2**	17	**1,412**	0,003	0,023	0,061	0,041	0,385	**0,773**	0,773
			0/1000	0/1000		0/1000						0/1000	
21	KE/TZ	43	**31**	**3**	9	**3,444**	0,009	0,014	0,080	0,108	0,177	**0,926**	0,926
			0/1000	0/1000		0/1000						0/1000	
22	RO	12	**12**	**2**	0	**∞**	0,016	0,024	0,072	0,218	0,325	**0,950**	0,950
			0/1000	0/1000		0/1000						0/1000	
23	SN	33	**33**	**1**	0	**∞**	0,018	0,033	0,075	0,240	0,438	**0,980**	0,980
			0/1000	0/1000		0/1000						0/1000	
24	SZ	26	**17**	**2**	7	**2,429**	0,002	0,005	0,082	0,029	0,063	**0,913**	0,913
			0/1000	0/1000		0/1000						0/1000	
25	SZ	15	**12**	**2**	3	**4,000**	0,003	0,020	0,054	0,051	0,361	**0,750**	0,803
			0/1000	0/1000		0/1000						0/1000	
26	SZ	87	**64**	**3**	11	**5,818**	0,003	0,045	0,074	0,036	0,612	**0,749**	0,766
			0/1000	0/1000		0/1000						0/1000	
27	SZ	33	**30**	**2**	3	**10,000**	0,002	0,021	0,059	0,041	0,352	**0,781**	0,781
			0/1000	0/1000		0/1000						0/1000	
28	SZ	14	**13**	**2**	1	**13,000**	0,007	0,015	0,057	0,126	0,262	**0,909**	0,909
			0/1000	0/1000		0/1000						0/1000	
29	SZ	20	**18**	**3**	1	**18,000**	0,015	0,029	0,060	0,250	0,485	**0,967**	0,967
			0/1000	0/1000		0/1000						0/1000	
30	SZ	10	**10**	**1**	0	**∞**	0,008	0,011	0,058	0,135	0,197	**0,894**	0,894
			0/1000	0/1000		0/1000						0/1000	
31	ZM	3605	**475**	**3**	281	**1,690**	0,014	0	0,115	0,120	0,000	**0,880**	0,880
			0/1000	53/1000		0/1000						76/1000	
33	ZM	153	**19**	**1**	9	**2,111**	0,003	0,037	0,089	0,028	0,411	**0,805**	0,888
			10/1000	321/1000		0/1000						158/1000	

Tableau 11. Valeurs associées à chaque critère pour tous les *phylotypes* significatifs observés lorsque *size* ≥ 10 avec DELTRAN.

Ce tableau contient la liste de tous les *phylotypes* obtenus lorsque *size* ≥ 10, *persistence* ≥ 1, *size/different* ≥ 1 et *support* ≥ 70% avec la parcimonie DELTRAN, ainsi que toutes les valeurs associées à chacun des critères que l'on peut choisir. Les valeurs en rose représentent les p-valeur obtenues par le *shuffling*. Seuls les *phylotypes* dont la p-valeur est strictement inférieure à 11/1 000 pour le critère *size* sont représentés. Les abréviations des pays sont les suivantes : AR, Argentine ; BE, Belgique ; BI, Burundi ; BR, Brésil ; BW, Botswana ; CD, République Démocratique du Congo ; CG, Congo ; CN, Chine ; CU, Cuba ; ER, Érythrée ; ET, Éthiopie ; IN, Inde ; KE, Kenya ; MM, Birmanie ; RO, Roumanie ; SN, Sénégal ; SZ, Swaziland ; TZ, Tanzanie ; ZA, Afrique du Sud ; ZM, Zambie. Les abréviations des titres sont les suivantes : P, identifiant du *phylotype* ; A, *annotation* ; Tt, *total* ; Sz, *size* ; Ps, *persistence* ; Df, *different* ; Sl, *local separation* ; Sg, *global separation* ; Dv, *diversity* ; Sp, *support* ; Spg, *global support*.

P	A	Tt	Sz	Ps	Df	Sz/Df	Sl	Sg	Dv	Sl/Dv	Sg/Dv	Sp	Spg
1	ZA	86	**76**	**2**	9	**8,444**	0,005	0,007	0,083	0,064	0,079	**0,855**	0,855
			0/1000	88/1000		0/1000						39/1000	
2	ZA	16	**15**	**1**	1	**15,000**	0,005	0,006	0,081	0,060	0,080	**0,858**	0,858
			8/1000	99/1000		0/1000						39/1000	
3	ZA	311	**265**	**3**	33	**8,030**	0,002	0,014	0,089	0,027	0,154	**0,831**	0,843
			0/1000	16/1000		0/1000						44/1000	
4	ZA	80	**33**	**2**	29	**1,138**	0,007	0,010	0,075	0,092	0,137	**0,876**	0,876
			0/1000	88/1000		30/1000						35/1000	
8	ZA	51	**20**	**2**	16	**1,250**	0,005	0,008	0,071	0,069	0,108	**0,874**	0,874
			0/1000	88/1000		22/1000						35/1000	
9	AR/BR	269	**260**	**2**	5	**52,000**	0,018	0,029	0,106	0,172	0,269	**0,992**	0,992
			0/1000	0/1000		0/1000						0/1000	
10	BE	11	**11**	**1**	0	∞	0,008	0,022	0,047	0,165	0,458	**0,858**	0,858
			0/1000	0/1000		0/1000						0/1000	
11	MM/CN/IN	356	**339**	**2**	14	**24,214**	0,004	0,022	0,081	0,044	0,265	**0,740**	0,882
			0/1000	0/1000		0/1000						0/1000	
12	MM/CN/IN	12	**10**	**2**	2	**5,000**	0,006	0,018	0,056	0,102	0,320	**0,886**	0,886
			0/1000	0/1000		0/1000						0/1000	
13	BW	32	**16**	**1**	9	**1,778**	0,017	0,030	0,076	0,224	0,399	**0,939**	0,939
			0/1000	0/1000		0/1000						0/1000	
14	BI	829	**300**	**3**	76	**3,947**	0,006	0,010	0,093	0,065	0,106	**0,861**	0,861
			0/1000	0/1000		0/1000						0/1000	
15	CG/CD	13	**12**	**1**	1	**12,000**	0,015	0,031	0,081	0,184	0,388	**0,885**	0,885
			0/1000	0/1000		0/1000						0/1000	
16	CU	18	**17**	**2**	1	**17,000**	0,009	0,024	0,097	0,097	0,247	**0,848**	0,848
			0/1000	0/1000		0/1000						0/1000	
17	ER/ET	39	**19**	**2**	19	**1,000**	0,005	0,012	0,044	0,111	0,282	**0,902**	0,902
			0/1000	0/1000		0/1000						0/1000	
18	ER/ET	24	**15**	**2**	9	**1,667**	0,008	0,013	0,080	0,102	0,159	**0,894**	0,894
			0/1000	0/1000		0/1000						0/1000	
19	ER/ET	47	**24**	**2**	19	**1,263**	0,003	0,023	0,061	0,041	0,385	**0,773**	0,773
			0/1000	0/1000		0/1000						0/1000	
20	KE/TZ	43	**31**	**3**	10	**3,100**	0,009	0,014	0,080	0,108	0,177	**0,926**	0,926
			0/1000	0/1000		0/1000						0/1000	
21	RO	12	**12**	**2**	0	∞	0,016	0,024	0,072	0,218	0,325	**0,950**	0,950
			0/1000	0/1000		0/1000						0/1000	
22	SN	33	**33**	**1**	0	∞	0,018	0,033	0,075	0,240	0,438	**0,980**	0,980
			0/1000	0/1000		0/1000						0/1000	
23	SZ	13	**13**	**2**	0	∞	0,009	0,022	0,061	0,140	0,364	**0,885**	0,885
			0/1000	0/1000		0/1000						0/1000	
24	SZ	87	**70**	**3**	13	**5,385**	0,003	0,045	0,077	0,035	0,594	**0,749**	0,766
			0/1000	0/1000		0/1000						0/1000	
25	SZ	33	**30**	**2**	3	**10,000**	0,002	0,021	0,059	0,041	0,352	**0,781**	0,781
			0/1000	0/1000		0/1000						0/1000	
26	SZ	14	**13**	**2**	1	**13,000**	0,007	0,015	0,057	0,126	0,262	**0,909**	0,909
			0/1000	0/1000		0/1000						0/1000	
27	SZ	20	**18**	**3**	2	**9,000**	0,015	0,029	0,060	0,250	0,485	**0,967**	0,967
			0/1000	0/1000		0/1000						0/1000	
28	SZ	10	**10**	**1**	0	∞	0,008	0,011	0,058	0,135	0,197	**0,894**	0,894
			0/1000	0/1000		0/1000						0/1000	
29	ZM	3605	**564**	**4**	492	**1,146**	0,014	0	0,114	0,120	0,000	**0,880**	0,880
			0/1000	0/1000		14/1000						16/1000	

Tableau 12. Valeurs associées à chaque critère pour tous les *phylotypes* significatifs observés lorsque *size* ≥ 5 avec ACCTRAN.

Ce tableau contient la liste de tous les *phylotypes* obtenus lorsque *size* ≥ 5, *persistence* ≥ 1, *size/different* ≥ 1 et *support* ≥ 70% avec la parcimonie ACCTRAN, ainsi que toutes les valeurs associées à chacun des critères que l'on peut choisir. Les valeurs en rose représentent les p-valeur obtenues par le *shuffling*. Seuls les *phylotypes* dont la p-valeur est strictement inférieure à 11/1 000 pour le critère *size* sont représentés. Les abréviations des pays sont les suivantes : BE, Belgique ; BI, Burundi ; BR, Brésil ; BW, Botswana ; CD, République Démocratique du Congo ; CG, Congo ; CU, Cuba ; ER, Érythrée ; ET, Éthiopie ; IN, Inde ; KE, Kenya ; MM, Birmanie ; MZ, Mozambique ; NO, Norvège ; PT, Portugal ; RO, Roumanie ; SE, Suède ; SN, Sénégal ; SZ, Swaziland ; TZ, Tanzanie ; UG, Ouganda ; ZA, Afrique du Sud ; ZM, Zambie. Les abréviations des titres sont les suivantes : P, identifiant du *phylotype* ; A, *annotation* ; Tt, *total* ; Sz, *size* ; Ps, *persistence* ; Df, *different* ; Sl, *local separation* ; Sg, *global separation* ; Dv, *diversity* ; Sp, *support* ; Spg, *global support*.

P	A	Tt	Sz	Ps	Df	Sz/Df	Sl	Sg	Dv	Sl/Dv	Sg/Dv	Sp	Spg
3	ZA	86	**76**	**2**	7	**10,857**	0,005	0,007	0,083	0,064	0,079	**0,855**	0,855
			0/1000	1000/1000		9/1000						935/1000	
5	ZA	311	**261**	**3**	25	**10,440**	0,002	0,014	0,089	0,027	0,155	**0,831**	0,843
			0/1000	385/1000		9/1000						957/1000	
6	ZA	80	**24**	**2**	16	**1,500**	0,007	0,010	0,073	0,095	0,141	**0,876**	0,876
			9/1000	1000/1000		759/1000						889/1000	
17	BE	11	**11**	**1**	0	∞	0,008	0,022	0,047	0,165	0,458	**0,858**	0,858
			0/1000	0/1000		0/1000						0/1000	
18	MM/IN	356	**335**	**2**	18	**18,611**	0,004	0,022	0,082	0,044	0,264	**0,740**	0,882
			0/1000	325/1000		3/1000						372/1000	
19	MM/IN	14	**11**	**2**	3	**3,667**	0,005	0,014	0,063	0,081	0,225	**0,843**	0,843
			3/1000	325/1000		6/1000						185/1000	
20	BW	6	**5**	**2**	1	**5,000**	0,002	0,013	0,056	0,043	0,231	**0,747**	0,747
			5/1000	5/1000		0/1000						5/1000	
21	BW	19	**5**	**2**	1	**5,000**	0,002	0,003	0,059	0,040	0,044	**0,755**	0,755
			5/1000	5/1000		0/1000						4/1000	
22	BW	5	**5**	**2**	0	∞	0,011	0,073	0,041	0,271	1,780	**0,892**	0,892
			5/1000	5/1000		5/1000						1/1000	
23	BW	5	**5**	**2**	0	∞	0,006	0,017	0,021	0,283	0,800	**0,879**	0,879
			5/1000	5/1000		5/1000						1/1000	
24	BW	52	**22**	**2**	9	**2,444**	0,002	0,007	0,074	0,033	0,097	**0,756**	0,756
			0/1000	5/1000		0/1000						4/1000	
25	BR	269	**252**	**2**	10	**25,200**	0,018	0,029	0,107	0,171	0,267	**0,992**	0,992
			0/1000	70/1000		1/1000						0/1000	
26	BI	829	**289**	**3**	54	**5,352**	0,006	0,010	0,093	0,065	0,106	**0,861**	0,861
			0/1000	11/1000		0/1000						69/1000	
27	CG/CD	13	**12**	**1**	1	**12,000**	0,015	0,031	0,081	0,184	0,388	**0,885**	0,885
			0/1000	1/1000		0/1000						1/1000	
28	CU	5	**5**	**2**	0	∞	0,025	0,032	0,046	0,555	0,701	**0,965**	0,965
			0/1000	0/1000		0/1000						0/1000	
29	CU	18	**17**	**2**	1	**17,000**	0,009	0,024	0,097	0,097	0,247	**0,848**	0,848
			0/1000	0/1000		0/1000						0/1000	
30	ER/ET	71	**33**	**2**	30	**1,100**	0,005	0,014	0,052	0,094	0,263	**0,906**	0,906
			0/1000	0/1000		0/1000						0/1000	
31	ER/ET	24	**15**	**2**	8	**1,875**	0,008	0,013	0,080	0,102	0,159	**0,894**	0,894
			0/1000	0/1000		0/1000						0/1000	
32	ER/ET	47	**24**	**2**	17	**1,412**	0,003	0,023	0,061	0,041	0,385	**0,773**	0,773
			0/1000	0/1000		0/1000						0/1000	
33	ER/ET	15	**7**	**2**	6	**1,167**	0,002	0,011	0,062	0,040	0,171	**0,725**	0,759
			0/1000	0/1000		0/1000						0/1000	
34	KE/TZ	43	**31**	**3**	9	**3,444**	0,009	0,014	0,080	0,108	0,177	**0,926**	0,926
			0/1000	0/1000		0/1000						0/1000	
35	MZ	12	**9**	**2**	3	**3,000**	0,002	0,005	0,072	0,033	0,075	**0,808**	0,808
			0/1000	1/1000		0/1000						1/1000	
36	MZ	8	**8**	**1**	0	∞	0,008	0,014	0,061	0,139	0,227	**0,845**	0,845
			0/1000	1/1000		1/1000						1/1000	
37	UG	6	**5**	**1**	1	**5,000**	0,005	0,014	0,057	0,090	0,250	**0,848**	0,848
			0/1000	0/1000		0/1000						0/1000	
38	PT	7	**7**	**1**	0	∞	0,041	0,051	0,044	0,941	1,160	**0,999**	0,999
			0/1000	0/1000		0/1000						0/1000	
39	RO	7	**7**	**1**	0	∞	0,005	0,010	0,089	0,058	0,113	**0,770**	0,770
			0/1000	0/1000		0/1000						0/1000	
40	RO	12	**12**	**2**	0	∞	0,016	0,024	0,072	0,218	0,325	**0,950**	0,950
			0/1000	0/1000		0/1000						0/1000	
41	SN	33	**33**	**1**	0	∞	0,018	0,033	0,075	0,240	0,438	**0,980**	0,980
			0/1000	1/1000		1/1000						0/1000	
42	SE	10	**9**	**2**	1	**9,000**	0,012	0,017	0,061	0,199	0,282	**0,763**	0,763
			0/1000	0/1000		0/1000						0/1000	
43	SZ	26	**17**	**2**	7	**2,429**	0,002	0,005	0,082	0,029	0,063	**0,913**	0,913
			0/1000	126/1000		7/1000						16/1000	

44	SZ	9	**8** 1/1000	**2** 126/1000	1	**8,000** 1/1000	0,006	0,009	0,053	0,110	0,175	**0,810** 72/1000	0,810
46	SZ	15	**12** 0/1000	**2** 126/1000	3	**4,000** 3/1000	0,003	0,020	0,054	0,051	0,361	**0,750** 128/1000	0,803
47	SZ	87	**64** 0/1000	**3** 4/1000	11	**5,818** 1/1000	0,003	0,045	0,074	0,036	0,612	**0,749** 129/1000	0,766
49	SZ	33	**30** 0/1000	**2** 126/1000	3	**10,000** 1/1000	0,002	0,021	0,059	0,041	0,352	**0,781** 90/1000	0,781
50	SZ	14	**13** 0/1000	**2** 126/1000	1	**13,000** 1/1000	0,007	0,015	0,057	0,126	0,262	**0,909** 18/1000	0,909
51	SZ	20	**18** 0/1000	**3** 4/1000	1	**18,000** 1/1000	0,015	0,029	0,060	0,250	0,485	**0,967** 3/1000	0,967
52	SZ	10	**10** 0/1000	**1** 147/1000	0	∞ 147/1000	0,008	0,011	0,058	0,135	0,197	**0,894** 29/1000	0,894
53	ZM	3605	**475** 0/1000	**3** 275/1000	281	**1,690** 348/1000	0,014	0	0,115	0,120	0,000	**0,880** 765/1000	0,880

Tableau 13. Valeurs associées à chaque critère pour tous les *phylotypes* significatifs observés lorsque *size* ≥ 5 avec DELTRAN.

Ce tableau contient la liste de tous les *phylotypes* obtenus lorsque *size* ≥ 5, *persistence* ≥ 1, *size/different* ≥ 1 et *support* ≥ 70% avec la parcimonie DELTRAN, ainsi que toutes les valeurs associées à chacun des critères que l'on peut choisir. Les valeurs en rose représentent les p-valeur obtenues par le *shuffling*. Seuls les *phylotypes* dont la p-valeur est strictement inférieure à 11/1 000 pour le critère *size* sont représentés. Les abréviations des pays sont les suivantes : BE, Belgique ; BI, Burundi ; BR, Brésil ; BW, Botswana ; CD, République Démocratique du Congo ; CG, Congo ; CU, Cuba ; ER, Érythrée ; ET, Éthiopie ; IN, Inde ; KE, Kenya ; MM, Birmanie ; MZ, Mozambique ; NO, Norvège ; PT, Portugal ; RO, Roumanie ; SE, Suède ; SN, Sénégal ; SZ, Swaziland ; TZ, Tanzanie ; UG, Ouganda ; ZA, Afrique du Sud ; ZM, Zambie. Les abréviations des titres sont les suivantes : P, identifiant du *phylotype* ; A, *annotation* ; Tt, *total* ; Sz, *size* ; Ps, *persistence* ; Df, *different* ; Sl, *local separation* ; Sg, *global separation* ; Dv, *diversity* ; Sp, *support* ; Spg, *global support*.

P	A	Tt	Sz	Ps	Df	Sz/Df	Sl	Sg	Dv	Sl/Dv	Sg/Dv	Sp	Spg
1	ZA	86	**76**	**2**	9	**8,444**	0,005	0,007	0,083	0,064	0,079	**0,855**	0,855
			0/1000	952/1000		23/1000						715/1000	
2	ZA	16	**15**	**1**	1	**15,000**	0,005	0,006	0,081	0,060	0,080	**0,858**	0,858
			8/1000	987/1000		23/1000						703/1000	
4	ZA	311	**265**	**3**	33	**8,030**	0,002	0,014	0,089	0,027	0,154	**0,831**	0,843
			0/1000	221/1000		23/1000						780/1000	
5	ZA	80	**33**	**2**	29	**1,138**	0,007	0,010	0,075	0,092	0,137	**0,876**	0,876
			0/1000	952/1000		883/1000						643/1000	
13	ZA	51	**20**	**2**	16	**1,250**	0,005	0,008	0,071	0,069	0,108	**0,874**	0,874
			0/1000	952/1000		818/1000						646/1000	
14	BE	11	**11**	**1**	0	**∞**	0,008	0,022	0,047	0,165	0,458	**0,858**	0,858
			0/1000	0/1000		0/1000						0/1000	
15	MM/IN	356	**335**	**2**	18	**18,611**	0,004	0,022	0,082	0,044	0,264	**0,740**	0,882
			0/1000	125/1000		2/1000						144/1000	
16	MM/IN	12	**10**	**2**	2	**5,000**	0,006	0,018	0,056	0,102	0,320	**0,886**	0,886
			1/1000	125/1000		7/1000						43/1000	
17	BW	32	**16**	**1**	9	**1,778**	0,017	0,030	0,076	0,224	0,399	**0,939**	0,939
			0/1000	2/1000		0/1000						0/1000	
18	BW	5	**5**	**2**	0	**∞**	0,011	0,073	0,041	0,271	1,780	**0,892**	0,892
			2/1000	2/1000		2/1000						0/1000	
19	BW	5	**5**	**2**	0	**∞**	0,006	0,017	0,021	0,283	0,800	**0,879**	0,879
			2/1000	2/1000		2/1000						0/1000	
20	BW	5	**5**	**1**	0	**∞**	0,014	0,015	0,056	0,254	0,266	**0,913**	0,913
			2/1000	2/1000		2/1000						0/1000	
21	BW	12	**9**	**2**	3	**3,000**	0,015	0,023	0,041	0,367	0,560	**0,974**	0,974
			0/1000	2/1000		0/1000						0/1000	
22	BR	269	**252**	**2**	11	**22,909**	0,018	0,029	0,107	0,171	0,267	**0,992**	0,992
			0/1000	34/1000		0/1000						0/1000	
23	BI	829	**300**	**3**	76	**3,947**	0,006	0,010	0,093	0,065	0,106	**0,861**	0,861
			0/1000	5/1000		1/1000						24/1000	
24	CD/CG	13	**12**	**1**	1	**12,000**	0,015	0,031	0,081	0,184	0,388	**0,885**	0,885
			0/1000	0/1000		0/1000						0/1000	
25	CU	5	**5**	**2**	0	**∞**	0,025	0,032	0,046	0,555	0,701	**0,965**	0,965
			0/1000	0/1000		0/1000						0/1000	
26	CU	18	**17**	**2**	1	**17,000**	0,009	0,024	0,097	0,097	0,247	**0,848**	0,848
			0/1000	0/1000		0/1000						0/1000	
27	ER/ET	9	**5**	**2**	3	**1,667**	0,003	0,004	0,045	0,061	0,099	**0,778**	0,778
			0/1000	0/1000		0/1000						0/1000	
28	ER/ET	39	**19**	**2**	19	**1,000**	0,005	0,012	0,044	0,111	0,282	**0,902**	0,902
			0/1000	0/1000		0/1000						0/1000	
29	ER/ET	24	**15**	**2**	9	**1,667**	0,008	0,013	0,080	0,102	0,159	**0,894**	0,894
			0/1000	0/1000		0/1000						0/1000	
30	ER/ET	47	**24**	**2**	19	**1,263**	0,003	0,023	0,061	0,041	0,385	**0,773**	0,773
			0/1000	0/1000		0/1000						0/1000	
31	KE/TZ	43	**31**	**3**	10	**3,100**	0,009	0,014	0,080	0,108	0,177	**0,926**	0,926
			0/1000	0/1000		0/1000						0/1000	
32	MZ	12	**9**	**2**	3	**3,000**	0,002	0,005	0,072	0,033	0,075	**0,808**	0,808
			0/1000	1/1000		0/1000						1/1000	
33	MZ	8	**8**	**1**	0	**∞**	0,008	0,014	0,061	0,139	0,227	**0,845**	0,845
			0/1000	1/1000		1/1000						1/1000	
34	UG	6	**5**	**1**	1	**5,000**	0,005	0,014	0,057	0,090	0,250	**0,848**	0,848
			0/1000	0/1000		0/1000						0/1000	
35	PT	7	**7**	**1**	0	**∞**	0,041	0,051	0,044	0,941	1,160	**0,999**	0,999
			0/1000	0/1000		0/1000						0/1000	
36	RO	7	**7**	**1**	0	**∞**	0,005	0,010	0,089	0,058	0,113	**0,770**	0,770
			0/1000	0/1000		0/1000						0/1000	
37	RO	12	**12**	**2**	0	**∞**	0,016	0,024	0,072	0,218	0,325	**0,950**	0,950
			0/1000	0/1000		0/1000						0/1000	
38	SN	33	**33**	**1**	0	**∞**	0,018	0,033	0,075	0,240	0,438	**0,980**	0,980
			0/1000	1/1000		1/1000						0/1000	

39	SE	10	**9** 0/1000	**2** 0/1000	1	**9,000** 0/1000	0,012	0,017	0,061	0,199	0,282	**0,763** 0/1000	0,763
40	SZ	13	**13** 0/1000	**2** 47/1000	0	∞ 57/1000	0,009	0,022	0,061	0,140	0,364	**0,885** 10/1000	0,885
41	SZ	7	**7** 0/1000	**1** 57/1000	0	∞ 57/1000	0,008	0,016	0,048	0,177	0,328	**0,846** 21/1000	0,846
42	SZ	6	**6** 6/1000	**1** 57/1000	0	∞ 57/1000	0,003	0,006	0,052	0,049	0,121	**0,736** 47/1000	0,736
44	SZ	87	**70** 0/1000	**3** 3/1000	13	**5,385** 1/1000	0,003	0,045	0,077	0,035	0,594	**0,749** 45/1000	0,766
45	SZ	33	**30** 0/1000	**2** 47/1000	3	**10,000** 1/1000	0,002	0,021	0,059	0,041	0,352	**0,781** 32/1000	0,781
46	SZ	14	**13** 0/1000	**2** 47/1000	1	**13,000** 1/1000	0,007	0,015	0,057	0,126	0,262	**0,909** 5/1000	0,909
47	SZ	20	**18** 0/1000	**3** 3/1000	2	**9,000** 1/1000	0,015	0,029	0,060	0,250	0,485	**0,967** 0/1000	0,967
48	SZ	10	**10** 0/1000	**1** 57/1000	0	∞ 57/1000	0,008	0,011	0,058	0,135	0,197	**0,894** 8/1000	0,894
49	ZM	3605	**564** 0/1000	**4** 0/1000	492	**1,146** 770/1000	0,014	0	0,114	0,120	0,000	**0,880** 483/1000	0,880