Phylodynamique des infections virales

Samuel Alizon, Emma Saulnier

Phylodynamique des infections virales

Phylodynamics of viral infections

Samuel Alizon et Emma Saulnier

MIVEGEC, CNRS, IRD, Université de Montpellier, France

samuel.alizon@cnrs.fr et emma.saulnier@ird.fr

Résumé (200 mots)

La facilité croissante avec laquelle les gènes, voire les génomes, viraux sont séquencés a entraîné l'essor d'une discipline appelée phylodynamique. Celle-ci vise à utiliser les données de séquences génétiques pour estimer des paramètres tels que le taux de croissance de la population virale, le nombre d'infections ou même leur durée moyenne. L'hypothèse de travail est que la manière dont les virus se propagent laisse des traces dans leur génome. Dans cette revue, nous introduisons d'abord l'originalité des inférences en phylodynamique par rapport aux approches « classiques » en phylogénétique. Puis, nous présentons la nouveauté qu'ont constitué les phylogénies d'infections virales par rapport aux phylogénies d'espèces tout en donnant des pistes pour inférer ces phylogénies. Ensuite, nous retraçons la naissance de la phylodynamique et ses premiers succès, afin de passer en revue les différentes questions auxquelles l'approche permet de répondre. Enfin, nous présentons quelques défis pour la discipline.

Abstract (150 mots)

Phylodynamics is a recent field that aims at using genetic sequence data to estimate epidemiological parameters such as the viral population growth rate, the number of infections in the population or even their duration. Its main underlying assumption is that the way viruses spread leaves marks in their genome. In this review, we first introduce the originality of phylodynamics inferences compared to 'classical' phylogenetic approaches. Then, we present the novelty of using phylogenies of infec-

tions compared to species trees, while giving some directions to infer of such objects. We discuss the birth of phylodynamics and its first successes, in order to present some of the questions the approach can address. Finally, we highlight some future challenges for the field.

Introduction

Comprendre et quantifier la propagation des maladies infectieuses est un des défis majeurs en santé publique [1]. Depuis la naissance de l'épidémiologie moderne, les approches se fondent généralement sur des données d'**incidence** [1] ou de **prévalence**. En combinant ces données à des modèles mathématiques issus de la dynamique des populations, on peut estimer des paramètres d'intérêt, en particulier le nombre de reproduction de base (ou R_0), qui correspond au nombre d'infections secondaires causées par un individu infecté dans une population d'hôtes entièrement susceptibles [1–3]. Des suivis longitudinaux ou des questionnaires permettent d'estimer d'autres paramètres tels que la durée de l'infection ou la durée d'incubation, mais aussi d'inférer des réseaux de transmission [voir par exemple 4, 5].

Les techniques de **séquençage** ont beaucoup évolué au cours des dernières années et permettent aujourd'hui d'avoir accès, presque en temps réel et à moindre coût, aux génomes de virus causant des infections. Elles sont maintenant utilisées en routine pour le suivi des épidémies virales. L'épidémie d'Ebola qui a sévit de 2013 à 2016 en Afrique de l'ouest l'illustre de manière frappante car elle fut le théâtre d'une véritable course au **séquençage**, avec au final plus de 1600 séquences couvrant l'épidémie de 2014 à 2016 [6–8]. Ce **séquençage** massif ne concerne pas que les explosions épidémiques mais aussi des virus **endémiques** tels que le virus de l'immunodéficience humaine (VIH). Ainsi, la cohorte suisse sur le VIH (SHCS) estime avoir réuni des informations génétiques portant sur 54.7 % de toutes les infections par le VIH déclarées à l'Office fédéral de la santé publique suisse depuis 1985 [9].

Une des applications directes, voire même parfois routinière, de ces séquences consiste à rechercher des mutations dans le génome qui soient connues pour être associées à des traits particuliers d'une infection. On peut ainsi détecter des résistances aux traitements, comme dans le cas du VIH [10]. Une autre application concerne les virus émergents : les séquences permettent de positionner une nouvelle souche ou espèce virale par rapport à diversité connue. Dans le cas des récentes épidémies

1. Les termes en gras sont explicités dans le Glossaire

de virus Zika ou Ebola par exemple, la phylogénie a permis d'obtenir des informations sur l'origine génétique (génotype) et géographique de l'épidémie [6, 11, 12].

Toutefois, la prise de conscience de la précision que peut apporter l'utilisation de ce type de données dans l'étude quantitative de la propagation des épidémies reste relativement récente [13].

Phylogénies virales

Une phylogénie ou arbre phylogénétique est une structure de classification hiérarchique qui représente des relations de parenté entre « objets » (individus, groupes d'individus, espèces) sous la forme d'un arbre (Figure 1). Celui-ci est composé de nœuds internes formant des branches vers des descendants qui peuvent être d'autres nœuds internes ou des nœuds externes, aussi appelés feuilles. Ces feuilles constituent les « objets » dont on cherche a caractériser les relations de parenté et chaque nœud interne correspond alors à un **ancêtre commun** des feuilles qui en descendent. Les phylogénies sont généralement construites par comparaison d'espèces représentées par un type de données. Initialement on utilisait des caractères morphologiques pour construire ces phylogénies, mais aujourd'hui on utilise quasi-exclusivement des **séquences moléculaires consensus** (nucléotidiques ou protéiques). Les étapes clés de l'inférence d'une phylogénie virale sont décrites dans le Tableau 1 et plus de détails sont disponibles dans un ouvrage récent sur la biologie de l'évolution [14].

Le **séquençage** des micro-organismes, en particulier les virus à ARN, a permis l'essor des phylogénies d'individus de la même espèce du fait de leur génome court et de leur évolution rapide. Ceci est lié au concept clé d'**horloge moléculaire**, qui décrit la vitesse à laquelle les mutations sont fixées dans un génome (on parle de **taux de substitution**). D'un point de vue pratique, si l'on sait à quelle vitesse tourne l'horloge moléculaire, on peut estimer le nombre de substitutions que l'on s'attend à trouver entre deux génomes sachant la date à laquelle ils ont divergé. Inversement, si l'on possède les séquences virales et leurs dates de collecte, on peut estimer cette horloge moléculaire [15] et ce de manière spécifique pour différentes **clades** de la phylogénie [16]. Les virus à ARN font partie des être vivants évoluant le plus rapidement avec des taux de substitution pouvant atteindre plus de 10^{-3} substitutions par position dans le génome par an pour le VIH [17]. Pour les virus à ADN, ces taux sont plus faibles (de l'ordre de 10^{-4} à 10^{-6} [18]) mais restent bien plus élevés que ceux des mammifères (10^{-8} chez l'homme). Du fait de ces taux élevés, des **séquences consensus** de virus infectant des patients différents sur des échelles de temps relativement courtes (quelques jours) permettent souvent

de générer une phylogénie d'infections suffisamment **résolue**.

Une des premières études a porté sur le virus de la grippe [19]. Puis, dans le cas du VIH et du virus de l'hépatite C (VHC), les phylogénies virales ont été sur le devant de la scène pour leur utilisation dans des affaires judiciaires portant sur des réseaux de transmission et des directionalités d'événements, par exemple pour déterminer si des événements de transmission du personnel médical vers des patients avaient eu lieu [20].

Dans la suite de l'article, nous nous intéresserons aux phylogénies virales datées et plus particulièrement à celles construites à partir de séquences, généralement nucléotidiques, consensus d'un même pathogène viral échantillonné à différentes dates chez différents patients infectés issus de la même épidémie. La date d'échantillonnage est un atout majeur car elle permet le calibrage dans le temps de la phylogénie. Sans lui, la phylogénie virale ne contient aucune information dynamique.

Dans le raisonnement phylodynamique, on fait généralement l'approximation qu'un nœud interne correspond à une transmission (non orientée) et qu'une feuille correspond à une fin d'infection observée, pour laquelle on a échantillonné puis séquencé le virus. Formellement, ce parallèle est faux car une phylogénie n'est pas strictement équivalente à un réseau de transmission. Ainsi, on ignore qui est l'infectant dans une phylogénie virale. De plus, l'**échantillonnage** des hôtes infectés dans une épidémie est généralement partiel et on a donc toujours un arbre de transmission incomplet. De plus, l'agrégation ou non des cas échantillonné est aussi déterminante dans la couverture de l'épidémie [21]. Toutefois, l'analogie n'en reste pas moins informative (Figure 1).

Naissance de la phylodynamique

Le terme « phylodynamique » est apparu dans une revue de Grenfell *et alii* en 2004 [13] pour caractériser l'étude simultanée de l'épidémiologie, la biologie évolutive et la dynamique immunitaire d'un pathogène. L'essor de ce domaine a suivi celui des techniques de **séquençage**, mais aussi l'accroissement des puissances de calcul, qui ont rendu possible l'intégration des séquences pour l'inférence phylodynamique Bayésienne grâce à la méthode de Monte Carlo par Chaînes de Markov (**MCMC**) [22]. Sans entrer dans les détails, ces inférences nécessitent de définir avant l'analyse une distribution dans laquelle on s'attend à trouver les valeurs des paramètres cibles (on parle de *prior*). Suite à l'analyse on obtient un distribution affinée du prior, qui est appelée distribution postérieure.

L'inférence phylodynamique se base sur des modèles visant à décrire la dynamique épidémio-

logique. Les premiers modèles utilisés étaient des modèles démographiques basés sur la théorie du **coalescent** de Kingman [23] et qui faisaient l'hypothèse d'une croissance épidémique exponentielle ou de variations plus flexibles du nombre d'infectés au cours du temps [24–26]. Ainsi, l'une des premières études phylodynamique par Pybus *et alii* a consister à comparer l'histoire épidémiologique des sous-types A et B du VIH à l'aide d'un modèle de coalescent [24]. Ce dernier permettait de retracer l'évolution de la taille de la population d'infectés au cours du temps sans faire d'hypothèse sur le modèle démographique paramétrique.

Les premières approches estimaient les paramètres par méthode de maximum de vraisemblance. En effet, la théorie du coalescent permet d'exprimer une fonction de vraisemblance, qui décrit la probabilité d'observer une phylogénie sous l'hypothèse d'un modèle donné et de valeurs de paramètres associées. Un obstacle majeur était que les techniques classiques de maximisation de la vraisemblance nécessitaient des temps de calcul très longs. C'est cette limite que les méthodes bayésiennes ont permis de dépasser en 2005 en combinant le coalescent et la méthode MCMC [16]. Des versions ultérieures de cette approche ont été implémentées dans les logiciels BEAST [27] et BEAST2 [28].

Essor de la phylodynamique

Des progrès récents ont permis de concevoir des modèles démographiques plus proches des modèles épidémiologiques classiques. En 2009, Volz *et alii* ont décrit les fonctions de vraisemblance de modèles de coalescent adaptés aux modèles épidémiologique de type SIR. Ces derniers décrivent la dynamique d'hôtes pouvant appartenir à trois classes : susceptibles à l'infection (S), infectés et infectieux (I) et retirés du système (car guéris et immunisés ou morts, R). Un des paramètres clés de ces modèles, que le modèle de coalescent permet d'estimer, est le nombre de reproduction de base (R_0). Cette approche a été généralisée pour être aujourd'hui applicable à des modèles épidémiologiques plus complexes [29].

Un autre pan de la phylodynamique se base sur les processus de naissance et de mort (BD, pour *birth death* en anglais) et a été développé par Stadler en 2009 [30]. Dans les modèles BD, les transmissions se produisent à un taux constant et correspondent à des naissances d'infection et chaque guérison ou décès d'un individu infecté correspond à une fin (mort) d'infection. Ce modèle a ensuite été étendu à d'autres modèles épidémiologiques et permet l'inférence du R_0 et des durées de latence et d'infection [31–33].

D'autres extensions de ces méthodes ont été réalisées. Par exemple, Rasmussen *et alii* ont utilisé une technique de physique statistique dite de filtre à particules afin d'inférer les paramètres épidémiologiques et la dynamique passée d'une épidémie sous l'hypothèse d'un modèle SIR stochastique à partir de données de surveillance et de phylogénies [34]. Cette méthode ne nécessite aucune hypothèse sur l'intensité de l'échantillonnage, permet d'utiliser des informations sur la provenance des séquences et fournit la fonction de vraisemblance associée à chaque modèle.

Au final, malgré sa définition initiale [13], la phylodynamique s'est surtout intéressée aux liens entre l'évolution et l'épidémiologie des pathogènes viraux et a laissé de côté la dynamique immunitaire. Il existe des exceptions, comme par exemple Koelle *et alii*, qui ont combiné dynamique des populations et immunité, afin de générer des phylogénies ressemblant le plus possible aux phylogénies de grippe A H3N2 [35].

Sujets d'étude

Il n'est pas possibles d'entrer dans les détails des quelques deux cents études listées à ce jour dans le Web of Science et qui mentionnent la phylodynamique. On peut cependant avoir un aperçu de leur diversité de par les questions qu'elles posent.

Estimations de dynamique de populations

Dans la lignée des études initiales [24, 25], la phylodynamique est maintenant fréquemment utilisée pour inférer des variations de **taille de population efficaces**. Cette approche présente l'avantage d'être relativement heuristique, dans le sens où elle ne fait pas de présupposé sur la manière dont le virus se propage.

Ainsi, on peut voir sur la Figure 2 que le nombre de cas d'infection par le virus de l'hépatite C en Égypte a cru exponentiellement entre 1920 et 1950 [36]. Cette explosion des cas est bien documentée [25] et est liée aux infections iatrogènes transmises au cours de campagnes de traitement de la bilharziose par voie injectable.

Mais les méthodes récentes fournissent des interprétation plus mécanistes en permettant d'estimer des paramètres épidémiologiques. Ainsi, toujours dans le cas du VHC en Égypte, on peut voir que le taux de reproduction de base de l'épidémie (R_0) a commencé à augmenter avant que la grande augmentation du nombre de cas ne soit détectable. La durée d'infection, définie par la période pen-

dant laquelle un hôte infecté est infectieux, a elle baissé fortement entre l'apparition du virus et sa stabilisation dans la population (Figure 2C).

Réseaux de transmissions

On a mentionné qu'il existe un parallèle intuitif entre une phylogénie et un réseau de transmission. Certaines études ont donc utilisé les séquences pour compléter les informations provenant des données épidémiologiques dans le but d'inférer plus précisément les réseaux de transmissions [37]. D'autres études se sont attachées à expliciter le lien entre phylogénie et réseau de transmission [38, 39]. À l'inverse, des études de clustering (regroupement) peuvent se faire à une échelle plus large nationale, internationale, voire même intercontinentale [40].

Une limite importante à ne pas perdre de vue est que les résultats sont toujours dépendants de l'échantillonnage de la population infectée. Ainsi, les valeurs de support de nœuds dans la phylogénie (ou « bootstrap » en anglais), apportant une information limitée puisqu'ils ne permettent que de conclure qu'il n'existe pas d'autres séquences proches dans le jeu de données. Pour conclure que des séquences sont proches, il faut prendre en compte la longueur de branches.

De manière générale, il existe deux approches pour détecter des « cluster » de transmission à partir de jeux de données de séquences génétiques virales : on peut soit se baser sur une distance génétique « brute », soit utiliser une distance phylogénétique. Une étude récente a comparé les deux, mettant encore en avant l'importance du plan d'échantillonnage dans les résultats [41].

En lien avec les réseaux de transmission, des modèles de phylodynamiques se sont intéressés à estimer la nature des réseaux de contact entre les hôtes. Pour le cas d'une infection sexuellement transmissible par exemple, ces réseaux correspondent à ceux des interactions sexuelles et des études de terrain ont montré qu'ils sont particulièrement hétérogènes (beaucoup d'individus ont peu de partenaires et certaines individus ont énormément de partenaires). Toujours selon l'hypothèse que la manière dont les virus se propagent laissent des traces dans leurs génomes, on peut espérer détecter des propriétés du réseau de transmission à partir des séquences. Les premiers résultats ont permis de détecter une telle hétérogénéité à partir de séquences issues de la cohorte suisse pour l'étude du VIH [42] (Figure 3) et une étude récente mettant en jeu une approche de coalescent offre encore plus de puissance dans l'analyse [43].

Phylogéographie

Dans les études de phylodynamique mentionnées jusqu'ici, les seules données qui étaient mobilisées étaient les séquences génétiques et les dates d'échantillonnage. On dispose souvent de plus d'information, en particulier sur le lieu de collecte des séquences. Lier phylogénies et géographie n'est pas nouveau. Beaucoup d'études descriptives se sont basées sur des méthodes de parcimonie pour inférer les états ancestraux, c'est à dire la localisation géographique des différents nœuds de la phylogénie virale [44]. Toutefois, la phylogéographie a connu un renouveau grâce à la phylodynamique, qui a rendu possible l'estimation plus précise de paramètres de dispersion géographique.

Comme l'illustre la Figure 4 dans le cas du virus de la rage en Afrique [45], on peut aujourd'hui estimer conjointement la dynamique d'une épidémie et sa propagation géographique. Ceci permet également d'inférer les états ancestraux (c'est-à-dire par quels pays l'épidémie a circulé) mais aussi les variations de taille de population virales (la courbe en arrière plan sur la Figure 4). Enfin, on peut aussi caractériser les routes de dispersion entre pays.

Une des limites importantes des modèles de phylogéographie dits de « mugration », qui traitent la migration comme une mutation, est liée, encore une fois, à l'échantillonnage original. Intuitivement, il est compliqué de faire des inférences sur des endroits non échantillonnés. Mais en plus, le fait d'avoir un déséquilibre dans le jeu de données avec beaucoup d'échantillons d'une même région risque de donner un poids exagéré à cette région dans l'histoire de l'épidémie. Des modèles plus récents tentent de limiter de tels biais [46].

Identification de sources de propagation

Toujours dans le but de mieux contrôler efficacement la propagation des épidémies il est parfois important de prendre en compte dans les modèles épidémiologiques une structuration de la population d'hôtes. En effet, pour le VIH par exemple, on sait que certains individus transmettent plus que d'autre, en fonction de leur sexe, de leur orientation sexuelle et du stade de leur infection.

Le projet PANGEA-HIV (pour *Phylogenetics And Networks for Generalised HIV Epidemics in Africa*) a pour but d'analyser la propagation de l'épidémie de VIH en Afrique sub-saharienne grâce à 20.000 génomes mis en relation avec des données cliniques (phase de l'infection), démographiques (âge, sexe, type de relation) et épidémiologiques (prévalence). Le projet cherche notamment à identifier des cibles pour une nouvelle campagne de traitement grâce aux inférences phylodynamiques qui

seront produites sur ces données, concernant notamment les taux de transmission des infectés dans les différentes phases de l'infection par le VIH [47]. Notons au passage que le concours lancé par le consortium PANGEA a permis de commencer à résoudre un des soucis lié à la phylodynamique, qui était le peu d'étude comparant différentes approches.

Limites et perspectives

En 2015, des experts du champ ont fait le point sur les 8 défis pour la phylodynamique [48]. Ceux-ci portent sur les aspects génétiques (inclure la sélection et la recombinaison), épidémiologiques (gérer les échantillonnages biaisés et la stochasticité), biologiques (introduire la diversité des hôtes, combiner évolution intra- et inter-hôtes) et techniques (comment suivre la production de données et inclure d'autres types d'information que les séquences).

Pour insister sur certains de ces défis, il est vrai que la majorité des études phylodynamiques font une hypothèse de **neutralité** des processus évolutifs. Celle-ci permet de simplifier l'étude des phylogénies virales mais elle est souvent fausse, en particulier pour les virus les contraintes génomiques sont telles que peu de mutations peuvent être qualifiées de purement neutres. Il y a donc un enjeu à permettre l'utilisation de modèles d'évolutions de séquences non neutres.

De plus, les modèles épidémiologiques sous-jacents de ces études phylodynamiques sont souvent très simplifiés et ignorent en particulier la structuration de la population d'hôtes, qui est parfois essentielle pour comprendre et contrôler la propagation des épidémies virales. Il faut donc développer des modèles permettant l'inférence de paramètres pour des modèles épidémiologiques plus complexes. Une des difficultés à cela est que, pour la grande majorité des méthodes que nous avons présentées, il est nécessaire d'exprimer la fonction de vraisemblance associée au modèle dont on souhaite inférer les valeurs de paramètres à partir d'une phylogénie virale datée. Ceci est parfois difficile voire impossible pour des modèles complexes. Et même une fois la fonction exprimée, sa maximisation entraîne très souvent des difficultés techniques et un coût non négligeable en terme de temps de calcul.

Une des pistes qui permettrait de résoudre plusieurs de ces défis consiste à créer des méthodes phylodynamiques qui se passent de la fonction de vraisemblance, par exemple en utilisant des approches de type Calcul Bayésien Approché (CBA). Ces approches sont basées sur la simulation de données selon un modèle et la comparaison entre données simulées et données réelles, pour inférer les valeurs de paramètres de ce modèle associées aux données. L'approche CBA appliquée à la phylody-

namique semble se révéler d'une précision comparable aux approches fondées sur la vraisemblance et présente des atouts en termes de rapidité [49].

Un point crucial qui transparaît aussi des différents axes de recherche présentés dans cette revue est l'importance de la proportion des infections échantillonnées. De manière évidente, plus cette proportion est élevée, plus les résultats sont robustes. Mais un second aspect à ne pas négliger est l'homogénéité de cet échantillonnage dans la population d'intérêt. En effet, même avec une proportion d'échantillonnage élevée, le fait de ne pas avoir de séquence de sous-populations clés (celles-ci pouvant être définies sur des critères comportementaux, biologiques ou géographiques) peux fortement biaiser les données [40, 46]. Comme pour toute analyse, la puissance de la technique ne peut compenser une bonne connaissance biologique du système étudié.

Toujours sur les versant biologique, il faut reconnaître que la phylodynémique reste fortement associée aux virus à ARN. Toutefois, le nombre d'études de phylodynamique d'infections bactériennes commencent à augmenter du fait des progrès des méthodes de séquençage qui permettent de compenser les taux de substitution plus faibles par des séquences de taille plus longue. Pour des organismes évoluant encore plus lentement (eucaryotes ou virus à ADN double brin), les phylogénies d'infections semblent soit encore trop coûteuses car il faut des séquences très longues, soit impossibles du fait de la taille des génomes. Ainsi, même pour myxomavirus qui a une taille génomique respectable (162 kilobases), sa vitesse d'évolution d'environ 10^{-7} substitutions par site par an fait qu'une phylogénie d'infection basée sur des génomes entiers permet certes de retracer l'histoire de l'épidémie à un niveau continental mais pas d'analyser des échelles géographiques plus fines ou de dater des événements avec précision [50].

Un dernier défi consiste à convaincre, au delà du cercle des biologistes de l'évolution et des modélisateurs, que les données de séquences génétique peuvent constituer un apport déterminant pour comprendre l'épidémiologie des virus et même la biologie des infections qu'ils causent. En phytopathologie ce type d'approches est déjà un peu utilisé et par exemple une équipe a retracé la propagation du virus de la panachure jaune du riz en Afrique, montrant au passage l'importance de l'écologie de l'hôte dans la dispersion du virus [51, 52]. La phylodynamique reste plus une exception dans le domaine clinique, alors que pourtant elle y ouvre encore plus de débouchés par exemple du fait de l'évolution intra-hôte, que nous avons choisi de ne pas détailler ici [53]. De même en épidémiologie animale où la phylodynamique pourrait par exemple être un excellent outil dans la compréhension et

la lutte des explosions épidémiques de grippe aviaire [54]. Toutefois, la récente publication de données de séquences de l'épidémie de VIH en France montre la phylodynamique y émerge avec même des problématique de suivi en temps réel de l'épidémie par le biais des séquences virales [55].

References

[1] Anderson RM, May RM. *Infectious Diseases of Humans. Dynamics and Control*. Oxford: Oxford University Press, 1991.

[2] Keeling MJ, Rohani P. *Modeling infectious diseases in humans and animals*. Princeton University Press, 2008.

[3] Rohani P, King AA. Never mind the length, feel the quality: the impact of long-term epidemiological data sets on theory, application and policy. *Trends Ecol Evol*, 2010; 25(10) : 611–618.

[4] WHO Ebola Response Team. Ebola virus disease in West Africa–the first 9 months of the epidemic and forward projections. *N Engl J Med*, 2014; 371(16) : 1481–95.

[5] Faye O, Boëlle PY, Heleze E, et al. Chains of transmission and control of Ebola virus disease in Conakry, Guinea, in 2014: an observational study. *Lancet Infect Dis*, 2015; 15(3) : 320–326.

[6] Gire SK, Goba A, Andersen KG, et al. Genomic surveillance elucidates Ebola virus origin and transmission during the 2014 outbreak. *Science*, 2014; 345(6202) : 1369–72.

[7] Simon-Loriere E, Faye O, Faye O, et al. Distinct lineages of Ebola virus in Guinea during the 2014 West African epidemic. *Nature*, 2015; 524(7563) : 102–104.

[8] Holmes EC, Dudas G, Rambaut A, Andersen KG. The evolution of Ebola virus: Insights from the 2013-2016 epidemic. *Nature*, 2016; 538(7624) : 193–200.

[9] The Swiss HIV Cohort Study. Cohort Profile: The Swiss HIV Cohort Study. *Int J Epidemiol*, 2010; 39(5) : 1179–1189.

[10] Altmann A, Däumer M, Beerenwinkel N, et al. Predicting the Response to Combination Antiretroviral Therapy: Retrospective Validation of geno2pheno-THEO on a Large Clinical Database. *J Infect Dis*, 2009; 199(7) : 999–1006.

[11] Faria NR, Azevedo RdSdS, Kraemer MUG, et al. Zika virus in the Americas: Early epidemiological and genetic findings. *Science*, 2016; 352(6283) : 345–349.

[12] Quick J, Loman NJ, Duraffour S, et al. Real-time, portable genome sequencing for Ebola surveillance. *Nature*, 2016; 530(7589) : 228–232.

[13] Grenfell BT, Pybus OG, Gog JR, et al. Unifying the epidemiological and evolutionary dynamics of pathogens. *Science*, 2004; 303(5656) : 327–32.

[14] Thomas F, Lefevre T, Raymond M. *Biologie évolutive*. De Boeck Superieur, 2nd ed.

[15] Drummond A, Pybus OG, Rambaut A. Inference of viral evolutionary rates from molecular sequences. *Adv Parasitol*, 2003; 54 : 331–58.

[16] Drummond AJ, Ho SYW, Phillips MJ, Rambaut A. Relaxed phylogenetics and dating with confidence. *PLoS Biol*, 2006; 4(5) : e88.

[17] Alizon S, Fraser C. Within-host and between-host evolutionary rates across the HIV-1 genome. *Retrovirology*, 2013; 10 : 49.

[18] Firth C, Kitchen A, Shapiro B, Suchard MA, Holmes EC, Rambaut A. Using time-structured data to estimate evolutionary rates of double-stranded DNA viruses. *Mol Biol Evol*, 2010; 27(9) : 2038–51.

[19] Saitou N, Nei M. Polymorphism and evolution of influenza A virus genes. *Mol Biol Evol*, 1986; 3(1) : 57–74.

[20] Ou CY, Ciesielski CA, Myers G, et al. Molecular Epidemiology of HIV Transmission in a Dental Practice. *Science*, 1992; 256(5060) : 1165–1171.

[21] Novitsky V, Moyo S, Lei Q, DeGruttola V, Essex M. Impact of Sampling Density on the Extent of HIV Clustering. *AIDS Res Hum Retroviruses*, 2014; 30(12) : 1226–1235.

[22] Beaumont MA, Rannala B. The Bayesian revolution in genetics. *Nat Rev Genet*, 2004; 5(4) : 251–261.

[23] Kingman JFC. The coalescent. *Stochastic processes and their applications*, 1982; 13(3) : 235–248.

[24] Pybus OG, Rambaut A, Harvey PH. An integrated framework for the inference of viral population history from reconstructed genealogies. *Genetics*, 2000; 155(3) : 1429–1437.

[25] Pybus OG, Charleston MA, Gupta S, Rambaut A, Holmes EC, Harvey PH. The epidemic behavior of the hepatitis C virus. *Science*, 2001; 292(5525) : 2323–5.

[26] Strimmer K, Pybus OG. Exploring the demographic history of DNA sequences using the generalized skyline plot. *Mol Biol Evol*, 2001; 18(12) : 2298–2305.

[27] Drummond AJ, Rambaut A. BEAST: Bayesian evolutionary analysis by sampling trees. *BMC Evol Biol*, 2007; 7 : 214.

[28] Bouckaert R, Heled J, Kühnert D, et al. BEAST 2: a software platform for Bayesian evolutionary analysis. *PLoS Comput Biol*, 2014; 10(4) : e1003537.

[29] Volz EM. Complex population dynamics and the coalescent under neutrality. *Genetics*, 2012; 190(1) : 187–201.

[30] Stadler T. On incomplete sampling under birth-death models and connections to the sampling-based coalescent. *J Theor Biol*, 2009; 261(1) : 58–66.

[31] Stadler T, Kouyos R, von Wyl V, et al. Estimating the Basic Reproductive Number from Viral Sequence Data. *Mol Biol Evol*, 2012; 29(1) : 347–357.

[32] Leventhal GE, Günthard HF, Bonhoeffer S, Stadler T. Using an epidemiological model for phylogenetic inference reveals density dependence in HIV transmission. *Mol Biol Evol*, 2014; 31(1) : 6–17.

[33] Kühnert D, Stadler T, Vaughan TG, Drummond AJ. Simultaneous reconstruction of evolutionary history and epidemiological dynamics from viral sequences with the birth-death SIR model. *J R Soc Interface*, 2014; 11(94) : 20131106.

[34] Rasmussen DA, Ratmann O, Koelle K. Inference for nonlinear epidemiological models using genealogies and time series. *PLoS Comput Biol*, 2011; 7(8) : e1002136.

[35] Koelle K, Cobey S, Grenfell B, Pascual M. Epochal evolution shapes the phylodynamics of interpandemic influenza A (H3N2) in humans. *Science*, 2006; 314(5807) : 1898–903.

[36] Stadler T, Kühnert D, Bonhoeffer S, Drummond AJ. Birth-death skyline plot reveals temporal changes of epidemic spread in HIV and hepatitis C virus (HCV). *Proc Natl Acad Sci USA*, 2013; 110(1) : 228–33.

[37] Morelli MJ, Thébaud G, Chadœuf J, King DP, Haydon DT, Soubeyrand S. A Bayesian inference framework to reconstruct transmission trees using epidemiological and genetic data. *PLoS Comput Biol*, 2012; 8(11) : e1002768.

[38] Volz EM, Frost SDW. Inferring the Source of Transmission with Phylogenetic Data. *PLoS Comput Biol*, 2013; 9(12) : e1003397.

[39] Kenah E, Britton T, Halloran ME, Longini IM Jr. Molecular Infectious Disease Epidemiology: Survival Analysis and Algorithms Linking Phylogenies to Transmission Trees. *PLoS Comput Biol*, 2016; 12(4) : e1004869.

[40] Wertheim JO, Brown L, J A, et al. The Global Transmission Network of HIV-1. *J Infect Dis*, 2014; 209(2) : 304–313.

[41] Rose R, Lamers SL, Dollar JJ, et al. Identifying Transmission Clusters with Cluster Picker and HIV-TRACE. *AIDS Res Hum Retroviruses*, 2017; 33(3) : 211–218.

[42] Leventhal GE, Kouyos R, Stadler T, et al. Inferring epidemic contact structure from phylogenetic trees. *PLoS Comput Biol*, 2012; 8(3) : e1002413.

[43] Rasmussen DA, Kouyos R, Gunthard HF, Stalder T. Phylodynamics on local sexual contact networks. *PLoS Comput Biol*, 2017; in press.

[44] Chevenet F, Jung M, Peeters M, de Oliveira T, Gascuel O. Searching for virus phylotypes. *Bioinformatics (Oxford, England)*, 2013; 29 : 561–570.

[45] Lemey P, Rambaut A, Drummond AJ, Suchard MA. Bayesian phylogeography finds its roots. *PLoS Comput Biol*, 2009; 5(9) : e1000520.

[46] De Maio N, Wu CH, O'Reilly KM, Wilson D. New Routes to Phylogeography: A Bayesian Structured Coalescent Approximation. *PLoS Genet*, 2015; 11(8) : e1005421.

[47] Ratmann O, Hodcroft EB, Pickles M, et al. Phylogenetic Tools for Generalized HIV-1 Epidemics: Findings from the PANGEA-HIV Methods Comparison. *Mol Biol Evol*, 2017; 34(1) : 185–203.

[48] Frost SD, Pybus OG, Gog JR, Viboud C, Bonhoeffer S, Bedford T. Eight challenges in phylodynamic inference. *Epidemics*, 2015; 10 : 88–92.

[49] Saulnier E, Gascuel O, Alizon S. Inferring epidemiological parameters from phylogenies using regression-ABC: A comparative study. *PLoS Comput Biol*, 2017; 13(3) : e1005416.

[50] Kerr PJ, Ghedin E, DePasse JV, et al. Evolutionary History and Attenuation of Myxoma Virus on Two Continents. *PLoS Pathog*, 2012; 8(10) : e1002950.

[51] Fargette D, Pinel A, Abubakar Z, et al. Inferring the Evolutionary History of Rice Yellow Mottle Virus from Genomic, Phylogenetic, and Phylogeographic Studies. *J Virol*, 2004; 78(7) : 3252–3261.

[52] Trovão NS, Baele G, Vrancken B, et al. Host ecology determines the dispersal patterns of a plant virus. *Virus Evol*, 2015; 1(1).

[53] Hartfield M, Murall CL, Alizon S. Clinical applications of pathogen phylogenies. *Trends Mol Med*, 2014; 20(7) : 394–404.

[54] Roche B, Drake JM, Brown J, Stallknecht DE, Bedford T, Rohani P. Adaptive Evolution and Environmental Durability Jointly Structure Phylodynamic Patterns in Avian Influenza Viruses. *PLoS Biol*, 2014; 12(8) : e1001931.

[55] Brand D, Capsec J, Chaillon A, et al. HIV surveillance combining an assay for identification of very recent infection and phylogenetic analyses on dried spots. *AIDS*, 2017; 31(3) : 407–416.

[56] Gouy M, Guindon S, Gascuel O. SeaView Version 4: A Multiplatform Graphical User Interface for Sequence Alignment and Phylogenetic Tree Building. *Mol Biol Evol*, 2010; 27(2) : 221–224.

[57] Martin DP, Murrell B, Golden M, Khoosal A, Muhire B. RDP4: Detection and analysis of recombination patterns in virus genomes. *Virus Evolution*, 2015; 1(1) : vev003.

[58] Paradis E, Claude J, Strimmer K. APE: analyses of phylogenetics and evolution in R language. *Bioinformatics*, 2004; 20(2) : 289–290.

[59] Ronquist F, Teslenko M, Van Der Mark P, et al. MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space. *Syst Biol*, 2012; 61(3) : 539–542.

[60] Lartillot N, Philippe H. A Bayesian Mixture Model for Across-Site Heterogeneities in the Amino-Acid Replacement Process. *Mol Biol Evol*; 21(6) : 1095–1109.

[61] Posada D. jModelTest: phylogenetic model averaging. *Mol Biol Evol*, 2008; 25(7) : 1253–6.

[62] Guindon S, Dufayard JF, Lefort V, Anisimova M, Hordijk W, Gascuel O. New Algorithms and Methods to Estimate Maximum-Likelihood Phylogenies: Assessing the Performance of PhyML 3.0. *Syst Biol*, 2010; 59(3) : 307–321.

[63] Stamatakis A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *2014*; 30(9) : 1312–1313.

[64] To TH, Jung M, Lycett S, Gascuel O. Fast dating using least-squares criteria and algorithms. *Syst Biol*, 2015; 65(1) : 82–97.

Glossaire

Ancêtre commun : Nœud interne à la phylogénie correspondant à une entité hypothétique à l'origine du groupe d'entités séquencées (descendants) représentées par l'ensemble de feuilles sous le nœud.

Dernier ancêtre commun : Ancêtre commun direct d'un ensemble de feuilles d'une phylogénie.

Endémique : Se dit d'une maladie dont la prévalence a atteint un état stable non nul (s'oppose à *épidémique*).

Epidémique : Se dit d'une maladie dont la prévalence varie fortement (s'oppose à *endémique*).

Généalogie : Représentation d'arbre binaire particulière où les feuilles correspondent aux ancêtres et les nœuds internes se forment par la coalescence de deux ancêtres et sont plus récents.

Horloge moléculaire : Hypothèse selon laquelle le temps depuis lequel deux entités biologiques ont divergé est lié au nombre de substitutions fixées dans leurs génomes.

Incidence : Nombre de nouvelles infections par unité de temps.

MCMC : technique d'exploration de l'espace des paramètres, qui permet en particulier d'incorporer des connaissances *a priori* (*prior* en anglais) sur les valeurs de paramètres.

Neutralité : Correspond, en évolution, à l'hypothèse selon laquelle les mutations n'ont pas d'effet sur la valeur sélective des individus qui les portent (elles ont donc toutes la même probabilité de se fixer).

Parcimonie : Principe utilisé pour la reconstruction phylogénétique et consistant à minimiser le nombre de substitutions entre séquences pour déterminer leur degré de parenté.

Phylogénie datée : Arbre phylogénétique calibré dans le temps, c'est-à-dire dont les dates des feuilles et des nœds internes sont connues.

Prévalence : Nombre total d'individus infectés dans une population à un moment donné.

Réseau phylogénétique : Objet décrivant les relations de parentés entre séquences mais qui, contrairement aux phylogénie, autorise les transferts horizontaux de matériel génétique.

Séquence consensus : Construite à partir d'un ensemble de séquences homologues alignées, elle indique la base (ou l'acide aminé) le plus fréquent à chaque position.

Série temporelle : Suite de valeurs ordonnées décrivant les changements d'une variable (par exemple la prévalence ou l'incidence) au cours du temps.

Taille de population efficace : Concept de génétique des populations décrivant le nombre d'individus génétiquement différents nécessaire pour expliquer des résultats de diversité génétique. Une taille de 1 signifie qu'un résultat pourrait être expliqué avec une population entièrement clonale.

Taux de substitution : Vitesse à laquelle les mutations se fixent dans un génome. Il est différent du taux de mutation, qui correspond à la vitesse à laquelle les mutations apparaissent. Au final, le taux de substitution incorpore toute la dynamique des populations, qui détermine le sort d'une mutation (fixation ou extinction).

Théorie du coalescent : Approche de génétique des populations consistant à remonter au dernier ancêtre commun en partant des feuilles d'une phylogénie et en calculant des probabilités de coalescence (ou fusion) entre branches.

Tableau 1 – **Étapes d'inférence d'une phylogénie virale datées.** Notez qu'à partir de l'étape 3 on peut suivre deux procédures (bayésien ou maximum de vraisemblance) qui ont chacune leurs avantages et inconvénients. Pour plus de détails, voir [14].

	Étape	**Logiciels**
1	Alignement des séquences	MAFFT, Muscle (SeaView [56])
2	Détection des recombinaisons : - *si peu de séquences concernées (<10 %)* : les supprimer et retourner à l'étape 1 - *si beaucoup de séquences concernées* : sélectionner une portion de la région séquencée et retourner à l'étape 1	RDP [57] SeaView [56], R et ape [58]
	Par approche bayésienne (peu de données)	Beast2 [28], Mr Bayes [59], PhyloBase [60]
3	Sélection du modèle d'évolution de séquences	
4	Sélection du modèle démographique	
5	Sélection du modèle de datation	
6	Inférence bayésienne	
7	Construction de l'arbre consensus	
	Par maximum de vraisemblance	
3	Sélection du modèle d'évolution de séquences	jModelTest [61]
4	Construction de la phylogénie par maximum de vraisemblance	PhyML [62], RaxML [63]
5	Datation	LSD [64]

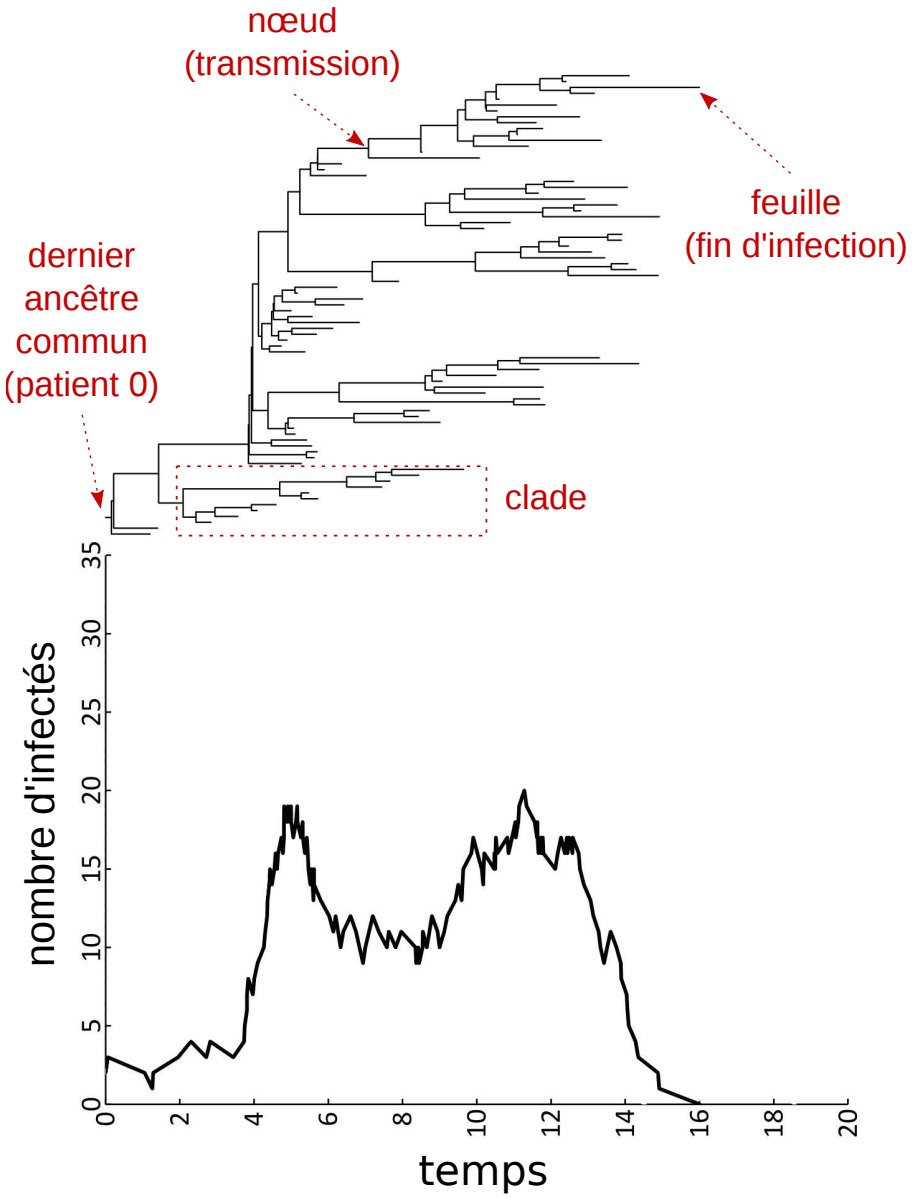

FIGURE 1 – **Phylogénies correspondant à une épidémie.** En haut, les éléments clés de la phylogénie sont décrits en rouge (voire aussi le Glossaire) et en bas une représentation « classique » d'une épidémie par une série temporelle de **prévalence**. On voit par exemple que l'augmentation du nombre de cas à partir de la 4e unité de temps s'accompagne d'un foisonnement de branches dans la phylogénies. À l'inverse, la fin de l'épidémie se traduit par une abondance de feuilles dans la phylogénie. Cette illustration a été créée par Samuel Alizon et est mise à disposition sous une licence CC-BY 4.0.

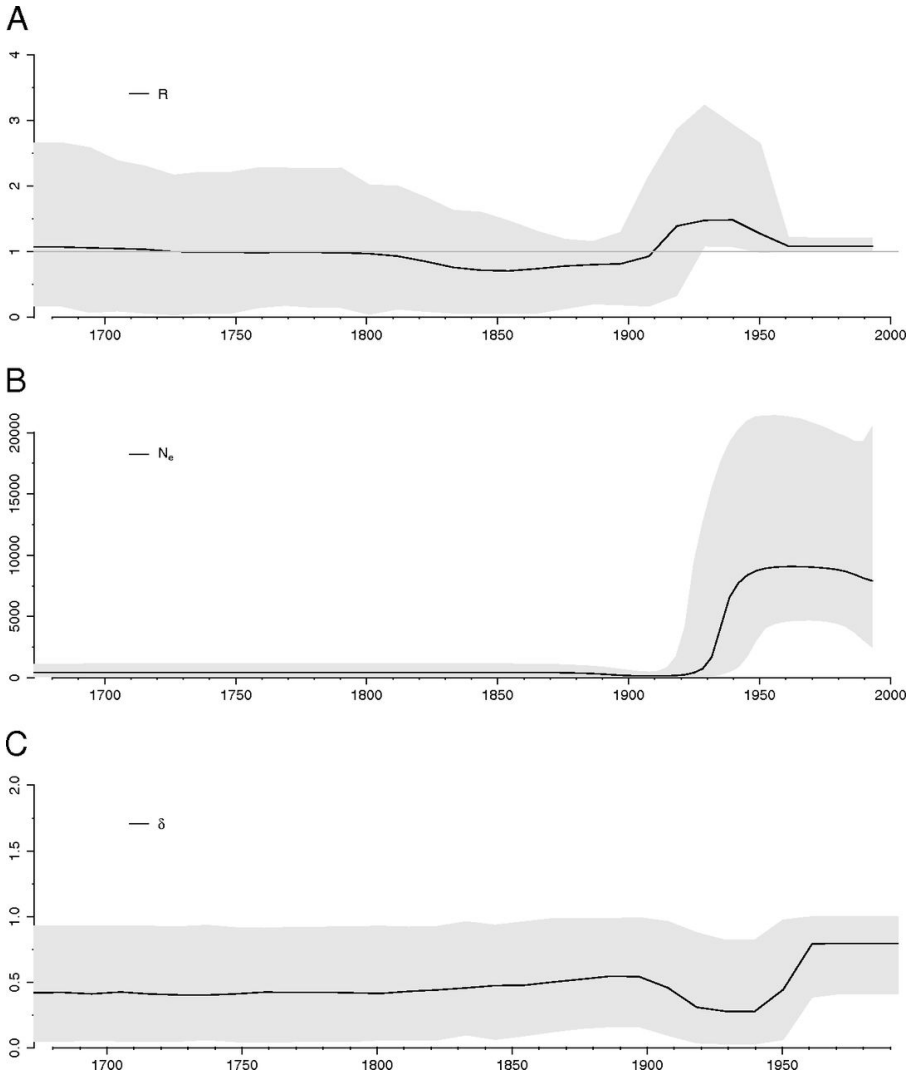

FIGURE 2 – **Inférence phylodynamique bayésienne de l'épidémie de VHC en Égypte.** A) Variation du taux de reproduction de base (R_0), B) variations de la taille de population efficace (N_E) et C) variations du taux de fin des infections (δ). Les lignes noires représentent les valeurs médianes inférées et les régions grisées les intervalles de crédibilité à 95 %. Cette figure correspond à la Figure 3 de [36].

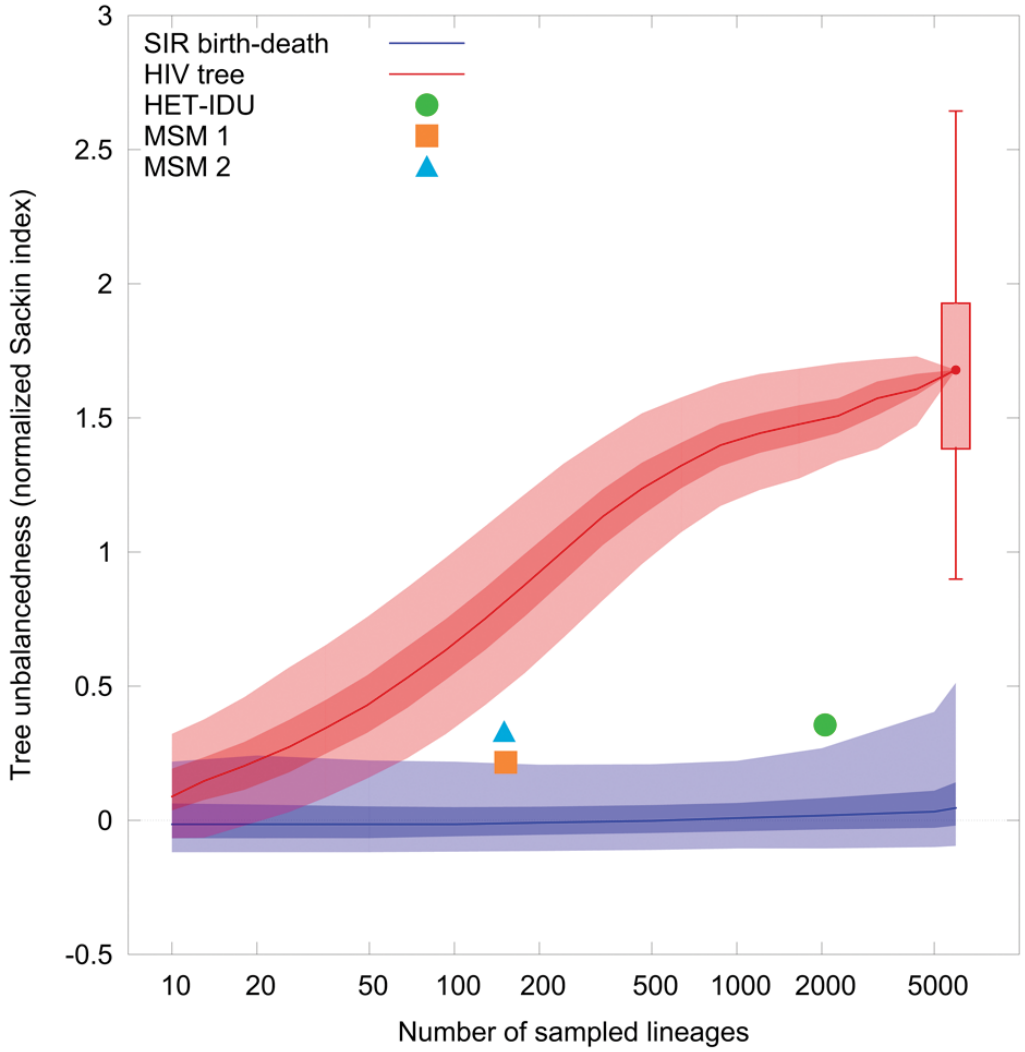

FIGURE 3 – **La forme des phylogénies démontre que les contacts entre individus ne se font pas au hasard.** L'indice utilisé mesure l'asymétrie de la phylogénie. La phylogénie inférée à partir de 5961 séquences de patients de l'Étude suisse de cohorte VIH (SHCS) en rouge dénote une asymétrie bien plus grande qu'une phylogénie de même taille simulée dans une population sans structure de contact (tout le monde peut être en contact avec tout le monde) en bleu. Le point à droite en rouge avec la boîte à moustaches indique la valeur obtenue sur la phylogénie de la SHCS entière et 100 arbres de bootstrap. La courbe en rouge a été obtenue en échantillonnant au hasard dans l'arbre complet du VIH. Les zones légèrement ombragées montrent les intervalles de crédibilité à 95 % et les zones sombres sombres ceux à 50 %. Les points de données individuels sont les valeurs pour les trois plus grands groupes de transmission : hétérosexuels et injection de drogues par voie intraveineuse (HET-IDU) et hommes ayant des relations sexuelles avec des hommes (MSM). Figure 7 de [42].

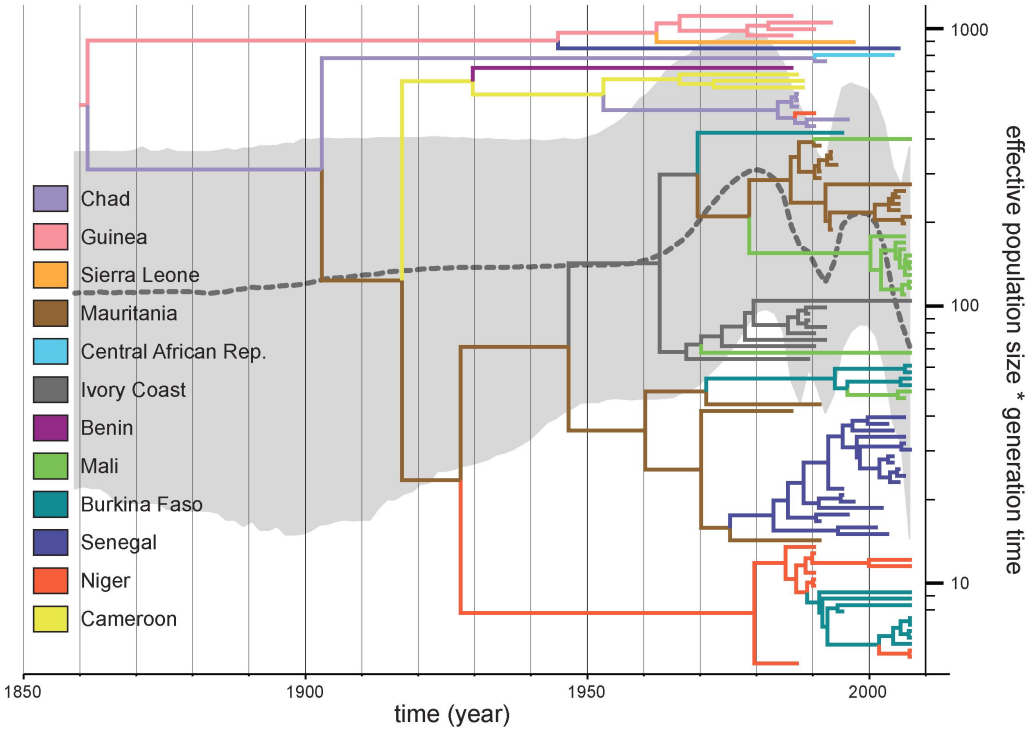

FIGURE 4 – **Phylogéographie du virus de la rage en Afrique.** Les couleurs de branches correspondent aux localisations ancestrales les plus probables du virus. En arrière-plan, la courbe en tirets et la zone grisée représentent la médiane et l'intervalle de crédibilité à 95 % de la taille de population effective. Cette figure correspond à la Figure 6A de [45].